UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

SISTEMA DE ALERTA TEMPRANA PARA ALUMNOS DE INGENIERÍA DE LA UNIVERSIDAD DE CHILE EN RIESGO DE REPROBAR UN RAMO

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN COMPUTACIÓN

DIEGO RODRIGO BAÑO RAMÍREZ

PROFESOR GUÍA:
JOCELYN SIMMONDS WAGEMANN
PROFESOR GUÍA 2:
WILLY MAIKOWSKI CORREA

MIEMBROS DE LA COMISIÓN:
JOSÉ PIQUER GARDNER
BARBARA POBLETE LABRA

SANTIAGO DE CHILE
2019

Resumen

El aumento en la cantidad de datos disponibles en las áreas de educación, gracias al uso de sistemas de administración del aprendizaje (LMS por sus siglas en inglés) ha ido incrementando con el paso de los años. Esto, mezclado con la motivación de mejorar las metodologías de enseñanza y las labores de docencia en las instituciones de educación superior, son los principales motivadores para la creación de sistemas que hagan uso de estos datos, con el objetivo de generar predicciones e indicadores de los resultados obtenidos por los estudiantes y permitir generar políticas institucionales para mejorar diversas métricas relacionadas.

Celis, Moreno, Poblete, Villanueva y Weber, en colaboración con el Centro Tecnológico Ucampus de la Universidad de Chile, desarrollaron el 2015 un Sistema de Alerta Temprana con el objetivo de predecir la probabilidad de doble reprobación para alumnos en su segundo semestre de estudio en el Plan Común de Ingeniería y Ciencias en la Facultad de Ciencias Físicas y Matemáticas (FCFM) de la Universidad de Chile, de modo que la institución les pudiera ofrecer apoyo o generar políticas al respecto. Este sistema logró resultados bastante buenos según los mismos autores, pero tenía el problema de ser demasiado limitado en el universo de estudiantes al cuál les es útil, consistiendo entre 100 y 200 personas de cada cohorte anual.

El objetivo del trabajo realizado en esta ocasión consistió en ampliar el espectro de estudiantes analizados por el sistema creado por Celis et al., analizar la posibilidad de integrar estadísticas de uso de las plataformas de LMS utilizadas y disponibilizar la información para los usuarios finales en interfaces fáciles de usar. La metodología de trabajo realizada consistió en ir procesando los datos de los estudiantes e ir ingresándolos a la heurística en dos iteraciones, incluyendo más cursos del Plan Común en la primera y luego integrando el resto de los datos, de modo que se pudiera evaluar que todo siguiera funcionando dentro de ciertos parámetros razonables o, de lo contrario, poder realizar modificaciones a la lógica lo antes posible para integrar más personas al alero del Sistema de Alerta Temprana.

Finalmente, se logró crear un modelo que analiza la probabilidad de reprobar nuevamente para cualquier alumno de pregrado en la FCFM. Este sistema predijo correctamente al 73 % de las personas que volvieron a reprobar pero obtuvo un 34 % de precisión en la predicción, lo que está dentro de los límites aceptables de rendimiento. No se logró incluir las estadísticas de uso del LMS debido a limitaciones en el tiempo de desarrollo. Finalmente, se generaron las primeras versiones de las interfaces, integradas como un módulo a la plataforma U-Cursos, donde se entregará esta información a profesores, académicos y funcionarios de la universidad. Estas interfaces obtuvieron una buena recepción por parte de los usuarios entrevistados.

A mi familia, por el apoyo que me han otorgado a lo largo de mi vida.

A Valentina, por motivarme a ser mejor cada día.

A las ratas y los manDCCos, por siempre subirme el ánimo.

A Ucampus y mis profesores, por creer en mis capacidades.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Introducción

Contexto y oportunidad a abordar

En la actualidad, existe un gran número de datos asociados al mundo de la educación gracias al uso cada vez más frecuente de plataformas computacionales, destinadas a la gestión y administración de las diversas actividades involucradas con el mantenimiento de una institución educativa. Este aumento en la información disponible motivó la creación de distintas comunidades en el mundo anglosajón dedicadas a estudiar y mejorar las labores de docencia, utilizando herramientas computacionales para el análisis y clasificación de los datos. Dichas comunidades reciben el nombre de *Learning Analytics* (LA) y *Educational Data Mining* (EDM) [14].

Este tema es de bastante relevancia para la Universidad de Chile, una de las instituciones de educación superior más importantes en el país y la casa de estudio de miles de personas tanto en planes de pregrado como postgrado [13]. La importancia del uso de estos datos nace en primer lugar del gran volumen de información que ellos manejan, gracias a que, en su día a día, los estudiantes de esta institución interactúan constantemente y de distintas maneras con las plataformas digitales creadas por el Centro Tecnológico Ucampus: *U-Cursos*[1] y *Ucampus*[2].

Dentro de los usos que le puede dar la Universidad de Chile a la información generada, uno de los principales objetivos que siempre mantiene en su horizonte es la mejora de la docencia y, por consecuencia, la mejora de los profesionales que egresan de la institución. Esto último forma parte de su misión, no solo porque ayuda a llevar adelante el país con las diversas mejoras que podrían generar sus egresados, si no que también sube el prestigio de la casa de estudio y se vuelve un lugar cada vez más interesante para estudiantes locales o extranjeros que buscan seguir sus estudios de educación superior.

Una de las formas en que esta universidad busca mejorar la docencia, además de la calidad y cantidad de egresados, es tratando de disminuir la deserción estudiantil de las distintas carreras que aquí se imparten. Según algunos estudios realizados internacional [20] y nacionalmente [11][16], uno de los factores que mayor impacto tiene en la deserción es el rendimiento académico obtenido por los estudiantes a lo largo de sus carreras. Por este motivo, la predicción del rendimiento académico, utilizando herramientas computacionales para el análisis

[1] `http://ucursos.cl`
[2] `https://ucampus.uchile.cl`

de los datos generados por los alumnos, puede ser de gran utilidad para ofrecer planes de apoyo, generar políticas institucionales, ajustar la forma en que se hace la docencia, entre otras medidas, que pueden ayudar a mejorar el rendimiento de los estudiantes y evitar su deserción.

En el capítulo 2 se habla más en detalle de las motivaciones detrás del trabajo realizado, junto al contexto que envuelve tanto este desarrollo como otros del mismo estilo.

La base del proyecto

El trabajo a realizar se centró en el estudio de la comunidad de estudiantes pertenecientes a la Facultad de Ciencias Físicas y Matemáticas (FCFM) de la Universidad de Chile. Esto es porque, en 2015, Celis, Moreno, Poblete, Villanueva y Weber, en colaboración con el Centro Tecnológico Ucampus de la Universidad de Chile, desarrollaron un Sistema de Alerta Temprana con el objetivo de predecir la probabilidad de doble reprobación para alumnos en su segundo semestre de estudio en el Plan Común de Ingeniería y Ciencias, carrera dictada en la FCFM, el cual sirvió como una base para comenzar con el trabajo y también una referencia ante la cuál se pudieron comparar los resultados obtenidos luego de realizar este proyecto [22].

El modelo de Celis et al. decidió enfocarse en la predicción de doble reprobación de los estudiantes, ya que las normas de la Universidad de Chile [8] establecen la reprobación de ramos cursados por segunda oportunidad como una causal de eliminación de la carrera, lo que va en contra del objetivo buscado de disminuir la salida de estudiantes de la institución. Por otro lado, el concentrarse solo en los alumnos de la FCFM permite dejar de lado ciertas diferencias que pueden derivarse por tratar con carreras de distintas áreas del conocimiento, o por tener perfiles de estudiantes demasiado diferentes entre sí, ya que la mayoría de los estudiantes de la FCFM tienen un nivel socio económico relativamente similar.

Por otro lado, Celis et al. realizaron la predicción de doble reprobación solo para alumnos del primer año del Plan Común de Ingeniería y Ciencias, ya que permitía analizar las variables que más impactaban en un conjunto relativamente uniforme de personas, las cuales debían tomar exactamente los mismos ramos y en los mismos periodos de tiempo. Esto no es tan cierto al avanzar la carrera, puesto que los estudiantes pueden atrasarse en su progreso o escoger diferentes líneas de especialización, con lo que se vuelve más difícil el procesamiento de los datos.

Con el modelo estadístico finalizado, Celis et al. consiguieron identificar correctamente al 86 % de los alumnos que finalmente reprobaron un ramo por segunda vez, lo que se puede considerar un resultado bastante bueno, ya que, de haber sido implementado para el conjunto utilizado para validar, hubiera permitido entregar apoyo, o generar planes de acción, para la gran mayoría de estudiantes que tuvieron dificultades para aprobar el ramo en su segunda oportunidad. Estos resultados abren la posibilidad de extender este modelo y las heurísticas utilizadas en el procesamiento de los datos, de modo que se pueda abarcar un público objetivo aún más general que el estudiado en esa ocasión.

La última parte del capítulo 2 consiste en una explicación más detallada de este modelo y cómo este se puede utilizar como una buena base para el desarrollo de una versión mejorada.

Trabajo a realizar

De acuerdo a lo anterior, el trabajo realizado por el estudiante en esta memoria, en conjunto con el Centro Tecnológico Ucampus, consistió en adaptar y escalar el modelo ya descrito, de modo que pudiera ser aplicado a cualquier alumno de pregrado en alguna de las carreras de ingeniería de la FCFM, independiente del año en que se encuentre y los ramos que posea o que haya o no reprobado en el pasado. Además de lo anterior, se decidió crear una plataforma web, integrada a las utilizadas en la Universidad de Chile, donde se entrega esta información procesada para que pueda ser utilizada por funcionarios, académicos e investigadores pertenecientes a la comunidad de la FCFM, siendo estos los usuarios finales a los que se apunta, con el fin de ayudar en la entrega de apoyo y la mejora de los planes educacionales ofrecidos por la institución.

El sistema creado podría ser una adición relevante para gran parte de los usuarios de las plataformas de la Universidad de Chile, ya sean alumnos, docentes, académicos o investigadores de la FCFM, ya que permitirá tomar medidas a tiempo para disminuir la reprobación de estudiantes identificados con un mayor riesgo, entregándoles el apoyo necesario para salir adelante o generar medidas a nivel institucional para disminuir la reprobación y deserción. Se espera que esto genere un impacto positivo en la vida de los estudiantes, puesto que – de tomar acciones según la información entregada – podría disminuir la carga emocional que significa reprobar un ramo y el tiempo que estos se demoran en terminar la carrera. Del mismo modo, la herramienta también podría ser útil para el equipo docente de cada ramo, mostrando información relevante con respecto a sus alumnos, para que puedan realizar los ajustes necesarios al curso y así mejorar el desempeño general. Por último, las instituciones mismas se verían beneficiadas por la existencia de este sistema, ya que los alumnos podrían mejorar su rendimiento académico general y evitarían así entrar en causal de eliminación, tanto como podrían evitar la deserción causada por malos resultados en la carrera.

Se decidió extender el modelo agregando información de todos los alumnos de pregrado de la FCFM, ya que estos poseen muchas características en común, tanto socio económicas como en los cursos que son impartidos en la facultad. Dentro de la malla curricular ofrecida en la institución, gran parte de los cursos comparten muchas similitudes en cuanto al tipo de evaluaciones que se realiza, incluso si la especialidad puede diferir. También, hay muchos cursos que comparten una buena cantidad de contenidos y que podrían mantener el tipo de datos utilizados en un margen de similitud razonable. Se descartó, por los mismos motivos, tratar de incluir en este estudio a personas cursando carreras en otras disciplinas o universidades, debido a que estos requerirían un análisis previo para estimar el efecto que podrían tener en los datos estadísticos con los cuáles se entrenarán los algoritmos.

Se puede encontrar una explicación más detallada de las decisiones tomadas y la justificación detrás de estas en el capítulo 3.

Objetivos del trabajo

Como objetivo general, se espera que el modelo a realizar sea capaz de predecir, con un rendimiento parecido al creado por Celis et al., la doble reprobación de estudiantes de pregrado en la FCFM, pero sin discriminar por año de ingreso, especialidad en curso u otras variables que sí consideraba el modelo original.

Para lograr esto, se propone realizar los siguientes objetivos específicos:

- Este modelo debe ser capaz de entregar predicciones para alumnos en cursos en progreso lo antes posible, ya sea utilizando la misma heurística que el modelo original o incorporando más información al procesamiento y generación del nuevo modelo.
- Analizar el efecto generado por la incorporación de algunas variables no estudiadas por Celis et al., como las estadísticas de uso de U-Cursos, en la predicción de doble reprobación o en el tiempo mínimo necesario para poder entregar un resultado confiable.
- Crear las interfaces necesarias para poder entregar esta información a los usuarios finales: docentes, académicos y funcionarios de la universidad, las cuales deben estar integradas en alguna de las plataformas creadas por el Centro Tecnológico Ucampus utilizadas en la FCFM y estas deben ser validadas por algunos de los usuarios, de modo que considere aceptable la información que se entrega para la implementación de políticas institucionales o sistemas de apoyo estudiantil.

Generación del modelo

Para la creación de la versión mejorada del modelo de Celis et al., se decidió trabajar en dos iteraciones, en cada cual se agregarían más datos al modelo y se evaluaría su desempeño. De obtener buenos resultados se pasaría a la siguiente iteración, en caso contrario, se realizaría un análisis del modelo hasta ese momento con el objetivo de incorporar nuevas variables o modificar las existentes, hasta generar un modelo aceptable que permita incorporar un mayor universo de estudio.

En la primera iteración del trabajo se incorporaron los datos de todos los cursos de primer año de Plan Común, en vez de solo los de primer semestre, pero sin discriminar en qué momento estos fueron tomados nuevamente. Esto implicó modificar la lógica de obtención de cursos y las comparaciones que eran realizadas. Los resultados obtenidos para esta iteración fueron bastante positivos a pesar, o en consecuencia, de que el conjunto de datos utilizado para el entrenamiento y validación presentaba un sesgo por la naturaleza de su origen. Producto de lo anterior se decidió seguir trabajando con la misma heurística y probar a ingresar el mayor conjunto posible para analizar el comportamiento del modelo.

La segunda iteración, por lo tanto, consistió en procesar todos los datos existentes de personas cursando ramos reprobados en el pasado, que se estuvieran dando en el contexto de un plan de pregrado en la FCFM, pero sin discriminar por el año en que se dieron los ramos, la especialidad a la que pertenece cada curso, el momento en que se decidió volver a tomar o la cantidad de veces que se hubiera reprobado en el pasado (siempre y cuando fuera

mayor o igual a una). Para esta iteración, el desempeño bajó en relación al modelo original de Celis et al. y de la primera iteración realizada, pero aun así se obtuvieron resultados bastante aceptables teniendo en cuenta el cambio que se realizó en el universo de datos.

Al terminar el desarrollo del nuevo modelo, se intentó integrar estadísticas del uso de U-Cursos por cada alumno de la FCFM a la lógica de procesamiento, con el objetivo de mejorar el desempeño obtenido al final de la última iteración y analizar el impacto que tiene el uso de la plataforma en el rendimiento académico. Lamentablemente, el procesamiento de la enorme cantidad de registros tomó más tiempo del esperado y esta parte del desarrollo tuvo que ser dejada de lado, sin embargo, esta parte no estaba considerada en el trabajo original, por lo que solo representaba una posible oportunidad a aprovechar y no un objetivo sin cumplir.

Para concluir el trabajo, se diseñaron e implementaron dos interfaces destinadas a mostrar la información generada por el modelo a los equipos docentes, académicos, investigadores y diversos actores involucrados en la administración de las mallas curriculares de la FCFM. Estas fueron validadas con una profesora coordinadora de ramos de Plan Común y un profesor de ramos de Plan Común y del Departamento de Ciencias de la Computación, ambos posibles usuarios finales del sistema, y obtuvieron una buena recepción, a pesar de recibir ciertas sugerencias de mejoras que se pueden realizar en el futuro.

Todo el proceso detallado de generación de los datos utilizados, entrenamiento de los algoritmos y desarrollo de las interfaces se puede encontrar en el capítulo 4, donde se van explicando todas las decisiones tomadas y pasos realizados por el autor para llegar a producir el modelo final. El capítulo 5 explica de forma más detallada del proceso de validación que fue realizado y que permite considerar a este proyecto como exitoso.

Conclusiones y trabajo futuro

El último capítulo de este informe hace un pequeño resumen de todo el trabajo realizado y los resultados obtenidos, concluyendo que se lograron cumplir satisfactoriamente gran parte de los objetivos establecidos. Por otro lado, también deja planteados diversos desarrollos que pueden ser realizados como trabajo futuro, tales como la incorporación de nuevas variables al modelo de datos, la modificación de las heurísticas implementadas, la consideración del contexto social, la incorporación de mejoras a las interfaces creadas y otros posibles trabajos que pueden ser realizados en paralelo pero que, sin lugar a dudas, podrían beneficiar a futuras iteraciones de este proyecto.

Capítulo 1

Marco teórico

En el capítulo a seguir se procederá a describir los conceptos importantes utilizados a lo largo de este informe como *Machine Learning* y algunos algoritmos como la regresión logística o *Support Vector Machine*. Además de esto se describen las herramientas que fueron utilizadas para el desarrollo de este trabajo como Jupyter y librerías como Numpy. Esto se hace para dar un poco más de contexto a lectores poco familiarizados con el área en que se enmarca el proyecto y el proceso que se explica en los capítulos a seguir.

1.1. Machine Learning

El Machine Learning, o Aprendizaje de Máquinas, corresponde al estudio científico de algoritmos y modelos estadísticos que son utilizados por sistemas computacionales, con el objetivo de realizar una tarea determinada de manera efectiva sin recibir instrucciones explícitas y haciendo uso, en cambio, de patrones e inferencias a partir de los datos utilizados, bautizados entonces como "datos de entrenamiento" [24].

Los algoritmos en torno al Aprendizaje de Máquinas son utilizados en una gran variedad de aplicaciones, donde no es posible generar un algoritmo preciso que pueda indicarle qué hacer a la máquina para completar la tarea, bajo cada posible escenario al que se pueda enfrentar. Algunas aplicaciones de esto son los filtros de correos electrónicos no deseados o el procesamiento de imágenes, ya que nunca se podrán abarcar exactamente todos los valores posibles a los que se podrían enfrentar los algoritmos durante su funcionamiento.

1.1.1. Algoritmos de Aprendizaje Supervisado

Existen muchas categorías y formas de realizar Aprendizaje de Máquinas, pero la más relevante para este trabajo corresponde a los algoritmos de aprendizaje supervisado. Estos se caracterizan por realizar la tarea de generar una función que mapee entre información de entrada y salida, basado en pares entrada-salida utilizados como ejemplo [23]. En el apren-

dizaje supervisado, cada ejemplo utilizado consiste en un par compuesto por información de entrada (usualmente un vector de datos) y el valor deseado de salida. Dichos algoritmos luego generan una función inferida, con la cual pueden crear, idealmente, un valor de salida para valores de entrada previamente desconocidos, por lo que requieren ser capaces de generalizar a partir de los datos de ejemplo nuevos valores de salida para situaciones que no se había observado antes de manera "razonable" [18].

Una de las labores más comunes en las que se aplica el Aprendizaje de Máquinas y el aprendizaje supervisado corresponde a la clasificación de elementos, la cual consiste en tratar de predecir con cierta certeza la clase a la que pertenece un elemento conociendo otras características del mismo. Por ejemplo, un algoritmo de este tipo podría decir que un animal de 45 kg de masa y 1.3 m de largo, que posee una cola y cuatro patas corresponde a un perro, si es que fue entrenado con la información suficiente para poder llegar a esa conclusión. De manera un poco más abstracta, un algoritmo clasificador también podría tratar de predecir si una persona va a pagar un préstamo o no basado en el comportamiento financiero de este comparado con el de otras personas.

La mayoría de los algoritmos de clasificación pueden dar como valor de salida tanto una clase en específico, como la probabilidad de pertenencia a cada clase. En los clasificadores binarios, se puede utilizar la probabilidad de pertenencia para establecer un "umbral de predicción" distinto al normal (50 %), a partir del cuál el sistema predice una clase sobre la otra.

Uno de los problemas más frecuentes y difícil de evitar en el entrenamiento y validación de algoritmos de aprendizaje supervisado es el sobreajuste en los datos. Dado que el proceso de preparación de un algoritmo de aprendizaje supervisado implica entregarle datos que utilizará como ejemplos para la clasificación, existen un gran número de causas que pueden llevar al modelo a ajustarse demasiado a la distribución de dichos ejemplos, con lo que dejaría de funcionar correctamente al tratar de predecir la clase de nuevos elementos que se diferencien de los datos utilizados como ejemplo.

Regresión Logística

Este consiste en un algoritmo de regresión simple que utiliza una función de regresión logística para modelar una variable dependiente binaria (con valores posibles 0 y 1, bautizados como "clase 0" y "clase 1" respectivamente). El modelamiento se logra realizando una combinación lineal de las variables independientes utilizadas, ponderadas por determinados coeficientes, como se señala en la ecuación 1.1, para obtener un valor llamado "log-odds" (el logaritmo de las probabilidades), que luego se procesa con una función logística predeterminada transformando el valor obtenido en la probabilidad de pertenecer a la clase 1. Los coeficientes utilizados para ponderar cada una de las variables independientes son usualmente obtenidos mediante el entrenamiento del modelo con datos de prueba.

$$log - o\text{dd}s = \beta_0 + \beta_1 * X_1 + ... + \beta_n * X_n \tag{1.1}$$

Redes Neuronales

Los algoritmos de redes neuronales corresponden a modelos computacionales de lenguaje supervisado para la clasificación, cuyo funcionamiento está levemente basado en el comportamiento observado en su homólogo biológico [3]. Estos algoritmos se componen de diversos nodos, llamados neuronas artificiales, que se organizan en varias "capas" de diferente cantidad de elementos, conectadas entre sí por enlaces de neurona a neurona por donde se propaga la información hasta generar un valor de salida con la clase elegida.

La información del algoritmo llega por la capa de entrada, compuesta por un nodo por cada valor en el vector de entrada, para luego pasar a la siguiente a través de los enlaces entre las neuronas de las distintas capas. Cada neurona está conectada con neuronas de la capa siguiente a través de unos enlaces. En estos enlaces el valor de salida de la neurona anterior es multiplicado por un "peso". Dichos pesos en los enlaces pueden "potenciar" o "inhibir" el estado de activación de la neuronas siguiente. Del mismo modo, a la salida de la neurona, puede existir una función limitadora o umbral, que modifica el valor obtenido o impone un límite que se debe sobrepasar antes de propagarse a otra neurona. Esta función se conoce como función de activación.

Las capas que componen una red neuronal se separan en tres categorías: capa de entrada (la primera que recibe el input), capa de salida (la que entrega el output) y capas ocultas (todas las capas que están entre la de entrada y salida). Si todos los nodos tienen un enlace a cada uno de las nodos de las capas adyacentes se dicen que la red es "fully connected".

El funcionamiento de una red neuronal se puede ajustar modificando parámetros como la función de activación utilizada, la cantidad de capas ocultas y la cantidad de nodos que las constituyen, la taza de aprendizaje con la que son entrenadas y otros valores varios de su funcionamiento interno. El valor de salida corresponde a la clase predicha o a un probabilidad de pertenecer a cada clase.

Support Vector Machine

Las máquinas de vectores de soporte (SVM por sus siglas en ingles) corresponden a otro tipo de algoritmo de aprendizaje supervisado para la clasificación de elementos. Dado un conjunto de datos de entrenamiento, cada uno etiquetado como perteneciente a una de dos categorías, una SVM corresponde a la representación de estos en el espacio, mapeados de tal forma que exista un hiper plano que genere la mayor separación posible entre las clases. Al recibir nuevos ejemplos, la SVM mapea su ubicación en el espacio y determina su clase dependiendo del lado del hiper plano en que quede posicionado [10].

Para realizar esta separación en conjuntos no triviales, los algoritmos de SVM utilizan un "truco del kernel", mediante el cual aumentan la dimensionalidad de los datos, utilizando una determinada función (o kernel) que recibe la información del vector original y genera nuevos valores, con la esperanza de generar una mejor división entre las clases. El hiper plano es determinado al tratar de maximizar la distancia a los elementos de cada clase más cercanos a este (bautizados como vectores de soporte).

Al igual que con las Redes Neuronales, existen varios parámetros de SVM que pueden ser modificados arbitrariamente para tratar de mejorar su rendimiento, tales como el umbral de tolerancia (C) y la tasa de aprendizaje (Gamma) o el kernel utilizado.

Al utilizar ciertas variaciones a la hora de determinar la salida del algoritmo, este se puede utilizar para generar un análisis de regresión y obtener más valores fuera de solo clases. Cuando se emplea de esta manera, la SVM pasa a llamarse SVR (Support Vector Regression)

1.1.2. Métricas de rendimiento

Al evaluar el rendimiento de un algoritmo de clasificación, se pueden utilizar distintas métricas derivadas de los resultados clasificados correcta o incorrectamente. Las descritas a continuación son algunas de estas, enfocadas en el análisis del rendimiento en clasificadores binarios, es decir, que solo poseen dos posibles respuestas a las que se refiere como positiva o negativa, eligiendo la etiqueta de cada clase de manera arbitraria según sea conveniente (e.g. referirse como positivo a los elementos clasificados como "enfermos" y negativo a los "sanos" según un clasificador para encontrar la presencia de una enfermedad en humanos). Estas métricas se pueden extender para abarcar casos de clasificación multiclase, pero no son el foco de este proyecto y, por lo tanto, no serán explicadas en este documento.

- *False Negatives* (FN): Corresponde a los elementos cuya clase era positiva, pero fueron clasificados como negativos.
- *False Positives* (FP): Corresponde a los elementos cuya clase era negativa, pero fueron clasificados como positivos.
- *True Negatives* (TN): Corresponde a los elementos cuya clase era negativa y fueron clasificados correctamente.
- *True Positives* (TP): Corresponde a los elementos cuya clase era positiva y fueron clasificados correctamente.
- *Recall*: Métrica compuesta que se calcula como la tasa de clasificaciones TP entre el total de elementos con clase positiva ($\frac{TP}{TP+FN}$). Esta métrica, también llamada como sensitividad o ratio de verdaderos positivos, da una idea de la cantidad de elementos positivos que fueron correctamente clasificados.
- *Precision*: Métrica compuesta que se calcula como la tasa de clasificaciones TP entre el total de elementos clasificados como positivos ($\frac{TP}{TP+FP}$). Esta métrica, también llamada especificidad o valor predictivo positivo (PPV por sus sigla en inglés), da una idea de la cantidad de elementos clasificados como positivos que realmente lo eran.

Matriz de Confusión

La matriz de confusión corresponde a una herramienta que permite la visualización del desempeño de un algoritmo al colocar de manera matricial los valores de TP, TN, FP y FN. Esto permite analizarlos todos a la vez y tener una pista del rendimiento del algorit-

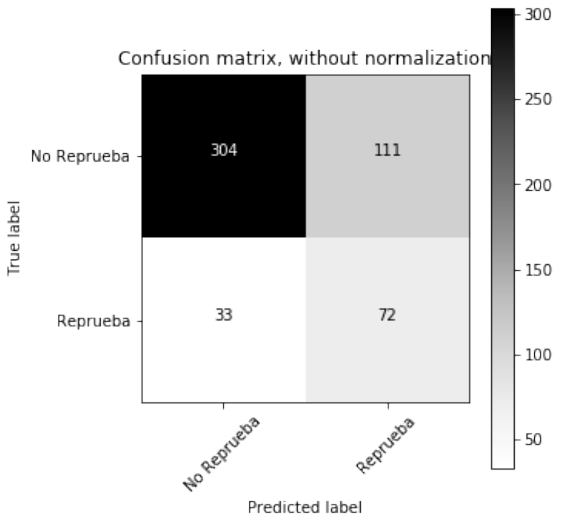

Figura 1.1: Matriz de confusión sin normalizar.

