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

DESARROLLO DE MODELOS PREDICTIVOS ENFOCADOS EN LA APROBACIÓN DE PLAN COMÚN PARA ESTUDIANTES DE LA FCFM

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTIÓN DE OPERACIONES
MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

IGNACIO OSVALDO CANTILLANO VERGARA

PROFESOR GUÍA:
CHARLES THRAVES CORTÉS-MONROY

MIEMBROS DE LA COMISIÓN:
MARCEL GOIC FIGUEROA
SERGIO CELIS GUZMÁN

SANTIAGO DE CHILE
2023

DESARROLLO DE MODELOS PREDICTIVOS ENFOCADOS EN LA APROBACIÓN DE PLAN COMÚN PARA ESTUDIANTES DE LA FCFM

La Facultad de Ciencias Físicas y Matemáticas (FCFM) de la Universidad de Chile imparte en la actualidad 13 programas de pregrado conducentes a diversas carreras profesionales y licenciaturas. En los primeros años de la carrera, deben cursar por obligación un plan de estudios compartido para cualquier carrera de la Facultad, conocido como Plan Común.

En este contexto, la presente tesis propone modelos que permitan predecir si un alumno aprobará o no Plan Común por motivos académicos. Es decir, se excluyen los casos en que un alumno abandona el plan de estudios sin tener causales de eliminación. Estos modelos en especifico buscan pronosticar la probabilidad de que un alumno termine este programa, usando diferentes variables del estudiante previo a su ingreso a la Facultad, como por ejemplo notas de enseñanza media, Ranking de egreso, colegio de egreso, entre otras. En particular, se analiza cuáles son las variables que tienen mayor importancia en la predicción, y a su vez cómo dichas variables impactan en la aprobación del Plan Común. Adicionalmente, se realizan pronósticos usando información de rendimiento del primer y segundo semestre del programa. Todo lo anterior, a través de modelos supervisados de Machine Learning tales como: Support Vector Machine, Árbol de Clasificación, Random Forest, Red Neuronal, K vecinos más cercanos y Regresión Logística.

Se logró identificar en los mejores modelos a un 66 % de alumnos que no completaron Plan Común y un 70 % que sí lo completaron, basándonos únicamente en variables previas al ingreso a la Facultad. Sin embargo, se observa una notable mejoría en la predicción en los modelos que incorporan variables sobre el rendimiento académico del primer y/o segundo semestre. Estos modelos lograron identificar aproximadamente al 85 % de los alumnos que no completaron Plan Común y al 88 % de los que sí lo hicieron.

Estos modelos pueden ser de utilidad para la focalización de diferentes iniciativas de apoyo por parte de la FCFM hacia aquellos estudiantes que poseen un mayor riesgo de desertar de Plan Común por motivos académicos.

Agradecimientos

Quiero agradecer a mi familia que a pesar de sufrir momentos difíciles en este último tiempo siempre estuvo presente, apoyándome en todo momento. Además, a todos mis amigos que fueron una parte fundamental para poder llegar hasta este punto.

Finalmente quiero agradecer a mi profesor guía Charles Thraves, que es una gran persona donde a través de su apoyo incondicional, pude aprender demasiado en este último tiempo y sobrellevar de la mejor manera este desafío.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

La Facultad de Ciencias Físicas y Matemáticas (FCFM) de la Universidad de Chile imparte en la actualidad trece programas de pregrado dirigido a diversas carreras, como también a licenciaturas. Para realizar cualquiera de estos programas, es necesario cursar un plan de estudios, conocido como Plan Común, que entrega una base importante en la formación de las ciencias básicas, con la finalidad de marcar una diferenciación en la formación de los alumnos egresados de la Facultad.

Este plan de estudio tiene una duración mínima de 4 semestres e ingresan aproximadamente 750 estudiantes cada año, los cuales cursan 120 créditos, que equivalen a 29 o 30 cursos, dependiendo del año de ingreso a la Facultad. En dichos cursos se abarcan tópicos teóricos como prácticos, de computación, economía, física, ingeniería de proyectos, matemáticas, y química; además se fomenta la comunicación efectiva, ética, innovación y trabajo en equipo, a través de los diversos cursos. Al terminar este plan de estudios los alumnos deben seleccionar alguno de los programas de pregrado disponibles en la Facultad para seguir con sus estudios y poder obtener un título universitario. Cabe destacar que no existen límites en los cupos en ningún programa de pregrado, lo cual hace que el Plan Común impartido en la FCFM sea diferente al de muchas otras universidades.

Para poder ingresar a Plan Común es necesario haber egresado de la educación media y postular por alguna de las vías de acceso disponibles, donde la más conocida es la admisión regular, que se basa en seleccionar según los mejores puntajes ponderados, que postulan a través del Sistema Único de Admisión (SUA), el cual es coordinado por el Departamento de Evaluación, Medición y Registro de la Universidad de Chile (DEMRE). Otra vía de acceso

es el Programa de Acceso a la Educación Superior (PACE), el cual es administrado por el Ministerio de Educación (MINEDUC) y busca que estudiantes de contextos vulnerables ingresen a la universidad, donde se destaca que quienes ingresen por medio de esta vía tendrán acompañamiento psicológico y pedagógico. Todas estas vías de acceso tienen en común que consideran el rendimiento del estudiante en el colegio a través de las variables NEM[1] y Ranking[2] , además de las pruebas estandarizadas que crea el DEMRE.

En los últimos 12 años, Plan Común ha sido aprobado aproximadamente por el 80 % de los alumnos que ingresan y concentra aproximadamente el 91 % de los estudiantes de pregrado que desertan de la FCFM, por diferentes causas. Cabe señalar que la deserción universitaria provoca pérdidas monetarias y problemas organizacionales para las universidades, como también problemas emocionales en los estudiantes que desertan; además "[. . .] desde el punto de vista social incide negativamente en los índices de pobreza, aumenta el desempleo, disminuye el aporte intelectual e incrementa el costo de la educación debido a la suboptimización de los de recursos" (Rodríguez y Zamora, 2014). Ejemplo de esto es lo que se pudo observar en la Universidad del Bio-Bio, donde se estimó que los gastos de arancel relacionados con los estudiantes que desertaron de esa institución entre los años 2016 al 2019 llegó a ser de USD 8.078 en promedio por estudiante y la mayoría de estos gastos los cubrió la beca de gratuidad (Améstica, King, Sanhueza y Ramírez, 2021). Ahora bien, este monto puede diferir con el de la Facultad, ya que, esta posee un arancel aún más costoso.

Es importante notar que existen distintos tipos de deserción, dentro de las cuales la más relevante para la Facultad es la que se produce por aquellos estudiantes que reprueban múltiples veces un mismo curso obligatorio de Plan Común. En otras palabras, las personas que no logran aprobar este plan de estudios por motivos académicos. Debido a que esta causa abarca sobre el 45 % de las deserciones totales de pregrado.

Debido a lo anterior, la Facultad ha estado impulsando distintas iniciativas que buscan contrarrestar la deserción de forma directa e indirecta. Un ejemplo de esto último es el "Taller Dos Relojes", que ofrece diversas actividades para los estudiantes, con el objetivo de apoyar el aprendizaje y mejorar el rendimiento principalmente en los cursos matemáticos de Plan Común, dado que son los cursos que mayores cantidades de alumnos reprueban múltiples veces. Un caso que permite ejemplificar lo que se menciona anteriormente es lo que sucede

[1] Notas Enseñanza Media.

[2] El Puntaje Ranking es un factor de selección que considera el rendimiento académico de un estudiante en relación a su contexto educativo.

con el curso Introducción al Cálculo, donde 1 de cada 25 estudiantes lo reprueba más de 2 veces. Sin embargo, y casi de manera paradójica, en dicho programa ("Taller Dos Relojes") no siempre participan los estudiantes que corren mayor riesgo de no terminar Plan Común por razones académicas, por ello es de suma relevancia poder identificar a estos estudiantes, para poder focalizar esta u otras iniciativas de apoyo y seguimiento de manera oportuna a quienes más las necesiten.

A partir de lo anterior, se elaboran diversos modelos predictivos que buscan identificar si un alumno aprobará Plan Común o desertará por motivos académicos, basándose en variables previas al ingreso, las cuales incluyen tipo de acceso, NEM, Ranking, los resultados de las pruebas PSU[3], características del colegio de egreso, entre otras, esto sustentado en la diversa literatura que ha encontrado que las variables previas al ingreso a la universidad pueden llegar a impactar en lo que se quiere estudiar. Además de examinar cómo evolucionan los modelos al ir considerando la tasa de aprobación de los estudiantes en los primeros semestres de la carrera.

Todo lo anterior con la finalidad de captar en parte los efectos que se producen una vez que los alumnos ingresan a la FCFM y tener herramientas para distintos períodos de tiempo que permitan comprender tempranamente qué factores afectan en tener éxito en el programa e identificar a las personas que corren mayor riesgo de no tenerlo, con el objetivo de focalizar los recursos y/o acciones preventivas por parte de la Facultad, que busquen minimizar la cantidad de estudiantes que no pudieron terminar Plan Común de manera eficaz por razones académicas o ser la base de la creación de estas acciones preventivas.

[3] Test estandarizados utilizado desde el 2003 hasta el 2019 para ingresar vía SUA, sus siglas provienen de Pruebas de Selección Universitaria.

Capítulo 2

Desarrollo

2.1. Antecedentes de Estudio

A grandes rasgos, se identifican trabajos relacionados en las dos siguientes categorías:

1. Estudios enfocados en entender y/o predecir la aprobación de un plan de estudio.
2. Estudios enfocados en entender y/o predecir la deserción universitaria.

La primera línea investigativa está directamente relacionada con el objetivo de la tesis, ya que, en estos estudios se conoce quién termina el plan de estudios y quién deserta, mientras que la segunda, a pesar de ser similar, tiene una pequeña salvedad, dado que, por lo general esta clase de estudios se enfoca en determinar si un estudiante deserta o se queda en la universidad en ciertos períodos de tiempo, por lo cual no se puede asegurar que los estudiantes terminen el plan de estudio. Estos últimos son importantes de considerar, ya que entregan antecedentes de los factores que inciden en que un estudiante no termine, es decir, no apruebe un plan de estudios.

Además, existen distintos tipos de deserción en la universidad, los cuales pueden ser voluntarios o involuntarios. El primer caso corresponde cuando el estudiante toma la decisión de desertar, mientras que el segundo es el resultado de una decisión que toma la universidad (Himmel, 2002). En este estudio sólo se abarca la deserción involuntaria, ya que se busca identificar a los estudiantes que terminan Plan Común y los que no terminan por razones académicas, estos últimos son quienes la universidad elimina por reprobar múltiples veces un mismo curso obligatorio de Plan Común.

A pesar de lo anterior, en la literatura estudian la deserción por lo general como una,

es decir, sin distinguir los tipos de esta o con las mismas variables explicativas, además, los principales modelos teóricos lo consideran así. Un ejemplo de esto es el modelo de Bean (1981) que intenta explicar la deserción en general, no distingue entre voluntario e involuntario, a través de un enfoque de rotación organizacional. Por otro lado, también está el modelo de Tinto (1975), el cual considera que las causas que llevan a la deserción, voluntaria e involuntaria, son distintas, pero se pueden estudiar con este mismo modelo longitudinal, donde la interacción inicial como la que se va dando con el tiempo entre el estudiante y el sistema académico y social de la universidad, conllevan a desertar.

En base a la primera línea investigativa, existe escasa literatura nacional, sin embargo destaca el estudio de Soria y Zuñiga (2014), donde se estudia qué factores permiten determinar qué estudiantes de la carrera de Ingeniería Comercial de la Universidad Católica del Norte en Coquimbo terminarán y cuáles no, donde se refleja que las variables NEM y PAA[4] de matemáticas inciden positivamente en terminar la carrera, mientras que una variable que mide la cantidad de años que se demora una persona en ingresar a una carrera, desde que egresa del colegio, afecta negativamente.

A nivel internacional se han encontrado que las variables de rendimiento previo a la universidad, como el rendimiento en la universidad, ayudan a predecir si un estudiante termina un programa o no. Una muestra de esto son los resultados encontrados en Estados Unidos, donde se aprecia que el High School Class Rank[5], el GPA[6] de la escuela secundaria y del primer semestre universitario, inciden positivamente en terminar un programa con honores en una universidad pública del Medio Oeste en Estados Unidos (Campbell y Fuqu, 2008); otra variable relacionada es la SAT[7] de matemáticas, que se aprecia como un indicador estadísticamente significativo para lograr terminar una carrera STEM en la Universidad de Binghamton (Kokkelenberg y Sinha, 2010). Es interesante notar de estos estudios que las variables mencionadas tienen variables similares en el contexto chileno, como por ejemplo el GPA de la escuela secundaria es equivalente al Nem.

Por otro lado, en la segunda línea de investigación, existe bastante literatura a nivel nacional, donde destacan Acuña (2012), Barrios (2011) y Larroucau (2013), donde demuestran

[4] Test estandarizados utilizado desde 1966 hasta el 2002 para ingresar a la universidad, sus siglas provienen de Prueba de Aptitud Académica.
[5] Factor que considera el rendimiento académico de un estudiante con respecto al de su clase en la escuela secundaria
[6] Escala que estandariza el rendimiento académico en Estados Unidos
[7] Examen estandarizado usado en la admisión universitaria en Estados Unidos

que las variables de rendimiento previo a la universidad, características socioeconómicas y demográficas, llegan a incidir en la deserción de un estudiante en la educación superior chilena. En específico, en el estudio de Barrios (2011) se destacan características como el género, si es mayor de edad, ingresos mensuales por grupo familiar y promedio PSU lenguaje y matemáticas, como indicadores significativos para predecir la deserción. En el caso de Larroucau (2013) se aprecia que, entre las características socioeconómicas y demográficas más relevantes para predecir la deserción, se encuentra el nivel ingreso económico del núcleo familiar y entre las variables de menor injerencia de esta categoría se encuentra la edad y el género, por el lado de las variables de rendimiento previos se aprecia que el Nem y Ranking, son mejores predictores de la deserción que el puntaje PSU de matemática y lenguaje. Además, se aprecia que las características relacionadas con la calidad del colegio de egreso son relevantes para predecir este fenómeno. Otra investigación significativa es la de Díaz (2009), donde se estudia la deserción en los estudiantes de ingeniería de la Universidad Católica de la Santísima Concepción, en esta se puede apreciar que provenir de un establecimiento educacional científico-humanista y obtener mayor puntaje de ingreso, bajan la probabilidad de desertar, mientras que los factores que aumentan las probabilidades de deserción son haber puesto la carrera en un menor orden de prioridad y el tener mayor edad de egreso de enseñanza media.