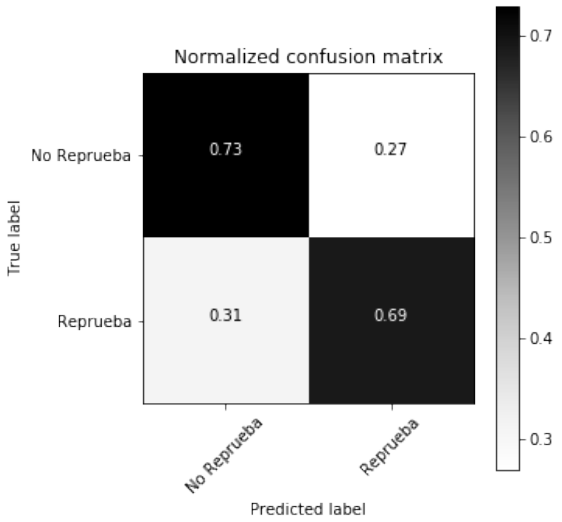

Figura 1.2: Matriz de confusión normalizada.

mo fácilmente, muchas veces mezclando los números con un mapa de calor para un mejor entendimiento.

En las figuras 1.1 y 1.2 se pueden observar una matriz de confusión sin normalizar y normalizada, respectivamente. En la primera se muestra la cantidad de elementos que pertenecen a cada grupo de predicción y su etiqueta real, el mapa de calor entonces muestra más oscuro el grupo donde hubo un mayor número de predicciones asociadas. Para la versión normalizada, en cada cuadro se muestra el porcentaje de predicciones con respecto a la etiqueta real, por lo que el mapa de calor pasa a señalar de qué manera se distribuyen los elementos de cada clase con respecto al total de elementos en dicha clase.

Curva ROC

Corresponde a una representación gráfica de la sensibilidad (Recall) frente a la especificidad (Precision) para un sistema de clasificación binaria según se varía el umbral de clasificación. Esta permite analizar y encontrar visualmente un estimado del umbral de clasificación óptimo para un determinado algoritmo, de modo que se pueda mantener la especificidad y sensitividad lo más altas posibles sin que ninguna sufra una caída muy fuerte.

La curva ROC se crea iterando por distintos valores del umbral de decisión sobre un conjunto de datos de entrenamiento y graficando, para cada valor del umbral, los valores de TP versus FP. Esto sirve, por una parte, para validar superficialmente el rendimiento del algoritmo: mientras más se aleje de la diagonal indica una mayor posibilidad de obtener buena sensitividad sin sacrificar demasiada especificidad (o viceversa), mientras que más cercano a la diagonal implica un comportamiento más parecido a un algoritmo aleatorio. Por otro lado, esta curva también sirve para encontrar el umbral óptimo (o un aproximado), correspondiente al punto de inflexión en que la curva deja de crecer tanto y se comienza a estancar, ya que

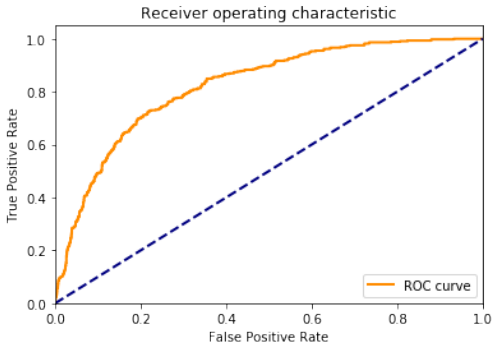

Figura 1.3: Curva ROC de ejemplo.

dicho punto marca el punto en que se deja de ganar mucha sensitividad a riesgo de perder especificidad. La figura 1.3 muestra una curva ROC de ejemplo con resultados mejores a un clasificador aleatorio al tener la curvatura hacia la izquierda y con un punto de inflexión cercano al 20 % de falsos positivos y 70 % de verdaderos positivos.

Validación cruzada

La técnica de validación cruzada se utiliza para obtener métricas del rendimiento de un algoritmo clasificador pero tratando de eliminar en la mayor medida el sobreajuste sobre los datos de entrenamiento. Este consiste en dividir el conjunto de datos en k grupos al azar de igual (o similar) número de observaciones y luego realizar k iteraciones, donde en cada una se elige un grupo distinto para dejar fuera del conjunto de entrenamiento, se entrena el modelo con los $k-1$ grupos restantes y luego se valida con el grupo extraído al inicio de la iteración, obteniendo las métricas deseadas para esa iteración. Una vez se han realizado k operaciones, dejando una vez a cada grupo como conjunto de validación, se promedian las métricas obtenidas en cada oportunidad para obtener los valores finales del algoritmo como tal.

La idea de este algoritmo de validación es que, al ir sacando un grupo de datos distinto en cada iteración, se puede disminuir el efecto del sobreajuste, ya que nunca será el mismo conjunto de datos sobre el que se van a estar haciendo las pruebas en cada vez.

1.1.3. Data Mining

La minería de datos es el proceso de descubrir útiles e interesantes patrones y relaciones en grandes volúmenes de información. Este campo de investigación combina herramientas de estadísticas e inteligencia computacional con el manejo de bases de datos para analizar grandes colecciones, conocidas como “data sets” [4].

Más allá del simple análisis, la minería de datos también involucra aspectos del manejo de bases de datos como el pre-procesamiento de la información, consideraciones en la inferencia y el modelamiento, la obtención de métricas de interés, consideraciones de complejidad, post

procesamiento de las estructuras encontradas, visualización de los datos y actualizaciones en línea. La diferencia entre el análisis de los datos y la minería de los datos radica en que el análisis de los datos se utiliza para probar modelos e hipótesis en un cierto conjunto de datos, mientras que la minería de datos utiliza aprendizaje de máquinas y modelos estadísticos para descubrir patrones ocultos o clandestinos en grandes cantidades de datos [19].

1.2. Tecnologías utilizadas

La principal herramienta de desarrollo utilizada durante esta memoria fue el lenguaje Python en su versión 3.6. Para agilizar el desarrollo se utilizó una distribución es específico, llamada Anaconda, que viene precargada con una gran cantidad de librerías orientadas a la investigación en la ciencia de datos.

1.2.1. Anaconda

"Anaconda es un administrador de paquetes, un administrador de ambientes, una distribución centrada en *Data Science* de Python/R y una colección de 1500+ paquetes de código abierto" [2]. Tal como mencionan en su página web, la distribución Anaconda está diseñada para facilitar el desarrollo centrado en el manejo de datos utilizando Python o R, otorgando fácil acceso a herramientas y librerías que ayudan en el proceso de prueba y error involucrado en la investigación de datos, la visualización de la información para mejor comprensión, el desarrollo y entrenamiento de modelos de aprendizaje de máquinas, entre otras varias tareas involucradas en el trabajo de un *Data Scientist* [1]. Esta fue la distribución utilizada en el desarrollo de gran parte del proyecto debido a la utilidad de las herramientas que vienen incluidas.

Jupyter Notebook

El Jupyter Notebook consiste en una aplicación web Open Source que permite crear y compartir documentos que contienen código en vivo, ecuaciones, visualizaciones y texto narrativo. Sus usos incluyen: limpieza y transformación de datos, simulaciones numéricas, modelamiendo estadístico, visualización de datos, aprendizaje de máquinas, entre otros [7]. Esta herramienta viene integrada con la distribución Anaconda y permite generar bloques de código que se pueden ejecutar de manera independiente, almacenando sus resultados y los valores de variables utilizadas, de modo que puedan ser invocadas en el futuro. Esto permite ir analizando rápidamente los efectos de pequeños trozos de código, acelerando el desarrollo al no tener que esperar a que se ejecuten constantemente secciones anteriores cuyo resultado ya es conocido, para en cambio enfocarse en el análisis de otras partes más relevantes.

1.2.2. Librerías utilizadas

NumPy

NumPy es una de las librerías más populares de Python para el procesamiento y manipulación de datos numéricos. Esta permite realizar operaciones con matrices multi dimensionales fácil y eficientemente junto a muchas de las operaciones más comunes en el uso de matrices, como la suma, multiplicación, producto punto, entre otras [5].

Scikit-learn

Scikit-learn consiste en una librería para Python enfocada en la fácil utilización e implementación de herramientas y modelos de aprendizaje de máquinas, entregando implementaciones de algoritmos populares con ciertos grados de libertad y personalización, listos para recibir los datos de entrenamiento y empezar a realizar clasificaciones en pocas líneas de código [9][21].

Pandas

Pandas es otra librería de código abierto para Python que provee estructuras de datos y herramientas de análisis de datos fáciles de usar y de alto rendimiento. Esta contiene algoritmos que facilitan la manipulación de datos, interactuando con distintos motores de bases de datos para intercambiar información, entregando funciones para agrupar, filtrar y transformar los datos fácil y eficientemente [6].

Capítulo 2

Situación Actual

En el capítulo a seguir, se explicará en más detalle el contexto alrededor del proyecto a realizar. Esto consiste en la descripción de las áreas de estudio en las cuáles se enmarca el trabajo realizado, seguido de una sección relatando la importancia de la reducción de la deserción estudiantil y como el trabajo realizado puede ayudar a esto. Por último, el capítulo termina explicando en detalle el modelo de Celis et al. [22] que sirve como la base para todo este desarrollo, el cual se utilizará como el principal referente a la hora de crear heurísticas y validar el rendimiento de los algoritmos implementados.

2.1. Áreas de exploración

En los últimos años el volumen de datos disponibles acerca de la población mundial ha aumentado de manera vertiginosa. El mundo de la educación no se ha quedado atrás, generando constantemente información sobre los millones de estudiantes pertenecientes a las miles de instituciones a nivel mundial [15], debido al uso cada vez más frecuente de plataformas computacionales, destinadas a la gestión y administración de las diversas actividades involucradas con el mantenimiento de una institución educativa.

Gracias al aumento en la información disponible y al deseo de estudiar y mejorar todos los aspectos relacionados con la docencia, nacieron en el mundo anglosajón diversas comunidades en torno a *Learning Analytics (LA)* y *Educational Data Mining (EDM)*. Ambos mundos hacen uso del análisis de datos utilizando algoritmos computacionales con el objetivo de mejorar los sistemas de evaluación, el entendimiento de los procesos educativos y la priorización y diseño de intervenciones educativas [14].

La diferencia entre estas comunidades radica principalmente en la metodología aplicada y los focos de investigación, donde el EDM se centra en el descubrimiento automatizado de patrones con poca intervención de juicio experto, mientras que LA busca fortalecer el juicio experto y testea hipótesis educacionales con la ayuda de modelos de descubrimiento automático. A pesar de lo anterior, las definiciones de ambos campos se mezclan constantemente, por lo que de aquí en adelante se utilizará el término *Learning Analytics* para referirse a

ambos indistintamente.

En Chile, una de las instituciones de educación superior más importante en el país es la Universidad de Chile, la cual alberga a miles de estudiantes cada año en sus distintas carreras de pre y postgrado [13]. Para administrar este gran número de estudiantes, en la universidad se hace uso de las plataformas creadas por el equipo del Centro Tecnológico Ucampus (Ucampus de aquí en adelante) para casi todo el manejo de la información pertinente a sus alumnos, docentes, funcionarios y administración en general.

Tanto para Ucampus, la Universidad de Chile y otras entidades similares, los campos de LA y EDM son de particular utilidad, ya que el uso de nuevas tecnologías permite realizar diversos estudios y aplicar medidas que ayuden a mejorar las labores de docencia impartidas en estas instituciones. En un mundo ideal, esto se traduce en mejores resultados en el desempeño de sus estudiantes, lo que los transforma en mejores profesionales, aumentando el prestigio de la casa de estudios y el valor que le pueden entregar al país cada uno de ellos.

2.2. Retención y rendimiento en educación superior

Dentro de los temas posibles de explorar haciendo uso de LA y EDM, está la predicción del rendimiento académico de los diversos estudiantes que ingresan cada año a la educación superior. Esto sirve para poder emplear diversas técnicas o metodologías de apoyo para alumnos con problemas de rendimiento o para realizar ajustes a la dificultad de las evaluaciones, por ejemplo.

Tal como comentan Celis et al. en su trabajo [22], la deserción estudiantil en educación superior se ha vuelto un tema bastante discutido en las últimas décadas, tanto a nivel de instituciones como de gobiernos, debido al impacto negativo que tiene en la educación superior. Dicho impacto se puede ver reflejado en aumentos en los aranceles y en cómo se afecta la imagen de quienes desertan, debido a la importancia social que hoy ha adquirido la educación superior como una instancia clave para el desarrollo personal, social, económico y cultural. En el contexto nacional, las cifras de deserción se estiman cercanas al 40 % al tercer año de estudios, aunque con una gran variabilidad según el tipo de institución (universitaria, centro de formación técnica, institutos profesionales) y áreas disciplinarias [27]. También, en promedio, solo un 65 % de los estudiantes del país permanece en su programa luego del primer año [25]. Por otro lado, si bien se ha mantenido una mejora en la retención estudiantil a lo largo de los años, de todas maneras se presentan periodos con comportamientos erráticos en donde aumenta la deserción por diversos factores, esto se observa en un aumento de la retención entre 2007 y 2010, desde el 67 % a un 71 %, pero con una disminución al 69 % luego en 2013 [26].

Por supuesto, la deserción no es solo un tema relevante a nivel nacional, si no que en diversas partes del mundo han propuesto diversas teorías y seleccionado distintas variables con el fin de explicar la deserción estudiantil. Dentro de los factores más explorados, comunes y más estadísticamente significativas se encuentran las características individuales de preingreso, tales como el estatus socioeconómico, la habilidad académica, grado de motivación

y expectativas de logro, considerándolos más importantes que características institucionales a la hora de evaluar persistencia y deserción [20]. Sin embargo de lo anterior, existen factores controlables, o que pueden ser incentivados, por las instituciones con un impacto observable sobre el abandono de los estudios. Entre estos se encuentran la selectividad de las instituciones, integración al campus, participación en actividades extra curriculares, actividades de introducción a nuevos estudiantes, becas para personas de menores ingresos, interacción con docentes fuera del ámbito de clases y la interacción entre pares.

En el ámbito nacional también se han realizado varias investigaciones enfocadas en descubrir las causas de la deserción, pero con un enfoque más localizado y específico al contexto que se da en el país. Acuña [11] y Larroucau [16], basándose en datos de la educación superior y universitaria disponible en el país concluyeron que el fenómeno se da por diferentes causas y que las variables analizadas en el ámbito internacional también tienen coherencia al concentrarse en el contexto local. En particular, Larroucau determinó que variables como el establecimiento de origen, el promedio de notas y el ranking en educación secundaria eran mejores predictores de la deserción que el puntaje de la Prueba Nacional de Selección Universitaria (PSU), comprobando que las variables de pre ingreso poseían un gran peso a la hora de determinar la permanencia en las diversas instituciones.

Según un estudio realizado mediante una encuesta por el Centro de Microdatos de la Universidad de Chile [12], las principales causas de la deserción universitaria en el primer año son problemas vocacionales (e.g., no quedar en la carrera elegida como primera opción), la situación socioeconómica del estudiante y el rendimiento académico dentro de la carrera. De estos tres fenómenos, el único donde parecería tener un impacto directo las acciones tomadas por las instituciones corresponde al último punto, ya que puede ofrecer sistemas de tutorías personalizados para personas con bajo rendimiento, ajustar la dificultad de la malla curricular o las evaluaciones, crear programas de ayuda colectiva, etcétera. Sin embargo, todo esto solo es de utilidad si es que dichos problemas son diagnosticados a tiempo, de modo que se puedan ejecutar los planes que se consideren apropiados y/o necesarios, con el objetivo de disminuir la deserción estudiantil

2.3. El modelo original

En 2015, Celis et al. [22] realizaron en conjunto con Ucampus un modelo analítico (enmarcado en LA) capaz de predecir el rendimiento académico de estudiantes de primer año de la Facultad de Ciencias Físicas y Matemáticas (FCFM) de la Universidad de Chile, con el objetivo de identificar a estudiantes de segundo semestre que hubieran reprobado al menos una asignatura y que tuvieran una alta probabilidad de reprobar un mismo ramo por segunda vez. Esto, debido a que las normas creadas por la FCFM establecen el reprobar dos veces un mismo curso como una causal de eliminación, lo que, como se discutió anteriormente, puede derivar en un aumento en la probabilidad de deserción por parte del alumno. Este estudio también se realizó con el objetivo de poder crear medidas a adoptar por la FCFM con el fin de ayudar a estos casos de alumnos en peligro de reprobar, de modo que se viera disminuida la tasa de deserción en la facultad (a pesar de ser cercana a un 5 %, lo que se considera baja en comparación con el resto del país [25]).

Teniendo el conjunto de datos, el modelo en concreto que se desarrolló consistió en un modelo de regresión logística en combinación con una metodología de selección de atributos, tomando esta elección debido a, en sus propias palabras, "la simplicidad de interpretación y utilización ampliamente aceptada" de estas herramientas.

La selección de atributos se llevó a cabo mezclando una técnica de *Forward Feature Selection* (FS), la que consiste en ir agregando variables de a uno al modelo, evaluar su evolución y dejarla solo si implica una mejora al estado anterior, con una técnica de *Backward Extraction* (BE), mediante la cual se van quitando variables que sean redundantes o empeoren la clasificación cada vez que se agrega una nueva variable. Además de lo anterior, debido a la reducida cantidad de datos en comparación con la inmensa cantidad de atributos puros y compuestos disponibles (i.e., información tal cual se encuentra en la base de datos versus comparaciones u operaciones entre atributos y/o constantes), se realizó una selección por frecuencias, lo que consistía en realizar muchas veces el proceso de obtención de atributos mediante FS y BE y quedarse con el conjunto de variables que fuera escogido más veces. Todo el proceso recién descrito dejó a Celis et al. con el siguiente conjunto de variables independientes a utilizar en su modelo:

1. El género - variable binaria
2. Tipo de establecimiento de enseñanza media - separado en dos variables binarias, particular y particular subvencionado, considerando el que ambas sean falsas como que viene de un colegio municipal
3. Ratio de créditos reprobados el semestre anterior - variable continua
4. Comparación entre variables: Promedio controles 1 del segundo semestre (C1IRS2) *menor* al promedio final del primer semestre en cursos reprobados (FIRS1) - etiqueta: C1IRS2<FIRS1, tipo: variable binaria
5. Diferencia con nota 4.0 del promedio actual de "controles 1" en cursos reprobados en segundo semestre - etiqueta: C1IRS2 - 4.0, variable continua
6. Comparación entre variables: Promedio controles 1 ramos no reprobados del segundo semestre (C1INRS2) *menor* al promedio final del primer semestre en cursos no reprobados (C1INRS1) - etiqueta: C1INRS2<C1INRS1, variable binaria
7. Comparación entre variables: El peor promedio "controles 1" de ramos no reprobados del 2do semestre (minC1INRS2) *menor* que el peor promedio final del 1er semestre en cursos no reprobados (C1INRS1) - etiqueta: minC1INRS2<C1INRS1, variable binaria

Como se puede notar en el listado anterior, quedaron seleccionadas dos variables puras (1 y 2), dos variables compuestas continuas (3 y 5) y el resto (4, 6 y 7) corresponde a variables compuestas binarias con comparaciones entre valores, donde se marca un 1 en cada una si se cumple la comparación para un estudiante en específico y 0 en el caso contrario.

Utilizando estas variables y el conjunto de datos antes descrito, se procedió a entrenar el modelo utilizando el conjunto de estudiantes que ingresaron entre 2010 y 2013, dejando a la cohorte del 2014 aparte para poder validar el poder de predicción del algoritmo. Los coeficientes obtenidos para cada variable independiente se pueden observar en la tabla 2.1, dentro de los cuales se identificó que los más relevantes en la decisión tomada por el predictor son los señalados por un asterisco (*).

Variable	Coeficiente
Género (hombre)	0.63*
colegio particular	-1.63
colegio subvencionado	-2.24
ratio créditos reprobados	4.41*
C1IRS2 <FIRS1	0.13
C1IRS2 - 4.0	-0.38*
C1INRS2 <C1INRS1	0.19
minC1INRS2 <C1INRS1	0.53

Tabla 2.1: Coeficientes obtenidos para el modelo de regresión logística de Celis et al. (extracto de [22], página 18, Tabla 2: Resultado de Regresión Logística: Doble Reprobación en Primer Año (n=830))

Por último, dado que el modelo entrega un valor entre 0 y 1 como la probabilidad de que un alumno vaya o no a tener doble reprobación (siendo 0 que no reprueba y 1 que sí lo hace), se debía escoger un umbral a partir del cual se podía interpretar un resultado como 0 o 1. Para esto, Celis et al. escogieron dicho umbral empíricamente como aquel valor donde se interceptan las curvas de sensitividad y especificidad (i.e., donde se optimiza la correcta clasificación de casos positivos y negativos), en este caso 19 %.

Según los autores, el modelo creado logró resultados bastante satisfactorios, obteniendo un recall del 86 %, lo que se traduce en que el 86 % de los alumnos que terminaron reprobando un ramo por segunda vez fueron identificados correctamente en la cohorte del 2014. Por otro lado, los resultados no fueron del todo positivos, puesto que también se obtuvo una precisión de tan solo el 37,5 %, lo que implica que cerca de dos tercios de los alumnos diagnosticados con una mayor probabilidad de reprobar en realidad terminaron aprobando sus asignaturas. Dentro de las conclusiones expuestas por Celis et al. descartan la baja precisión como un factor a preocuparse, ya que declaran que "el número de falsos positivos es tolerable para el tipo de intervenciones y decisiones a tomar en base a los resultados del modelo", por lo que los resultados se pueden considerar un éxito moderado de todas maneras. En la figura 2.1 se puede observar una representación gráfica de los resultados obtenidos por el modelo predictivo de Celis et al.

Producto de lo anterior, este modelo abre la posibilidad de ser extendido, de modo que los datos utilizados para su funcionamiento puedan abarcar un público objetivo aún más general que el estudiado en esa ocasión siempre y cuando los resultados obtenidos por el modelo nuevo pueda mantener o, mejor aun, mejorar los porcentajes de recall y precisión logrados por el modelo original. Por este motivo, el trabajo a realizar en esta memoria, en conjunto con Ucampus, consistirá en adaptar y escalar el modelo ya descrito, de modo que se pueda aplicar a cualquier alumno de pre grado en alguna de las carreras de ingeniería de la FCFM, independiente del año en que se encuentre y los ramos que posea o que haya o no reprobado en el pasado. Además de lo anterior, se creará una plataforma web, integrada a las ya ofrecidas por Ucampus, donde se entregue esta información procesada para que pueda ser utilizada por funcionarios, académicos, investigadores y docentes pertenecientes a la comunidad de la FCFM con el fin de ayudar en la mejora de los apoyos y planes educacionales ofrecidos por la institución.

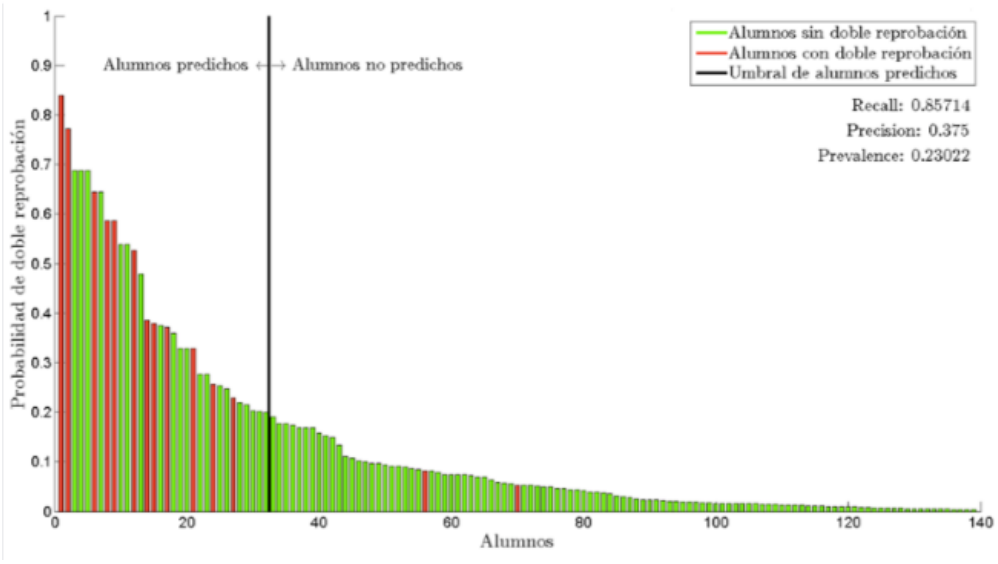

Figura 2.1: El poder predictivo del modelo (Celis et al. [22])

El sistema a realizar podría ser una adición relevante para gran parte de los usuarios de las plataformas de Ucampus, ya sean alumnos, docentes, académicos o investigadores de la FCFM, ya que permitirá tomar medidas a tiempo para evitar la reprobación de estudiantes identificados con un mayor riesgo a esto, entregándoles el apoyo necesario para salir adelante o generar medidas a nivel institucional para disminuir la reprobación y deserción de estudiantes. Esto generará un impacto positivo en la vida de los estudiantes, puesto que – de tomar acciones según la información entregada – podría evitar la carga emocional que significa reprobar un ramo y el tiempo que estos se demoran en terminar la carrera. Del mismo modo, la herramienta también podría ser útil para el equipo docente de cada ramo, mostrando información relevante con respecto a sus alumnos, para que puedan realizar los ajustes necesarios al curso y así mejorar el desempeño general. Por último, las instituciones mismas se verían beneficiadas por la existencia de este sistema, ya que menos alumnos entrarían en causal de eliminación y por tanto disminuiría el porcentaje de deserción en general.

Una de las mayores limitaciones del modelo creado por Celis et al. es que solo abarca un pequeño universo de los estudiantes, compuesto por personas en el segundo semestre de su primer año y que hubieran reprobado al menos un ramo en el primer semestre. Este universo se traduce a entre 195 a 255 estudiantes de la cohorte de cada año (según se informa en su publicación), un número bastante pequeño considerando que cada año entran más de 800 personas tan solo a la FCFM y que, en 2017, se registraron 343.703 matrículas de primer año y 1.162.306 matrículas totales a lo largo de todas las instituciones de educación superior del país [13]. Con el objetivo de aumentar el universo objetivo, pero manteniendo la dificultad del proyecto acotada, se decidió tratar de incluir en este modelo al resto de los estudiantes de ingeniería de la FCFM y dejar para el futuro la opción de seguir ampliando hasta abarcar el resto de las disciplinas impartidas en la Universidad de Chile y el país.

El motivo por el cual se prefirió extender el modelo dentro de la FCFM en vez de, por ejemplo, tratar de aplicarlo a alumnos de primer año de otras carreras, se debe a que los estudiantes de la FCFM poseen grandes similitudes entre sí en el ámbito socio-económico

y de desempeño previo al ingreso a la institución (considerando que todos los alumnos de esta facultad pertenecen al 3 % con mejor rendimiento a nivel país en la PSU). Tal como se discutió en la sección anterior, dentro de los factores que alteran el porcentaje de deserción se encuentran las características del alumno previo al ingreso y el área de la disciplina elegida, por lo que resulta más fácil tratar de modelar a todos los alumnos de ingeniería en la FCFM independiente de la generación, que tratar de hacer un modelo adaptable a distintas realidades, presentes en otro tipo de disciplinas o instituciones con características distintas a las observadas. Lo anterior no quita, por supuesto, la posibilidad de seguir extendiendo el modelo como trabajo futuro, pero sí establece el alcance buscado por el trabajo a realizar.

2.4. Sistemas de Gestión del Aprendizaje en la FCFM

La FCFM está compuesta por más de 5.000 estudiantes de pregrado, donde todos aquellos estudiantes que ingresan a la facultad primero pasan por un Plan Común de dos años de duración, después del cuál se ven en la libertad de elegir seguir por una de las múltiples líneas de especialización que están disponibles: Astronomía, Ciencias de la Computación, Física, Geofísica, Geología, Ingeniería Civil, Ingeniería de Minas, Ingeniería Eléctrica, Ingeniería Industrial, Ingeniería Matemática, Ingeniería Mecánica e Ingeniería Química, Biotecnología y Materiales. Además de los alumnos de pregrado, existen más de 1.000 alumnos pertenecientes a planes de postgrado en diversas disciplinas.

Para la gestión del gran volumen de estudiantes concurrentes en la facultad, estos alumnos utilizan de manera habitual una plataforma catalogada como "Sistema de Gestión del Aprendizaje" (LMS por sus siglas en ingles) llamada U-Cursos[1], creada por el Centro Tecnológico Ucampus. Esta tiene funcionalidades enfocadas en sociabilizar, entregar información de manera más expedita y facilitar la comunicación entre los integrantes de la FCFM. Por otro lado, también se hace uso de la plataforma Ucampus[2], la cual funciona como un sistema de gestión curricular, con su símil más cercano siendo los sistemas de ERP del mundo financiero. Esta última está enfocada en el manejo de los recursos administrativos y curriculares de una institución de educación superior tales como la toma de ramos, entrega de varios documentos y certificados, solicitudes de cambio de carrera, consultas del estado de avance curricular en el plan, entre otras facilidades.

La plataforma U-Cursos, en particular, cuenta con información bastante interesante desde el punto de vista de LA y EDM, ya que es una plataforma enfocada en el uso frecuente para la conversación, la entrega de información rápida y la sociabilización entre los distintos integrantes de la comunidad de la FCFM. Para los estudiantes y docentes, esta cuenta con una página dedicada a cada uno de los cursos en los cuales forman parte, además de una página a nivel de institución y otras dedicadas a comunidades (opcionales) de actividades extra curriculares o departamentos de especialidades. En cada una de estas páginas se pueden encontrar distintos módulos donde pueden interactuar los integrantes como un foro, páginas para la carga y descarga de material docente digital, información de notas parciales, evaluaciones en línea, registro de asistencias, enlaces de interés, novedades, afiches, etcétera. Todos

[1]URL: https://www.ucursos.cl
[2]URL: https://www.ucampus.uchile.cl

estos módulos son usados por los estudiantes, ya sea diariamente o de manera ocasional por eventos puntuales, y toda interacción que sea realizada con el sitio queda guardada en los registros de Ucampus, generando varios gigas de información por cada estudiante y cientos de gigas de información con cada año de uso.

Una investigación realizada Celis et al. [17] compara el impacto en el rendimiento académico que tienen cuatro factores distintos para alumnos de primer año de ingeniería de la Universidad de Chile y alumnos de primer año de educación de la Universidad Católica. Estos factores analizados corresponden a los siguientes:

- **Información Pre-College**: Correspondiente a toda aquella información numérica que exista sobre el alumno antes de entrar a la institución. Esto incluye al género, promedio de notas en educación secundaria, información socioeconómica y puntaje PSU.
- **Desempeño académico**: Nota final para cada ramo de primer año, además de las tasas de aprobación para los ramos de primer semestre.
- **LEARN+**: Un cuestionario creado por los mismos investigadores para indagar más sobre los métodos de estudio y otra información sobre la estadía del estudiante en la institución.
- **Interacciones con el LMS**: Abarca todas la actividad registrada por el LMS (U-Cursos para el caso FCFM).

En grandes rasgos, dicha investigación consistió en procesar los datos de ambas instituciones por separado, entrenar y validar un modelo de "Support Vector Regression" (SVR) que recibiera dichos datos y calculara la nota final promediada del semestre para cada alumno y finalmente extrapolar a partir de este resultado cuáles variables tuvieron un mayor impacto en la predicción del rendimiento de los estudiantes. Dentro de las conclusiones preliminares obtenidas, ellos consideraron que, para la FCFM, algunas de las variables con mayor impacto en la predicción del rendimiento correspondían a aquellas derivadas de interacciones con el LMS, tales como la desviación estándar y el centro de masa en la frecuencia de uso semanal de la plataforma, tal como se ve en la figura 2.2.

Aunque este último estudio fue realizado analizando tan solo un subconjunto de la cohorte 2017, con un universo cercano a 270 personas estudiadas, los resultados obtenidos son bastante interesantes, ya que dan el puntapié inicial a un gran número de conjeturas relacionadas con la utilización de datos de uso del LMS con el fin de predecir el rendimiento. En particular, una de las interrogantes más grandes que conciernen a este trabajo, es si las métricas obtenidas por ellos –en conjunto con algunas otras derivadas de la comparación en el uso entre un semestre y otro– permiten mejorar los resultados obtenidos por el modelo original de Celis et al.

2.5. Discusión

Según el contexto y la situación expuesta en el capítulo actual, es posible llegar a la conclusión de que la utilización de datos relacionados con los estudiantes, su rendimiento pasado y el uso del LMS, pueden ayudar a la predicción del rendimiento académico futuro.

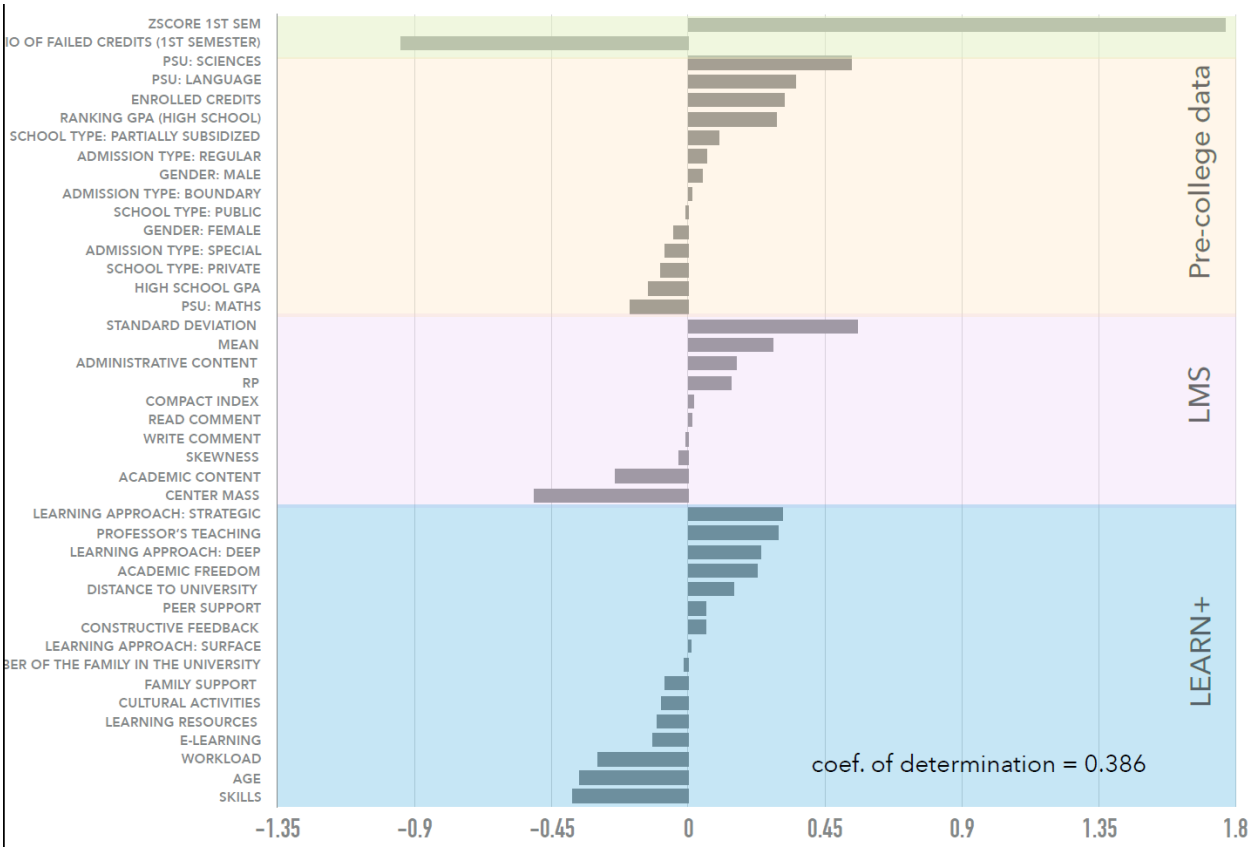

Figura 2.2: Resultados obtenidos por Celis, López y Silva [17]. Cada barra (etiqueta en el eje y) representa un "coeficiente" indicando el impacto en la predicción (valor en el eje x). Mientras más grande sea la barra, mayor el impacto. Una barra hacia la derecha indica que afecta positivamente el resultado y una barra a la izquierda indica una influencia negativa.

Esto puede ser una gran adquisición para las diversas instituciones de educación superior que existen en el país, ya que permitirá implementar planes de acción con el objetivo de disminuir la deserción estudiantil. Producto de esto, en el capítulo siguiente se explicará cómo puede ser modificado el modelo de Celis et al. para la predicción de un universo mucho más grande de estudiantes, de modo que se pueda incluir a la mayor cantidad de gente posible en la búsqueda e implementación de planes de apoyo para evitar la doble reprobación de ramos para estudiantes de la FCFM.

Capítulo 3

Análisis y diseño

Teniendo en cuenta las investigaciones relatadas en las secciones anteriores, el trabajo a realizar consistió en intentar recrear y ampliar el modelo generado por Celis et al. [22] para la predicción de doble reprobación en alumnos de primer año, de modo que pueda ser utilizada de manera activa en la mayor cantidad de alumnos posibles. Para lograr este objetivo, se siguieron una serie de pasos (descritos a continuación) para mantener la calidad de predicción a la vez que se van agregando más datos y variables paulatinamente, creando un modelo confiable y que pueda ser utilizado por un largo tiempo.

3.1. Objetivos del proyecto

El principal objetivo que se buscaba lograr consiste en crear un Sistema de Alerta Temprana que pueda permitir la predicción de la probabilidad de reprobación para alumnos de pregrado de la FCFM, cursando ramos ya reprobados en el pasado. Para cumplir este objetivo, el desarrollo del proyecto debió ir logrando ciertos objetivos más específicos que permitieran considerar como satisfactorio el trabajo realizado.

En primer lugar, se espera que el modelo realizado fuera capaz de predecir, con un rendimiento parecido al creado por Celis et al., la doble reprobación de estudiantes de pregrado en la FCFM, pero sin discriminar por año de ingreso, especialidad en curso u otras variables que sí consideraba el modelo original. Para esto se debió ir aumentando el universo analizado de manera paulatina, siempre y cuando se pudiera mantener los niveles establecidos por Celis et al. o, de lo contrario, debía evaluarse la incorporación de nuevas variables o la reestructuración de la heurística utilizada.

Por otro lado, este modelo debió ser capaz de entregar predicciones para alumnos en cursos en progreso lo antes posible en el semestre, ya sea utilizando la misma heurística que el modelo original o incorporando más información al procesamiento y generación del nuevo modelo. Esto significa poder entregar una predicción al tener la nota de las primeras pruebas de sus ramos (como lo hace el modelo original) o antes de esto utilizando otras variables.

Como otro objetivo específico, también sería útil poder analizar el efecto generado por la incorporación de algunas variables no estudiadas por Celis et al., como las estadísticas de uso de U-Cursos, en la predicción de doble reprobación. Pero, del mismo modo, es valioso ver su impacto en la disminución en el tiempo mínimo necesario para poder entregar un resultado confiable.

Como último objetivo, se deberán crear las interfaces necesarias para poder entregar esta información a los usuarios finales, las cuales deben estar integradas en alguna de las plataformas creadas por el Centro Tecnológico Ucampus utilizadas en la FCFM y estas deben ser validadas por algunos de los usuarios, de modo que consideren aceptable la información que se entrega para la implementación de políticas institucionales o sistemas de apoyo estudiantil.

3.2. Procedimiento a seguir para generar el nuevo modelo predictivo

El primer paso para lograr los objetivos fue recrear el modelo en las condiciones establecidas originalmente por Celis et al., con el fin de validar que aun fueran válidas las conclusiones obtenidas por ellos y, también, decidir si se va seguir con el modelo tal como se planteó o si se debe realizar el proceso de selección de variables nuevamente. Una vez tomada esta decisión, se evaluaron también los resultados comparando con otros algoritmos conocidos de aprendizaje de máquinas, como Redes Neuronales o Support Vector Machine, para evaluar si se pueden obtener mejores resultados con estos.

Lo anterior significó generar, con un conjunto de datos entregado por Ucampus, las mismas variables escogidas por Celis et al. (tabla 2.1) y realizar el entrenamiento y validación para evaluar el rendimiento del modelo.

Posteriormente, se procedió a ampliar el conjunto de datos utilizados en el modelo con el objetivo de incluir a todos los cursos del Plan Común y no tan solo aquellos del primer semestre. El motivo de hacer esta pequeña ampliación del modelo es primero integrar cursos con bastante similitud a los que ya existían en el modelo, puesto que la gran mayoría de cursos en Plan Común siguen un estándar de tres pruebas en el semestre y con ponderaciones similares en la nota final.

De seguir presentando resultados satisfactorios, se procederá a ingresar datos incluyendo información de otros años y especialidades en la carrera, de modo que se pueda abarcar el mayor público objetivo posible. Con este cambio se pasaría de utilizar información de tan solo Plan Común, a integrar a los alumnos de las diversas líneas de especialización disponibles en la FCFM. Esto significa encontrarse con cursos con distintos tipos de evaluaciones al no existir un estándar global definido, teniendo cursos con el mismo formato que los de Plan Común, mientras que en otros solo existen tareas y/o presentaciones sin ninguna evaluación escrita. En este punto se debieron tomar ciertas decisiones sobre qué hacer con la variedad de evaluaciones existente, pudiendo directamente dejar de lado cursos que no cumplan con cierto estándar o adecuando el pre-procesamiento para, de alguna manera, transformar los datos o el modelo a un formato utilizable.

Teniendo lo anterior terminado, se requerirá que los resultados obtenidos sean al menos tan buenos como los obtenidos por Celis et al., es decir, que se obtenga un recall similar o mayor al 86 % y una precisión mayor o igual al 37,5 %. En caso contrario, se seguirá iterando sobre el modelo creado, intercambiando variables, modificando el pre-procesamiento de los datos, probando otros algoritmos, etcétera, de modo que se logren los resultados deseados.

Una vez que se de por terminada esa parte del desarrollo se procederá con el procesamiento de los logs de U-Cursos, de modo que se puedan extraer métricas del uso de la plataforma para cada curso de cada estudiante y después se puedan agrupar entre cursos reprobados y no reprobados, obteniendo los estadísticos que estimaron más relevantes en el estudio de Celis et al. [17] sobre las múltiples variables que impactan en el rendimiento académico de primer año. Con los datos ya procesados, se añadirán variables a las utilizadas en el modelo anterior relacionadas con el uso de U-Cursos, con el objetivo de analizar y comparar los resultados de este modelo con el anterior para determinar si estos datos ayudan a la predicción de doble reprobación.

En caso de que se obtengan buenos resultados al agregar las variables (incluso siendo un poco peores que el modelo hasta ese momento), se procederá a realizar una selección de características de la misma manera que realizaron Celis et al. [22] cuando crearon el modelo original mediante varias iteraciones de *Forward Feature Selection* (FS) y *Backward Elimination* (BE) junto a una selección de atributos por frecuencia según lo explicado en la sección 2.3.

Claramente, al terminar este nuevo modelo, su desempeño deberá ser evaluado contra el modelo original de Celis et al. y todos los modelos propios creados a lo largo de este trabajo. Finalmente, el modelo que mejor desempeño presente será el utilizado para la implementación del Sistema de Alerta Temprana al finalizar el desarrollo.

3.3. Arquitectura de la plataforma

Dado que se quiere montar el trabajo realizado en las plataformas creadas por el Centro Tecnológico Ucampus, es necesario explicar brevemente cómo esta está construida y de qué manera se incorporará el modelo generado dentro de la lógica utilizada en estas.

La arquitectura de U-Cursos se basa en una metodología modular en donde existe un "núcleo" que administra los distintos módulos. Dichos módulos se organizan en tres niveles: Usuario, Curso e Institución, dependiendo de la visibilidad que se quiera o la relación existente con los integrantes del nivel escogido. Dado que el Sistema de Alerta Temprana estaba asociado de manera más directa a los alumnos dentro de un curso, el nivel escogido en esta ocasión fue de un módulo de curso.

A nivel de desarrollo, un modulo está altamente desacoplado de los demás, por lo que la lógica puede estar implementada en cualquier lenguaje. En este caso se decidió utilizar Python junto a las librerías mencionadas en el marco teórico. A pesar de lo anterior, se decidió hacer la "presentación" de la información siguiendo los lineamientos de Ucampus,

donde se utiliza PHP bajo un framework desarrollado por el mismo centro.

Teniendo en consideración los puntos recién expuestos, la lógica desarrollada en Python alimentará una base de datos, a partir del procesamiento de la información contenida en los servidores de Ucampus. Esta será finalmente consumida utilizando PHP bajo el framework de U-Cursos y presentada en las diversas interfaces que se desarrollen llegando al final del proyecto.

3.4. Planificando las interfaces para entregar la información del nuevo modelo

Si bien hasta ahora solo se ha hablado de las mejoras y expansiones del modelo predictivo, este no sirve de nada si la información que genera no es utilizada para su objetivo original: Identificar a los estudiantes en riesgo de doble reprobación lo antes posible, para poder implementar sistemas de apoyo y tratar de evitar su caída en causal de eliminación. Por este motivo, otra parte importante del desarrollo es la creación de scripts que procesen los datos de manera periódica y generen constantemente predicciones para las personas pertinentes, presentándolas en una interfaz web integrada con los servicios ofrecidos por el Centro Tecnológico Ucampus.

Estas interfaces deberían entregar la información necesaria para los docentes, coordinadores o investigadores, para que puedan conocer de antemano el contexto de cada estudiante y se puedan poner en contacto con estos en caso de querer realizar programas de apoyo. A su vez, estas estarán integradas en las páginas dedicadas a cada curso dentro de la plataforma U-Cursos y con permisos de lectura restringido para profesores, académicos, investigadores, secretarías docentes, coordinadores y administradores, de modo que puedan analizar la situación para cada uno de sus cursos por separado.

Para la validación de estas interfaces, se realizarán entrevistas a usuarios de U-Cursos que cumplen distintos roles dentro de la institución, de modo que se pueda analizar el punto de vista tanto de los docentes, coordinadores, secretarias docentes e investigadores. Se planea entrevistar al menos a un persona dentro de cada grupo que interactúe regularmente con U-Cursos sobre su opinión acerca de la información entregada y la forma en que esta se muestra, considerando la aprobación de estos usuarios como la prueba de validación.

Antes de realizar las interfaces como tal, se crearon diversos bocetos para validar rápidamente si estas tendrían, al menos, la información mínima para mostrar los datos generados por el modelo. Dichos bocetos se hicieron utilizando el software Balsamiq[1] y tratan de mostrar cómo se vería la integración de la información a entregar con el estilo y las interfaces actuales de U-Cursos.

Originalmente se pensó ubicar y acceder a este módulo desde la vista de curso y desde la vista de perfil del usuario. La idea de colocarlo en la vista de curso es que un profesor pueda ver rápidamente la información de los estudiantes que tiene directamente bajo su cargo

[1] https://balsamiq.com/

en cada curso, mientras que el módulo a nivel personal permitiría observar la información de todos sus cursos a la vez. Esto último estaría especialmente pensando para los usuarios coordinadores que tienen muchos cursos a su cargo a la vez.

Con respecto a las interfaces mismas, se pensó en generar al menos dos vistas: una general que muestre información resumida de todos los estudiantes pertenecientes al curso o institución y otra que se utilice para obtener información más detallada de un alumno en específico.

De este modo, en la figura 3.1 se puede observar la primera versión de la vista general en un curso en específico. La idea de esta es mostrar gráficamente la probabilidad de reprobación de los distintos alumnos en el curso junto a su información básica (nombre y RUT). En la figura 3.2 se puede observar que el seleccionar una barra en específico haría aparecer un tooltip con más información y un botón para ir a la vista de detalle del estudiante.

Las vistas representadas en las figuras 3.3 y 3.4 serían las correspondientes a la vista del perfil del usuario coordinador. Estos usuarios suelen estar a cargo de muchos cursos o muchas secciones del mismo curso a la vez, por lo que es importante entregarles herramientas para visualizar los alumnos en riesgo para todos estos cursos a la vez.

En la figura 3.5 se puede observar la idea inicial que se tuvo para la vista del detalle de un alumno. Aquí se mostraría más información con respecto al contexto del estudiante, como la cantidad de cursos que está cursando, cuántos de esos los está cursando por segunda vez, cómo le fue en iteraciones pasadas de los mismos ramos, el horario que tiene, etcétera. Si bien todo eso no aparece reflejado en el mockup, la idea era dejar el espacio de abajo vacío para poder definir la información a colocar al crear las interfaces finales.

Aunque en un principio se pensó generar las interfaces solo a nivel de curso y perfil, en la figura 3.6 se puede observar una vista a nivel de institución, de modo que se pudieran ver muchos cursos de la FCFM sin tener que estar dictados o coordinados por la misma persona. Además, en esta vista se entrega otra opción para la disposición de la información en la interfaz general, transformando el gráfico de barras en una tabla con el objetivo de entregar más información sobre cada estudiante en la misma vista, mezclándolo con un sistema de semáforo para destacar con color a los alumnos en mayor riesgo.

Como ya se dijo, estos mockups fueron validados entre el autor y ambos profesores guía y se consideró que poseían la información necesaria para realizar la primera versión de las interfaces finales, pero dejando la versión de la vista general en una tabla en vez de la que utiliza un gráfico de barras.

3.5. Resumen

En resumen, en el capítulo anterior se plantearon los requisitos y pasos a seguir para la realización de una mejora al modelo creado por Celis et al., pero con un espectro de usuarios ampliado a todos los estudiantes de pregrado de la FCFM y con la posible incorporación de datos de uso de U-Cursos. Para mostrar esta información se creará, además, una interfaz

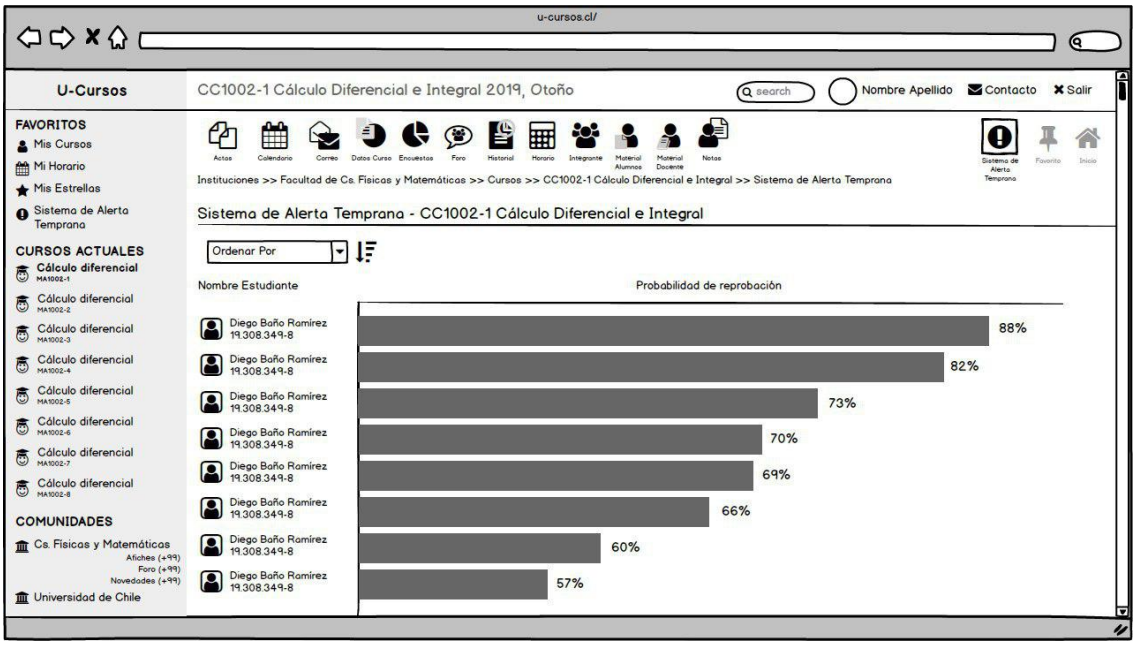

Figura 3.1: Interfaz de usuario coordinador en la vista de curso.

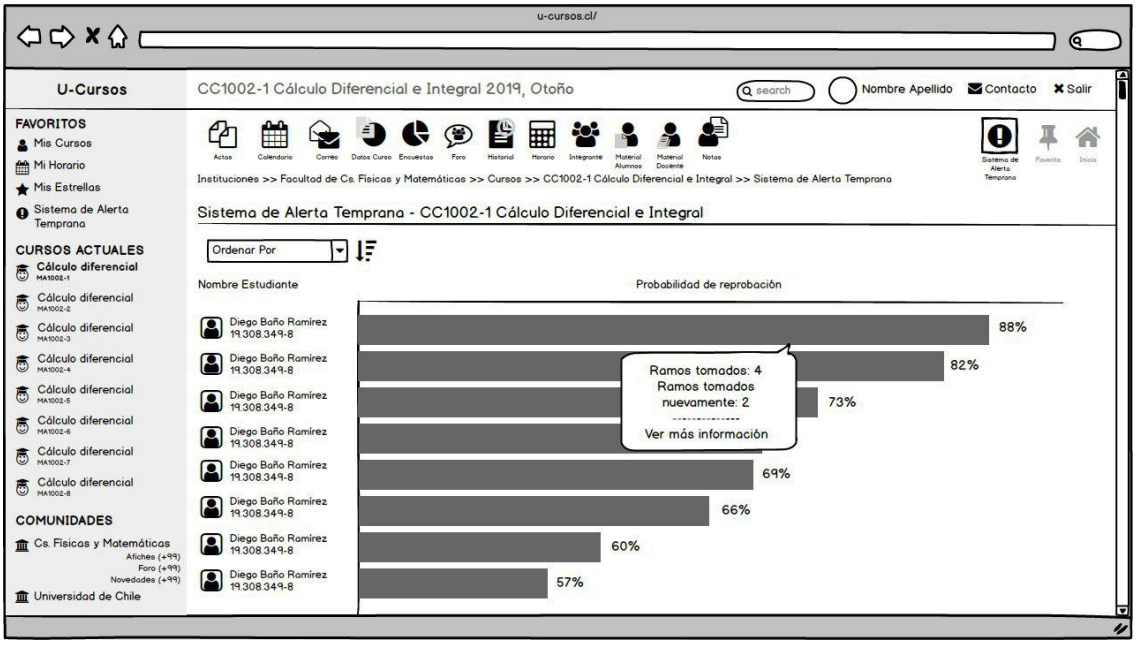

Figura 3.2: Interfaz de usuario coordinador en la vista de curso.

Figura 3.3: Interfaz de usuario coordinador en la vista de perfil.