En el contexto de la FCFM es necesario mencionar el estudio de Celis, Moreno, Poblete, Villanueva y Weber (2015), que encuentra efectos significativos en la deserción creados por las variables asociadas al rendimiento de los primeros semestres y género.

Finalmente, también se puede observar en el estudio de Rilling (2022), que la tasa de retención en los primeros años en la Universidad de Chile son diferentes dependiendo de la vía de ingreso, factor que es relevante en los porcentajes de deserción, ya que con esta variable se pueden captar de manera prematura los posibles alumnos que deserten. Además, en la literatura argentina se aprecia que trabajar de manera paralela mientras se estudia una carrera universitaria, puede llegar a incidir negativamente en el rendimiento académico (Fazio, 2010), lo cual está directamente relacionado con el éxito o deserción de un programa de estudio.

En resumen, las variables relacionadas con el rendimiento en la universidad y previo a ello, características del colegio de egreso, socioeconómicas y demográficas, son importantes de explorar y considerar en este estudio.

Finalmente, es importante mencionar que los modelos de aprendizaje automático, que

usualmente se utilizan para este tipo de problemas o similares son los más sencillos de interpretar, como son el modelo Logit utilizado como por ejemplo en los estudios de Celis, Moreno, Poblete, Villanueva y Weber (2015) y Barrios (2011) y Arboles de Decisión (DT) ocupados en Kuna, García, y Villatoro, (2010) y Daza (2016). A pesar de lo anterior, se ha apreciado que Redes Neuronales (RN) puede llegar a obtener resultados similares o mejores que los dos modelos dichos anteriormente en este tópico (Čukušić, Garača y Jadrić, 2010), en la misma línea se observa que Support Vector Machine (SVM) y K vecinos más cercanos (KNN) presentan mejores resultados que Logit y DT cuando se consideran variables académicas y socioeconómicas (Julca, Larios, Navarro y Valero, 2022). Por otro lado, se apreció en un estudio que compara el rendimiento de diversos modelos que buscan identificar a los estudiantes de ingeniería que desertan en el primer año en la Universidad Técnica Federico Santa María, que entre las técnicas Random Forest (RF), KNN, SVM y Gradient boosting (GB), la que obtuvo mejores resultados fue RF (Dombrovskaia, del Rio y Rodriguez, 2020). Por todo lo anterior, se puede concluir que para este tipo de problemas no se puede identificar el mejor modelo a ocupar.

2.2. Metodología

Con la finalidad de lograr herramientas que permitan comprender tempranamente qué factores afectan en tener éxito en Plan Común e identificar a las personas que corren mayor riesgo de no tenerlo por razones académicas, se trabajó bajo la siguiente metodología. Esta consistió en 7 pasos, los cuales se describen a continuación:

1. **Definir requerimiento y alcances**: con el objetivo de enfocar el trabajo, se definió qué tipo de información es requerida para lograr el objetivo, además de limitar hasta qué punto se quiere profundizar.

2. **Recolección de datos**: Los datos necesarios para la investigación fueron solicitados al DEMRE, los cuales poseen la información de los estudiantes antes de ingresar a la universidad; a la FCFM, que contiene la información de cuando éstos ya están en la casa de estudio; y a las bases abiertas del MINEDUC, que manejan la información de los colegios de los cuales egresan los estudiantes que ingresan a la Facultad.

3. **Definir variable a pronosticar**: Se determinó que la variable objetivo a pronosticar

en los modelos, sería la siguiente:

$$y_i = \begin{cases} 1 \text{ Si el estudiante } i \text{ termina Plan Común} \\ 0 \text{ Si el estudiante } i \text{ abandona Plan Común} \end{cases}$$

Cabe mencionar que la variable y_i es independiente del tiempo que se demore el alumno en aprobar Plan Común, $y_i = 0$ sólo abarca los casos de alumnos que abandonan Plan Común por motivos asociados al bajo rendimiento académico. Se profundizarán sus alcances en las siguientes secciones.

4. **Limpieza de datos y construcción de variables de interés**: Se determinó qué datos estaban disponibles y cuáles eran los relevantes para la investigación, además de eliminar las variables con una gran cantidad de datos faltantes, en base a lo anterior se crearon nuevas variables que podrían ser útiles para el desarrollo de los modelos.

5. **Análisis exploratorio de datos**: Con el fin de entender el comportamiento de las variables y cómo podrían llegar a estar correlacionadas entre ellas, se realizó un análisis exploratorio.

6. **Desarrollo y estimación de los modelos predictivos**: Se ocuparon herramientas estadísticas que utilizan aprendizaje automático, con el fin de obtener clasificadores que puedan lograr identificar a las personas que corren mayor riesgo de no terminar Plan Común por razones académicas.

7. **Análisis y evaluación de los resultados**: Se analizan los resultados de los diferentes modelos y se evalúa cuáles son las mejores técnicas, además de identificar cómo las distintas variables ocupadas pueden llegar a impactar en lo estimado e *insights* que resulten de lo desarrollado.

2.3. Limitaciones

Dada la estructura de trabajo y los datos obtenidos se debe tener en consideración que:

– Existe población objetivo que no será considerada, debido a problemas de identificación, ya que, no se puedo crear la relación 1:1 de todos los de estudiantes al unir las distintas bases.

– Las pruebas PSU fueron remplazadas por las pruebas PAES[8]. Se espera que estas últimas permitan medir las habilidades y competencias que necesitan los estudiantes para la universidad (MINEDUC, 2022), por lo cual se cree que la PAES será una mejor variable explicativa que la PSU.

– No se consideran factores psicosociales, como el nivel de interacción social que tiene el estudiante una vez que ingresa a la universidad, lo cual puede ser interesante y beneficioso de considerar, ya que desde la perspectiva de los modelos teóricos de Tinto (1975) y Bean (1981), es un factor relevante.

– Existen parámetros inexplorados en los diversos modelos de *machine learning* ocupados.

2.4. Datos

Como se mencionó anteriormente, se utilizaron 3 fuentes de datos, las cuales son: DEMRE, FCFM y MINEDUC.

2.4.1. Descripción de las Bases de Datos

2.4.1.1. DEMRE

Se obtuvo informe de los períodos de admisión a la educación superior del año 2010 al 2022, en específico 3 bases de datos por período, los cuales son:

– **Base de inscritos:** contiene toda la información que las personas entregan al momento de inscribirse para rendir las pruebas de selección universitaria, lo cual abarca variables socio-demográficas, características del establecimiento de egreso, variables socioeconómicas y situación educacional del grupo familiar.

– **Base de Puntaje PSU por individuo:** contiene principalmente la información de los puntajes, tanto del proceso del año asociado, como del proceso anterior, en caso de que la persona haya rendido las pruebas PSUs en ese período, además del NEM y Ranking.

– **Base de postulación selección y matrícula:** Contiene toda la información relacionada con las postulaciones por vía de ingreso regular, BEA y PACE, además de los matriculados debido a estos procesos.

[8] Test estandarizados utilizado desde 2022 hasta la actualidad para ingresar a la universidad, sus siglas provienen de Prueba de Acceso a la Educación Superior.

Es importante mencionar que las bases vienen con un identificador único por individuo, creado por la institución para anonimizar los datos.

2.4.1.2. FCFM

Se obtuvo la información anonimizada de los estudiantes de la Facultad entre los años 2012 al 2022, que se divide en 2 bases de datos, que son los siguientes:

- **Base de antecedentes:** Contiene la información de los alumnos que ingresan a la Facultad en el período mencionado, además del tipo de ingreso, nombre del colegio de egreso, puntaje de ingreso, NEM, Ranking y todos los cambios en el estado de matrícula y carrera.
- **Base de Rendimiento:** Contiene la información académica de los estudiantes durante los períodos solicitados, lo cual abarca ramos inscritos, notas, semestres en que se cursaron y si fueron aprobados o reprobados.

2.4.1.3. MINEDUC

A través de los datos abiertos del centro de estudios del MINEDUC, se obtuvo la información de los distintos colegios que funcionaron en Chile entre los períodos 2012 al 2022. Los datos comprenden desde el nombre y Rol Base de Datos (RBD), hasta su localidad, rama educacional y dependencia del establecimiento.

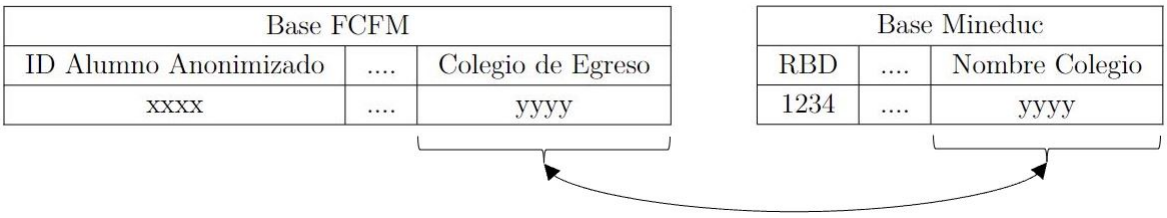
2.4.2. Cruce de Bases de Datos

Para el desarrollo del proyecto solamente se consideraron a las personas que ingresan directamente a Plan Común y por vías de ingreso especial que requieren ciertas pruebas del DEMRE, por lo cual se eliminaron a los estudiantes con estudios medios en el extranjero y quienes ingresan por Bachillerato, estos últimos debido a que cursan un programa que los prepara en ciencias básicas y posteriormente ingresan a la Facultad, por ello no se cuenta con el resultado de estos estudiantes en dichas pruebas.

Dado que las bases del DEMRE y la FCFM están anonimizadas, no se pueden cruzar directamente, por ello se utilizó el siguiente procedimiento:

1. Eliminar a los alumnos que ingresaron vía bachillerato y estudiantes extranjeros de las bases de la FCFM.

2. Cruce de información entre las Bases de Antecedentes de la FCFM con las del MINEDUC, a través del nombre del colegio y año, con la finalidad de obtener el RBD. Cabe mencionar que también se aplicaron técnicas de análisis de texto, con el objetivo de estandarizar todas las llaves, principalmente el nombre del colegio. Ejemplo:

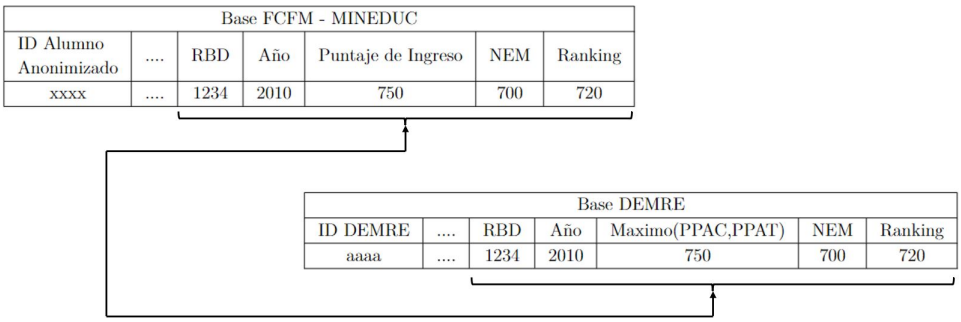


Base FCFM - MINEDUC				
ID Alumno Anonimizado	....	Colegio de Egreso	RBD	....
xxxx	....	yyyy	1234	....

3. Creación de la variable puntaje de ingreso en la base del DEMRE: consiste en crear el puntaje ponderado que hubiese tenido la persona al postular a Plan Común en la Universidad de Chile en ese proceso de admisión, ya que el DEMRE no registra todos los ingresos a la educación superior, por lo cual no posee todos los puntajes de ingreso a esta. Además, si la persona postula con puntajes del proceso anterior y el correspondiente, se selecciona la ponderación más alta, debido a que es la regla que ocupa la institución. Ejemplo:

Base DEMRE					
ID DEMRE	Año	....	Puntaje Ponderado con Datos del Proceso Actual (PPAC)	Puntaje Ponderado con Datos del Proceso anterior (PPAT)	Maximo(PPAC,PPAT)
aaaa	2010		770	750	770
bbbb	2015		0	760	760
cccc	2016		720	0	720

4. Cruce de las bases del DEMRE y FCFM - MINEDUC a través de las llaves RBD, puntaje ingreso, NEM y Ranking. Ejemplo:



Base FCFM - MINEDUC - DEMRE							
ID Alumno Anonimizado	ID DEMRE	....	RBD	Año	Puntaje de Ingreso	NEM	Ranking
xxxx	aaaa	....	1234	2010	750	700	720

5. Se verifica que la información sea coherente, en específico que cada persona que ingresó a la Facultad mediante vía especial haya postulado a Plan Común y que no registre matrícula en otras carreras, igualmente las personas que ingresan por las vías que ve el DEMRE, deben registrar matrícula en Plan Común U. de Chile. Cabe destacar que pueden llegar a existir casos puntuales que se salgan a lo planteado.

De los 8.457 estudiantes que ingresaron entre el 2012 y 2021, existentes en las bases de la FCFM, quedaron 7.590. De los faltantes, 432 fueron por Bachillerato, 18 por estudios en el extranjero y 417 por no poder crear una relación 1:1 en los cruces entre bases.

2.4.3. Creación de variables

Se crearon diversas variables que buscan capturar efectos, que no se pueden apreciar con la información disponible. Estas variables son:

1. **Razón aprobación pandemia**: Variable que busca medir el nivel de exposición de los estudiantes en la pandemia, a través de una relación entre la cantidad ramos de Plan Común rendidos en pandemia y su dificultad; y el total de ramos de Plan Común rendidos. Se calcula para cada estudiante $i \in E$ (donde E es el set de estudiantes) como:

$$\text{Razón aprobación pandemia}_i = \frac{\sum_{r \in R_i} \text{Tasa de aprobación del ramo}_r}{\sum_{t \in T_i} 1 - \sum_{r \in R_i} \text{Tasa de aprobación del ramo}_r}$$

$$R_i = \{\text{Ramos de Plan Común rendidos en pandemia por el alumno } i\}$$

$$T_i = \{\text{Ramos de Plan Común rendidos por el alumno } i\}$$

$$\forall i \in Estudiantes$$

Los datos de entrada para calcular esta variable provienen de las bases de datos de la FCFM. Con esta variable se busca incluir el efecto respecto a la cantidad de cursos que tomaron los estudiantes en pandemia. El motivo de considerar esta variable es la diferencia significativa en las tasas de aprobación de los cursos dictados en pandemia respecto a sus mismas versiones en un semestre normal.