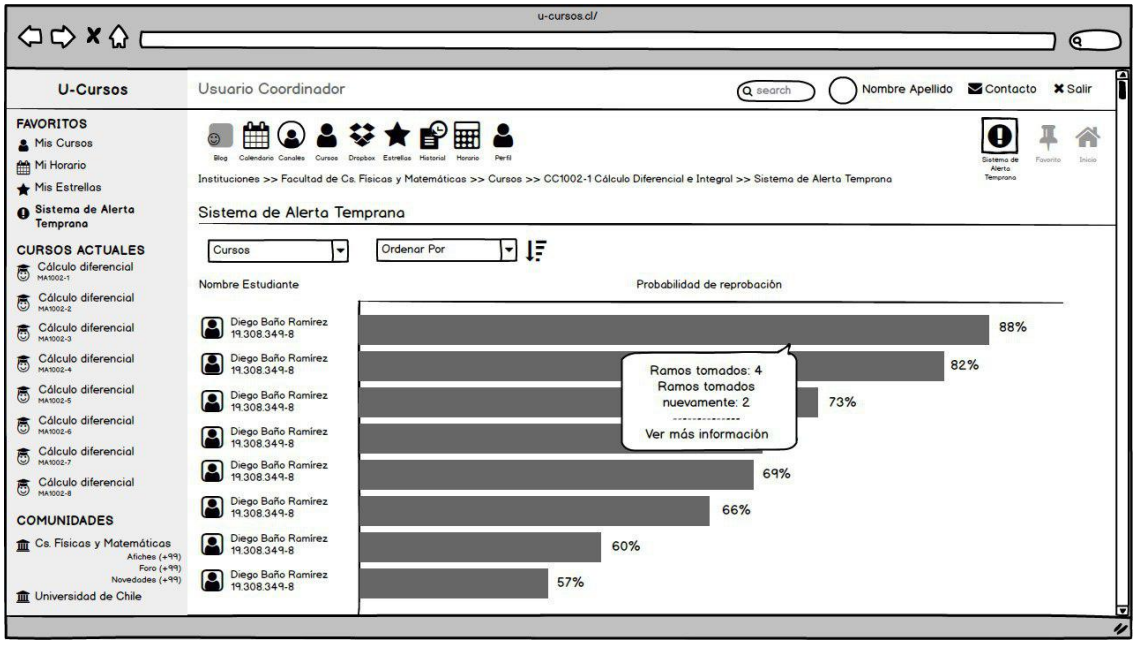

Figura 3.4: Interfaz de usuario coordinador en la vista de perfil.

Figura 3.5: Interfaz de detalle para un alumno.

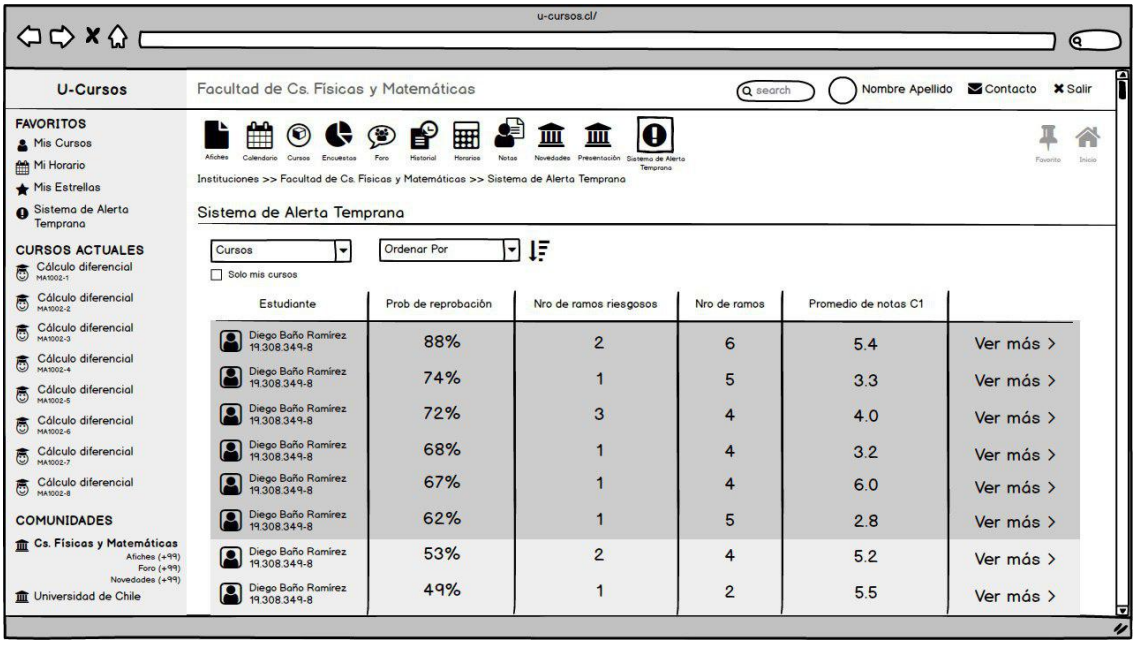

Figura 3.6: Interfaz de usuario coordinador a nivel institución.

de usuario dentro de la página de cada curso en U-Cursos, donde se podrán consultar los resultados para los estudiantes cursando el ramo por segunda vez apenas exista información disponible para realizar los cálculos. Por lo anterior, el capítulo siguiente servirá como una explicación de todas las decisiones tomadas durante la ejecución del plan recién descrito y las dificultades encontradas en el proceso.

Capítulo 4

Implementación

En el capítulo a seguir se explicará paso a paso el trabajo realizado con el fin de cumplir el objetivo de este desarrollo: crear un Sistema de Alerta Temprana que permita la predicción de la probabilidad de reprobación para alumnos de pregrado de la FCFM, cursando ramos ya reprobados en el pasado. Se comenzará explicando las iteraciones en las que se modificó el modelo de Celis et al. en la sección 4.1, mostrando el procesamiento realizado a los datos, las restricciones tomadas y los resultados obtenidos. Posteriormente, en la sección 4.2 se hablará del desarrollo de las interfaces donde se mostrará la información generada por el modelo final.

4.1. Extendiendo el modelo

La metodología utilizada para el desarrollo del nuevo modelo consistió en ir realizando modificaciones al procesamiento de datos de manera que se fuera aumentando el universo de personas estudiadas paulatinamente hasta llegar a lo más amplio posible manteniendo los estándares establecidos. Esto permitiría determinar en qué punto el modelo dejaría de ser aplicable y, por lo tanto, requeriría realizar un nuevo proceso de selección de variables o de restricciones para la información procesada o alguna otra medida que permitiera volver a aumentar el rendimiento.

4.1.1. Integrando el primer año y las ventanas entre semestres

Debido a lo anterior, se decidió partir integrando más información de estudiantes en Plan Común antes de comenzar a interactuar con los cursos de especialidad, ya que no se tenía pensado en este momento cómo tratar con cursos sin controles o con distintas ponderaciones. Por estos motivos, se amplió el universo estudiado de dos maneras: primero considerando la reprobación de cualquier curso de primer año, en vez de solo utilizar el primer semestre, y segundo tomando observaciones de estudiantes que volvieron a tomar los cursos incluso si no fue el semestre siguiente al reprobarlos.

La idea de analizar todos los cursos de primer año fue aumentar la cantidad de cursos y observaciones estudiadas para obtener un mayor conjunto de entrenamiento y validación, pero sin modificar demasiado la lógica de procesamiento de los datos, gracias a que gran parte de los cursos existentes en el primer año siguen la misma lógica de tres controles en el semestre con similar ponderación en la nota final.

Por otro lado, también se quiso ampliar el espectro de estudiantes analizados considerando a quienes estuvieran tomando un ramo nuevamente, pero que no lo hicieran en el semestre inmediatamente después a reprobarlo, diferenciándose así del modelo original de Celis et al. Estos cambios implicaron varias modificaciones que se tuvieron que realizar a la lógica de procesamiento de los datos, debido a la dependencia de las variables C1IRS2<FIRS1, C1INRS2<C1INRS1, C1IRS2 - 4.0 y minC1INRS2<C1INRS1 (tabla 2.1) a un semestre en específico y a que un alumno podría reprobar más de una vez un mismo ramo, tomándolo en múltiples semestres.

Exploración de los datos

La base de datos entregada para realizar la primera parte de este trabajo consistía en 8 tablas de información acerca de unas 6.000 personas distintas (como se observa en la figura 4.1), entregando información relevante como lo son notas parciales y finales, información y antecedentes personales, inscripción de cursos y estado final de cada ramo (aprobado o reprobado siendo los relevantes para el análisis). Algo importante de destacar es que esta base de datos estaba compuesta por un subconjunto del total de estudiantes de la FCFM caracterizados por haber solicitado al menos una Inscripción Académica Excepcional (IAE) a lo largo de su carrera, debido a que era una base de datos que ya estaba preprocesada y disponible rápidamente, debido a su utilización en otro estudio, para comenzar a trabajar en ella. En las tablas existían algunos atributos comunes que permitían hacer relaciones entre una tabla y la otra, dentro de los cuales destacan:

- PERSONA: Número que identifica de manera única y anónima a cada persona en la base de datos
- SEMESTRE: Número compuesto por el año en que se está dando un determinado ramo seguido de un 1 si se dio durante el semestre de otoño o un 2 en caso de haber sido durante el semestre de primavera
- CURSO (o RAMO): Código identificador de un ramo, escrito como *LLXXXX*, con L una letra y X un dígito, y seguido de la sección específica que tomó la persona de la forma *-XX*

Procesamiento inicial de los datos

En primer lugar, se decidió trabajar solo con los datos de alumnos cuyo ingreso a la facultad haya sucedido posterior al año 2007 (información contenida en la tabla "ANTECEDENTES_RAND"), ya que dicho año se llevó a cabo el último cambio importante en la malla curricular, e inferior al 2018 por ser el año en curso, dejando un total cercano a 1.800

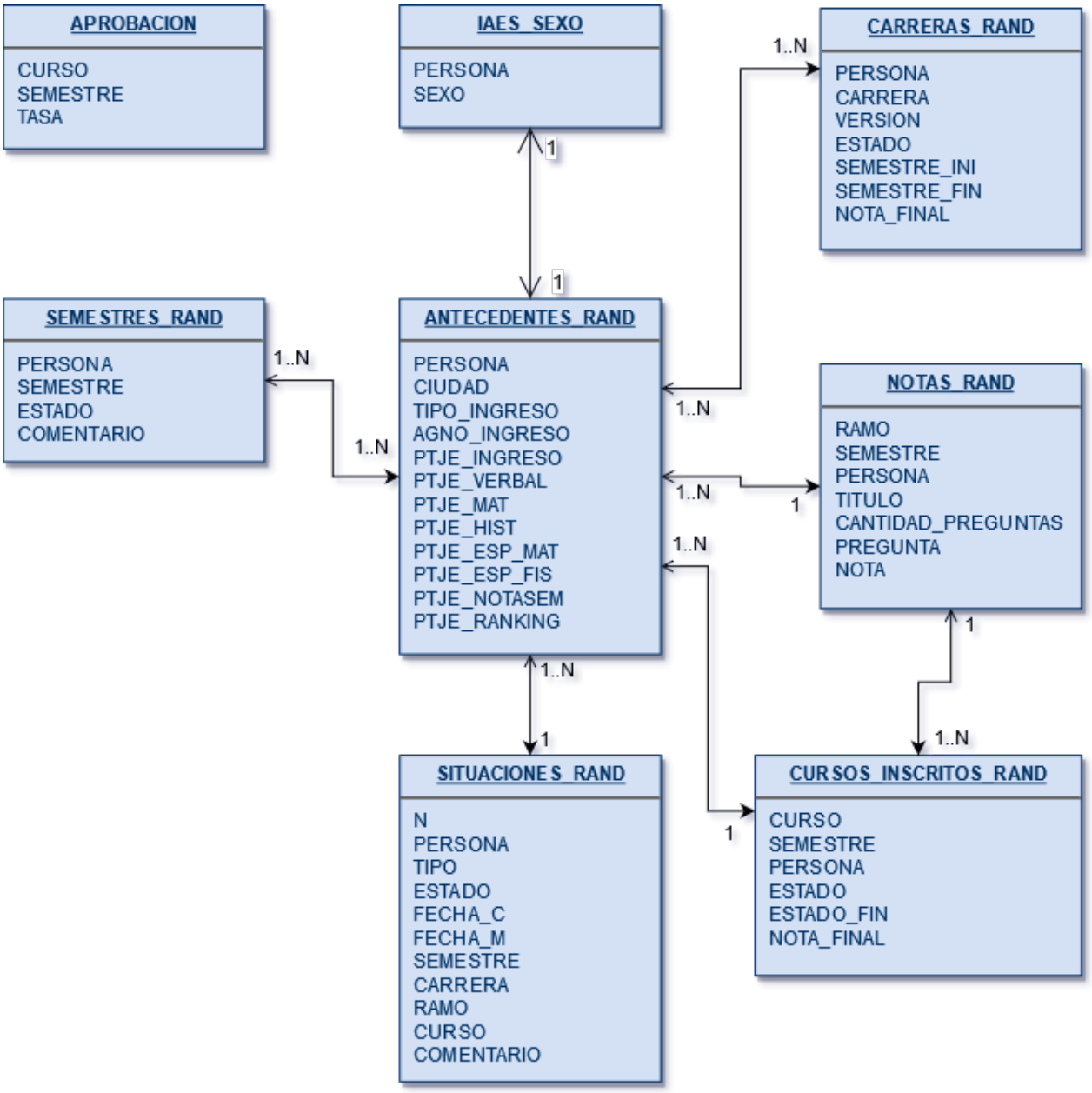

Figura 4.1: Tablas entregadas por Ucampus

personas para trabajar. Una vez obtenido este subconjunto, se hizo una unión con la tabla "CURSOS_INSCRITOS_RAND" para obtener la información de cursos aprobados y reprobados para cada alumno desde el 2007 al 2017, llamaremos a esta nueva tabla "ESTADOS" de aquí en adelante.

De manera paralela se trabajó en el procesamiento de la tabla de notas, la que originalmente contenía más de 4 millones de observaciones para 4.238 personas distintas. Para obtener las variables independientes a utilizar en el modelo, se necesitaba extraer de esta tabla el promedio de notas final de los cursos reprobados y el promedio de notas de los controles 1 de ramos reprobados y no reprobados. La tabla de notas (llamada "NOTAS_RAND") contiene las siguientes columnas:

- RAMO: Código del ramo al que pertenece la nota, usualmente de la forma LLNNNN-SS, donde LL representa el código de dos letras del departamento a cargo de dicho ramo, NNNN es el código numérico del ramo y SS es el código de la sección a la que pertenecía el alumno. Generalmente, el primer dígito del código numérico indica el año en que se espera que el estudiante lo tome en su carrera.
- SEMESTRE: Semestre en que se estaba cursando el ramo, en la forma AAAAS, donde AAAA es el año y S indica el semestre de otoño (1) o primavera (2).
- PERSONA: Código único para identificar de manera anónima a la persona que obtuvo dicha nota
- TITULO: Nombre asociado a la nota. Por ejemplo "Control 1".
- CANTIDAD_PREGUNTAS: Cantidad de preguntas que componen la evaluación.
- PREGUNTA: Número de la pregunta asociado a la nota. En caso de ser el promedio de la evaluación este campo adquiere el valor 0.
- NOTA: Nota obtenida en escala 1 a 7 amplificada por 1.000 (no se explicó el motivo para tener las notas amplificadas por este factor).

Como se mencionó anteriormente, dado que el modelo de Celis et al. solo consideraba los datos de alumnos en su primer año y se deseaba ampliar este espectro al extender el modelo, se decidió trabajar con todas las notas asociadas a cursos del primer año de Plan Común pero sin importar el momento en que los ramos reprobados hubieran sido tomados por segunda vez. En términos prácticos, esto implica que solo se consideraron las notas de ramos cuyo nombre (campo "RAMO") empezara con CC1, MA1, EI1, CM1 y FI1 seguido de otros tres números pero sin importar la sección, ya que solo los ramos de primer año cumplían con esta característica y cada alumno solo puede pertenecer a una sección en un semestre determinado. Además de esto, como llave primaria para identificar cada fila de manera independiente se utilizó la tupla ("PERSONA", "SEMESTRE", "RAMO"), donde "RAMO" corresponde al campo del mismo nombre, pero sacando la parte que identifica a la sección tomada.

Extracción de las notas

La siguiente tarea a desarrollar consistió en obtener las notas del Control 1 de cada uno de los ramos que quedaban. Esto fue un gran desafío debido a que la única información de

la tabla que se podía utilizar para obtener una evaluación en particular era el "TITULO" de cada una. Desafortunadamente, este campo era inconsistente a lo largo de la tabla, usando muchos nombres distintos para referirse a la misma evaluación (e.g. "Control 1", "Prueba 1", "C1", "Control 1 Final", etc). Para resolver este problema, se decidió utilizar la expresión regular descrita en la expresión 4.1.

$$\wedge ((C(ontrol)?)|(Prueba))(n^o)?1[\wedge 0-9]*\$ \tag{4.1}$$

La idea detrás de esta es que permite identificar todas aquellas notas donde su "TITULO" comience con "C", "Control" o "Prueba", esté seguido de un 1 (con o sin símbolo n^o) y pueda terminar en otra cosa que no contenga más números. Esto permite calzar expresiones como "Control 1", "Prueba 1", "C1" y "Control 1 Final". Además de esta expresión regular, se utilizaron solo las notas correspondientes a la pregunta 0, lo que corresponde al promedio de dicha evaluación. Se obtuvo una nota para cada tupla ("PERSONA", "SEMESTRE", "RAMO") y, en caso de calzar con más de 1 nota para una tupla (e.g., existen para la misma tupla las notas "Control 1" y "Control 1 Final"), se decidió simplemente mantener el promedio entre estos valores y así evitar tener que revisar caso por caso al no tener un estándar claro.

Para obtener la nota final de los ramos reprobados se tuvo que recurrir a una estrategia muy similar a la recién descrita pero haciendo uso de la expresión regular que se muestra en la expresión 4.2, producto de que la nota final de un ramo no es registrada en el sistema si este fue reprobado. La lógica de este método radica en que usualmente (sobretodo en ramos de primer año) la nota final de un estudiante en un ramo se calcula como el promedio simple entre los controles rendidos o una combinación lineal entre dicho promedio y la nota obtenida en el examen (en caso de que lo haya rendido obligatoria o voluntariamente). Por lo anterior, en vez de obtener las notas cuyo título calce con algo similar a "Control 1", se tomaron todas aquellas que calzaban con alguna de las pruebas o controles realizados y se calculó el promedio simple entre todas estas. Así, para una tupla ("PERSONA", "SEMESTRE", "RAMO") se podrían encontrar las notas con título "Control 1", "Control 2 Final" y "Prueba 3" y la nota final se calcularía como el promedio simple entre estas.

$$\wedge ((C(ontrol)?)|(Prueba))(n^o)?[0-9]+[\wedge 0-9]*\$ \tag{4.2}$$

Una vez obtenidas estas notas para cada tupla ("PERSONA", "SEMESTRE", "RAMO"), se separó estos datos en dos tablas realizando dos joins con la tabla "ESTADOS" – la cual poseía información del estado final de cada ramo –, para obtener una tabla con los ramos reprobados y otra con los ramos aprobados para cada persona entre los años 2007 y 2017.

Consolidación de la información

Como ya se ha mencionado, el modelo original de Celis et al. consideraba en el análisis solo la realización de ramos por alumnos en su primer y segundo semestre del Plan Común, por lo que las variables con sufijo S1 y S2 hacían referencia directa a los semestres de otoño y primavera, respectivamente, de su primer año en la universidad. En el caso del modelo que

fue construido en esta oportunidad, se amplió el espectro para considerar todos los ramos de primer año, pero sin importar el semestre en específico en que se tomó. Este cambio implica dos cosas:

1. En un mismo semestre, se pueden estar tomando varios ramos reprobados en semestres distintos. Por ejemplo, se reprobó un ramo en otoño 2016, otro en primavera 2016 y se tomaron ambos por segunda vez en otoño 2017. Esto implica revaluar el significado de S1 y S2 en las variables utilizadas.
2. Un mismo ramo puede ser reprobado más de una vez. Esto significa que para estos alumnos pueden haber múltiples observaciones de un mismo ramo en distintos semestres. Por ejemplo, un estudiante que tomó el mismo ramo en otoño 2016, primavera 2016 y otoño 2017 va a tener dos instancias para el entrenamiento, cuando lo tomó por segunda vez en primavera 2016 y cuando lo tomó por tercera vez en otoño 2017.

Así, cuando se habla de ramos reprobados, el sufijo S2 pasa a ser una referencia al semestre para el cual se quiere predecir si un estudiante va a reprobar o no un ramo nuevamente y S1 corresponde al último semestre en que el ramo se tomó (y reprobó). Con esto en mente, variables como «el promedio de notas de control 1 de los ramos reprobados en el primer semestre» (C1IRS1) se transforman en «el promedio de notas de control 1 de los ramos reprobados la última vez que se tomaron».

Para entender mejor lo anterior, si por ejemplo un estudiante reprueba el ramo FI1001 en su primer semestre y el ramo MA1102 en su segundo semestre, pero toma ambos por segunda vez en su tercer semestre, entonces la variable C1IRS1 se calcularía como el promedio entre la nota del primer control del ramo FI1001 en el semestre de otoño con la nota del primer control del ramo MA1102 en el semestre de primavera y la variable C1IRS2 sería el promedio del control 1 de ambos ramos cuando los tomó juntos por segunda vez.

Para el segundo punto, respecto a las personas que reprueban más de una vez el mismo ramo, lo que se va a considerar como S1 para los ramos reprobados va a ser la instancia inmediatamente anterior en que se tomó el ramo, para cada vez que se volvió a tomar el mismo ramo. Por ejemplo, si una persona reprobó el ramo MA1001 en el primer y segundo semestre del año 2010 y lo aprobó el semestre de otoño 2011, esta generaría dos observaciones en la tabla con comparaciones entre el primer y el segundo semestre 2010 y otra entre el segundo semestre 2010 y el primer semestre 2011.

En cuanto a los ramos aprobados, el sufijo S2 de las variables corresponde al semestre que se quiere predecir y el S1 es, simplemente, el semestre cursado inmediatamente anterior. Esto es porque los ramos que se están cursando nuevamente pueden haber sido reprobados en diferentes semestres entre sí y no es fácil determinar cuál semestre es el que se debería utilizar en cálculos relacionados con cursos no reprobados. También, a diferencia de los ramos reprobados, no existe una relación tan clara y directa entre los ramos de un semestre y otro como el estar tomando exactamente el mismo ramo nuevamente producto de su reprobación en un semestre anterior.

La función objetivo de este modelo fue obtenida mediante operaciones con la tabla "ESTADOS" descrita anteriormente: para cada fila donde hubiera un ramo reprobado, se buscaba el siguiente semestre donde la misma persona tomara el mismo ramo, para luego marcar una

nueva columna – "REPRUEBA" – como verdadero si es que esta nueva realización del ramo terminaba en reprobación y falso en caso contrario.

Posterior a esto, para todas las tablas obtenidas hasta el momento se agruparon los datos por persona y semestre, ya que este es el identificador de tupla que se quería mantener. Para la función objetivo se marcó la columna "REPRUEBA" como verdadero para aquellos grupos donde al menos un ramo hubiera sido reprobado y falso para ellos donde ninguno hubiera sido reprobado. En el caso de las notas calculadas para los controles 1 y las notas finales, estas se dejaron como un promedio entre las notas de cada grupo. De este modo, para cada tupla ("PERSONA", "SEMESTRE") se mantuvo el promedio de sus notas en controles 1 en dicho semestre, el promedio de las notas finales obtenidas en ese semestre y un campo indicando si se reprobó o no un ramo por segunda vez.

Por último, la tasa de créditos reprobados se tuvo que reemplazar por la tasa de ramos reprobados ya que no se pudo conseguir la información de créditos de cada ramo antes de la finalización del primer ciclo de desarrollo. El valor de esta tasa para cada vector de características corresponde a la tasa de reprobación del semestre inmediatamente anterior al que se quiere predecir, por los mismos motivos que se explicaron para el cálculo de notas de los ramos aprobados.

Una vez obtenidos todos estos datos, el obtener las variables independientes señaladas por Celis et al. consistió en realizar varios joins entre las distintas tablas creadas de tasas y notas – y varias veces de tablas consigo mismas para obtener información sobre el semestre inmediatamente anterior –, para luego componerlas tal como se indica en la descripción de las variables incluida cuando se explicó el modelo original.

Riesgos del procesamiento de los datos

Existieron dos variables binarias que no se pudieron obtener a tiempo en esta ocasión, correspondientes a si el alumno venía de un colegio particular y si venía de un colegio subvencionado, ya que no venían incorporadas en la base de datos entregada para esta ocasión.

También es importante notar que la forma en que se calculó o se obtuvo las notas en esta oportunidad, tanto para los controles 1 como las notas finales, suponen un gran riesgo para el modelo como fue realizado en esta ocasión, ya que pueden haber reglas particulares de cada curso que no se están tomando en cuenta en este análisis, como cursos con otro tipo de evaluaciones (e.g. laboratorios, tareas y ejercicios), cursos donde cada control sea ponderado de una manera distinta o que se hayan hecho modificaciones a notas producto de evaluaciones recuperativas o de otra índole, ejemplo de esto son los ramos con notas del tipo "Control X final" donde esta solía ser la nota original de un control X, pero alterada producto de una evaluación adicional. Tampoco se alcanzó a idear un algoritmo o expresión regular para considerar las notas de los exámenes en el cálculo de la nota final de cada ramo e, incluso de haberse realizado, el factor por el que se pondera esta nota también puede cambiar según el curso o el semestre en que se calcula.

Entrenamiento y validación inicial del modelo

Una vez procesados los datos a utilizar, y generados los vectores de características que se utilizarían finalmente para el entrenamiento y validación del modelo, se procedió con algunas pruebas iniciales para validar el funcionamiento del mismo.

En una primera instancia se realizó una validación cruzada sobre todo el conjunto de datos, arrojando resultados bastante malos en comparación con los que obtuvieron Celis et al., generando un *recall* de 57 % y un *precision* de 69 %, lo que significa que, en promedio, solo se pudo identificar correctamente a cerca de la mitad de los alumnos con doble reprobación y que tan solo dos tercios de los que fueron predichos como que iban a reprobar efectivamente lo hicieron.

Ajustando el entrenamiento

Dados los malos resultados de este acercamiento inicial, se propuso analizar el conjunto de datos que se le entregó al modelo para realizar la validación cruzada, con lo cual se descubrió que de las 1.516 observaciones que quedaron finalmente, tan solo 203 correspondían a personas que reprobaron nuevamente un ramo, por lo que existía una desigualdad notoria en el equilibrio de las clases utilizadas para entrenar. Para solucionar esto, se decidió elegir un subconjunto aleatorio de cada clase, tal que la diferencia de número entre estas no fuera tan grande (solo dos veces la cantidad de personas que sí reprobaron), para evitar que el modelo prefiriera escoger una clase sobre la otra solo porque había una mayor presencia en el conjunto de entrenamiento.

Una vez realizado este cambio, la mejora en las estadísticas de predicción aumentaron considerablemente, generando un *recall* promedio de 77 % y un *precision* promedio de 76 %. Si estos resultados se comparan con los obtenidos por Celis et al. es claro que, si bien se obtuvo un *recall* menor, el *precision* aumentó considerablemente, por lo que igual pueden ser evaluados de manera favorable.

Cabe destacar que, como se mencionó en la sección 2.3, el modelo de Celis et al. utilizaba un umbral de predicción más bajo que el que se suele usar habitualmente (correspondiente al 50 %, valor usado en la validación cruzada) ya que la experimentación realizada por ellos determinó que un umbral del 19 % permitía optimizar de mejor manera los resultados generados por el modelo. Por este motivo se decidió probar los resultados entregados por el nuevo modelo pero con el umbral propuesto para el modelo original.