2. **Semestres pandemia**: Cantidad de semestres en pandemia que el estudiante cursó, mínimo un ramo de Plan Común. Los datos de entrada para calcular esta variable provienen de las bases de datos de la FCFM. Con esta variable se busca capturar el período de exposición de los estudiantes a la pandemia y cómo esto repercute en la variable estudiada.

3. **Prom Leng Mat**: Promedio de las pruebas PSU de lenguaje y matemáticas del colegio de egreso en el año que el estudiante termina su enseñanza media. Se calcula para cada colegio y año como:

$$\text{Prom Leng Mat}_{j,t} = \frac{\sum_{m \in M_{j,t}} \text{Puntaje PSU Matemáticas}_m + \sum_{l \in L_{j,t}} \text{Puntaje PSU Lenguaje}_l}{\sum_{m \in M_{j,t}} 1 + \sum_{l \in L_{j,t}} 1}$$

$$M_{j,t} = \{\text{Estudiantes del colegio } j \text{ en el año } t \text{ que rindieron la PSU de Matemáticas}\}$$

$$L_{j,t} = \{\text{Estudiantes del colegio } j \text{ en el año } t \text{ que rindieron la PSU de Lenguaje}\}$$

$$\forall j \in Colegios, t \in Años$$

Los datos de entrada para calcular esta variable provienen de las bases de datos del DEMRE y posteriormente se le asignan a cada estudiante que ingresan a la Facultad,

dado su colegio y año de egreso. Con esta variable se busca identificar el nivel de preparación de los alumnos por colegio, con la finalidad de agregar variables que midan la calidad del colegio de egreso, ya que, según la literatura es relevante considerar.

4. **Ratio NEM sobre Prom Leng Mat**: Variable que busca captar las discrepancias que se crean entre el rendimiento de los estudiantes y la PSU en los diversos colegios, se calcula para cada colegio y año como:

$$\frac{\text{Promedio NEM}_{j,t}}{\text{Prom Leng Mat}_{j,t}}, \forall j \in Colegios, t \in Años$$

donde:

$$\text{Promedio NEM}_{j,t} = \frac{\sum_{n \in N_{j,t}} \text{NEM estudiante}_n}{\sum_{n \in N_{j,t}} 1}$$

$$N_{j,t} = \{\text{Estudiantes del colegio } j \text{ que egresaron en el año } t\}$$

$$\forall j \in Colegios, t \in Años$$

Los datos de entrada para calcular esta variable provienen de las bases de datos del DEMRE y posteriormente se le asignan a cada estudiante que ingresan a la Facultad, dado su colegio y año de egreso. Con esta variable se busca agregar otra característica que mida la calidad del colegio de egreso, ya que, según la literatura es relevante de considerar.

5. **Egreso**: Clasifica el tiempo desde que un estudiante salió del colegio e ingresó a la universidad. Se define para cada estudiante $i \in E$ (donde E es el set de estudiantes) como:

$$\text{Egreso}_i = \begin{cases} 0 \text{ Si desde que egresó se demoró más de 1 año en ingresar a la universidad} \\ 1 \text{ Si desde que egresó se tardó 1 año en ingresar a la universidad} \\ 2 \text{ Si ingresó inmediatamente a la universidad} \end{cases}$$

$$\forall i \in Estudiantes$$

Los datos de entrada para calcular esta variable provienen de las bases de datos del

DEMRE. Con esta variable se busca captar el efecto que produce el tiempo entre egresar de la educación media e ingresar a la universidad, debido a que en otros estudios similares fue una variable relevante.

6. **Opción**: Lugar en que posicionó Plan Común - U. de Chile, al momento de postular. Esta variable es un número entero que va desde el 1 al 10. Los datos de entrada para calcular esta variable provienen del orden en que se registran los código de las carreras al momento de postular en la bases de datos del DEMRE. Con esta variable se busca captar la motivación del estudiante por pertenecer al Plan Común de la FCFM.

7. **Ing. Opción**: Indica si el estudiante solicitó una carrera de ingeniería en 1º posición al postular vía SUA. Se define para cada estudiante $i \in E$ (donde E es el set de estudiantes) como:

$$\text{Ing. Opción}_i = \begin{cases} 1 \text{ Si el estudiante solicitó una carrera de ingeniería en 1º posición} \\ 0 \text{ Si el estudiante no solicitó una carrera de ingeniería en 1º posición} \end{cases}$$

$$\forall i \in Estudiantes$$

   Los datos de entrada para calcular esta variable provienen de las bases de datos del DEMRE. Con esta variable se busca captar la motivación del estudiante por estar estudiando algo relacionado con sus gustos.

8. **Ingreso especial**: Indica si el estudiante ingresó por alguna vía distinta a la regular. Se define para cada estudiante $i \in E$ (donde E es el set de estudiantes) como:

$$\text{Ingreso especial}_i = \begin{cases} 1 \text{ Si el estudiante ingresa por una vía diferente a la regular} \\ 0 \text{ Si el estudiante ingresa mediante vía regular} \end{cases}$$

$$\forall i \in Estudiantes$$

   Los datos de entrada para calcular esta variable provienen de las bases de datos de la FCFM. Con esta variable se busca captar el efecto que tiene ingresar por una vía diferente a la regular, ya que como se vio en Rilling (2022), la tasa de deserción es diferente en las vías especiales.

9. **Delta puntaje**: Diferencia entre el puntaje ponderado con el que entra un estudiante y el puntaje de corte en el año de ingreso. Se calcula para cada estudiante $i \in E$ (donde E es el set de estudiantes) como:

$$\text{Puntaje ponderado}_i - \text{Puntaje de corte al ingreso}_i, \forall i \in Estudiantes$$

Los datos de entrada para calcular esta variable provienen de las bases de datos de la FCFM e información declarada por la universidad. Con esta variable se busca medir el impacto de tener puntajes altos o bajos con respecto al puntaje de corte en terminar Plan Común.

10. **Delta cero**: Indica la cantidad de puntaje que le faltó al estudiante, para ingresar de forma regular a la Facultad. Se calcula para cada estudiante $i \in E$ (donde E es el set de estudiantes) como:

$$\text{Delta cero}_i = \begin{cases} 0 \text{ Si Puntaje ponderado}_i \geq \text{Puntaje de corte al ingreso}_i, \forall i \in Estudiantes \\ \text{Puntaje ponderado}_i - \text{Puntaje de corte al ingreso}_i, \forall i \in Estudiantes, \text{ Si no} \end{cases}$$

Los datos de entrada para calcular esta variable provienen de las bases de datos de la FCFM e información declarada por la universidad. Con esta variable se busca identificar el nivel de injerencia de estar lejos de los requisitos para ingresar de forma regular, en terminar o no Plan Común.

11. **Mayor de edad**: Indica si un estudiante es mayor de edad al inicio del año de ingreso a la universidad. Se define para cada estudiante $i \in E$ (donde E es el set de estudiantes) como:

$$\text{Mayor de edad}_i = \begin{cases} 1 \text{ Si es mayor de edad al inicio del año de ingreso a la universidad} \\ 0 \text{ Si es menor de edad al inicio del año de ingreso a la universidad} \end{cases}$$

$$\forall i \in Estudiantes$$

Los datos de entrada para calcular esta variable provienen de las bases de datos del DEMRE. Con esta variable se busca medir el nivel impacto de ser mayor de edad desde inicio del año de ingreso en terminar o no Plan Común.

12. **Segmento**: Indica el segmento socio-económico al que pertenece el estudiante. Esta variable fue una redistribución de la variable original existente en la base del DEMRE. Se define para cada estudiante $i \in E$ (donde E es el set de estudiantes) como:

$$\text{Segmento}_i = \begin{cases} 0 \text{ Si el ingreso bruto mensual del grupo familiar declarado, es inferior a los \$356.000 pesos chilenos} \\ 1 \text{ Si el ingreso bruto mensual del grupo familiar declarado, está entre los \$356.000 y los \$938.999 pesos chilenos} \\ 2 \text{ Si el ingreso bruto mensual del grupo familiar declarado, es superior a los \$938.999 pesos chilenos} \end{cases}$$

$$\forall i \in Estudiantes$$

Esta variable busca capturar el impacto del nivel de ingreso en terminar o no Plan Común.

2.4.3.1. Creación de la variable a pronosticar

Se creó la variable a predecir y_i, la cual se mencionó en la sección 2.2. Se debe destacar que sólo se consideran los ramos obligatorios de Plan Común para construir esta variable, lo cual implica que será 1, si la persona ha aprobado todos los ramos obligatorios de Plan Común, independiente si le quedan electivos por cursar. Esta decisión se basa en que los cursos no obligatorios, como son los electivos de formación, tienen una tasa de aprobación que rodea el 100 % y muchos alumnos los dejan para el final de su carrera, y_i será 0 en caso contrario. Es importante mencionar que el objetivo de esta variable es permitir comprender qué factores afectan la finalización de Plan Común y cuáles llevan a la deserción involuntaria por mal rendimiento. En la siguiente sección se limitan los casos donde y_i=0 es deserción involuntaria por mal rendimiento.

2.4.4. Preprocesamiento

Se consideraron sólo a los alumnos que ingresaron entre el 2012 al 2018, esto se debe principalmente a que Plan Común tiene una duración mínima de cuatro semestres y los estudiantes que ingresan posterior al 2018 no alcanzaron a rendir de manera normal sus primeros

semestres, debido a la pandemia, estallido social y la reestruccturacion de la malla de Plan Común. También se eliminaron los alumnos que hasta el año 2022 seguían cursando ramos de Plan Común, pues hasta el momento en que se realizaron los análisis, dichos estudiante siguen cursando Plan Común, por lo que podrían tanto aprobarlo como desertarlo.

Se eliminó de la base a todos los estudiantes:

i. Que rindieron menos de 2 ramos en la Facultad, dado que son casos anormales.

ii. Que no pasaron Plan Común (y_i =0), y no reprobaron un mismo ramo más de 2 veces. Esto se debe a que en el reglamento de la FCFM, el mínimo requisito para ser eliminado del programa por razones académicas es reprobar un mismo ramo 2 veces, por lo cual a estos estudiantes no se le puede asociar directamente no aprobar Plan Común con razones académicas. Cabe mencionar que entre los estudiantes excluidos por este motivo existen quienes desertaron debido a su bajo rendimiento antes de que reprobaran Plan Común, pero no se pueden identificar con total seguridad, ya que, comparten características con personas que desertaron por motivos distintos y además existen casos en los cuales estudiantes similares pudieron aprobar el plan de estudios.

De tal forma, que de 7.590 estudiantes que había en la base, quedaron 4.522. De los eliminados, 2.630 fueron por haber ingresado posterior al 2018, 46 por estar cursando ramos de Plan Común en el 2022, 66 casos anómalos (rindieron menos de 2 ramos en la Facultad) y 326 por haber desertado de Plan Común debido a razones distintas a las académicas.

Finalmente, se excluyeron de los próximos análisis todas las variables que contienen grandes cantidades de datos faltantes, que son subconjunto de otra variable, que poseen un alto grado de complejidad y aquellas que de acuerdo a la literatura no son relevantes.

2.4.5. Variables disponibles para ser ocupadas en los diversos modelos

Las variables disponibles para ser ocupadas en los diferentes modelos son:

Nombre de la variable	Base de origen	Descripción
1.- Puntaje de ingreso	DEMRE/FCFM	Puntaje de ingreso a la Facultad
2.- NEM	DEMRE/FCFM	NEM del estudiante
3.- Ranking	DEMRE/FCFM	Ranking del estudiante
4.- Opción	Creada	Lugar en que colocó la opción de Plan Común - U. de Chile al postular
5.- P022	DEMRE	Número de horas que trabaja el estudiante, al momento de egresar de la educación media
6.- Puntaje de Mat	DEMRE/FCFM	Puntaje de matemáticas con el que ingresa a la Facultad
7.- Puntaje de Len	DEMRE/FCFM	Puntaje de lenguaje con el que se ingresa a la Facultad
8.- Puntaje Ciencias	DEMRE/ FCFM	Puntaje de ciencias con el que se ingresa a la Facultad
9.- Rural	MINEDUC	1 si el estudiante egresó de un colegio rural según los estándares del MINEDUC, 0 si no
10.- Tipo de colegio*	DEMRE	Indica la dependencia del colegio, lo cual pueden ser: - Particular Pagado - Particular Subvencionado - Municipal
11.- Segmento	Creada	Indica el segmento socio-económico del estudiante, en específico toma los valores: - 0, si el ingreso bruto mensual del grupo familiar declarado en los formularios del DEMRE, es inferior a los $356.000 pesos chilenos - 1, si el ingreso bruto mensual del grupo familiar declarado en los formularios del DEMRE, está entre los $356.000 y los $938.999 pesos chilenos - 2, si el ingreso bruto mensual del grupo familiar declarado en los formularios del DEMRE, es superior a los $938.999 pesos chilenos

Nombre de la variable	Base de origen	Descripción
12.- Delta puntaje	Creada	Diferencia entre el puntaje de corte con el puntaje de ingreso al momento de ingresar a la Facultad
13.- Mayor de edad	Creada	1 si es mayor de edad al incio del año de ingreso a la universidad, 0 si no
14.- Mujer	DEMRE	Variable binaria que es 1 si es mujer, 0 si no
15.- P076*	DEMRE	Rama educacional del establecimiento de egreso, que puede ser: - Humanista Científico Diurno - Humanista Científico Nocturno - Humanista Científico – Validación de estudios - Humanista Científico – Reconocimiento de estudios - Técnico Profesional Comercial - Técnico Profesional Industrial - Técnico Profesional Servicios - Técnico Profesional Agrícola - Técnico Profesional Marítima
16.- Tipo de ingreso*	FCFM	Indica el tipo de ingreso, donde se consideraron: - Ingreso regular - Equidad de género - 5 % superior colegios municipalizados - Sistema de Ingreso Prioritario de Equidad Educativa (SIPEE) - Deportista - PACE

Nombre de la variable	Base de origen	Descripción
17.- Egreso	Creada	Variable que considera el tiempo entre el egreso de la educación media e ingreso a la educación superior, que toma los valores: - 2, si ingresó a la educación superior al año siguiente de haber egresado de la educación media - 1, si ingresó a la educación superior al segundo año de haber egresado de la educación media - 0, si ingresó a la educación superior después de 3 o más año de haber egresado de la educación media
18.- Delta cero	Creada	Indica la cantidad de puntaje que le faltó al estudiante, para ingresar de forma regular a la Facultad. Es 0 en caso de que el puntaje ponderado de ingreso sea mayor al puntaje de corte
19.- Ing. Opción	Creada	1 si el estudiante solicitó una carrera de ingeniería en 1º posición al postular vía SUA, 0 si no
20.- Semestres pandemia	Creada	Cantidad de semestres en que cursó ramos de Plan Común en pandemia
21.- Ingreso especial	Creada	1 si ingresa por medio de una vía especial, 0 si no
22.- Prom Leng Mat	Creada	Promedio de las pruebas PSU de lenguaje y matemáticas, del colegio de egreso en el año que el estudiante termina su enseñanza media
23.- Ratio NEM sobre Prom Leng Mat	Creada	Promedio NEM colegio dividido en promedio lenguaje matemáticas colegio
24.- Razón aprobación pandemia	Creada	Aprobación de los ramos de Plan Común que rinde el estudiante en pandemia relacionados con la cantidad total de ramos de Plan Común que rindió
** Las variables categóricas se transformaron en dummies*		

Tabla 2.1: Variables ocupadas en el análisis

2.4.6. Desbalance de clases

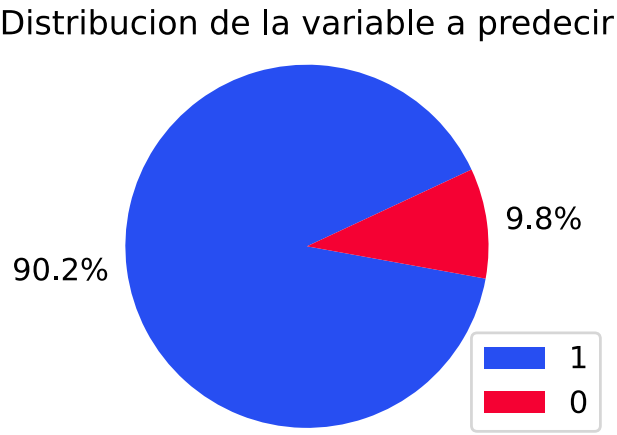


Figura 2.1: Distribución de la variable y_i en los datos, N=4.522.

La variable dependiente en los diversos modelos será y_i, que ocupará el resto de información como variables independientes, para poder predecir su valor. Es importante destacar que se aprecia un desbalance significativo en los datos de la variable dependiente, lo cual fue considerado para el desarrollo de la investigación. Además, en la Figura 2.2 se puede apreciar que este desbalance está presente en cada uno de los años a estudiar.

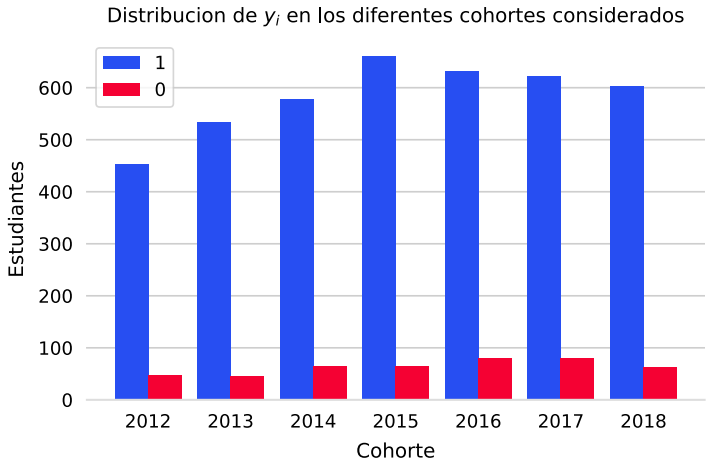


Figura 2.2: Distribución de y_i en los diferentes cohortes considerados.

Cabe mencionar, que las personas que ingresan por las vías especiales son una cantidad baja en consideración con las que ingresan de maneras regulares.

Tipo de ingreso	Cantidad
Ingreso Regular	4062
Equidad de Género	169
5 % Colegios Municipalizados	125
Sistema de Ingreso Prioritario de Equidad Educativa (SIPEE)	96
Deportista	69
PACE	1

Tabla 2.2: Detalle de los Tipos de ingreso en la Base de Datos (2012-2018).

Basado en lo anterior, se debe destacar que la variable *Delta puntaje* toma valores negativos sólo para las personas que ingresan por Equidad de Género, 5 % Colegios Municipalizados, Sistema de Ingreso Prioritario de Equidad Educativa y Deportista. De manera similar, la variable Delta Cero es mayor a cero sólo para las vías de ingreso mencionadas anteriormente, por lo cual implica que estas variables entreguen información relevante de la vía de acceso indirectamente. Finalmente, es importante mencionar que existen variables altamente correlacionadas con otras, lo cual será considerado en la construcción de los modelos que no permitan esto. Se puede observar más información descriptiva de la base de datos en el Anexo A.

2.5. Modelos

Con el objetivo de predecir y_i, se ocuparon 6 algoritmos de aprendizaje supervisado, los cuales buscan crear la mejor función que considere como input la información de los estudiantes y prediga y_i.

Específicamente se ocupó la librería de Python scikit-learn, para desarrollar todos los modelos. Asimismo, con la finalidad de tratar el desequilibrio existente en los datos, se aplicaron las siguientes tres técnicas:

- **Sobre-muestreo aleatorio**: se duplican de manera aleatoria los datos pertenecientes a la clase minoritaria presente en la base. Esta técnica fue aplicada en todos los modelos y se debe tener en consideración que puede generar overfitting en la clase minoritaria y problemas de generalización.

- **Peso de las clases equilibrado**: consiste en determinar class_weight= “balanced”, en los parámetros de los modelos, lo cual utiliza los valores de y_i en la base de entrenamiento,

para ajustar automáticamente los pesos de cada clase y así el modelo pueda tener en consideración que es relevante lograr una “buena” predicción de la clase minoritaria. Esta técnica fue aplicada en los modelos 2.4.1, 2.4.2, 2.4.3 y 2.4.6. Se debe considerar que puede generar overfitting en la clase minoritaria y problemas de generalizacion.

– **Sobre muestro sintético SMOTE (Synthetic Minority Oversampling Technique)**: "Genera nuevas instancias de la clase minoritaria interpolando los valores de las instancias minoritarias más cercanas a una dada"(Moreno, Riquelme, Rodríguez, Ruiz y Sicilia, 2009, p. 2). Esta técnica fue aplicada en los todos los modelos y se debe tener en consideracion que puede provocar problemas en la naturaleza de la clase minoritaria, generando dificultades al clasificar.

Cabe mencionar que estas técnicas se aplicaron en paralelo, nunca en simultáneo, es decir, ninguna de las técnicas se aplicó en conjunto con otra.

Los modelos predictivos utilizados se seleccionaron en base a los más usados en la literatura y los que presentaron mejores resultados en los estudios revisados, por lo cual se decidió ocupar los siguientes modelos:

2.5.1. Support Vector Machine (SVM)

El modelo de Support Vector Machine se crea “correlacionando los datos en un espacio de características de grandes dimensiones de forma que los puntos de datos se puedan categorizar” (IBM, 2021), con lo cual se detecta una función que separe los datos según sus categorías. Tras ello, se puede predecir el grupo al que pertenece un nuevo dato en base a sus características (IBM, 2021).

Para este modelo se regularizaron los parámetros:

– Kernel: tipo de kernel utilizado en el modelo.

– Degree: determina el grado de la función, cuando se ocupa kernel = poly.

– C: parámetro de regularización, que permite flexibilizar la función que separa los datos.

– Coef0: término que afecta el cálculo del Kernel cuando se ocupa poly o sigmoid como kernel.

2.5.2. Árbol de Clasificación (DT)

El Árbol de Clasificación es un algoritmo que tiene una estructura de árbol jerárquica, que consta de un nodo raíz y las ramas que poseen los nodos internos y nodos hoja (IBM, s.f., s.p.), donde a través de decisiones binarias en los nodos internos y raíz, se busca clasificar los datos de la mejor manera posible al tipo de dato que corresponde, lo cual es representado en los nodos hojas.

Para este modelo se regularizaron los parámetros:

- Max_depth: profundidad máxima del árbol.
- Min_samples_split: número mínimo de datos necesarias para que se pueda dividir un nodo.
- Max_features: número de variables a considerar al momento de buscar la mejor división.
- Criterion: función bajo la cual se mide la calidad de la división de los nodos.

2.5.3. Random Forest (RF)

Random Forest, combina una cantidad determinada de árboles de clasificación individuales que operan como un solo conjunto, donde cada árbol entrega una predicción de la clase y la clase con más votos se convierte en la predicción del modelo (Yiu, 2019, s.p.).

Para este modelo se regularizaron los parámetros:

- N_estimators: número de árboles con el cual se construye el modelo.
- Max_depth: profundidad máxima para cada árbol.
- Max_feature: número de variables a considerar al momento de buscar la mejor división.
- Min_samples_split: número mínimo de datos necesarios para dividir un nodos.
- Min_samples_leaf: número mínimo de datos necesarios en los nodos hoja de cada árbol.

2.5.4. Red Neuronal (RN)

La Red Neuronal es un "[...] modelo simplificado que emula el modo en que el cerebro humano procesa la información [...] " (IBM, 2021, s.p.), donde se utilizan nodos que simulan

las neuronas y se agrupan por capas. Los nodos de cada capa reciben información de los nodos de la capa anterior, la cual transforma y envían a la siguiente capa. Este proceso sigue hasta llegar a la capa de salida, que es en donde el modelo toma la decisión del pronóstico en base a la información recibida, formando así una estructura, que se parece al cerebro humano.

Para este modelo se regularizaron los parámetros:

- Solver: determina el solver utilizado para la optimización de pesos.
- Activation: función de activación.
- Hidden_layer_sizes: tupla que indica la cantidad de neuronas por capa.
- Alpha: fuerza de la regularización.
- Learning_rate_init: tasa de aprendizaje inicial utilizada y sirve soló cuando se utiliza sgd o adam como solver.

2.5.5. K vecinos más cercanos (KNN)

K vecinos más cercanos, es un modelo clasificador de aprendizaje supervisado no paramétrico, donde cada nuevo dato se clasifica, según la distancia de este a cada uno de los datos existentes, y se ordenan dichas distancias de menor a mayor (Merkle Company, 2020), donde se consideran los K vecinos cercanos para tomar la decisión del modelo. Cabe mencionar que K es entregado como parámetro.

Para este modelo se regularizaron los parámetros:

- Algorithm: algoritmo utilizado para calcular los vecinos más cercanos
- N_neighbors: cantidad de vecinos considerados para tomar la decisión del modelo.
- Leaf_size: tamaño de hoja a ocupar cuando se utilizan los algoritmos BallTree o KDTree.
- Weights: función que edita la importancia de los nodos considerados en la predicción.
- P: parámetro que permite editar la forma en que se calcula la distancia.

2.5.6. Regresión Logística (Logit)

La regresión Logística es modelo matemático de clasificación que predice variables dicotómicas ajustando una función sigmoide, en base a distintas variables que se le entregan como input.

2.5.7. Ajustes Generales

La base se dividió en 70 % de los datos para entrenamiento y el 30 % restante, para testeo. En todos los modelos se mantuvo el umbral de decisión predefinido, es decir, si la probabilidad es menor a 0.5, $\widehat{y_i}$ es 0, en caso contratrio $\widehat{y_i}$ es 1.

Para los modelos SVM, DT, RF, RN, y KNN, se ocuparon todas las variables dependientes para ajustar los modelos. Además, se ocupó validación cruzada (cross-validation) para seleccionar los mejores modelos de entrenamiento, con la finalidad de evitar el sobreajuste y mantener independiente la construcción del modelo con los datos de testeo. Para la validación cruzada se ocuparon 5 pliegues (K) y se consideraron las métricas recall, accuracy, f1-score y roc_auc para evaluar el rendimiento del modelo en su construcción.

Para el modelo Logit, se hizo un filtro de variables, a través de la regularización Lasso (L1), que penaliza las variables más irrelevantes con la finalidad de omitirlas en proceso de la búsqueda de la mejor regresión logística. Es importante destacar que se aplicó validación cruzada para todos los modelos posibles, donde cada uno se obtiene en base a una combinación que cumpla los supuestos de la regresión logística y sea una combinación de las variables que resultaron de la penalización L1, con lo cual se crearon modelos con todas las combinaciones posibles, los cuales fueron evaluados en base a sus resultados en la validación cruzada, que posee la misma configuración que en los otros modelos. Cabe decir, que para los modelos RN y KNN, no se ocupó la técnica Peso de las clases equilibrado, dado que la librería a ocupar no lo permite, debido al funcionamiento y estructura de estos algoritmos que en primera instancia deberían poder contrarrestar el desequilibrio, por ello fue remplazada, por ajustar los modelos sin controlar el desequilibrio.

Capítulo 3

Resultados

Antes de empezar, es necesario recordar que se desconoce qué tipo de modelo se ajusta de mejor manera a los datos, por lo cual es relevante estudiar los resultados al momento de predecir y la probabilidad con la cual predicen, esto último con el objetivo de reconocer si existen datos que son difíciles de clasificar y estudiar con que seguridad clasifican.

También se requiere mostrar todos los posibles resultados que se pueden obtener al variar el umbral de corte, ya que, cada iniciativa tiene diferentes requerimientos y limitaciones que a priori se desconocen, por lo cual es importante entregar distintas herramientas que puedan adecuarse a cada una de las iniciativas, para que así, los estudiantes que corren mayor riesgo de no terminar Plan Común participen de estas.

3.1. Predicciones con datos previos al ingreso a Plan Común

Considerando como input los datos descritos con anterioridad (sección 2.4.5.) se obtuvieron 3 modelos por cada algoritmo de aprendizaje supervisado (detalles en el Anexo C), donde cada uno representa el resultado de la técnica utilizada para tratar el desequilibrio, de los cuales se escogió un modelo por algoritmo basado en sus diferentes métricas de rendimiento, priorizando el F1-score de $y_i = 0$, para así poder obtener y reconocer la categoría $y_i = 0$ de la mejor manera, sin perder significativamente poder predictivo de la clase $y_i = 1$, ya que, los $y_i = 0$ representan a las personas que no son capaces de terminar Plan Común debido a razones académicas e identificarlos en este período, es decir antes de que inicien su primer

año académico en la FCFM, ayuda a focalizar las acciones preventivas desde antes.