Validación final de la iteración

Se procedió a dividir el conjunto de datos en conjuntos de entrenamiento y validación de una manera similar a la utilizada por Celis et al.: se utilizó un subconjunto (con las clases más balanceadas) de las observaciones anteriores al 2017 para el entrenamiento del modelo y las observaciones del año 2017 (sin balancear) para la validación de los resultados. De este modo,

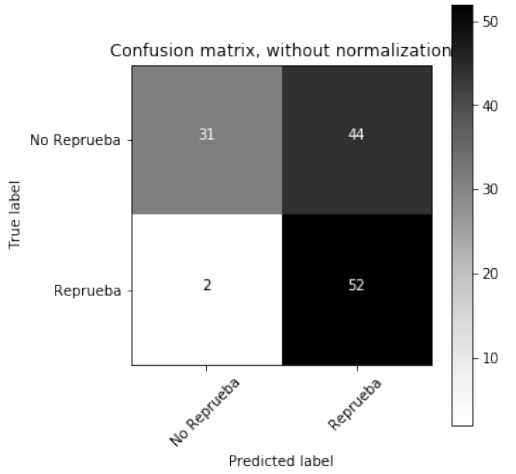

Figura 4.2: Matriz de confusión sin normalizar, utilizando un umbral del 19 %

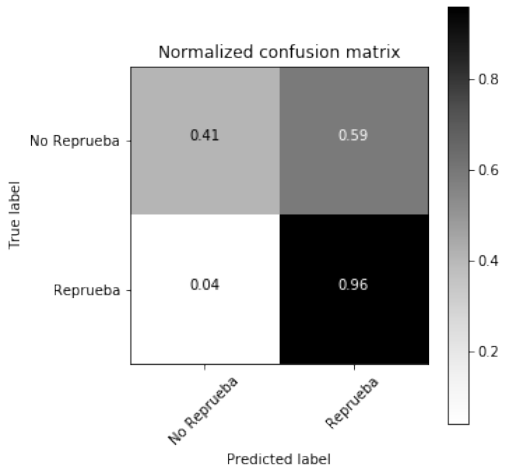

Figura 4.3: Matriz de confusión normalizada, utilizando un umbral del 19 %

para entrenar se utilizaron 591 datos, con un tercio de estos siendo aquellos que reprobaban nuevamente, y 129 datos para la validación, donde 54 de estos reprobaron nuevamente un ramo.

Se creó una nueva Regresión Logística utilizando el conjunto de entrenamiento para entrenar el modelo y luego se validó su funcionamiento con el conjunto de testing, pero utilizando el mismo umbral que Celis et al. en lugar del umbral por defecto de la librería. Los resultados obtenidos por este nuevo modelo fueron bastante mejores que los anteriores e incluso que el modelo original, pudiendo observar la matriz de confusión obtenida en las figuras 4.2 y 4.3. En este caso se logró un *recall* del 96 % y *precision* del 55 %, superando con creces los resultados obtenidos por Celis et al.

En la figura 4.4 se puede observar una representación de los resultados obtenidos por el modelo al predecir la probabilidad de doble reprobación para el conjunto de datos de prueba. Cada columna de dicho gráfico representa la probabilidad estimada por el modelo (eje y) para un alumno en un semestre en específico, en verde salen marcados aquellos estudiantes que no reprobaron por segunda vez en el semestre y en rojo aquellos que sí lo hicieron. La línea negra representa el umbral que fue escogido por Celis et al. al aplicarlo sobre los resultados del nuevo modelo, de este modo, las columnas que quedan a la izquierda de esta linea son las que el modelo marca como probable doble reprobación, mientras que las de la derecha son marcadas con la opción contraria. La figura 2.1 corresponde a un representación del mismo tipo que la recién descrita, pero creada por Celis et al. para mostrar los resultados del modelo original.

Análisis de los resultados

Al observar y comparar las figuras 2.1 y 4.4, una de las primeras cosas que cabe destacar es el volumen de datos presentes de cada clase. Se observa claramente que en el modelo original, de los cerca de 140 alumnos analizados, tan solo 12 de estos efectivamente reprobó por segunda

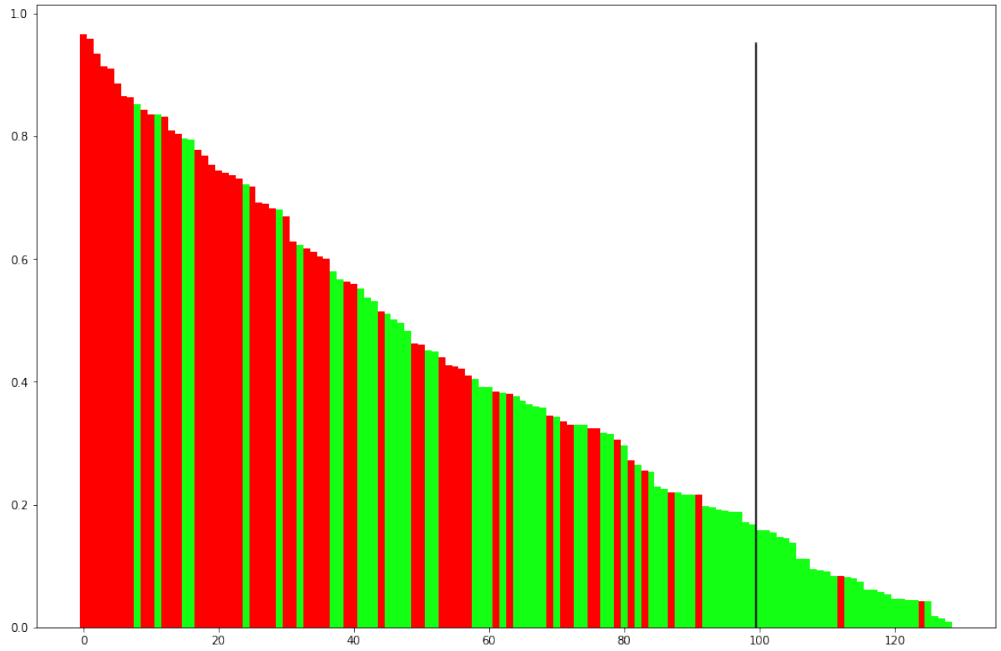

Figura 4.4: Representación visual de la clasificación del algoritmo y sus etiquetas reales. El eje x corresponde a los alumnos y el eje y a la probabilidad de reprobación asignada por el modelo a cada estudiante.

vez alguna asignatura, pero en el conjunto de datos del nuevo modelo, probablemente debido a la naturaleza de este conjunto de datos, existía un número mucho más grande de personas que efectivamente reprobaron por segunda vez, correspondiente a un poco más del 40 %. Como se mencionó al inicio de la sección "Procesamiento de los datos", los datos entregados por Ucampus para esta experiencia correspondían a alumnos con al menos una IAE en su historial y, dado que uno de los motivos por los que se puede tener que solicitar una IAE es reprobar dos veces un mismo ramo, es probable que se haya presenciado un sesgo en los datos utilizados para validar. En el entrenamiento este sesgo no es tan evidente, ya que de todas maneras se hizo un proceso de balance de clases que ayuda a prevenir este efecto.

Otra diferencia clara entre ambas imágenes es que en 4.4 el umbral se encuentra mucho más a la derecha que en 2.1, esto probablemente se dio por algo similar a lo que recién se explicó. Dado que hay un mayor número de observaciones presentes con doble reprobación que en el modelo original, el umbral refleja esto etiquetando un mayor número de datos como posible reprobación. Esto se ve apoyado por el hecho de que, a pesar de la diferencia en la posición del umbral, ambos modelos clasificaron incorrectamente solo dos elementos como que no iban a reprobar nuevamente.

4.1.2. Incorporando el resto de pregrado

Dados los resultados positivos obtenidos por el modelo antes relatado, se decidió mantener las variables utilizadas y el modelo tal y como se dejó para comenzar a integrar el resto de información de pregrado, considerando a todas las personas cuyo ingreso a la facultad haya

sido posterior a 2010, pero sin importar si los cursos analizados corresponden a Plan Común o a alguna de las especialidades impartidas, pero siempre y cuando se esté dando en el contexto de un plan de pregrado.

De aquí en adelante, la base de datos utilizada corresponde, en realidad, al conjunto de datos completo utilizado por los sistemas del Centro Tecnológico Ucampus, compuesto de una gran cantidad de bases de datos distintas cada una con distinto número de tablas y elementos. En particular, se va a trabajar con tres bases de datos: "MUFASA", que contiene información relacionada con los alumnos, sus antecedentes, planes, cursos inscritos y otras cosas relevantes; "UC_NOTAS", que almacena la información relacionada con las notas parciales de los estudiantes y, finalmente, "UCURSOS", que hace como un puente para relacionar la información contenida entre "MUFASA" y "UC_NOTAS".

Exploración de las nuevas bases de datos

El nuevo universo de datos evita el sesgo existente en los datos trabajados anteriormente al utilizar todo el conjunto de datos disponible en los sistemas de U-Cursos y UCampus. De las tres bases de datos utilizadas, la más extensa en términos de tablas y atributos existentes es, sin dudas, "MUFASA". Esta base de datos SQL contiene 90 tablas de datos y cerca de 12 millones de observaciones a lo largo de todas estas con información pertinente a gran parte de las facultades y estudiantes bajo el alero de la Universidad de Chile. Las 10 tablas más relevantes para este trabajo se pueden observar en la figura 4.5 y se describen a continuación:

- "INSTITUCIONES": Contiene la información de las instituciones, departamentos de la FCFM y otras entidades que interactúan con esta base de datos, conteniendo su id, el nombre y la sigla.
- "ANTECED_ALUMS": Contiene los antecedentes de todos los estudiantes de la facultad, con datos como el código identificador, la ciudad de residencia, año de egreso de la educación media, puntajes de la PSU, promedio de notas y ranking de educación media, tipo de colegio de egreso, año de ingreso a la facultad, "INSTITUCIONES" a la que ingresó y un poco más de información correspondiente a datos previos al ingreso a la universidad.
- "CARRERAS": Corresponde a la información de las distintas carreras impartidas en las INSTITUCIONES de la facultad.
- SEMESTRE: Esta tabla contiene información de los distintos semestres impartidos en la FCFM, incluyendo la fecha de inicio y fin estimadas y real, el código del semestre, el año al que corresponde y el periodo en que se dictó (anual, otoño, primavera o verano).
- "RAMOS": Corresponde a los distintos ramos impartidos en cada "INSTITUCION". Cada "RAMO" posee un código alfanumérico único y legible, un nombre, una cantidad de créditos asignada y otra información de ámbito administrativo.
- "CURSOS": Elementos correspondientes a una instancia de un "RAMO" dado en un "SEMESTRE" en específico con información de la sección a la que corresponde, los cupos que posee y otra información de carácter administrativo.
- "CURSOS_INSCRITOS": Tuplas de "CURSOS" inscritos por "PERSONAS" en un "SEMESTRE" en específico y con el estado final de dicho curso, que puede ser aprobado,

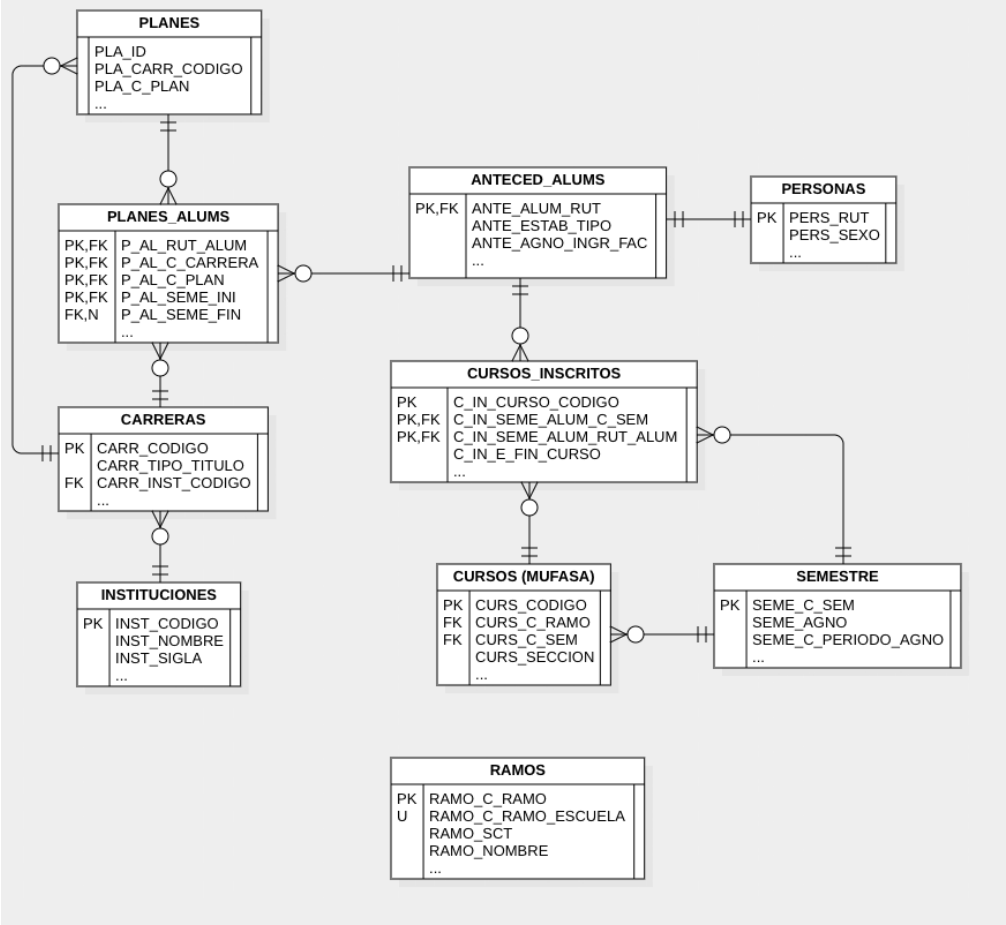

Figura 4.5: Estructura general de las tablas relevantes usadas en "MUFASA". El uso de "..." representa columnas omitidas.

reprobado, eliminado o en curso y la nota final (de estar disponible).

- "PERSONAS": Tabla con la información de cada una de las personas que estudian o trabajan en cada una de las instituciones de la FCFM y que tengan acceso a las plataformas del Centro Tecnológico Ucampus. Aquí se guardan los nombres, el rut o número identificador, nacionalidad, sexo, fecha de nacimiento y otra información personal.
- "PLANES": En esta tabla se almacena la información relacionada con los distintos planes de estudio que se imparten en cada una de las "INSTITUCIONES" para cada una de las "CARRERAS". Estos se diferencian de las carreras ya que se utilizan para notar cambios en la malla curricular o distintas versiones de la misma "CARRERA".
- "PLANES_ALUMS": Contiene tuplas de "PERSONAS" y "PLANES", correspondientes a alumnos que ingresan a distintos planes de estudios disponibles en las "INSTITUCIONES", guardando información del semestre en que se ingresó al plan, del semestre en que se salió (cualquiera sea el motivo) y otras columnas con información administrativa.

Otra base de datos importante para la realización del modelo corresponde a "UC_NOTAS", donde se incluye toda la información de notas parciales de los distintos cursos y alumno de la facultad. En esta base de datos se encuentran 8 tablas distintas con más de 48 millones de

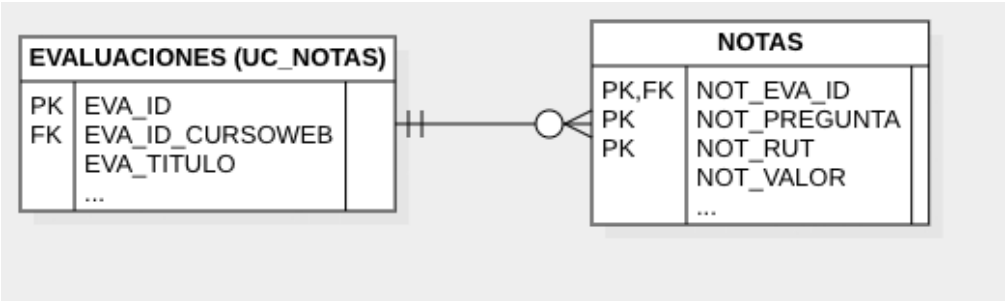

Figura 4.6: Estructura general de las tablas relevantes usadas en "UC_NOTAS". El uso de "..." representa columnas omitidas.

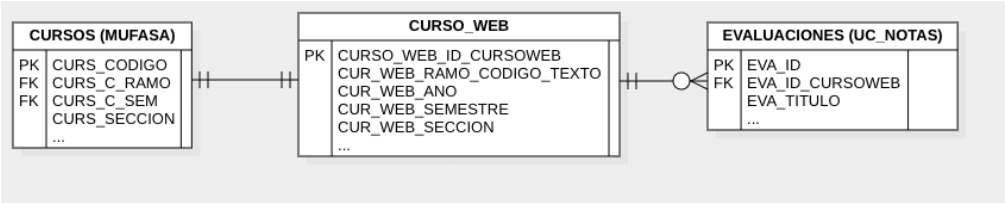

Figura 4.7: Estructura general de las tablas relevantes usadas en "UCURSOS". El uso de "..." representa columnas omitidas.

observaciones, pero solo se utilizan 2 de estas, las cuales concentran cerca de 42 millones de estas observaciones. Dichas tablas se describen a continuación:

- "EVALUACIONES": Contiene la información asociada a cada evaluación existente, almacenando el nombre de la evaluación (e.g. "Control 1", "Examen", "Tarea 3"), la fecha en que se publicó, el curso al cual está relacionada y otra información no tan relevantes en esta ocasión.
- "NOTAS": Cada fila de la tabla corresponde a la puntuación obtenida por un alumno en una determinada pregunta para una determinada evaluación, donde la pregunta 0 corresponde a la ponderación de las preguntas restantes y es, por tanto, la nota final de cada estudiante en la evaluación correspondiente.

Estas bases de datos son imprescindibles a la hora de crear el modelo predictivo, ya que es de aquí de donde se extraen todas las variables utilizadas. Sin embargo, los identificadores de los cursos a los que pertenecen las evaluaciones contenidas en la tabla "EVALUACIONES" de la base de datos UC_NOTAS, son distintos a los identificadores de los CURSOS en la base de datos "MUFASA", ya que estos corresponden a los identificadores únicos de cada curso que utilizados en U-Cursos y almacenados en la base de datos "UCURSOS", por lo que se hace necesario interactuar con esta tercera base de datos compuesta por 22 tablas distintas y unos 3 millones de observaciones, de las cuales solo interesa una:

- "CURSOS_WEB": Contiene la información de los cursos existentes en la plataforma U-Cursos, los cuales tienen un identificador único al cuál hacen referencia las "EVALUACIONES" y un código alfanumérico que hace referencia a los "RAMOS" de "MUFASA". Además, incorpora información relacionada a la sección correspondiente, el año y semestre en que se dictó junto a otros datos poco relevantes en esta oportunidad.

Utilizando estas 13 tablas distintas se deben realizar distintas operaciones de intersección, unión, filtrado y mapeado, para poder obtener finalmente todas las observaciones necesarias para poder hacer el entrenamiento y validación con el nuevo universo de estudiantes a analizar.

Consideraciones iniciales del procesamiento de los datos

Antes de seguir, es bueno recordar el objetivo al cuál se apunta con el procesamiento de los datos: obtener una tabla donde cada fila contenga la información necesaria para poder realizar una predicción de doble reprobación utilizando el modelo creado. Cada fila de esta tabla debe representar a un alumno en un semestre donde se encuentre tomando al menos un ramo que haya reprobado en el pasado y debe contener los atributos seleccionados por Celis et al. (presentes en la tabla 2.1) y cuya validez fue verificada en la primera parte de este desarrollo, es decir, tanto información varía sobre el semestre que se quiere predecir (referido como S2) y el semestre ficticio compuesto por ramos reprobados y aprobados en el pasado (referido como S1, recordar la explicación dada en el procesamiento de datos de la sección anterior).

Para comenzar con el procesamiento de los datos, se decidió solo utilizar solo la información de cursos tomados por personas que estuvieran cursando un plan de pregrado durante los años estudiados, es decir, cuyo ingreso al plan fuera durante o posterior al semestre de otoño de 2010. Para esto se partió trabajando con la tabla "ANTECED_ALUMS" para obtener a todas las personas cuyo ingreso a la facultad fuera en dicho periodo y que vinieran de un colegio nacional municipal, subvencionado o privado. Esto último se realiza porque las circunstancias de fondo de una persona extranjera pueden ser muy distintas a las de un estudiante nacional y el tipo de institución o educación recibida antes de entrar a la universidad también puede tener grandes variaciones, lo que, de considerar a estudiantes extranjeros en el estudio, puede jugar en contra al tener en cuenta que el tipo de institución de educación media de egreso del estudiante es una de las variables consideradas en el modelo original. Adicionalmente, estos datos se unen con la información de "PERSONAS" para obtener el sexo de cada uno de los alumnos y así completar la información de pre ingreso de cada estudiante necesaria para el modelo.

Procesamiento de los cursos inscritos

Por otro lado, se hace una intersección entre las tablas "PLANES_ALUMS" y "CARRERAS" de modo que se obtengan los planes relacionados con carreras de pregrado –lo que se traduce en un valor de Plan Común, Licenciatura o Título Profesional en el tipo de carrera–. Por su parte, se hace una intersección de las tablas "CURSOS", "RAMOS" y "SEMESTRES" de modo que se puedan obtener todos los cursos dictados en la facultad en un semestre de otoño o primavera (dejando afuera cursos dados en semestres de verano o con modalidad anual) con toda la información necesaria para utilizarlos más adelante. De manera similar, también se filtra la tabla "CURSOS_INSCRITOS" para obtener todas los cursos inscritos por los estudiantes de la FCFM durante o posterior al 2010, cuyo estado final haya sido aprobado

o reprobado y se procede a unirla con la información de los cursos obtenida anteriormente. Finalmente, se extrae la tabla de "CURSOS_WEB" de "UCURSOS" y se obtienen todos los cursos existentes desde el 2010 y que hayan sido dictados en la FCFM, para luego hacer la intersección con la información consolidada anteriormente de cursos y antecedentes, generando una gran tabla de datos con información de cursos, ramos, cursos inscritos y personas.

La última intersección a realizar, entre la tabla compuesta de cursos y la que contiene los planes de pregrado, tiene como objetivo mantener solo aquellos registros de cursos inscritos en el contexto de un plan de pregrado, para esto se procede primero a agrupar todos los planes existentes por alumno y a determinar cuál sería el semestre de inicio y el semestre de fin de cada uno de estos grupos. Esto se realiza porque una misma persona podría haber comenzado un plan de pregrado en la facultad y después seguido uno de postgrado, por lo que interesa dejar afuera los ramos inscritos posterior al terminó de su plan de pregrado. También, gran parte de los estudiantes en esta tabla poseen más de un plan de pregrado inscrito, debido a que el Plan Común corresponde a un plan distinto a la especialidad que después se elija. En estos casos se considera como semestre de inicio el primer semestre en que el estudiante aparece inscrito en un plan de pregrado y como semestre de fin al último semestre de sus planes existentes o un valor nulo, en caso de que haya algún plan aun en curso.

Al finalizar este proceso, se obtiene una tabla que llamaremos "CURSOS_PREGRADO" con todos los datos de cursos tomados durante un plan de pregrado que haya comenzado durante o después del 2010, junto a la información personal de cada alumno que tomo esos cursos y el estado final con el que salió del curso. En términos de dimensionalidad, esta tabla posee 23 columnas de información para 242.795 filas de cursos inscritos, compuestas por 5230 alumnos distintos eligiendo entre 1.586 ramos existentes en los 18 semestres disponibles en la base de datos (2 semestres por año desde 2010 a 2018 inclusive).

Aplicando la nueva heurística de agrupación de semestres

Aunque obtener la tabla "CURSOS_PREGRADO" tomó bastante tiempo de desarrollo mediante prueba y error producto de las múltiples llaves foráneas y bases de datos distribuidas, lo que verdaderamente es relevante para la generación del modelo no son los cursos inscritos por personas de la Facultad, sino que los ramos inscritos más de una vez por la misma persona producto de haber sido reprobados en el pasado. Para generar estas relaciones, se hace una intersección de esta última tabla con sí misma de modo que en cada fila de la intersección se encontraran dos inscripciones de un mismo ramo por la misma persona, es decir, que los valores de persona y ramo coincidieran entre las filas, pero dejando afuera aquellas intersecciones donde el semestre de la primera instancia (S1 de aquí en adelante) sea más reciente que el de la segunda (S2 de aquí en adelante, el semestre que se espera predecir), para así tener en cada fila dos inscripciones del mismo ramo por la misma persona pero en semestres distintos y con orden creciente (i.e. S2 posterior a S1). Además de lo anterior, para evitar duplicar filas de personas que hayan tomado más de dos veces un mismo ramo, se agruparon los datos obtenidos en la intersección por la tupla (persona, ramo, S1) y se dejó una sola fila por cada grupo, escogiendo esta como aquella en que el S2 de dicha fila fuera el menor entre los que existieran en el grupo.

Para ejemplificar el enredado proceso anterior se va a considerar una persona X que tomó el ramo Y en los semestres 2010-1, 2010-2 y 2011-1, donde en 2010-1 y 2010-2 reprobó y en 2011-1 finalmente aprobó. En la tabla original esto corresponde a tres filas representadas por las tuplas (X, Y, 2010-1), (X, Y, 2010-2) y (X, Y, 2011-1), ignorando el resto de los datos que no son representativos como llave primaria. Si se hace la intersección de esta tabla consigo misma uniendo las filas que compartan el valor del alumno y el ramo, este par (X, Y) generaría 9 combinaciones distintas donde las filas se pueden representar por los pares de semestres correspondientes a S1 y S2: (2010-1, 2010-1), (2010-1, 2010-2), (2010-1, 2011-1), (2010-2, 2010-1), (2010-2, 2010-2), (2010-2, 2011-1), (2011-1, 2010-1), (2011-1, 2010-2), (2011-1, 2011-2). Como existen pares con los mismos valores para S1 y S2 y distintas filas con los mismos valores de S1 y S2 pero en distinto orden, esto genera información redundante, por lo que se decide utilizar solo aquellas filas en que S1 sea anterior a S2, quedando con los pares (2010-1, 2010-2), (2010-1, 2011-1) y (2010-2, 2011-1). Es importante notar que si la persona X hubiera aprobado el curso Y en 2010-2 y no lo hubiera vuelto a tomar en 2011-1, la única fila que quedaría luego de la intersección sería (2010-1, 2010-2), ya que 2011-1 no existiría como fila en primer lugar, por lo que el procesamiento para personas que solo tomaron dos veces cada ramo terminaría en este punto.

Para las personas que tomaron más de dos veces algún ramo (como el estudiante X) aun se deben filtrar filas, debido a que tanto (2010-1, 2011-1) y (2010-2, 2011-1) comparten el valor de S2 y por tanto intentarían predecir el resultado para el mismo semestre (considerando que el semestre de la derecha en cada par es la instancia a la que se quiere predecir la doble reprobación). Por este motivo se realiza una segunda agrupación de los datos dada por la tupla (persona, ramo, S1), con lo que se generarían dos grupos para el estudiante X: (X, Y, 2010-1) y (X, Y, 2010-2), donde el primer grupo tendría los valores de S2 2010-2 y 2011-1 y el segundo grupo solo tendría la fila con S2 correspondiente a 2011-1. Entonces, en cada grupo se deja la fila correspondiente al menor valor de S2 y se eliminan las otras, dejando así las filas deseadas: (X, Y, 2010-1, 2010-2) y (X, Y, 2010-2, 2011-1). De este modo, cada fila representa una instancia en que se tomó un ramo reprobado y la instancia inmediatamente anterior en que se tomó el mismo ramo.

Todo el proceso realizado hasta este punto genera una tabla que llamaremos "CURSOS_PLANES" con las instancias de personas que tomaron un mismo ramo más de una vez, donde cada fila se puede identificar de manera única por la tupla (alumno, ramo, S1 y S2). Dicha tabla posee 14.179 filas de personas tomando ramos reprobados en el pasado, con información de 3.238 alumnos distintos en 317 ramos diferentes.

Agrupando los cursos por semestres

En este punto quedan dos pasos por realizar: obtener los cursos correspondientes a los "aprobados" y "reprobados" de cada S1 y S2 para todos los alumnos tomando ramos reprobados, según los criterios explicados en la sección "Procesamiento de los datos" de 4.1.1, y obtener las notas asociadas a dichos cursos para calcular notas parciales, notas finales y los distintos atributos necesitados para el desarrollo del modelo.