Algunos ejemplos de iniciativas donde se pueden ocupar estos modelos son en determinar qué alumnos deben cursar un semestre de preparación antes de rendir los cursos obligatorios de Plan Común o quienes deben asistir a tutorías académicas durante el 1ºsemestre.

A continuación, se muestran los resultados de testeo de los modelos escogidos:

Modelo	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
KNN **	0.93	0.68	0.79	0.16	0.55	0.25	0.67
SVM *	0.95	0.67	0.78	0.19	0.69	0.29	0.67
DT **	0.93	0.69	0.79	0.16	0.53	0.25	0.67
RF **	0.93	0.65	0.77	0.16	0.58	0.25	0.64
RN **	0.95	0.70	0.80	0.20	0.66	0.30	0.69
Logit ***	0.94	0.73	0.82	0.19	0.57	0.29	0.72
* *El Modelo escogido es con la técnica de peso de las clases equilibrado*							
** *El Modelo escogido es con la técnica de sobre-muestreo aleatorio*							
*** *El Modelo escogido es con la técnica de sobre-muestreo sintético SMOTE*							

Tabla 3.1: Resumen del rendimiento de los diversos modelos en Test

Se puede observar en la Tabla 3.1 que, independiente del modelo, el F1-score en la categoría y_i= 0 es bajo. A pesar de lo anterior, los modelos son capaces de identificar sobre el 53 % de esta categoría, llegando en el mejor caso a identificar el 69 % de esta clase. Esto, en otras palabras, significa que en todos los modelos se puede identificar sobre el 53 % de las personas que no son capaces de terminar Plan Común debido a razones académicas.

Cabe destacar, que la base está desbalanceada, ya que la cantidad de alumnos que son categorizados como y_i= 0 por generación son pocos en comparación con el resto, lo cual implica que pequeñas variaciones en el recall de y_i= 1, muevan una gran cantidad de datos en comparación con la otra categoría. Además, se debe tener en consideración que los modelos tienen una baja precisión para la categoría y_i= 0, lo cual indica que estos no son capaces de identificar de forma consistente porqué predicen $\widehat{y_i}$= 0. Por ello, al ocupar estas herramientas, se debe tener en consideración que entre los $\widehat{y_i}$= 0 que se pronostiquen, existirá una gran cantidad de falsos negativos, lo que significa que cualquier iniciativa y/o programa que seleccione su público objetivo en base a estos modelos, incluirá en su mayoría personas que son y_i= 1, donde, en el mejor de los casos, existirá una razón aproximada de cada cinco predicciones, una será un y_i= 0 mientras las otras cuatro serán falsos negativos.

Complementando lo anterior, se puede observar en la Figura 3.1 el comportamiento de

cada modelo al momento de clasificar y la fuerza con que clasifica los datos, como una función de distribución acumulada, en donde el eje x representa el porcentaje acumulado, mientras que el eje y representa el valor a predecir menos la probabilidad obtenida en el modelo, lo cual hace que los valores menores a -0.5 sean datos y_i= 0 mal clasificados, los datos de -0.5 a 0 datos y_i= 0 bien clasificados, de 0 a 0.5 datos y_i= 1 bien clasificados y mayores a 0.5 datos y_i= 1 mal clasificados.

Con lo anterior se puede decir que el modelo SVM es el modelo que mejor clasifica los y_i= 0 seguido por RN, mientras que Logit es el modelo que mejor clasifica los y_i= 1 seguido de RN. Asimismo, se puede apreciar que las curvas de los modelos SVM y RF, crecen menos que el resto cuando sobrepasan el umbral del eje y igual 0.5, lo que significa que estos modelos predicen una probabilidad que difiere por menos con la probabilidad correcta que permitiría clasificar bien a los y_i= 1 mal clasificados, mientras que en el caso de las curvas bajo el umbral del eje y igual -0.5 se puede ver que el modelo SVM es el que más se acerca a éste, por lo cual es el que a priori se equivoca por menos en la probabilidad de los y_i= 0 mal clasificados.

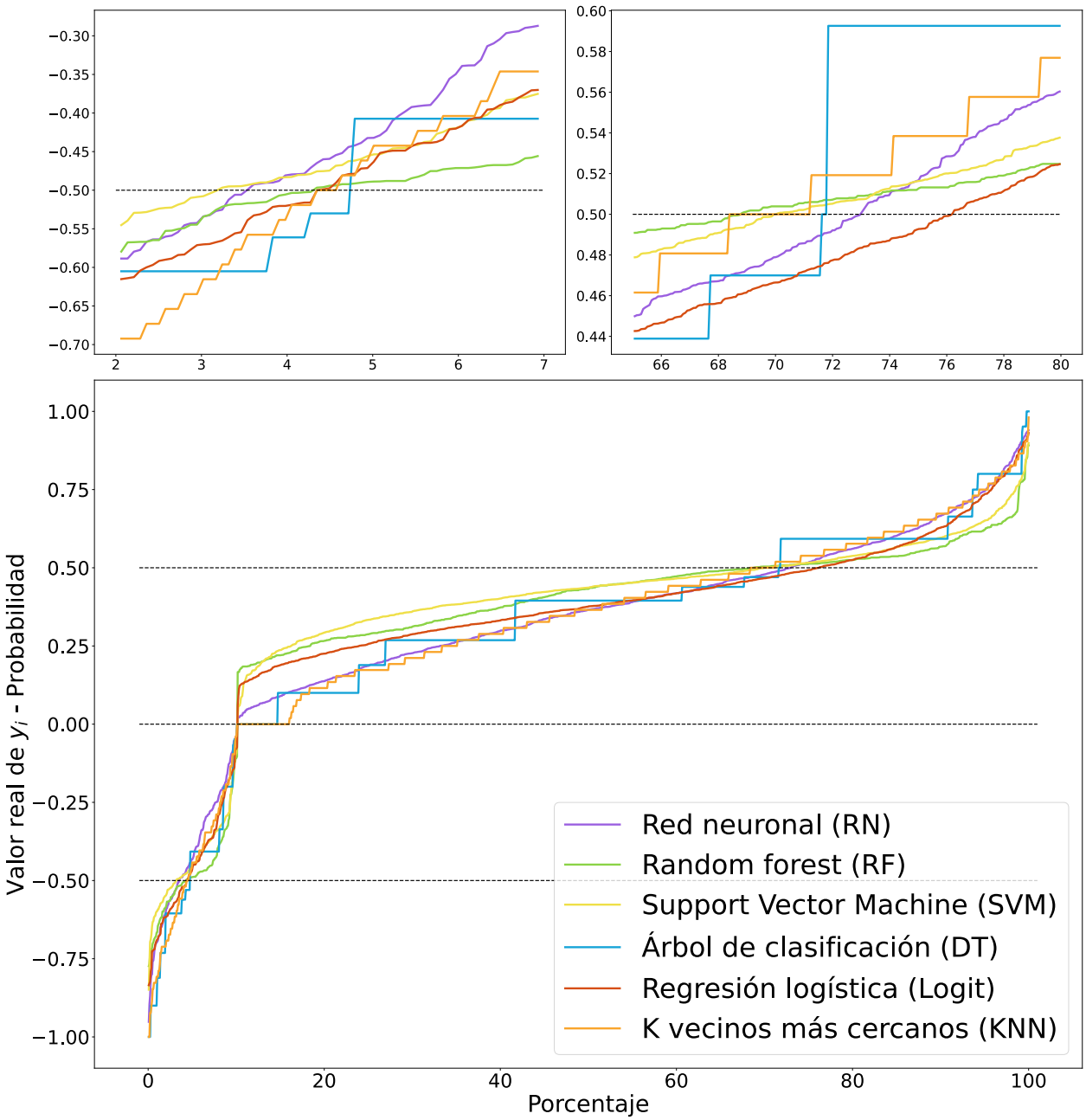


Figura 3.1: Porcentaje de datos vs probabilidad.

Modelo	Máximo error $y_i = 1$	Máximo error $y_i = 0$	Mínimo error $y_i = 1$	Mínimo error $y_i = 0$	Promedio error $y_i = 1$	Promedio error $y_i = 0$	Promedio de todos los errores
KNN **	0.481	0.499	0.019	0.001	0.145	0.188	0.151
SVM *	0.399	0.349	0.001	0.001	0.085	0.080	0.084
DT **	0.499	0.499	0.001	0.030	0.150	0.189	0.155
RF **	0.392	0.274	0.001	0.002	0.063	0.093	0.067
RN **	0.439	0.452	0.001	0.005	0.147	0.144	0.146
Logit ***	0.431	0.335	0.000	0.000	0.135	0.123	0.133

* *El Modelo escogido es con la técnica de peso de las clases equilibrado*

** *El Modelo escogido es con la técnica de sobre-muestreo aleatorio*

*** *El Modelo escogido es con la técnica de sobre-muestreo sintético SMOTE*

Tabla 3.2: Detalle del error en lo datos mal clasificados en Test por modelo.

Se puede observar de la Tabla 3.2 que SVM y RF, son los modelos que menos se equivocan en promedio al determinar la probabilidad de la clase y_i, lo cual va en línea con lo anterior, además de que son los modelos con los errores máximos más bajos. Por el contrario, KNN y DT son los modelos que se equivocan por más al determinar las distintas clases y poseen los errores máximos más altos.

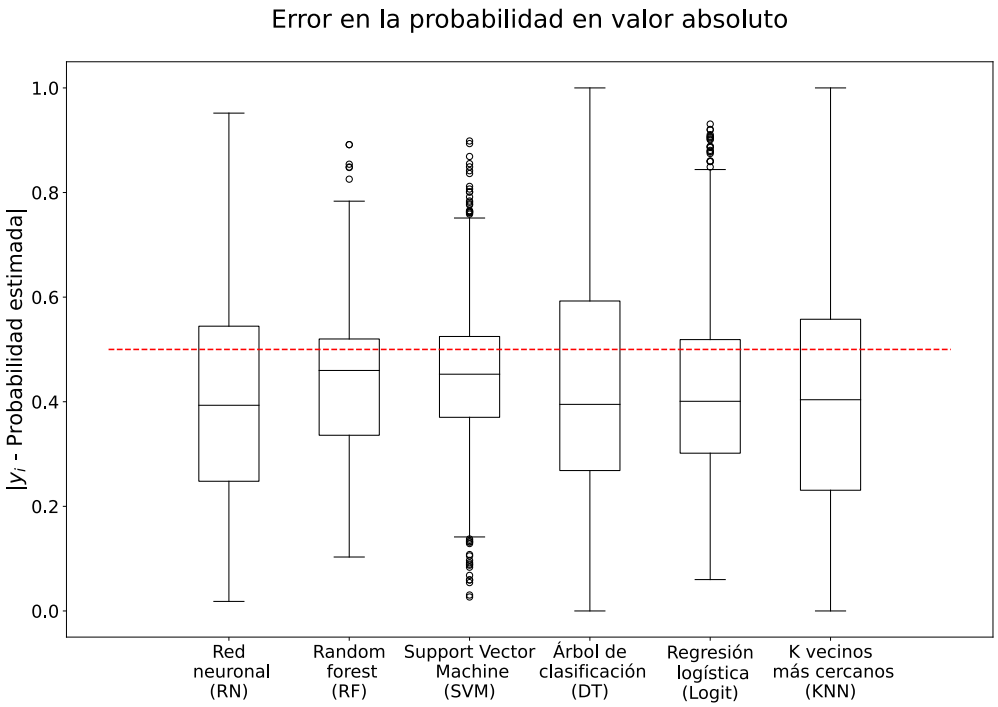


Figura 3.2: Error en la probabilidad en valor absoluto.

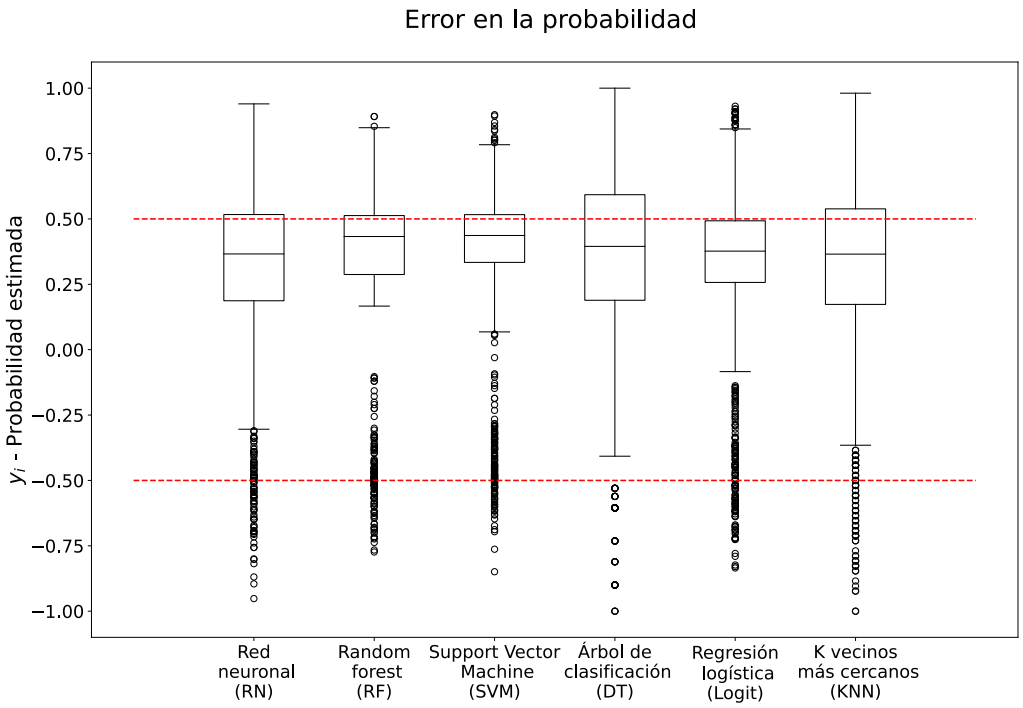


Figura 3.3: Error en la probabilidad.

Se puede apreciar de la Figura 3.3 y 3.4 que los modelos RF y SVM, tienen una menor dispersión al pronosticar la probabilidad y tiene una mediana muy cerca a los límites de eje y igual 0.5, lo cual se traduce en que pronostican probabilidades muy al límite. De igual manera, se aprecia que el modelo de DT es el modelo con mayor dispersión. Complementando con lo expuesto en la tabla de probabilidad, se puede concluir que es el modelo que más difiere entre la probabilidad mínima para clasificar correctamente con la probabilidad estimada. Asimismo, se debe destacar que los modelos RN y Logit son los modelos que mayor cantidad de datos tienen dentro de los diversos límites y, además, son de los que tienen su mediana más cerca al 0, lo cual se traduce en que sus probabilidades tienden a ser más "fuertes" al momento de predecir.

Dado que el rendimiento de los modelos es diferente y su forma de clasificar y equivocarse igual, es necesario mostrar cómo cambian sus resultados en los diferentes umbrales de decisión, con la finalidad de explicitar todos los rendimientos posibles, para que así las diferentes iniciativas puedan elegir el resultado que más se acomode a sus requerimientos, limitaciones y cantidad $y_i = 0$ que deseen abarcar, lo cual se puede deducir de la Figura 3.4.

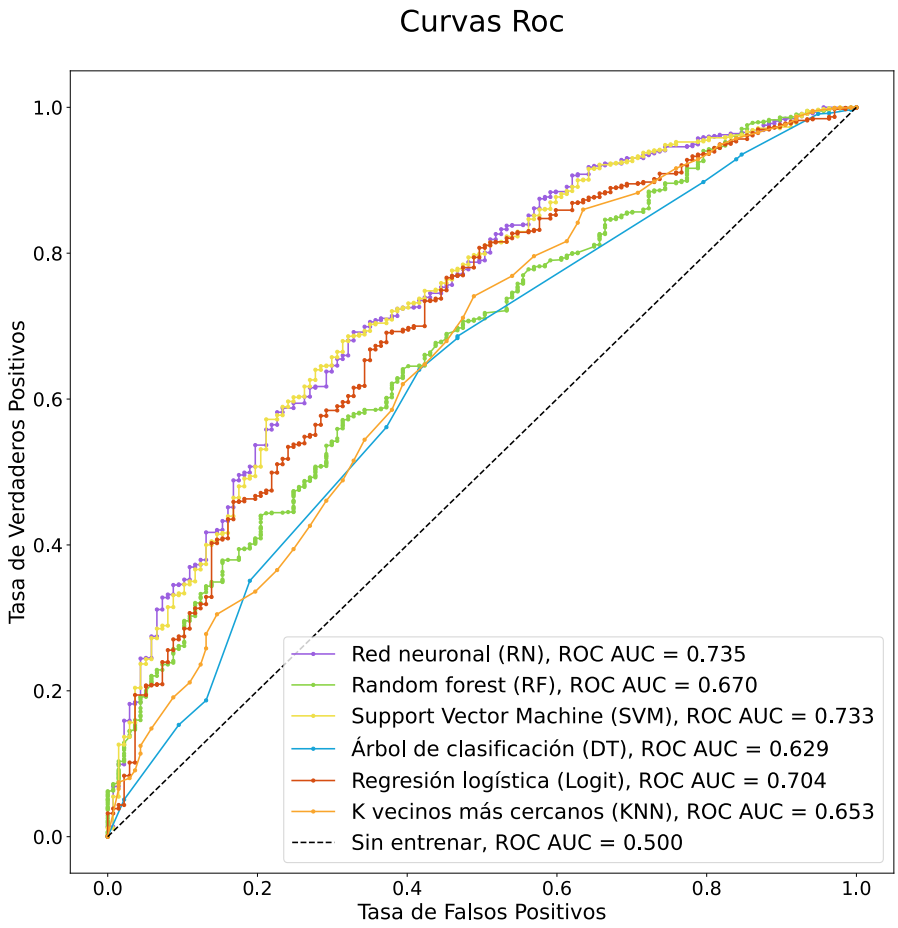


Figura 3.4: Curvas ROC de los diversos modelos.

Complementando lo anterior, en la Figura 3.4 se puede observar que el modelo con mayor AUC es RN, el cual tiene un valor de 0.735. Esto se traduce en que es el modelo con mejor poder discriminatorio, lo que significa que puede distinguir mejor entre un alumno que pasará Plan Común y uno que no. Además, se aprecia que si se quiere modificar el umbral con el objetivo de aumentar la cantidad de $y_i = 0$ (verdaderos negativos) se puede llegar a identificar correctamente cerca del 80 % de esta población manteniendo un recall de $y_i = 1$ superior al 50 %. De igual manera, se debe recordar que la base está desbalanceada, lo cual significa que esto llevará a clasificar incorrectamente una cantidad mayor a la que se clasificará correctamente. Además, se aprecia que el modelo SVM posee una curvas ROC similares con RN, despues viene Logit en similitud. Por el contrario, los modelos KNN, RF y DT, difieren por mucho con RN. Con base en todo lo expuesto anteriormente, se determina que el modelo más

completo es RN.

3.1.1. Variables importantes en la predicción

Con el objetivo de comprender cómo afectan las diversas variables en la probabilidad de predecir si una persona terminará Plan Común o no, se estudiaron sus efectos en el modelo RN dado que fue el modelo más completo entre todos, y el modelo Logit, ya que, es el más interpretable, posee el mayor accuracy, es el que mejor identifica los $y_i = 1$ y es el cuarto mejor modelo en identificar $y_i = 0$, donde difiere por muy poco con el tercero.

3.1.1.1. Análisis RN

Para comprender cómo cada variable afecta este modelo se ocupó la librería SHAP de Python, donde su nombre proviene de SHapley Additive exPlanations y permite dar un enfoque en base a teoría de juegos, ocupando los Valores de Shapley para poder explicar comó afecta cada variable en los modelo de aprendizaje automático (SHapley Additive exPlanatios, 2018, s.p.).

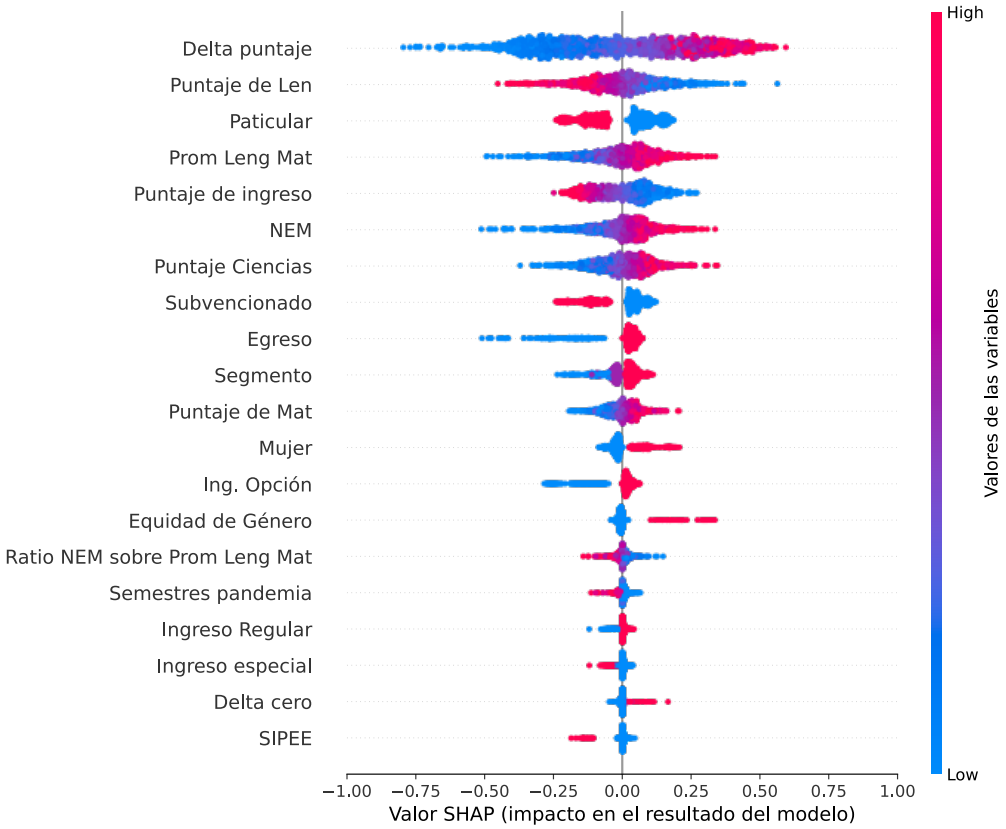


Figura 3.5: Efecto de las variables en el modelo RN.

En la Figura 3.5 las variables están ordenadas de mayor a menor efecto en la predicción y se aprecia cómo sus valores más altos (rojo) y bajos (azul) afectan en el pronóstico, donde el lado negativo significa que aumenta la probabilidad de desertar y el lado positivo lo contrario. Se rescata que las 5 variables más importantes para este modelo son Delta puntaje, Puntaje de Len, Particular, Promedio Leng Mat y Puntaje de ingreso. Además, se ve que las variables Delta puntaje, Prom Leng Mat, NEM, Puntaje Ciencias, Egreso, Segmento, Puntaje de Mat, Mujer, Ing. Opción, Equidad de Género, Ingreso especial y Delta cero aparece una relación directamente proporcional entre sus valores y la probabilidad de predecir $\widehat{y_i}= 1$. Mientras que en las variables Puntaje de Len, Particular, Puntaje de Ingreso, Subvencionado, Ratio NEM sobre Prom Leng Mat, Semestres pandemia, Ingreso especial y SIPEE presentan una relación inversamente proporcional entre sus valores y la probabilidad de predecir $\widehat{y_i}= 1$.

Es importante mencionar que las variables que no aparecen en el gráfico no son relevantes para el modelo y los efectos descritos anteriormente pueden diferir con la realidad, lo cual se puede deber a las interacciones que se crean en el modelo con la finalidad de predecir correctamente, por ejemplo, se aprecia que el puntaje de ingreso afecta negativamente la probabilidad de terminar Plan Común ($y_i= 1$), lo cual es contrario a lo que se esperaría. Además, entre la variable a predecir y el puntaje de ingreso existe una correlación positiva, lo que lleva a pensar que este efecto, fuera de los cálculos y relaciones creadas por el modelo, no se puede apreciar. A pesar de lo anterior, sirve para dar una visión importante de qué variables juegan un rol importante.

Es valioso mencionar que la vía de ingreso toma un rol importante en el análisis de forma indirecta, ya que sólo las personas que ingresan por Equidad de Género, 5 % Colegios Municipalizados, Sistema de Ingreso Prioritario de Equidad Educativa o Deportista tienen Delta puntaje menor a 0, lo que conlleva a interpretar que las personas que ingresan por estas vías de acceso ven afectada negativamente la probabilidad de que $\widehat{y_i}$ sea 1, es decir, Delta puntaje explica mayoritariamente el efecto de la vía de ingreso. Además, este efecto podría estar relacionado con la interpretación de Puntaje de ingreso, ya que están altamente correlacionadas.

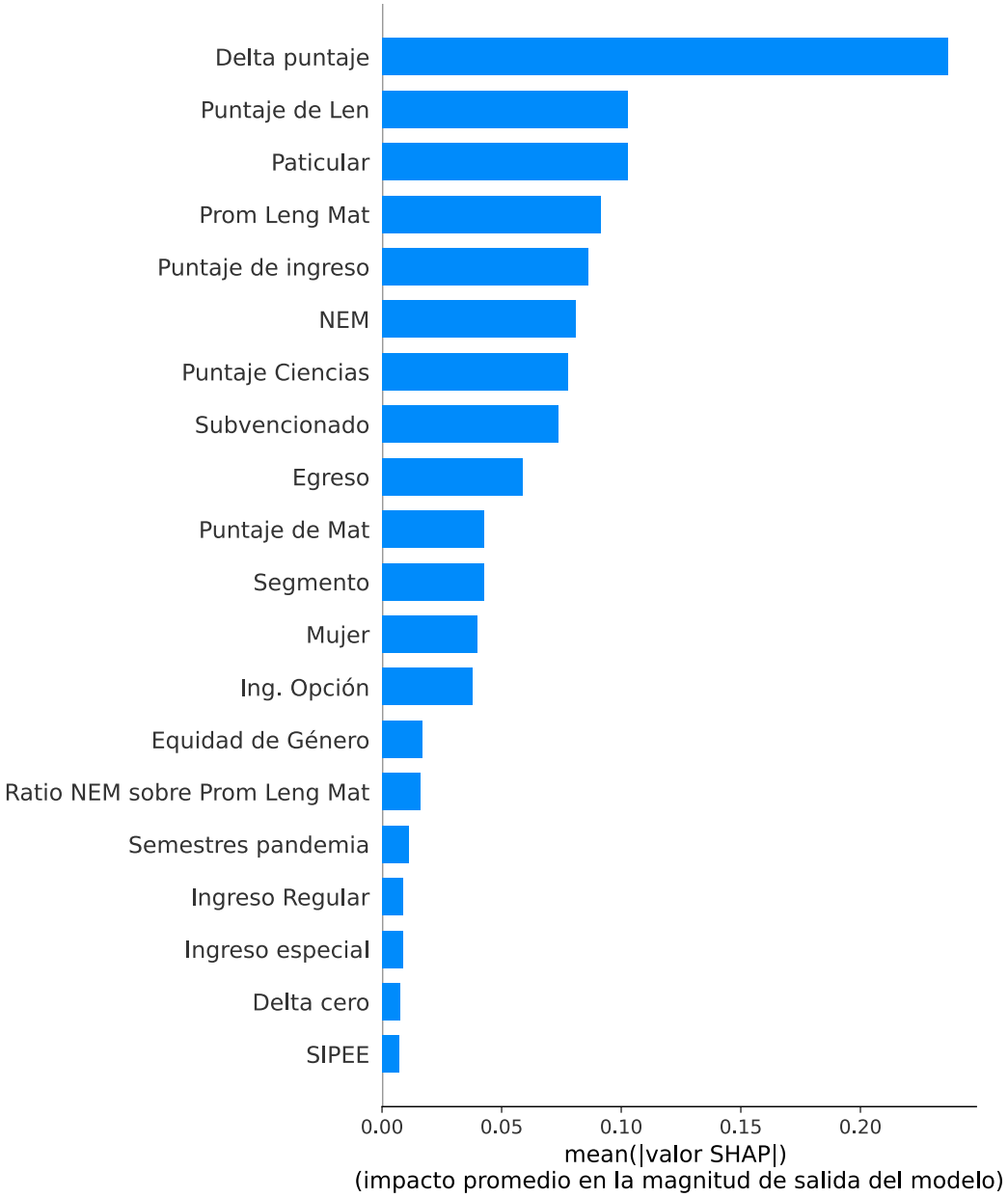


Figura 3.6: Impacto de las variables en el modelo RN.