En primer lugar, para obtener los cursos correspondientes al S1 y S2 de cada alumno, se

procede a agrupar las filas de la tabla generada "CURSOS_PLANES" por los valores de la tupla (alumno, S2). En cada uno de estos grupos, los cursos inscritos en S2 corresponden a los ramos "reprobados" en S2, mientras que los cursos inscritos en S1 (donde el semestre en específico en que se tomaron estos ramos puede variar) corresponden a los ramos "reprobados en S1". Con esto, solo queda obtener los cursos "aprobados" para S1 y S2.

Los cursos "aprobados en S1" corresponden a todos los cursos tomados el semestre inmediatamente anterior a S2 y que no fueron reprobados en esa ocasión, mientras que los cursos "aprobados en S2" son en verdad, todos los cursos que se están tomando en S2 pero que es primera vez que son inscritos, i.e. que nunca habían sido reprobados en el pasado. Ambos grupos de cursos se obtienen recorriendo la tabla "CURSOS_PLANES" agrupada por (alumno, S2), de modo que solo se busquen los cursos para alumnos y semestres relevantes para el modelo, y luego utilizando la tabla creada anteriormente, "CURSOS_PREGRADO", para obtener los cursos que cumplen con los criterios recién descritos.

Teniendo estos grupos de cursos definidos para cada estudiante, se decidió crear una tabla intermedia que fuera almacenada directamente en "MUFASA" para guardar esta información de manera más segura y que no desapareciera al término de la ejecución de un script. Esto se decidió hacer por varios motivos: en primer lugar, para agilizar el trabajo futuro, ya que el procesamiento de todos los datos para obtener estos grupos de cursos toma un tiempo considerable y no es conveniente estar ejecutando todo el proceso de nuevo cada vez que se quiera cambiar algo en el modelo, cambiar la lógica, corregir un error, etcétera; por otro lado, al ser datos históricos es muy poco probable que vayan a ser modificados en el futuro, por lo que tampoco tiene mucho sentido estar gastando poder de cálculo con datos que siempre se van a mantener iguales; lo anterior, de todas maneras, implica el riesgo de que cambie la información para algún momento dado, sin embargo, estas actualizaciones podrían ser tan pocas que no importaría tener que ejecutar todo el proceso de nuevo. En esta tabla, bautizada "ALERTA_TEMPRANA_CURSOS", se almacenan entonces las siguientes columnas:

- "PERSONA": ID del alumno en la base de datos de "MUFASA". Para la mayoría de los alumnos nacionales corresponde a su rut sin dígito verificador.
- "SEMESTRE": Semestre al que corresponde la observación y al cual se quiere analizar la probabilidad de doble reprobación. En la forma YYYYP, donde YYYY corresponde al año y P al periodo (1 otoño y 2 primavera).
- "CRS2": Cursos "reprobados en S2".
- "CNRS2": Cursos "No reprobados en S2".
- "CRS1": Cursos "reprobados en S1".
- "CNRS1": Cursos "No reprobados en S1".
- "Reprueba": Función objetivo. Su valor es 1 si alguno de los ramos en CRS2 fue reprobado al finalizar el semestre y 0 en caso contrario

Al finalizar el procesamiento y carga de estos datos a "MUFASA", quedaron cerca de 9.000 observaciones distintas sobre más de 3.000 alumnos en algún semestre de sus carreras de pregrado, donde 1.721 de estas filas corresponden a alumnos que volvieron a reprobar un mismo ramo en el semestre analizado.

Obtención de las notas

Posterior a la creación de la tabla "ALERTA_TEMPRANA_CURSOS" se procede con la obtención y procesamiento de las notas parciales y finales que se necesitan para el desarrollo del modelo. En una primera instancia se quiso procesar todas las notas de todos los cursos necesitados a la vez, pero las 40 millones de filas y más de 1.5 Gb de información contenida en la base de datos "UC_NOTAS" no hicieron fácil esta tarea.

Al ser tan grande la base de datos, fue imposible almacenarla completa en memoria principal para procesarla rápida y eficientemente, mientras que el intentar realizar una consulta a la base de datos por toda la información necesaria se demoraba horas, en las cuales algo podría pasar que cancelara el cálculo o impidiera el traspaso de datos entre el programa y el servidor, obligando a tener que realizar absolutamente todo el cálculo de nuevo. Estos son otros motivos para haber creado la tabla intermedia "ALERTA_TEMPRANA_CURSOS", ya que permite recorrer los pares (alumno, semestre) y obtener para cada uno las notas correspondientes a los grupos de cursos en CRS1, CRS2, CNRS1 y CNRS2. Así, el proceso de obtención de las notas se realizó siguiendo los siguientes pasos:

1. Se guarda la información de la tabla "ALERTA_TEMPRANA_CURSOS" en memoria principal y se recorren una a una las filas de la tabla
2. Para cada fila (correspondiente a un alumno en un semestre) se buscan las notas de "control 1" para cada uno de los grupos. Esto se realiza con una consulta a la base de datos para cada grupo, de modo que primero se consultan todas las notas de "control 1" para los cursos de CRS1 del estudiante, luego para CRS2, CNRS1 y, finalmente, CNRS2.
3. Los valores de notas C1 obtenidos para cada grupo son sumados y divididos por la cantidad de cursos en el grupo, de modo que se calcule el promedio de notas para cada uno.
4. Se repite el mismo proceso pero para obtener las notas finales de cada grupo.

Algo importante de aclarar es que se utilizaron las mismas expresiones regulares mostradas en la sección anterior (las expresiones 4.1 y 4.2) para obtener las notas de control 1 y notas finales, respectivamente. Se decidió seguir utilizando solo las notas de controles y no las de tareas u otras evaluaciones, puesto que evaluar el impacto de estas últimas implica un trabajo investigativo y de procesamiento de los datos bastante grande, ya que se debe decidir qué hacer si existen tareas y no controles o viceversa, si estas se deben ponderar de la misma manera que los controles, cómo trabajar con cursos que cambian la modalidad de evaluaciones, entre otras cosas. Para poder tener un modelo funcional con la mayor cantidad de alumnos de pregrado posible, pero dentro del plazo destinado para hacerlo, se decidió simplemente utilizar las notas de control 1 y dejar fuera del estudio a cursos que no tuvieran estas notas. De manera similar, también se dejan afuera a personas que en CNRS1 o CNRS2 no tengan ningún curso inscrito, es decir, que no hayan aprobado ningún ramo el semestre anterior o que no hayan tomado ningún ramo en S2, ya que no tendrán la información suficiente para ser procesados por el modelo y ya se encuentran en una situación de riesgo distinta de por sí.

Otra decisión que se tomó en torno a las notas, fue la de mantener el cálculo de la nota final como el promedio de notas de controles a lo largo del semestre. Esto se mantuvo así porque

de esta manera se simplifica el procesamiento de los datos al dejar fuera variables como la ponderación de los controles en la nota final, la existencia de cursos con distinta cantidad de controles, cursos en que el examen pondera distinto y personas que mejoraron o empeoraron mucho su rendimiento a la hora de dar el examen final. También, este procedimiento permite mantener el mismo estándar entre cursos aprobados y reprobados, ya que, como se explicó anteriormente, en la base de datos solo se encuentra la nota final del estudiante para un curso si este fue aprobado, almacenando un valor "NULL" en caso contrario.

Todo el proceso realizado hasta este momento demora cerca de 3 o 4 horas en entregar los datos procesados, por lo que se creó otra tabla intermedia para almacenar cada una de las filas procesadas con los valores listos para ser utilizados para el entrenamiento y validación del modelo. Tal como se explicó al justificar la tabla "ALERTA_TEMPRANA_CURSOS", esta tabla posee información histórica que es bastante poco probable que cambie, pero que si lo hace, pasaría tan poco que no sería tan grave tener que realizar el proceso de nuevo.

Resultados del procesamiento

La tabla creada final, bautizada "ALERTA_TEMPRANA_DATOS", posee 10 columnas, donde la llave primaria de cada fila serían las columnas "PERSONA" y "SEMESTRE", la columna "REPRUEBA" representa la función objetivo, siendo 1 si la persona volvió a reprobar un ramo y 0 si logró aprobar todos y las otras 7 representan los atributos que necesita el modelo para funcionar:

- SEXO: Género del estudiante.
- TASA: Tasa de créditos reprobados en el semestre anterior.
- PARTICULAR: Es 1 si el alumno egresó de un colegio particular, 0 en caso contrario.
- SUBVENCIONADO: Es 1 si el alumno egresó de un colegio subvencionado, 0 en caso contrario.
- C1IRS2diff4: Valor calculado como el promedio de los controles 1 de los ramos reprobados en el semestre analizado menos la nota de aprobación del curso (4.0).
- C1IRS2ltFIRS1: Valor booleano obtenido de la comparación entre la nota de los controles 1 de los ramos reprobados en el semestre a analizar contra las notas finales de estos ramos la vez pasada que fueron reprobados. El valor de la columna es 1 si el primero es menor y 0 en caso contrario.
- C1INRS2ltC1INRS1: Valor booleano obtenido de la comparación entre la nota de los controles 1 de los ramos no reprobados en el semestre a analizar y la nota de los controles 1 del semestre anterior. El valor de la columna es 1 si el primero es menor y 0 en caso contrario.

Cabe destacar que en el nuevo set de datos si existía la información del establecimiento de educación media de egreso de los estudiantes y que, para poder guardar los nombres de las etiquetas como nombres de columnas, se cambiaron los signos "<" por "lt" y el signo "-" por "diff" con respecto a las etiquetas creadas por Celis et al. (tabla 2.1). Con esta información procesada y lista para ser utilizada, solo queda entrenar el modelo y validar si aun sigue siendo vigente.

Generando los conjuntos de entrenamiento y validación

Para el entrenamiento del modelo se obtuvieron los datos almacenados en la tabla "ALERTA_TEMPRANA_DATOS" y se separaron en conjuntos de entrenamiento y validación. En un inicio, el conjunto de entrenamiento fue conformado por todas las filas correspondientes a semestres comprendidos entre los años 2010 y 2017 (incluidos), conformado por 11.675 observaciones, mientras que los datos del año 2018 fueron utilizados para la validación del modelo, siendo estos 520 del total de 12.202.

Antes de ingresar los datos del 2010 al 2017 al modelo se balancearon parcialmente las clases, ya que la proporción de datos de estudiantes que reprobaron nuevamente en cada semestre era cercana a tan solo un séptimo del total de los datos utilizados (1.817 observaciones del total de 12.202) e ingresarlos tal cual al modelo podría generar un sesgo indeseado hacia la predicción de la clase "no reprueba" debido a su predominancia en los datos. Se dice que es un balanceo parcial ya que no se ingresaron la misma cantidad de datos para cada uno como se haría normalmente, si no que, para mantener un poco esta relación de presencia en los datos, los elementos de la clase "reprueba" se ingresaron completos mientras que se eligió una cantidad arbitraria de valores de la clase "no reprueba".

Para determinar cuántos elementos de la clase predominante ingresar al modelo, se realizó una validación cruzada utilizando el conjunto de entrenamiento (dejando afuera al conjunto de validación) con distintas proporciones de elementos. Se probó ingresando la misma, dos, tres, cuatro y cinco veces la cantidad de datos "no reprueba" con respecto a los "reprueba", además de mantener el conjunto de entrenamiento sin alteraciones, y se determinó que los mejores resultados se obtenían al tener cinco veces más elementos de la clase "no reprueba" versus "reprueba". Esto significó que en el conjunto de entrenamiento final se encontraban 1.712 representantes de la clase "reprueba" y 8.560 representantes de la clase "no reprueba" escogidos al azar del total, en vez de las 9.963 que existían originalmente.

Otro punto importante de acotar es que tanto los datos de entrenamiento como los de validación fueron normalizados para las variables no binarias, con el fin de evitar un sesgo producto de "outliers" que podrían estar presentes en cada conjunto. Dicha normalización se llevó a cabo utilizando el método "min_max_scaler", incluido con la librería Scikit-learn, y realiza la transformación siguiendo la ecuación 4.3 para obtener un valor entre 0 y 1 para cada elemento x de un conjunto X.

$$x_{scaled} = \frac{x - min(X)}{max(X) - min(X)} \tag{4.3}$$

Entrenando los modelos

Una vez realizadas las últimas alteraciones al conjunto de entrenamiento se procedió a ingresar este al algoritmo de Regresión Logística, a una Red Neuronal y a una SVM. Se decidió probar con los tres algoritmos a la vez para poder identificar cuál entregaba un mejor

resultado y comparar directamente el comportamiento de los tres frente al modelo original de Celis et al. y a la primera expansión que se hizo del modelo. Al tener los datos procesados y listos para ingresar en la tabla "ALERTA_TEMPRANA_DATOS", el único tiempo extra utilizado para entrenar estos otros algoritmos es el que demora cada uno en generar el modelo predictor. En las figuras 4.8, 4.9 y 4.10 se pueden observar las curvas ROC obtenidas por los tres algoritmos, indicando que, al menos, el resultado obtenido por cada algoritmo es mejor que un algoritmo de predicción aleatoria.

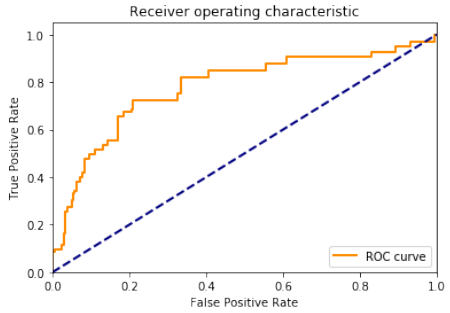

Figura 4.8: Curva ROC Regresión Logística

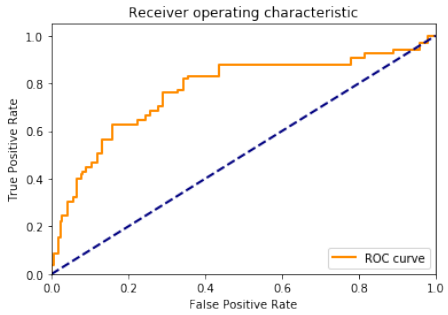

Figura 4.9: Curva ROC Red Neuronal

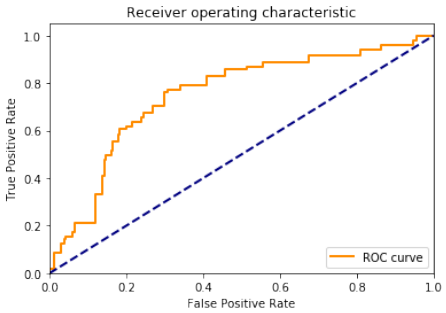

Figura 4.10: Curva ROC SVM

El conjunto de validación utilizado, correspondiente a las observaciones de los semestres de otoño y primavera de 2018, contiene 520 observaciones, de las cuales 415 corresponden a la clase "no reprueba" y 105 a "reprueba". Al igual que en la primera expansión del modelo, las pruebas con estos algoritmos se realizaron obteniendo la probabilidad de pertenencia a la clase "reprueba" y estableciendo el umbral de decisión en el 19 % por los motivos ya explicados en la sección 2.3.

Resultados de la Regresión Logística

Para el algoritmo de Regresión Logística, las figuras 4.11 y 4.12 muestran los resultados obtenidos en la matriz de confusión sin normalizar y normalizada, respectivamente. En estas se observa que el 72 % de las personas que iban a reprobar fueron identificadas correctamente, mientras que tan solo el 24 % de las personas que lograron aprobar sus ramos fueron incorrectamente clasificadas. El mapeo de colores entre ambas matrices permite observar que, si bien la mayoría de los elementos fueron clasificados correctamente como "no reprueba", en

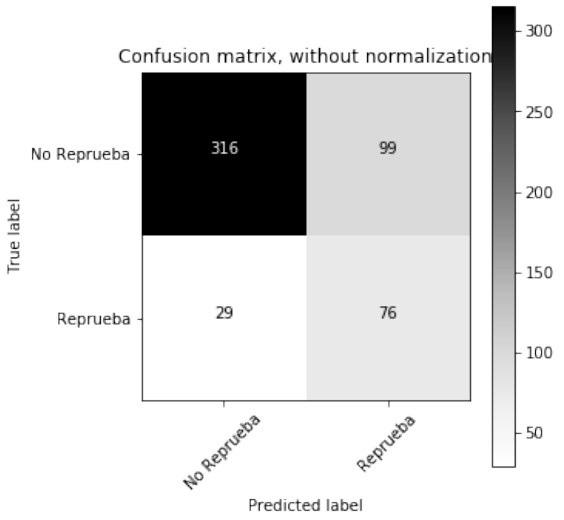

Figura 4.11: Matriz de confusión sin normalizar para Regresión Logística

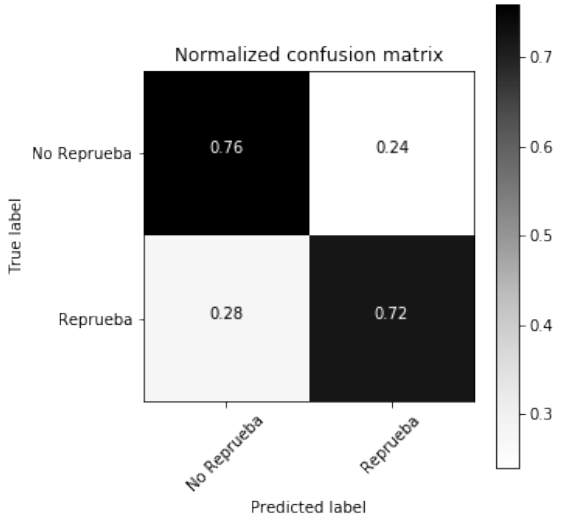

Figura 4.12: Matriz de confusión normalizada para Regresión Logística

Variable	Precision	Recall
Reprueba	0,43	0,72
No Reprueba	0,92	0,76

Tabla 4.1: Métricas de rendimiento obtenidas para la Regresión Logística.

términos porcentuales no es menor el impacto de la correcta predicción de la clase "reprueba" debido a la menor cantidad de datos asociadas a esta clase.

En la tabla 4.1 se pueden observar los resultados obtenidos para las clases "reprueba" y "no reprueba", siendo la primera de estas la más relevante para la experiencia. En la figura 4.13 se pueden observar de forma gráfica los resultados del algoritmo. Aquí se puede observar cómo se distribuyen los datos a través de la distribución de probabilidades asignadas por el modelo según las etiquetas originales y la ubicación relativa del umbral entre los datos clasificados.

Resultados de la Red Neuronal

En el caso de la Red Neuronal, para el entrenamiento de esta se utilizó una configuración de dos capas ocultas "fully connected" de 30 y 15 neuronas cada una, utilizando una función de activación "RELU" (*Rectified Linear Unit*), un solver optimizador basado en gradientes estocástico, una tasa de aprendizaje constante y un valor alpha de 0,0001.

Las matrices de confusión obtenidas por este algoritmo se pueden observar en las figuras 4.14 y 4.15, mientras que las métricas de rendimiento se encuentran en la tabla 4.2 y la figura 4.16 contiene la distribución de los elementos de cada clase en las probabilidades de reprobación asignadas por la red.

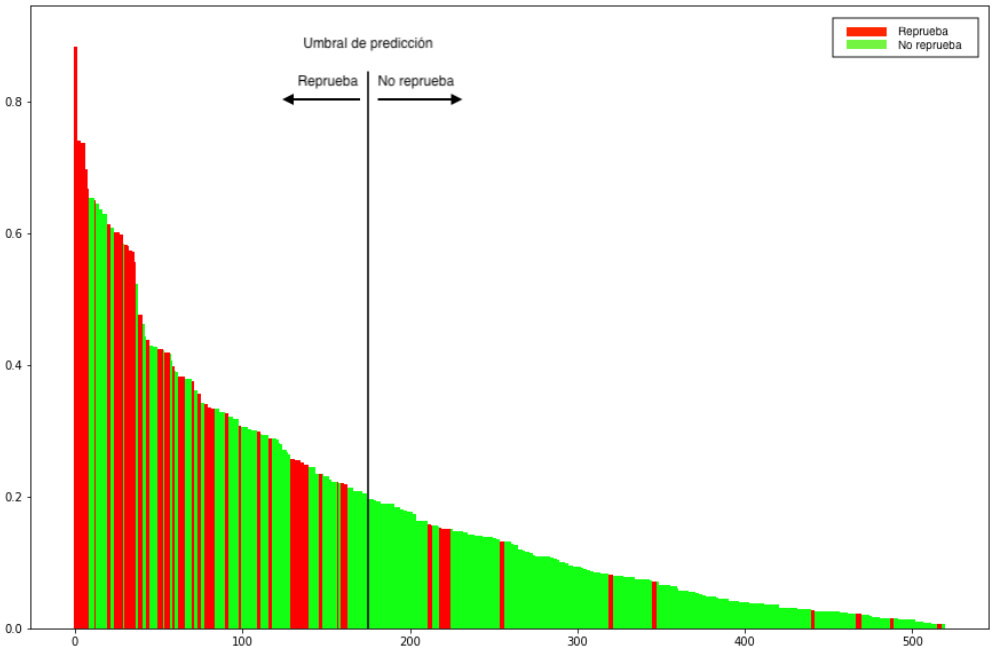

Figura 4.13: Resultados obtenidos para Regresión Logística.

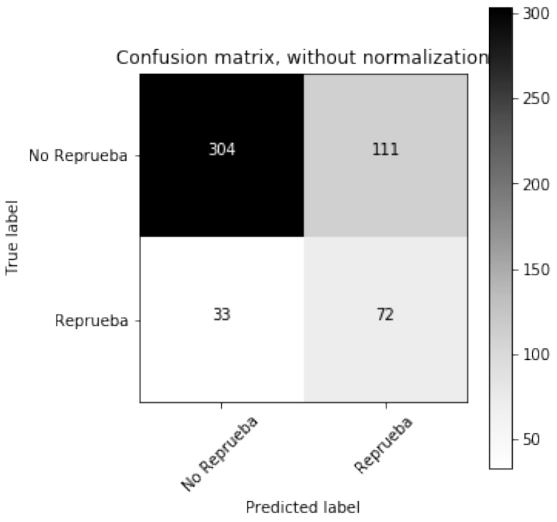

Figura 4.14: Matriz de confusión sin normalizar para Red Neuronal

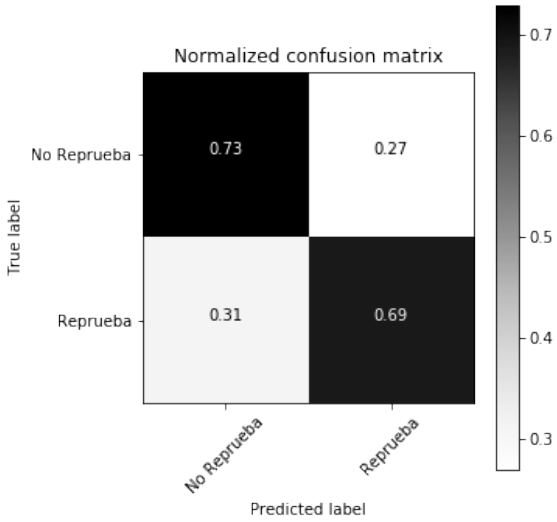

Figura 4.15: Matriz de confusión normalizada para Red Neuronal

Variable	Precision	Recall
Reprueba	0,39	0,69
No Reprueba	0,90	0,73

Tabla 4.2: Métricas de rendimiento obtenidas para la Red Neuronal.

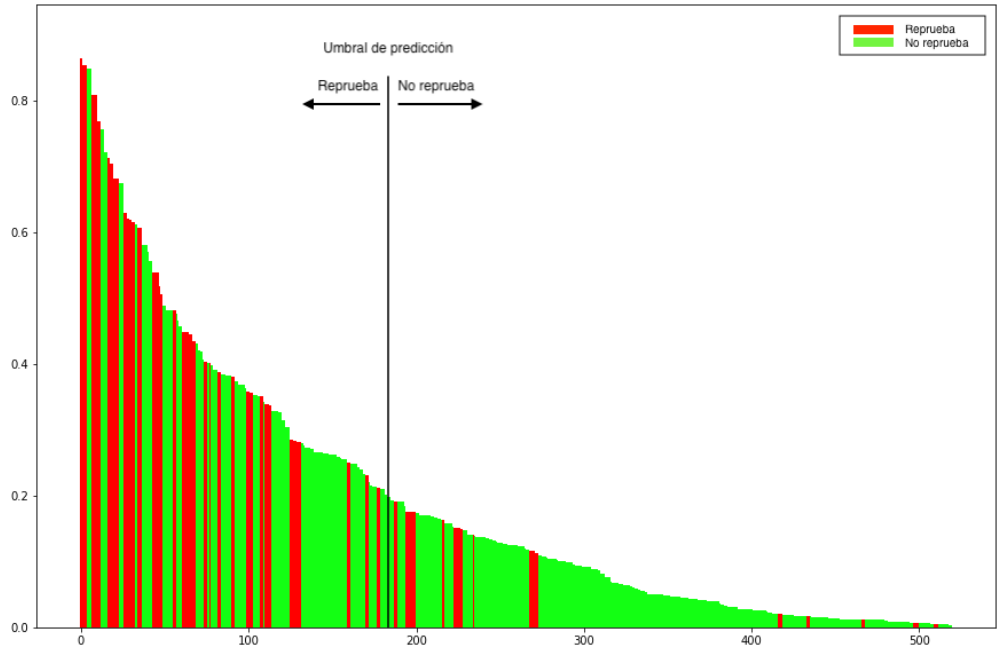

Figura 4.16: Resultados obtenidos para Red Neuronal.

Variable	Precision	Recall
Reprueba	0,40	0,21
No Reprueba	0,82	0,92

Tabla 4.3: Métricas de rendimiento obtenidas para la SVM.

Resultados de la SVM

Por último, el algoritmo de SVM fue probado con varias combinaciones de valores para C (1, 10, 100 y 1000) y Gamma (0,0001, 0,001, 0,01 y 0,1) utilizando validación cruzada sobre el conjunto de entrenamiento, eligiendo la que entregó mejores resultados entre todas con un valor C de penalización de 100, Gamma de 0.01 y un kernel RBF. Entre todos los algoritmos probados, este fue el que más tiempo demoró en el entrenamiento, tardando cerca de un minuto en realizar todo el proceso de entrenamiento y validación mientras que a los otros dos no les tomó más de 10 segundos.

Las matrices de confusión generadas por el modelo SVM escogido corresponden a las figuras 4.17 y 4.18, notando inmediatamente que este es el único algoritmo hasta ahora que entrega un mayor porcentaje de elementos de la clase "reprueba" clasificados incorrectamente que aquellos clasificados como deberían. La tabla 4.3 contiene los resultados para *recall* y *precision*, mientras que la figura 4.19 contiene la distribución de los elementos según su probabilidad de pertenencia a la clase "reprueba" (también muy distinto a los otros gráficos).

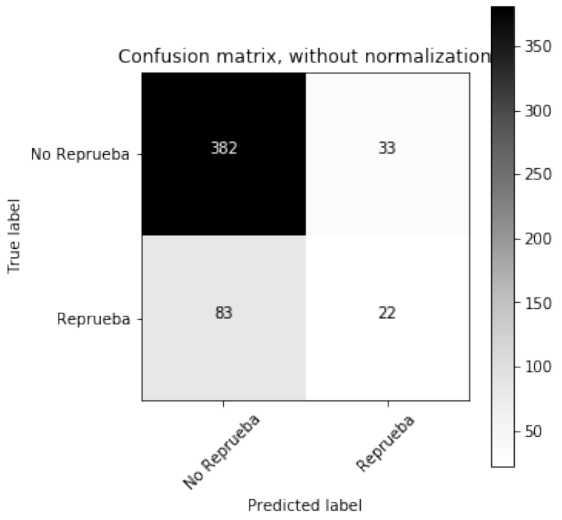

Figura 4.17: Matriz de confusión sin normalizar para SVM

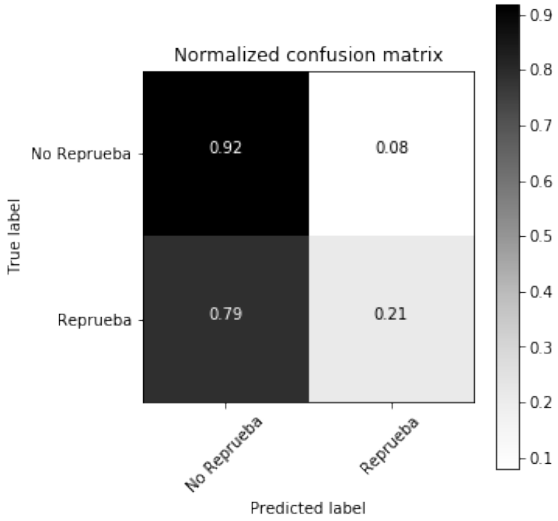

Figura 4.18: Matriz de confusión normalizada para SVM

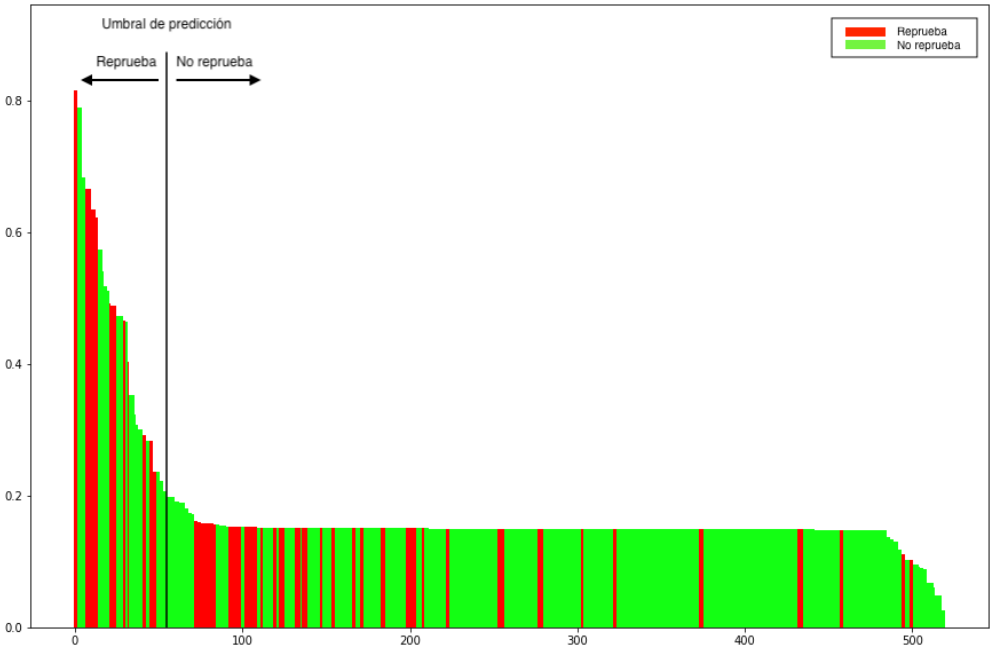

Figura 4.19: Resultados obtenidos para SVM.

Comparando los resultados de los algoritmos entrenados

Tal como se observa a primera vista en los resultados obtenidos, los resultados actuales fueron peores que aquellos obtenidos por el modelo original de Celis et al. y los obtenidos en la primera modificación que se hizo en este desarrollo. Gran parte de este comportamiento se puede explicar por las modificaciones realizadas al universo de datos utilizados y a las diferencias existentes en el procesamiento de la información.

En cuanto a los algoritmos utilizados, se puede observar en las tablas 4.1, 4.2 y 4.3 que de los tres modelos probados, aquel que mostró levemente mejores resultados fue la Regresión Logística, seguido de cerca por la Red Neuronal. Entre estas dos, distando en tan solo 3 puntos porcentuales entre sí para la métrica de *recall*, se puede observar que la distribución de las probabilidades predichas fueron bastante similares, pero la Regresión Logística logró concentrar a una mayor cantidad de observaciones "reprobado" en el lado alto del espectro de probabilidades. También se puede notar que la Regresión Logística mantuvo una menor cantidad de elementos sobre el umbral de predicción pero llegando hasta (aproximadamente) el mismo valor máximo predicho, lo que indicaría una mayor precisión a la hora de elegir qué elementos pertenecen a cada clase, al entregar una menor varianza entre las probabilidades de cada elección.

Por otro lado, los malos resultados obtenidos por la SVM parecen indicar que no existe un hiperplano generalizable para los datos que se están tratando de predecir, dado que la increíble mayoría de estos se clasifica como "no reprueba" de manera indistinta. Estos resultados, mezclados con la baja en rendimiento en los otros modelos, podría indicar que los datos pasados no son tan buenos predictores en el largo plazo o de manera tan generalizada como se pensó en un principio o, por otro lado, que no existe una manera definitiva de separar las dos clases y que mantienen una unión intrínseca a su naturaleza.

Análisis de los resultados del modelo ganador (y cómo se comparan con el modelo original)

La tabla 4.1 indica que el *recall* obtenido para estos resultados corresponde al 72 % mientras que el *precision* fue de 43 %. Si esto se compara con los resultados originales de Celis et al. es fácil notar que el *recall* de 86 % obtenido por ellos supera el obtenido en esta ocasión y la *precision* de Celis et al. de 37,5 % se encuentra levemente por debajo del 43 % obtenido en esta ocasión.

Dadas las increíbles diferencias entre los conjuntos de datos utilizados y la naturaleza de los mismos, esta discrepancia en los resultados puede no ser tan considerable como parece a primera vista. Si se tienen en cuenta factores como que el conjunto de validación de Celis et al. estaba conformado por un tercio de los datos utilizados en esta ocasión y que los cursos analizados fueron los 5 cursos que se dictan en primer semestre, versus todos los cursos dictados en la FCFM, no es difícil imaginar que el modelo iba a disminuir su poder predictivo si se mantenían las mismas variables analizadas.

A pesar de lo anterior, el aumento en la tasa de *precision* con respecto al modelo de Celis

et al. parece indicar que una mayor cantidad de datos y la generalización del modelo podrían ayudar a identificar de mejor manera a aquellas personas que se encontraban en peligro de reprobar pero lograron revertir su situación.

Comparación de los resultados con la iteración anterior

Al comparar los resultados actuales con los obtenidos en la primera iteración, los de esta ocasión son claramente inferiores, perdiendo en *recall* con un 96 % contra un 72 % y en *precision* con un 55 % versus un 43 %. La decisión de mantener el modelo tal cual estaba se tomó al observar los buenos resultados obtenidos en la primera extensión del modelo, pero si se comparan las figuras 4.4 y 4.13 se puede notar un claro sesgo de los datos utilizados para la primera iteración del desarrollo. Por ejemplo, se nota claramente que el umbral de decisión se encuentra mucho más a la derecha en la primera imagen, puesto que existen, de por sí, muchos más datos de la clase "reprueba" en el subconjunto de datos utilizado. Si bien se consideró en su momento que esto podría ser un problema al estar mostrando una realidad distinta a la existente, no se tenían las herramientas e información necesarias para demostrar o refutar la hipótesis planteada, derivando en los resultados que se muestran ahora.

Al existir una clara distorsión de la realidad en los datos utilizados para la primera iteración, se puede descartar que sus resultados sean verdaderos predictores de la calidad que debería presentar el algoritmo, por lo que el "deterioro" en las métricas de rendimiento entre ambas iteraciones se puede considerar poco relevante y no trae consigo consecuencias para la creación de nuevos modelos en el futuro.

4.1.3. Integrando estadísticas de uso de U-Cursos

Luego de la creación de este último modelo, se decidió estudiar y tratar de procesar e incorporar estadísticas de uso de U-Cursos según el estudio realizado por López et al. [17]. Si bien esto no estaba considerado en el análisis inicial del proyecto, se estimó que podría resultar beneficioso para el sistema y se procedió a intentar obtener información que podría ayudar a mejorar el modelo creado.

El 14 de mayo de 2019 se llevó a cabo una reunión entre López, Silva (co-autores de [17]) y los participantes de este desarrollo, para así entender un poco mejor cuál era el objetivo del estudio realizado y cómo iba progresando desde que se presentaron resultados de manera pública por primera vez, aparte de aclarar algunas dudas con respecto a la fiabilidad de los resultados mostrados.

Este estudio consistía en analizar una variedad de variables que podrían afectar el rendimiento académico de alumnos de primer año en distintas instituciones. Dentro de las variables analizadas se encontraban algunas representando el uso que le dan los estudiantes a las herramientas de LMS como U-Cursos.

En la reunión se mostraron algunos resultados diferentes a los seleccionados para la presentación realizada en el Departamento de Ciencias de la Computación de la FCFM [17],

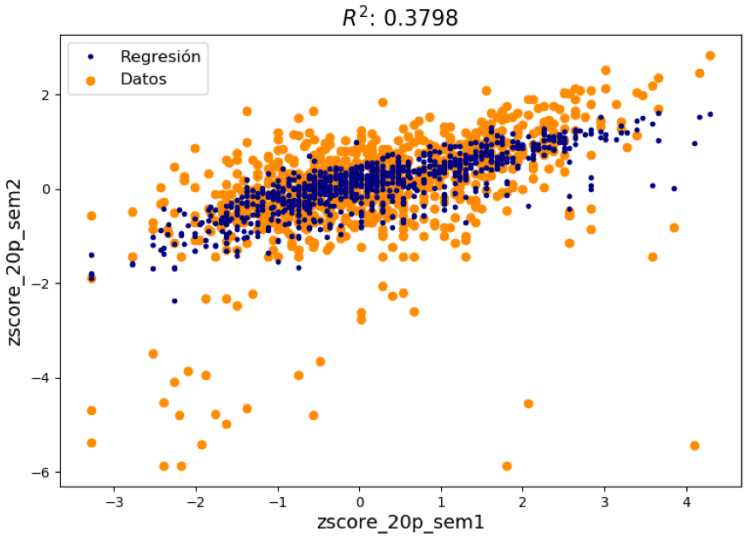

Figura 4.20: Resultados de la Regresión Logística

que permitían observar el verdadero poder predictivo de las variables obtenidas por ellos al comparar la correlación entre un desempeño académico generado a partir de las estadísticas de uso del LMS (junto a otras variables) y el desempeño real obtenido por los estudiantes. En la figura 4.20 se puede observar el resultado de una de las regresiones realizadas por ellos, donde se puede encontrar una baja dispersión de los datos predichos sobre los reales, consolidando la idea de que las variables identificadas por ellos en la figura 2.2 podrían ser una incorporación positiva para el modelo, por lo que se consideró relevante tratar de procesarlos de la misma manera e integrarlos a la lógica de predicción realizada.

Exploración de los datos

El conjunto de datos entregado por el Centro Tecnológico UCampus para poder hacer las primeras pruebas de procesamiento de los datos e implementarlos en el modelo, correspondía al registro de todos los movimientos y acciones realizadas en la plataforma U-Cursos durante el año 2018 y una breve parte del 2019 incluidos en la base de datos "UC_LOGS". Dicha base de datos estaba compuesta por 1.322 tablas distintas, donde cada una contenía los registros asociados a un módulo (o conjunto de) en específico en uno de los meses de 2018. Para cada mes de 2018 existían más de 80 tablas distintas, conformando en total más de 145 Gb de información contenida en más de 570 millones de filas.

La mayoría de las tablas estaban compuestas por las siguientes columnas:

- "GRU_ID": Identificador del módulo sobre el que se realizó la acción. El más relevante en este caso son los de la forma "curso.xxxx", donde "xxxx" representa el id de curso web donde se realizó la acción.
- "PERS_ID": Identificador del usuario que realizó la acción. Usualmente es el RUT sin el dígito verificador.

- "OPERACION": Tipo de operación realizada.
- "OBJ_ID": Objeto sobre el que se interactuó en el módulo.
- "SESSION": Identificador de la sesión bajo la cual se realizó la acción
- "FECHA": Fecha de la acción.

La información más relevante que se puede obtener de acá es el volumen de acciones realizadas por cada módulo. Tal como se explicó al inicio de la sección, lo que se buscó era recrear las variables generadas por López et al., las cuales se obtenían a partir de la frecuencia de uso semanal a lo largo del semestre, es decir, la cantidad de acciones registradas en la aplicación agrupadas por la semana en que se realizaron para cada alumno estudiado. Las variables a extraer de aquí serían entonces:

- Desviación estándar: Desviación estándar en las frecuencias de acceso por semana a lo largo del semestre.
- Centro de masa: Cálculo utilizado para obtener una estimación de la semana en el semestre en que más se concentra la actividad del estudiante.
- Media: Media de las frecuencias de acceso semanal.
- RP: Indicador que busca identificar las semanas de "trabajo real" en el semestre.

Procesamiento de los datos

Claramente, toda esta información en la base de datos "UC_LOGS" no se encontraba lista para utilizar directamente, por lo que se tuvo que realizar un pre-procesamiento de la información para obtener las frecuencias por semana de uso. Al ser una cantidad de datos mucho más grande que todos los usados hasta este momento, fue claro que la realización de una tabla intermedia para guardar la información relevante era indispensable, ya que de otro modo no se podría almacenar toda la información en memoria o se perdería y se debería procesar constantemente. A diferencia de las otras tablas intermedias, esta nunca debiera cambiar, ya que los registros creados en el pasado nunca van a ser alterados (por más que se quiera, las acciones del pasado no se pueden deshacer).

La primera parte de este procesamiento, entonces, consistió en utilizar la tabla "ALERTA_TEMPRANA_CURSOS" generada anteriormente para poder identificar los cursos sobre los cuales se debería consultar a la base de datos por su uso. Al recorrer las filas de dicha tabla se iba consultando a la base de datos de logs por la información relacionada al estudiante en cuestión para cada uno de sus cursos. Dado que se necesitaba realizar el conteo de uso por semanas, para cada fila de la tabla "ALERTA_TEMPRANA_CURSOS" se realiza una consulta a la tabla "SEMANAS" (de la base de datos MUFASA) para obtener las fechas de inicio de cada una de las 17 semanas que componen un semestre académico en la FCFM (15 de clases y 2 de exámenes).

Para cada semana del semestre correspondiente a la fila analizada se consultó por los logs existentes en cada uno de los cursos que posee, se contaron, sumaron y se fueron guardando en una tabla llamada "PROCESSED_LOGS" para luego acceder rápidamente. Lamentablemente, debido a la increíble cantidad de datos existentes, el tiempo que se tenía planificado

para realizar este procesamiento y la posterior selección de atributos junto a los utilizados en los modelos anteriores era mucho menor que el tiempo real que estaba tomando realizar todo este trabajo, por lo que se tuvo que abandonar esta parte del trabajo a medio camino.

Abandonando el uso de logs

El principal motivo por el que se abandonó el uso de logs para intentar mejorar el modelo creado fue la excesiva cantidad de tiempo que tomaría procesar toda la información disponible y realizar el proceso posterior de selección de atributos. Para dar una referencia del tiempo utilizado, la obtención de los logs para unas 100 filas de la tabla "ALERTA_TEMPRANA_CURSOS" llevó más de dos horas, representando estos datos cerca del 1 % del total a procesar. Esto quiere decir que, de seguir la misma tendencia, hubiera tomado más de 10 días seguidos de procesamiento para obtener la información de todas las observaciones en la tabla si es que no se encontrara algún problema en el camino.

Por otro lado, incluso terminando el procesamiento de los logs el trabajo estaría lejos de haber terminado, ya que se hubiera tenido que generar las métricas indicadas en la sección anterior (desviación estándar, centro de masa, media, RP) y luego se tendrían que agrupar de cierta manera para generar variables, como las utilizadas en el modelo actual, que permitan comparar un semestre con otro. Posterior a esto, las variables generadas deberían agregarse al modelo y comparar sus resultados con los obtenidos anteriormente. En caso de no tener resultados satisfactorios, o de querer mejorar los obtenidos, probablemente habría que realizar un proceso de selección de atributos como el que realizó Celis et al. cuando hicieron el modelo original, entre otras tareas que tomarían una cantidad considerable de tiempo. Solo en este punto se podría considerar como terminada esta parte del desarrollo, por lo que la posibilidad de utilizar estos datos quedó como un posible trabajo futuro a realizar debido al tiempo acotado de desarrollo del proyecto.

Si bien, según el estudio de López et al., existe una relación entre el uso del LMS y el rendimiento académico, no se sabe de qué manera se pueden usar mejor estos datos, si se deben usar de manera pura, comparando estadísticos, comparando entre semestres, comparando entre cursos reprobados y no reprobados, etcétera. A pesar de esto, el uso de logs como información externa a la utilizada por Celis et al. podría haber entregado la información necesaria para subsanar la baja en rendimiento, producida por el aumento en la cantidad de datos distintos recibidos por el modelo con respecto al universo original, sobretodo considerando que entregaría información más específica en cursos que no se ajustan completamente a los estándares que espera el modelo para funcionar correctamente.

Teniendo en cuenta todos los motivos antes expuestos, se decidió priorizar la realización de las interfaces para entregar la información generada. Incluso si se lograra hacer un modelo perfecto, que permitiera predecir exactamente a todas las personas que van a reprobar un ramo, este no serviría de nada si la información no se mostrara y entregara a las personas apropiadas para tomar acciones al respecto. Sin embargo, la adición de estos datos sería absolutamente interesante de agregar y será considerado como un trabajo futuro que se podría realizar en otra iteración de este proyecto.

4.2. Elaboración de las interfaces

El proceso de elaboración de las interfaces se realizó en dos partes. En primer lugar, se realizaron mockups de las interfaces que se planeaban realizar con la información que debería estar contenida en estas. Dichos mockups fueron validados utilizando opiniones de gente cercana al proyecto y sirvieron como la base para crear las interfaces finales, correspondiendo esto a la segunda parte.

Se decidió crear las interfaces pensando en su implementación como un módulo de U-Cursos, de modo que la información se encuentre disponible para profesores, coordinadores, académicos, investigadores y otras personas relacionadas con la administración de docencia de la institución, siendo estos los usuarios finales pensados para las interfaces realizadas.

4.2.1. Creación de las interfaces finales

La primera decisión que se tomó fue implementar solo las interfaces a nivel de curso, ya que esta poseería la mayor cantidad de información para generar la tabla a mostrar. Una vez que se tuviera terminada y validada esta interfaz se podría proceder a crear una versión general para la vista de institución y de perfil personal, ya que no significaría mucho trabajo extra al realizado hasta ese punto.

Por otro lado, se decidió mantener la lógica de mostrar la información en forma de tabla en vez de un gráfico, ya que así se puede entregar más información de manera simultanea sin ser demasiado molesta o sobrecargada a la vista. En la figura 4.21 se puede observar cómo se ve esta interfaz implementada en la versión de desarrollo de U-Cursos. Aquí, la tabla muestra el nombre de la persona (censurado), porcentaje predicho de reprobación, el número de ramos en riesgo de doble reprobación, el número de ramos total en curso, la nota promedio del primer control y un link que redirige a la vista de detalle. Esta corresponde a la interfaz diseñada en la figura 3.6, mostrada en el capítulo 3, pero cambiando la ubicación desde el módulo institución al de cada curso, por lo que no fue necesario agregar los selectores de curso pensados originalmente.

Para la vista de detalle, se decidió mantener la misma lógica de tablas utilizada en la vista general, de forma que no se rompiera con la estética general de U-Cursos y mantener coherencia entre una página y otra del módulo. La figura 4.22 muestra cómo quedó dicha interfaz, mostrando la misma información que la vista general en la parte superior y agregando el detalle del historial de cursos tomados por el estudiante, de forma que entregue más detalle del rendimiento académico que ha llevado hasta el momento. La figura 3.5 muestra el boceto pensado originalmente para esta interfaz, el cual fue modificado por los motivos antes descritos.

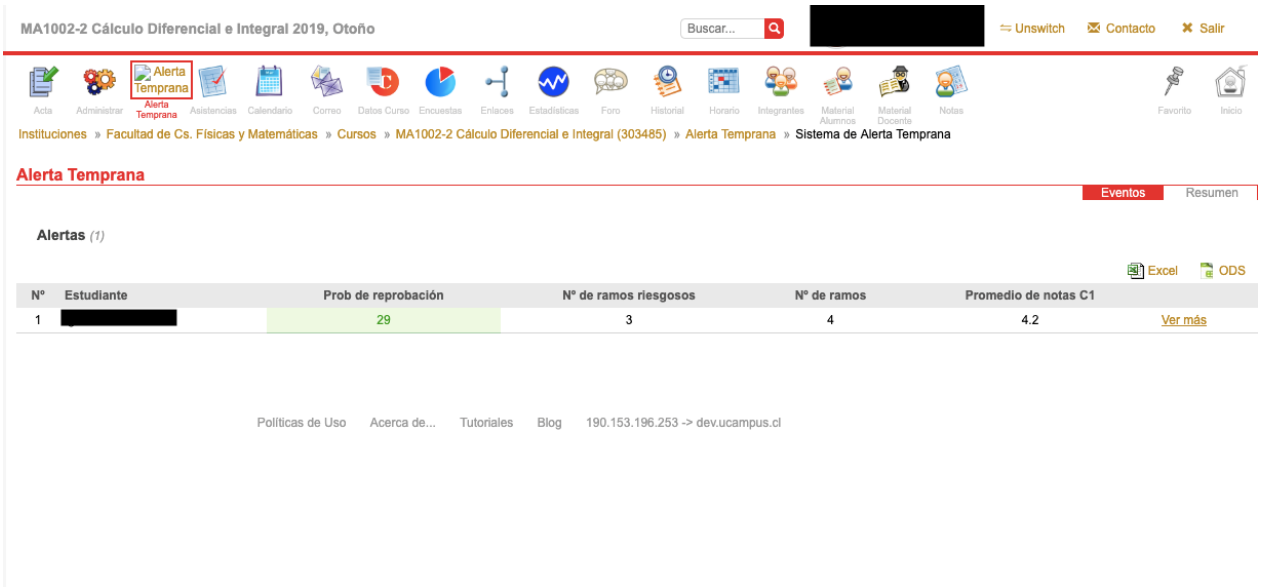

Figura 4.21: Vista general de un curso con un alumno en riesgo.

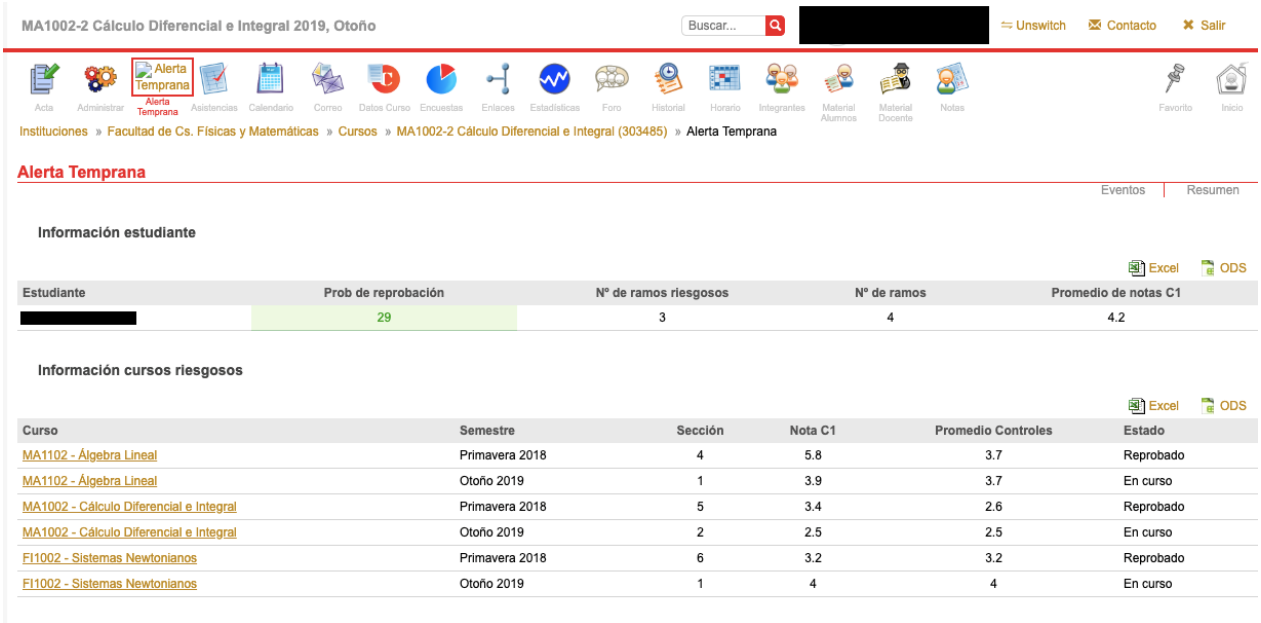

Figura 4.22: Vista detalle de un estudiante.

4.3. Resumen

Como resumen del capítulo anterior, este consistió en una explicación paso a paso de todas las decisiones tomadas durante el desarrollo de este proyecto. Se partió relatando cómo se realizó la primera iteración, donde se agregaron los datos de más cursos del primer año de Plan Común, y comentando que se obtuvieron resultados bastante favorables. A continuación se explicó el trabajo realizado para incorporar el resto de los datos de pregrado al modelo, para obtener resultados levemente inferiores al modelo de Celis et al., pero que aun así eran aceptables. El capítulo entonces termina contando el proceso de creación de las interfaces de usuario y mostrando las primeras versiones de estas. Esto da el puntapié para que en el siguiente capítulo se pueda discutir sobre la validación realizada a las distintas partes de este trabajo.

Capítulo 5

Validación

En el capítulo a seguir se procederá a explicar el proceso de validación realizado a las distintas componentes del proyecto. En la sección 5.1 se parte mostrando la validación final realizada al modelo obtenido hacia la última parte del desarrollo, mientras que en la sección 5.2 se describen las entrevistas realizadas a posibles usuarios finales con el fin de validar las interfaces creadas.

5.1. Validación del modelo final

El algoritmo escogido para la realización del modelo consistió en la Regresión Logística utilizando las mismas variables escogidas por Celis et al. y utilizadas en todas las iteraciones hasta ahora. Para la validación final, en vez de utilizar solo los datos de un año, se extrajo el 25 % de las observaciones para cada semestre, se unieron en un conjunto y se utilizó este para realizar la validación. El 75 % de datos restante se utilizó para realizar el entrenamiento del modelo siguiendo la misma lógica que se explicó en la sección 4.1.2, equilibrando parcialmente las clases antes de entrenar.

El conjunto de entrenamiento consistió en 8.346 observaciones y el de validación en 3.053. En la figura 5.1 se puede observar la curva ROC generada por el algoritmo, la cuál es mucho más suave que las otras gracias al mayor volumen de datos de validación.

En cuanto a los resultados obtenidos, en las figuras 5.2 y 5.3 se puede ver que se mantuvo bastante pareja la relación entre las tasas de predicción para elementos "reprueba" y "no reprueba", a pesar de que los volúmenes de las clases hubieran sido tan diferentes, siendo 444 y 2.609 observaciones respectivamente.

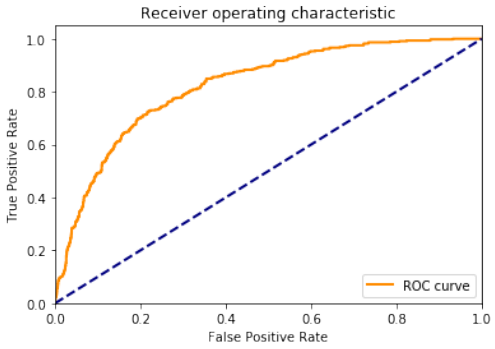

Figura 5.1: Curva ROC obtenida para el modelo final.