En la Figura 3.6 se puede apreciar que la variable Delta puntaje aporta alrededor de 2.2 veces, lo que aporta la segunda variable más importante, con lo cual se puede decir que esta variable puede llegar a marcar en parte las decisiones que tome el modelo al momento de predecir. De igual manera, entre las 5 variables más importantes existen 2 que hacen referencia al colegio de egreso (Particular y Prom leng Mat) y el resto están relacionadas con el rendimiento del alumno en las pruebas estandarizas.

3.1.1.2. Análisis Logit

Parámentro	Coeficiente	Desviación estándar	P-valor	Factor de inflación de la varianza
Constante	-9.646	0.771	$< 10^{-6}$	-
Puntaje Lenguaje	-0.003	0.001	$< 10^{-6}$	1.099
Puntaje Ciencias	0.011	0.001	$< 10^{-6}$	1.166
Egreso	1.072	0.093	$< 10^{-6}$	1.009
Prom Leng Mat	0.004	0.001	$< 10^{-5}$	1.100

Tabla 3.3: Detalle de las variables ocupadas en el modelo Logit.

Todas las variables del modelo Logit están contenidas en el modelo RN. Es más, estas variables aportan en el mismo sentido a la probabilidad de predecir el valor de y_i. Además, en los 2 modelos existen variables que son determinadas por el colegio de egreso y afectan fuertemente la decisión del valor a predecir.

Parámentro	Interpretación	Promedio de los efectos marginales
Puntaje de Lenguaje	Un punto más en la PSU de Lenguaje afecta negativamente la probabilidad de terminar Plan Común	-0.0007
Puntaje Ciencias	Un punto más en la PSU de Ciencias afecta positivamente la probabilidad de terminar Plan Común	0.0023
Egreso	Mientras menor sea la diferencia entre el año de egreso de la educación media y el ingreso a la universidad mayor será la probabilidad de terminar Plan Común	0.2324
Prom Leng Mat	Un punto más en el puntaje promedio de las pruebas PSU de matemáticas y lenguaje del colegio de egreso afecta positivamente la probabilidad de terminar Plan Común	0.0009

Tabla 3.4: Interpretación del modelo Logit.

En el modelo existe correlación moderada (Factor de inflación de la varianza), por lo cual, no se rompe el supuesto de multicolinealidad, todas las variables del modelo son significativas, y su interpretación, al igual que en el modelo anterior, pueden diferir con la realidad. Se puede observar de la tabla 3.3 y 3.4, que tener mayor Puntaje Ciencias es más importante que tener

mayor Puntaje de Lenguaje y que un año menos en la diferencia entre el año de ingreso a la Facultad y año de egreso del colegio, aumenta la probabilidad de terminar Plan Común en un 23,34 %. El efecto de la variable Egreso podría ser foco en próximos estudios, debido a que existen muchos factores que afectan el valor de esta variable.

Las variables consideradas en este modelo, contemplan 3 aspectos: rendimiento alumno antes de la universidad, rendimiento colegio y tiempo entre el egreso del colegio e ingreso a la universidad. Asimismo, 2 de las 4 variables son consideradas por la universidad al momento de ingresar. Además, tanto el modelo Logit como el modelo RN, tienen entre sus principales variables las de rendimiento previo y rendimiento colegio.

3.2. Predicciones con datos de rendimiento parcial de Plan Común

Con el fin tener herramientas para focalizar las iniciativas en contra de la deserción involuntaria por razones académicas en distintos períodos de tiempo, se añade al estudio el desarrollo académico de los estudiantes en su primer año universitario, a través de las Tasas de aprobación del 1º y 2º Semestre (detalles descriptivos de estas variables en el Anexo B) . Cabe mencionar que no se considera el promedio de notas, debido a que en la base entregada por la FCFM no aparecen las notas finales de los cursos reprobados.

En el cálculo de la tasa de aprobación se excluyeron los ramos deportivos, humanistas y del área de Ingeniería e Innovación, debido a que son ramos donde la aprobación que bordean el 100 %, además de todo ramo que no sea de Plan Común. En base a lo anterior tasa de aprobación se define como:

$$\text{Tasa de aprobación}_{i,s} = \frac{\sum_{a \in A_{i,s}} 1}{\sum_{c \in C_{i,s}} 1}$$

$$\text{A}_{i,s} = \{\text{Ramos considerados y aprobados por el estudiante i en el semestre s }\}$$

$$\text{C}_{i,s} = \{\text{Ramos considerados y cursados por el estudiante i en el semestre s }\}$$

$$\forall i \in Estudiantes, s \in Semestres$$

3.2.1. Modelos considerando tasa de aprobación 1º Semestre

Se obtuvieron 3 modelos por cada algoritmo de aprendizaje supervisado (detalles en el Anexo D), donde cada uno representa el resultado de la técnica utilizada para tratar el desequilibrio, de los cuales se escogió un modelo por algoritmo para ser analizado, basado en el criterio mencionado anteriormente.

Dado el momento en que se pueden calcular estos modelos, deberían ser las herramientas a ocupar en el 2º semestre para buscar a los estudiantes que deben participar de las iniciáticas en contra de la deserción.

Modelo	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
KNN **	0.96	0.82	0.88	0.29	0.69	0.41	0.80
SVM *	0.97	0.84	0.90	0.34	0.73	0.47	0.83
DT **	0.97	0.80	0.88	0.30	0.80	0.44	0.80
RF *	0.95	0.88	0.92	0.38	0.63	0.47	0.86
RN **	0.97	0.84	0.90	0.35	0.74	0.48	0.83
Logit **	0.96	0.87	0.92	0.39	0.72	0.50	0.86

** El Modelo escogido es con la técnica de peso de las clases equilibrado*

*** El Modelo escogido es con la técnica de sobre-muestreo aleatorio*

**** El Modelo escogido es con la técnica de sobre-muestreo sintético SMOTE*

Tabla 3.5: Resumen del rendimiento de los diversos modelos en Test considerando la Tasa de Aprobación del 1º Semestre.

Se puede ver en la Tabla 3.5 una mejora considerable en todos los modelos, ya que ahora, independiente del método ocupado, se puede captar sobre el 63 % de la categoría y_i=0 y sobre el 80 % de la categoría $y_i = 1$, sumado a esto, la precisión de cada modelo mejoró con respecto a su versión anterior, lo que provocó un aumento en el f1-score y acurracy.

Estas mejoras considerables permiten mejorar tempranamente la focalización de los programas existentes que busquen minimizar la cantidad de personas que no pueden aprobar Plan Común, debido a que sólo se necesita la información del rendimiento del primer semestre para poder ocupar estos modelos, además se abre la posibilidad de ocupar estas herramientas como evaluadores del impacto de las intervenciones contra la deserción que aplica la FCFM durante el primer semestre de carrera, ya que se espera que los estudiantes evaluados con los modelos anteriores y que participaron en dichas intervenciones, se le estime ahora una menor probabilidad de desertar que a comienzo del semestre, sustentado en que la variable agregada

(Tasa de aprobación 1º semestre), toma gran relevancia en la predicción y las intervenciones deberían afectarla.

Complementando lo anterior, con la Figura 3.7 se puede decir que el modelo DT es el modelo que mejor clasifica los $y_i = 0$ seguido por RN, mientras que RF es el modelo que mejor clasifica los $y_i = 1$ seguidos de Logit. Además, se puede apreciar que las curvas de los modelos SVM y RF crecen menos que el resto cuando sobrepasan el umbral del eje e igual 0.5, lo que significa que estos modelos predicen una probabilidad que difiere por menos con la probabilidad correcta que permitiría clasificar bien a los $y_i = 1$ mal clasificados.

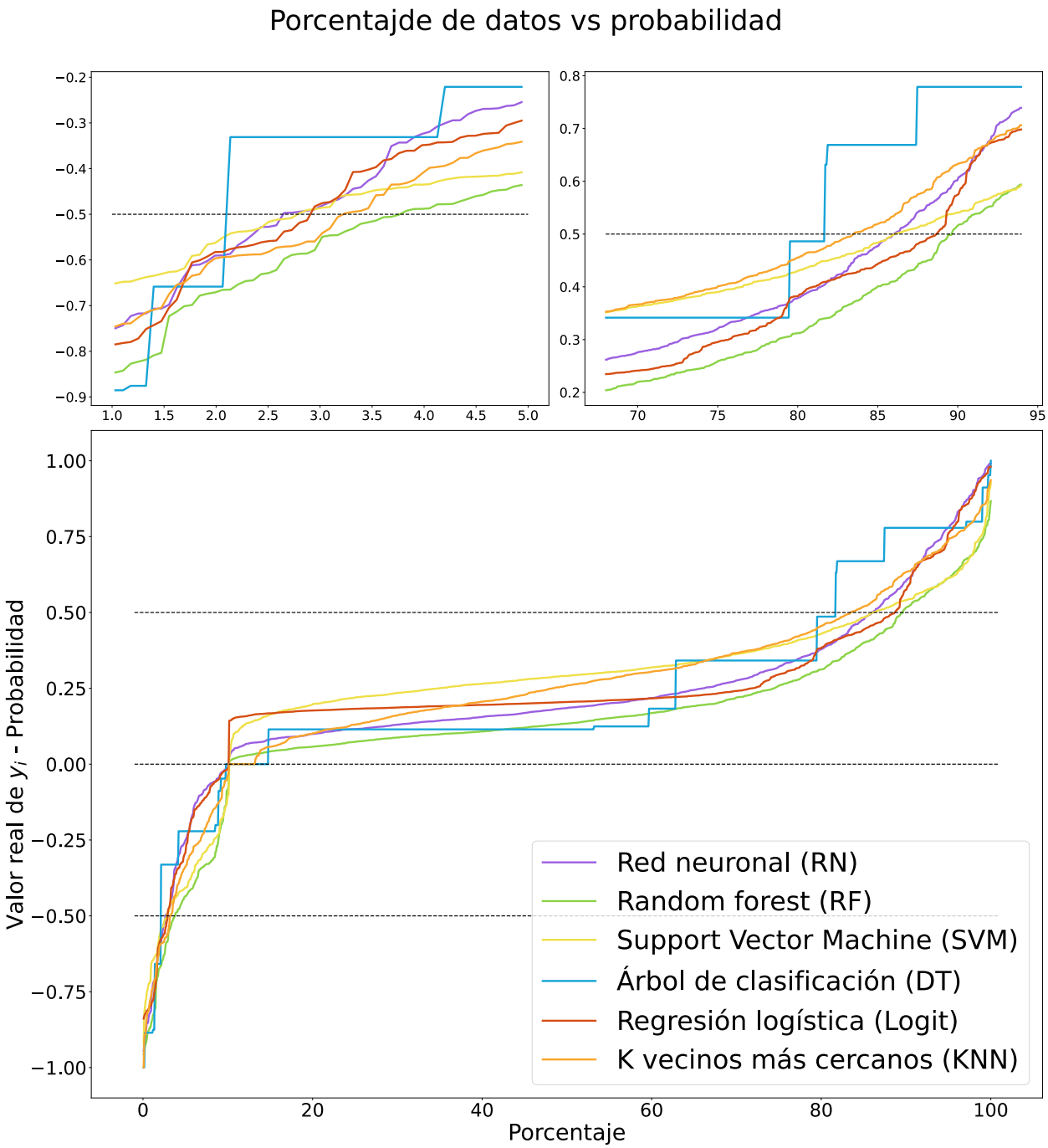


Figura 3.7: Porcentaje de datos vs probabilidad, considerando la Tasa de Aprobación del 1º Semestre.

Modelo	**Máximo error $y_i = 1$**	**Máximo error $y_i = 0$**	**Mínimo error $y_i = 1$**	**Mínimo error $y_i = 0$**	**Promedio error y_i =1**	**Promedio error $y_i = 0$**	**Promedio de todos los errores**
KNN **	0.436	0.500	0.001	0.006	0.170	0.187	0.173
SVM *	0.428	0.420	0.001	0.008	0.105	0.144	0112
DT **	0.500	0.500	0.131	0.158	0.255	0.311	0.261
RF *	0.367	0.455	0.005	0.001	0.126	0.210	0.148
RN **	0.490	0.445	0.002	0.024	0.222	0.218	0.222
Logit **	0.481	0.339	0.001	0.013	0.245	0.192	0.234
** El Modelo escogido es con la técnica de peso de las clases equilibrado*							
*** El Modelo escogido es con la técnica de sobre-muestreo aleatorio*							
**** El Modelo escogido es con la técnica de sobre-muestreo sintético SMOTE*							

Tabla 3.6: Detalle del error en lo datos mal clasificados en Test por modelo, considerando la Tasa de Aprobación del 1º Semestre.

Se puede observar de la Tabla 3.6 que SVM y RF, son los modelos que en promedio se equivocan por menos cuando la clase es y_i=1, lo cual va en línea con lo anterior, mientras que SVM y KNN son los que se equivocan en promedio por menos cuando la clase es y_i=0.

En consecuencia, es el modelo de SVM el que se equivoca en promedio por menos. Igualmente, se debe mencionar que los errores aumentaron con respecto a los modelos anteriores. Esto se debe principalmente a que ahora son menos los datos mal clasificados y estos son más difíciles de identificar correctamente, ya que no poseen el conjunto de características importantes consideradas en cada modelo para ser clasificados correctamente. Lo mencionado se puede apreciar en las Figuras 3.8 y 3.9, donde aparecen muchos datos outliers, los cuales reflejan en gran parte a los datos mal clasificados que poseen características muy marcadas del grupo opuesto al que pertenecen, lo que hace que tengan una probabilidad "fuerte" y una diferencia significativa entre y_i y la probabilidad.

Un ejemplo de lo anterior se puede apreciar en el modelo DT (Detalles en la sección 3.2.1.1.3), donde existen estudiantes a los cuales se le pronostica $\widehat{y}_i$=1 con probabilidad de 88 %, cuando en realidad son y_i=0, esto se debe a que poseen las características del y_i contrario, las cuales son Tasa de aprobación del 1º semestre sobre 91.7 %, Puntajes de matemática sobre 720 puntos y NEM sobre 707.