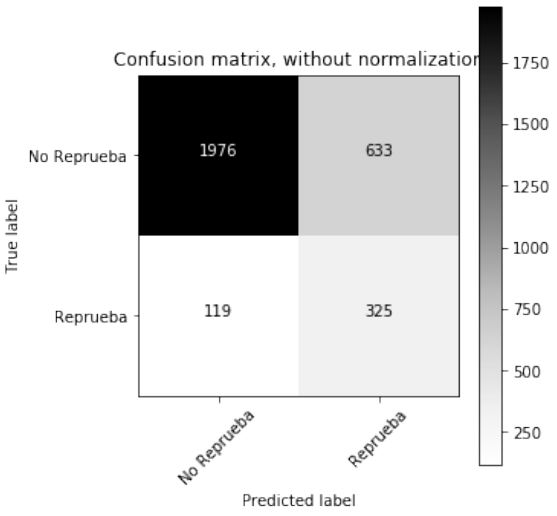

Figura 5.2: Matriz de confusión sin normalizar para el modelo final.

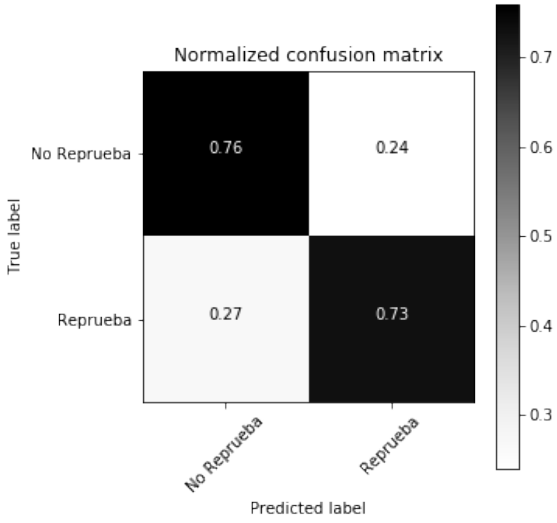

Figura 5.3: Matriz de confusión normalizada para el modelo final.

Por último, en la tabla 5.1 se puede ver que el rendimiento del modelo en esta última versión alcanzó un *recall* de 73 % y una *precision* de 34 %. Esto fue un poco peor en términos de *precision* que el último entrenamiento y validación que se hizo, pero se puede considerar marginal ya que se mantiene cercano a los parámetros establecidos por el estudio inicial de Celis et al.

Variable	Precision	Recall
Reprueba	0,34	0,73
No Reprueba	0,94	0,76

Tabla 5.1: Métricas de rendimiento obtenidas para el modelo final.

Variable	Coeficiente
Género (hombre)	0,26
colegio particular	-0,19
colegio subvencionado	0,01
ratio créditos reprobados	3,00
C1IRS2 <FIRS1	-0,32
C1IRS2 - 4,0	-5,98
C1INRS2 <C1INRS1	0,16

Tabla 5.2: Coeficientes obtenidos para el modelo de regresión logística final.

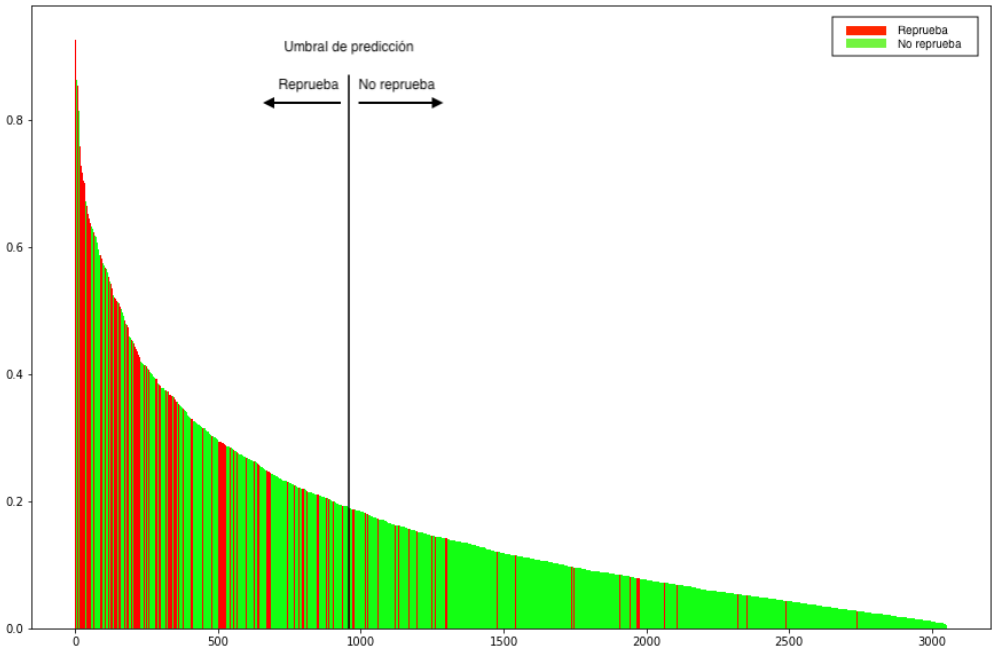

Figura 5.4: Resultados obtenidos para el modelo final.

Otro tema interesante de observar son los valores obtenidos para los coeficientes del último modelo generado, los cuales se encuentran en la tabla 5.2. Aquí se puede ver el efecto de integrar más datos al modelo sobre el poder explicativo que tienen cada una de las variables; la mayoría de estas se quedaron cercanas a los valores obtenidos por Celis et al. (tabla 2.1) o, por lo menos, mantuvieron el signo.

Uno de los coeficientes que sufrió el mayor cambio corresponde a la diferencia entre la nota de los primeros controles con la nota de aprobación, implicando que al considerar ramos más avanzados en la carrera, un mal desempeño en las primeras pruebas pareciera jugar un papel mucho más importante en la doble reprobación que cuando se centró el modelo para los alumnos de primer año. Otro cambio interesante es en el poder explicativo del tipo colegio de origen y género de los estudiantes, reduciendo la influencia de las variables género, "colegio particular" y "colegio subvencionado" (cambiando a positivo esta última), pareciendo indicar una baja en la influencia de la información de pre-ingreso al avanzar en la carrera.

5.2. Validación de las interfaces

Para realizar la validación de las interfaces, se entrevistó a dos posibles usuarios finales para indagar sobre su opinión respecto a la información entregada por las vistas generadas. La primera persona entrevistada correspondió a Natacha Astromujof, quien tiene el cargo de "profesor coordinador" para los ramos matemáticas que se dan en Plan Común. Posteriormente, se entrevistó a Patricio Poblete, profesor coordinador del ramo de primer semestre de computación (CC1000) y profesor de algunos ramos de especialidad para el Departamento de Ciencias de la Computación (DCC). Su labor como profesores coordinadores es administrar la manera en que se dictan los contenidos, mantener el ritmo a través de varias secciones de mismo ramo y ayudar con la creación de las evaluaciones de modo que el ramo sea lo más parecido posible para todos los alumnos independiente del profesor que lo dicte. Si bien ambos trabajan en ambientes distintos (una para Plan Común y el otro para una línea de especialización), las opiniones y comentarios que emitieron con respecto al sistema a partir de las preguntas planteadas fueron bastante similares.

En un inicio se les consultó sobre su opinión general con respecto a la existencia de un sistema de alerta temprana y las medidas que se deberían tomar al tener esta información. Ambos se mostraron entusiasmados por la existencia de este sistema pero se vieron preocupados por la forma en que se informaría/contactaría a los estudiantes que sean identificados con un alto riesgo de doble reprobación. "Si se le dice al alumno el porcentaje, se lo va a echar" dijo uno de ellos al referirse a la posibilidad de informarle el valor de la predicción a los estudiantes afectados, para luego elaborar argumentando que la presión de recibir una alta probabilidad de doble reprobación volvería inútil cualquier esfuerzo posterior del estudiante debido a los nervios y baja en la auto confianza de recibir tal noticia.

Por otro lado, también ambos declararon que mientras antes se pudiera realizar esta predicción, antes se podría intervenir en los hábitos del estudiante y así se podrían obtener mejores resultados. Teniendo eso en cuenta, ambos coincidieron con Celis et al. al considerar al control 1 como un buen momento para empezar a tomar acciones, aunque al coordinador del DCC le preocupó un poco más este tema, producto que en la especialidad son más frecuentes ramos que tienen menos controles que los ramos de Plan Común y que, por tanto, tienen la información de la primera nota bastante avanzado el semestre.

En cuanto a las interfaces se refiere, ambos consideraron buena y necesaria la información entregada en las interfaces, agradecieron, por ejemplo, una vista general donde se pueda ver la información de todos los estudiantes en peligro y la opción de descargarla en formato excel para poder analizarla en mayor detalle. De la vista detallada, encontraron interesante y útil que se mostrara información de las veces anteriores en que se tomaron los ramos reprobados.

Con respecto a las críticas recibidas, se habló de la posibilidad de mostrar más de un curso a la vez o de mostrar todos los cursos asociados al coordinador al mismo tiempo. Esto permitiría que ellos puedan ver fácilmente la situación global de los cursos a su cargo, quizás haciendo medidas a nivel generación más que persona por persona.

Otro punto a mejorar es la vista de detalle, donde señalaron que, a pesar de entregar una buena cantidad de información para el estudiante, el tener que ver esto persona por persona

para poder analizar sus situaciones haría poco práctico el uso de la plataforma, ya que existen cursos con más de 100 personas y los profesores coordinadores deben ver varias secciones de estas a la vez, lo que implicaría realizar decenas de clicks para poder analizar la situación de todos los estudiantes a su cargo. Esto fue particularmente cierto para Natacha, considerando que tiene cerca de 900 alumnos de los que preocuparse cada semestre.

Además de lo anterior, se sugirió incorporar un botón o formulario de contacto directo con el área de "calidad de vida" de la universidad, de modo que se puedan informar rápidamente los casos delicados, de personas con problemas externos a la universidad a las personas adecuadas para ayudar con estos. También, mostraron interés en que se pueda saber si algún otro profesor ya ha reportado o identificado el problema ingresando una observación al respecto.

Por último, luego de haber validado las interfaces, se les preguntó su opinión con respecto al rendimiento obtenido por el modelo actual, comparándolo con el modelo original de Celis et al. y considerándolo por sí solo. Ambos profesores consideraron que una predicción del 70 % de las personas que terminaron reprobando es bastante buena, coincidiendo también con Celis et al. al decir que la baja *precision* no es un problema grave, ya que de todas maneras permite implementar planes de acción con estudiantes que de por sí se encuentran en riesgo por haber reprobado ramos en el pasado.

Como un último detalle, también se pidió a los profesores entregar sugerencias, si es que las tenían, de cosas que podrían ayudar a mejorar el modelo o la interfaz en general. Dentro de las consideraciones que estos profesores entregaron para la elaboración del modelo estaban sugerencias como darle más peso a observaciones más recientes sobre las más antiguas, ya que según ellos, el contexto de los estudiantes cambia mucho entre una generación y otra, por lo que la generación 2010 puede no ser tan representativa para predecir a los alumnos ingresados el 2018, pero la generación 2017 sí podría entregar mejores pistas al respecto. También, el considerar más el contexto del estudiante y no solo su rendimiento académico, ya que muchas veces los problemas de rendimiento vienen generados por una mala base educativa o por dificultades externas a la universidad (e.g., enfermedades, problemas económicos).

5.3. Resumen

A lo largo de este capítulo se mostraron las distintas técnicas de validación implementadas para garantizar el funcionamiento del trabajo desarrollado. Se comenzó hablando de los resultados positivos obtenidos por el modelo final y se terminó con las opiniones positivas recibidas para las interfaces creadas.

El capítulo siguiente servirá como un cierre para este trabajo, entregando las conclusiones del trabajo realizado y proponiendo trabajos futuros que se podrían realizar a este sistema, o en torno a mejoras que pueden impactar en futuras mejoras del modelo creado.

Conclusión

A lo largo de esta memoria se ha trabajado en construir un modelo que utilizara el creado por Celis et al. como base e incorporara a este la mayor cantidad de gente posible. Para llevar a cabo esta tarea, primero se incorporó más información respecto a ramos pertenecientes a Plan Común, obteniendo muy buenos resultados comparado con el modelo original, pero afectando gravemente la predicción producto del bias en la base de datos que fue entregada para esta iteración.

La siguiente iteración consistió en incorporar todos los datos existentes de alumnos de pregrado, independiente de si los cursos eran de Plan Común o alguna de las especialidades impartidas en la FCFM. Para esto se tuvo que realizar un largo trabajo de procesamiento de los datos, de modo que se pudiera extraer tan solo la información necesaria para generar las variables que debían ser ingresadas al modelo.

Una vez que se tuvieron los datos procesados, se pudieron realizar varias pruebas de entrenamiento y validación para verificar que el nuevo modelo siguiera funcionando bajo los parámetros establecidos por el modelo original. Se entrenó primero con los datos de 2010 a 2017 y se validó con los datos de 2018, generando un *recall* de 72 % y una *precision* de 43 %; luego se extrajo un 75 % de los datos de cada semestre y se utilizó para el entrenamiento, dejando al otro 25 % como el conjunto de validación, obteniendo un *recall* de 73 % y *precision* de 34 %. Si bien ambos resultados estuvieron levemente bajo lo establecido por Celis et al., estos se encuentran dentro de los parámetros aceptables, ya que logran predecir un gran número de casos de doble reprobación aunque también pongan en la misma bolsa a personas que lograron aprobar.

Para finalizar el desarrollo, se crearon dos interfaces en las cuales se pretende mostrar la información de los alumnos en riesgo de cada curso del semestre vigente y se validaron con una profesora coordinadora de Plan Común y un profesor coordinador del DCC, obteniendo una buena recepción por parte de estos. La información necesaria para mostrar en estas interfaces se obtendrá a partir de una serie de scripts creados que actualizarán la información de manera periódica en los servidores de Ucampus.

En conclusión, tomando en cuenta los objetivos del proyecto en conjunto con el trabajo realizado y los resultados obtenidos, se puede decir que se logró crear satisfactoriamente un Sistema de Alerta Temprana que pueda permitir la predicción de la probabilidad de reprobación para alumnos de pregrado de la FCFM, cursando ramos ya reprobados en el pasado.

Para esto, se generó un modelo que, utilizando aprendizaje supervisado de máquinas, fuese capaz de predecir con un buen nivel de certeza, según lo establecido por Celis et al., la doble reprobación de cualquier alumno cursando un plan de pregrado en la FCFM. Además, aunque este modelo no mejoró el tiempo de demora necesario para realizar las primeras predicciones, al menos mantuvo el estándar del modelo original y puedo generar predicciones solo teniendo la nota de las primeras pruebas.

En cuanto a la creación de interfaces, se crearon dos vistas validadas con usuarios finales donde se puede observar la información y hacer uso de esta para la formulación e implementación de planes de acción con el objetivo de disminuir la reprobación estudiantil.

A pesar de lo anterior, no se pudo cumplir con el objetivo de analizar el efecto generado por la incorporación de variables no estudiadas por Celis et al. (tales como las estadísticas de uso de U-Cursos), ya que no alcanzaron los plazos planificados para llevar a cabo todo el trabajo que implicaba agregar una variable desconocida: desde el procesamiento de los datos para la generación de valores concretos, hasta el análisis de la incorporación de estos a los modelos generados y la toma de decisiones que podrían derivar de los resultados obtenidos.

Por otro lado, este desarrollo también sirve como un primer puntapié para una serie de estudios que pueden ser realizados para analizar el rendimiento académico y las maneras de hacer docencia efectiva en la FCFM. Uno de los puntos más significativos de este trabajo, es que el modelo realiza predicciones basado en qué tanto o no progresó el estudiante en su rendimiento, en vez de fijarse en estadísticas generales y tratar de predecir basado en lo que otras personas han podido o no realizar.

Para entender un poco mejor lo anterior, se puede pensar que de ingresar al modelo directamente cosas como las notas obtenidas en cada ramo en todas sus instancias anteriores, la tasa de reprobación histórica del ramo, la distribución de notas general de cada generación u otras métricas referidas al historial de gente que ha pasado antes por los cursos, a la hora de realizar predicciones el modelo entendería esto como: "si a estas personas les fue similar, es bastante probable que esta persona obtenga el mismo resultado". Lo anterior toma al individuo y generaliza completamente su comportamiento, dejando de lado totalmente las diferencias que pueden tener las personas, ya que no porque a muchas personas se les haya hecho fácil un ramo esto significa que también le irá bien a este individuo (o viceversa). En vez de esto, el modelo recibe comparaciones del rendimiento del mismo estudiante en dos semestres distintos, por lo tanto, lo que realmente está siendo analizado a la hora de realizar una predicción es la evolución que realiza cada persona, lo que se podría verbalizar como: "dado que estas otras personas alteraron su rendimiento de manera similar y obtuvieron tal resultado, es bastante probable que esta persona también lo haga de la misma forma". Esto es una manera mucho más humana de analizar el rendimiento, debido a que considera la superación o caída frente a obstáculos de cada estudiante y trata de analizar con esta base qué tanto se va a poder recuperar un estudiante en cuestión.

El enfoque más personal del que se hace uso en este trabajo, el cuál se ve poco en trabajos relacionados con educación y, en cambio, se utiliza más en el análisis de comportamiento en torno a redes sociales y otras cosas de ese estilo, debería, sin lugar a dudas, ser empleado en el futuro para otros desarrollos de este tipo: que busquen mejorar la calidad de enseñanza y de profesionales que existe en el país mediante el uso de datos que estos van dejando atrás.

Trabajo Futuro

Si bien el trabajo realizado en esta oportunidad compone un proyecto auto contenido, con resultados verificables y validado teóricamente y por usuarios finales, de este se desprenden diversas mejoras que podrían ser realizadas en futuras iteraciones del Sistema de Alerta Temprana o de otros desarrollos similares. Dentro de las modificaciones que se podrían realizar, estas estarían en torno al modelo mismo, a las interfaces creadas o como parte de trabajos relacionados que no alteren directamente el funcionamiento del sistema creado en esta ocasión.

Mejoras al modelo

Una de las primeras mejoras que vienen a la mente a la hora de modificar el modelo, o de hacerlo más adaptable para situaciones distintas, es la incorporación del rendimiento académico en otro tipo de evaluaciones. Lo anterior es porque el modelo actual solo considera el rendimiento obtenido a través de pruebas teóricas escritas que lleven ciertos patrones de nombre reconocidos, mientras que hay un gran número de cursos que hacen uso de otro tipo de evaluaciones, tales como tareas, ejercicios semanales, laboratorios prácticos, proyectos semestrales, entre otros.

Además de la diversidad en evaluaciones, hay cursos en que las pruebas teóricas corresponden a gran porcentaje de la nota final y de aprobación, mientras que en otros también las tareas tienen un porcentaje importante u otros en que solo existe un proyecto que se realiza en varias iteraciones a lo largo del semestre. Por esto, sería interesante considerar la ponderación de cada tipo de evaluación en la nota final a la hora de incorporarlos a las estadísticas.

Otro punto en el que se podría trabajar es otorgarle mayor peso a las observaciones utilizadas en el entrenamiento mientras más recientes sean. En las conversaciones con los usuarios que validaron las interfaces, se comentó que el contexto social afectaba mucho el estado base en que entraban los estudiantes nuevos a la FCFM, por lo que observaciones más recientes, de estudiantes que hayan pasado por procesos similares, podrían tener mejor poder predictivo que aquellas de años anteriores.

Del mismo modo, también podría ser útil considerar el año al que corresponde la observación o, por ejemplo, cuántos años pasaron entre que el estudiante ingresó a la universidad (o egresó de la educación media) hasta el momento de la observación analizada, ya que, según los resultados obtenidos en esta ocasión, la información de pre ingreso parece perder poder explicativo sobre datos de alumnos que ya llevan un tiempo en la carrera.

Por último, pero no por eso menos importante, queda pendiente ingresar los datos de uso de U-Cursos en el análisis realizado, debido a que el trabajo de procesamiento de los datos tomó más tiempo del esperado y no se alcanzó a incorporar en esta ocasión. La obtención de métricas que puedan ser utilizadas en la predicción de reprobación de los estudiantes, de modo que comparen la cantidad de actividad realizada consigo mismos en el pasado, podrían

entregar información útil y de rápido acceso, eliminando posiblemente la dependencia de datos de evaluaciones para entregar las primeras predicciones y que luego sean refinadas al existir más información.

Mejoras a las interfaces

En cuanto a las interfaces se refiere, uno de las mejoras más necesarias y fáciles de realizar es la de incorporar el módulo a nivel institución, en vez de dejarlo sujeto a cada curso en particular. Esta mejora permitiría observar la situación de muchos alumnos a la vez, lo que sería bastante útil para profesores coordinadores o profesores que tengan a cargo varias secciones del mismo curso, ya que facilita el uso de la interfaz para observar mejor el panorama global. Por otro lado, también ayudaría a académicos e investigadores para poder tomar acciones a nivel facultad, en vez de estar haciendo medidas curso por curso.

Junto con la mejora de implementar el módulo a nivel institucional, debería realizarse una mejora que permita filtrar los cursos mostrados, ya que no siempre es útil ver todos a la vez. Con esto se podrían hacer filtros por departamentos, para que cada especialidad pueda o no tomar medidas, u observar una secuencia de cursos con dependencia entre sí, por ejemplo, para analizar el impacto a través de esta línea.

Por otro lado, otro de los cambios solicitados corresponde a la posibilidad de ver el detalle de varios estudiantes a la vez, ya que estar viendo uno por uno a los estudiantes, cuando se tienen varias secciones de 100 estudiantes cada una a cargo, puede resultar bastante tedioso. Si bien no es una buena idea mostrar toda la información a la vez, porque la interfaz quedaría demasiado cargada y terminaría entorpeciendo su uso, sí sería una buena opción ofrecer la opción de descargar la información detallada en formato de planilla Excel, de modo que cada persona pueda realizar sus propios análisis sin problemas.

Como un último detalle que se podría implementar en las interfaces, es la posibilidad de comunicarse con el área de calidad de vida fácilmente en caso de identificar casos preocupantes dentro del estudiantado. Además de esto, también podría servir para enterarse si otros profesores ya han reportado problemas antes y, quizás, poder tomar iniciativas coordinadas como profesores.

Otros trabajos futuros relacionados

Un trabajo que puede salir derivado de este desarrollo, es el de la generación de estadísticas de uso o de métricas de uso sobre el uso de U-Cursos, de modo que su incorporación al modelo sea más fácil cuando se quiera volver a iterar sobre el algoritmo y las variables consideradas.

De manera paralela, también se podría trabajar en utilizar estas variables del uso de U-Cursos para realizar categorizaciones del tipo de usuarios que existen en la plataforma y ocupar esto como una guía para modificar lógicas del funcionamiento general, por ejemplo, la frecuencia o visibilidad con que las notificaciones que son mostradas, el orden de los temas

en los foros según el interés mostrado o hasta sugerencias de ramos a tomar según el historial y la caracterización del usuario que se realice.

En cuanto a la arquitectura misma de las aplicaciones de Ucampus, se podría trabajar en llevar a la nube algunas de sus bases de datos o a una plataforma que permita mejor manejo de grandes volúmenes de información. Esto ayudaría tanto a la generación como al procesamiento de los datos generados diariamente con el uso de las diversas plataformas, y podría ayudar a la generación de mejoras y políticas de las instituciones en base al comportamiento observado en los diversos procesos que se llevan a cabo ocupando U-Cursos o Ucampus.

Bibliografía

[1] Anaconda Distribution. https://www.anaconda.com/distribution/. Acceso: 2019-07-04.

[2] Anaconda Distribution documentation. https://docs.anaconda.com/anaconda/. Acceso: 2019-07-04.

[3] Artificial Neural Networks as Models of Neural Information Processing. Acceso: 2019-07-16.

[4] Data Mining: Encyclopedia Britannica. https://www.britannica.com/technology/data-mining. Acceso: 2019-07-16.

[5] NumPy. https://www.numpy.org/. Acceso: 2019-07-04.

[6] Pandas. https://pandas.pydata.org/. Acceso: 2019-07-04.

[7] Project Jupyter. https://jupyter.org/. Acceso: 2019-07-04.

[8] Reglamento General de Estudiantes de la Universidad de Chile. http://escuela.ingenieria.uchile.cl/dam/jcr:43500e07-5ea9-47bb-ba59-1db813e72327/016-estudiante-007586-reglamento-general-de-estudiantes-de-la-universidad-de-chile.pdf. Acceso: 2019-07-16.

[9] Scikit-learn. https://scikit-learn.org/stable/index.html. Acceso: 2019-07-04.

[10] ACM Workshop on Computational Learning Theory (5th 1992 Pittsburgh, Pennsylvania). *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, July 27-29, 1992, Pittsburgh, Pennsylvania*. Association for Computing Machinery, New York, NY, 1992.

[11] C. Acuña Veliz. Acceso y deserción en la educación superior: caso aplicado a chile. Master's thesis, Universidad de Chile, 2012. Disponible en http://repositorio.uchile.cl/handle/2250/112062.

[12] Centro de Microdatos, Departamento de Economía, Universidad de Chile. Estudio sobre causas de la deserción universitaria. https://www.oei.es/historico/pdf2/causas-desercion-universitaria-chile.pdf, 2008. Acceso: 2019-08-25.

[13] Consejo Nacional De Educación. Indices tendencias educación superior - pregrado 2017. https://www.cned.cl/sites/default/files/tendencias_matricula_pregrado_2017.pdf, 2017. Acceso: 2018-10-09.

[14] G. Siemens, R. Baker. Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. In *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 252–254, New York, NY, USA, 2012. ACM. http://doi.acm.org/10.1145/2330601.2330661.

[15] Instituto de estadística de la UNESCO. *Compendio mundial de la educación 2006*. 2006. https://www.oei.es/historico/quipu/estadisticas_unesco2006.pdf. Acceso: 2019-10-09.

[16] T. Larroucau De Magalhaes-Calvet. Estudio de los factores determinantes de la deserción en el sistema universitario chileno. Master's thesis, Universidad de Chile, 2013. Disponible en http://repositorio.uchile.cl/handle/2250/114843.

[17] D. López, J. Silva, and S. Celis. Analyzing the influence of online behaviors and learning approacheds on academic performance in first year engineering. Trabajo en progreso presentado en el Departamento de Ciencias de la Computación, Abril 2019.

[18] M. Mohri, A. Rostamizadeh, A. Talwalkar. *Foundations of Machine Learning*. Adaptive computation and machine learning series. MIT Press, Cambridge, MA, 2012.

[19] D. Olson. Data mining in business services. *Service Business*, 1:181–193, 09 2007.

[20] E. T. Pascarella and P. T. Terenzini. *How college affects students*. Jossey-Bass, San Francisco, CA, 3 edition, 2005.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[22] S. Celis, L. Moreno, P. Poblete, J. Villanueva, R. Weber. Un modelo analítico para la predicción del rendimiento académico de estudiantes de ingeniería. *Revista Ingeniería de Sistemas*, XXIX, 2015.

[23] S. J. Russell, P. Norvig, E. Davis. *Artificial intelligence: a modern approach*. Prentice Hall, Upper Saddle River, New Jersey, 3 edition, 2010.

[24] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, pages 71–105, 1959. http://www.cs.virginia.edu/~evans/greatworks/samuel1959.pdf. Acceso: 2019-07-04.

[25] SIES Ministerio de Educación. Retención de primer año en educación superior. programas de pregrado. https://www.mifuturo.cl/wp-content/uploads/2018/SIES/publicaciones/estudios/retencion_primer_ao_carreras_de_pregrado_2014.pdf, 2014. Acceso: 2019-07-17.

[26] SIES Ministerio de Educación. Retención en educación superior con perspectiva de género. https://www.mifuturo.cl/wp-content/uploads/2018/SIES/publicaciones/estudios/retencion%20en%20educacion%20superior%20en%20perspectiva%20de%20genero_2014.pdf, 2014. Acceso: 2019-07-17.

[27] V. Santelices, X. Catalán, C. Horn, D. Kruge. Determinantes de deserción en la educación superior chilena, con énfasis en efecto de becas y créditos. *Proyecto FONIDE*, (F611103), 2013.