De igual manera, se aprecia que la mediana y la caja en los distintos gráficos se acercan a 0, en conjunto con que mayor cantidad de datos queda dentro de los márgenes, lo cual demuestra que estos modelos pronostican con mayor seguridad los valores, donde destacan DT, RF y RN, que son los modelos con mediana más cercana a 0.

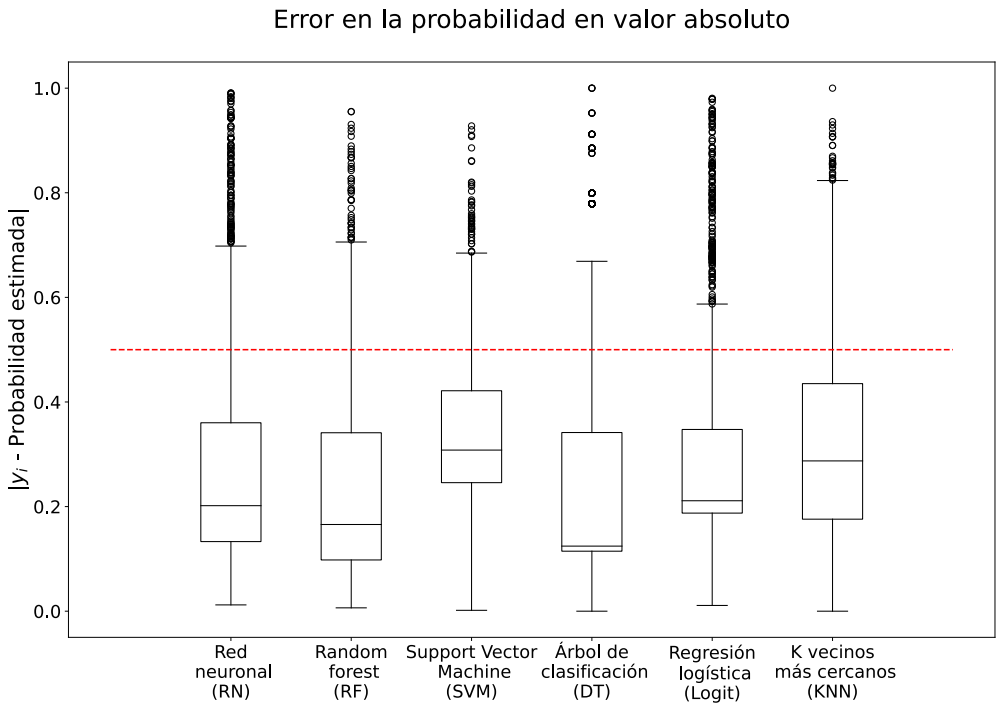


Figura 3.8: Error en la probabilidad en valor absoluto, considerando la Tasa de Aprobación del 1º Semestre.

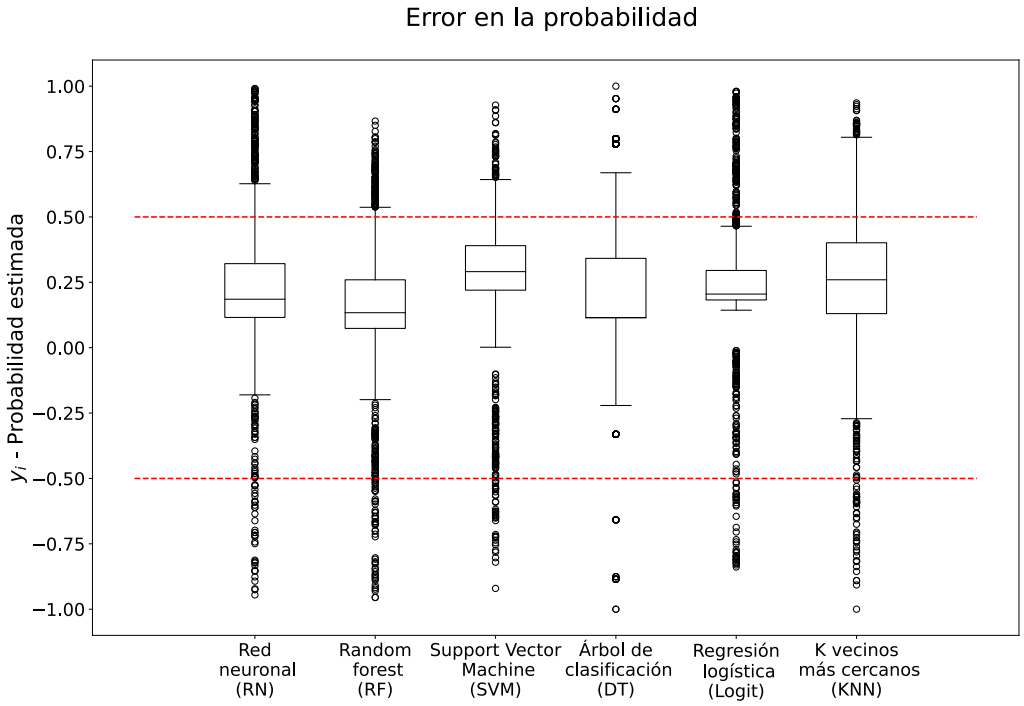


Figura 3.9: Error en la probabilidad, considerando la Tasa de Aprobación del 1º Semestre.

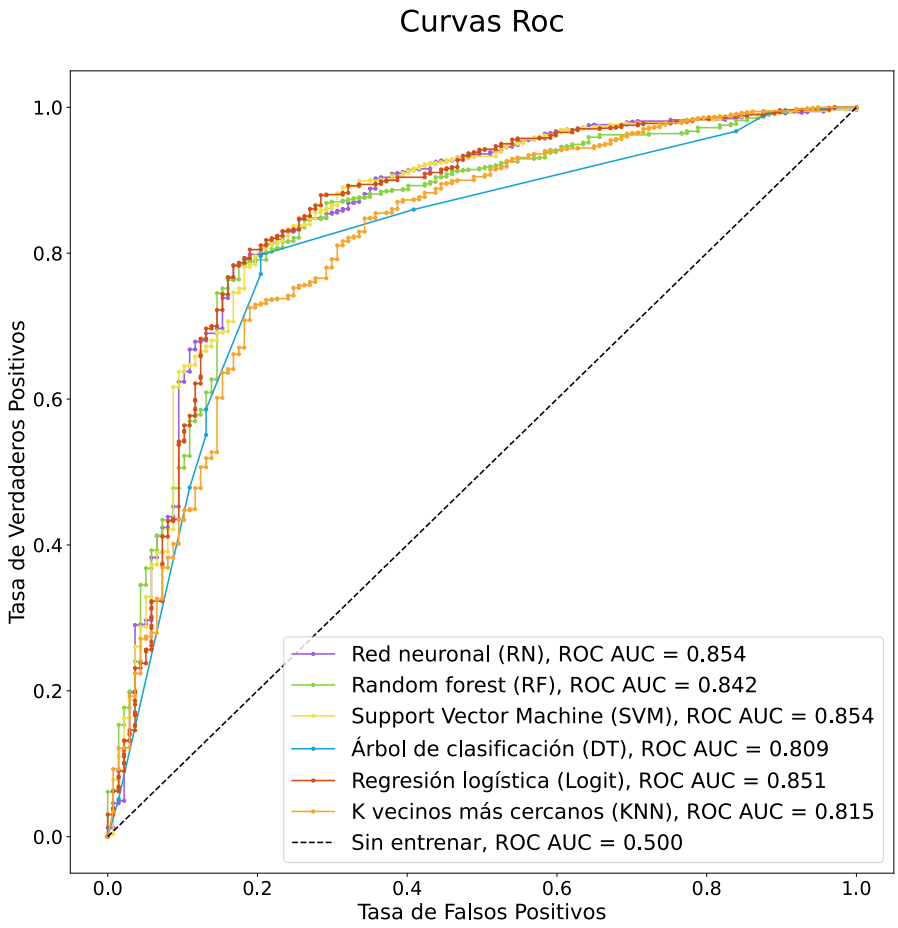


Figura 3.10: Curvas ROC de los diversos modelos, considerando la Tasa de Aprobación del 1º Semestre.

De la Figura 3.10 se puede observar que los modelos con mayor AUC son RN y SVM, los cuales tienen un valor de 0.854. Esto se traduce en que es el modelo con mejor poder discriminatorio, significando que puede distinguir mejor entre un alumno que pasará Plan Común y uno que no. Además, se aprecian todos los diferentes resultados que se pueden obtener al variar el umbral de decisión, con la finalidad de explicitar todos los rendimientos posibles, para que así las diferentes iniciativas de este período puedan elegir el resultado que más se acomode a sus requerimientos.

En específico, si se quiere modificar el umbral con el objetivo de aumentar la cantidad de $y_i = 0$ (verdaderos negativos), se puede llegar a identificar correctamente cerca del 90 % de esta población manteniendo un recall de $y_i = 1$ superior al 50 %. Se debe recordar que la base

está desbalanceada, lo que significa que esto llevará a clasificar incorrectamente una cantidad mayor a la que se clasificará correctamente. Asimismo, se aprecia que los modelos RN, SVM, RF y Logit poseen curvas ROC similares.

Todo lo anterior refleja una mejora considerable con respecto a los primeros modelos, no obstante, refleja que existen datos difíciles de clasificar correctamente. Por otra parte, nos entrega que el modelo más equilibrado es RN, ya que abarca más área bajo la curva ROC, es el segundo mejor clasificador de $y_i = 0$ y tercer mejor clasificador de $y_i = 1$. A pesar de lo anterior, si quieren identificar más $y_i = 0$ es recomendable ocupar DT, pero éste llevará una cantidad importante de falsos positivos, lo cual provocaría un gasto mayor de recursos al hacer algún programa basado en este modelo. Y si se quiere identificar más $y_i = 1$ se recomienda Logit, ya que es el segundo que más abarca de esta clasificación, pero con una diferencia mínima, con un valor mucho más alto para la clasificación $y_i = 0$.

3.2.1.1. Variables importantes para los modelos considerando la tasa de aprobación del 1º Semestre

Con el objetivo de comprender cómo afectan las variables considerando la tasa de aprobación 1º Semestre en la probabilidad de predecir si una persona terminará Plan Común o no, se estudiaron los efectos en los modelos RN, Logit y DT, dado lo mencionado anteriormente.

Es importante recordar que la variable tasa de aprobación 1º Semestre, es una característica en la cual la Facultad puede injerir forma directa o indirecta a diferencia de las otras variables.

3.2.1.1.1. Análisis RN

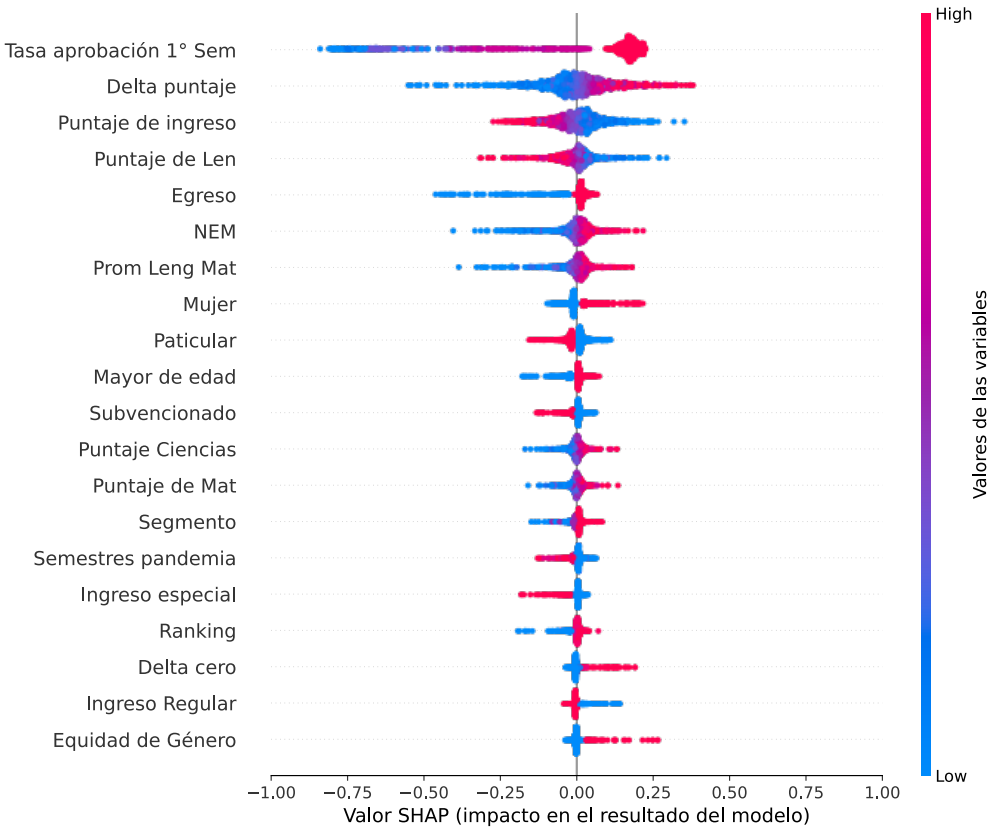


Figura 3.11: Efecto de las variables en el modelo RN, considerando la Tasa de Aprobación del 1° Semestre

En la Figura 3.11 se observan las 20 variables que ocupa el modelo RN, comparte 17 con el modelo que no considera la tasa de aprobación del 1° semestre, donde sólo 2 de estas cambian su comportamiento con respecto al sentido en que aportan a la probabilidad de terminar Plan Común, las cuales son Ingreso regular y Mayor de edad. Sumado a lo anterior, ahora la variable agregada Tasa de aprobación 1° semestre es la variable más importante, donde a mayor cantidad de ramos aprobados, mayor será la probabilidad de pasar Plan Común y se aprecia que los valores más altos se diferencian del resto. Posterior a esta variable, le sigue la variable más importante en el modelo anterior. Por otro lado, aparecen las variables Mayor de edad y Ranking, que aportan positivamente a la probabilidad.

Se destaca que la vía de ingreso toma un rol importante en el análisis de forma indirecta, ya que sólo las personas que ingresan por Equidad de Género, 5 % Colegios Municipalizados, Sistema de Ingreso Prioritario de Equidad Educativa o Deportista tiene Delta puntaje menor

a 0 y poseen los puntajes de ingreso más bajo. Complementando con lo que se aprecia en las Figuras 3.11 y 3.12, se puede deducir que existe un efecto en conjunto entre estas variables, donde predomina el efecto de Delta Puntaje.

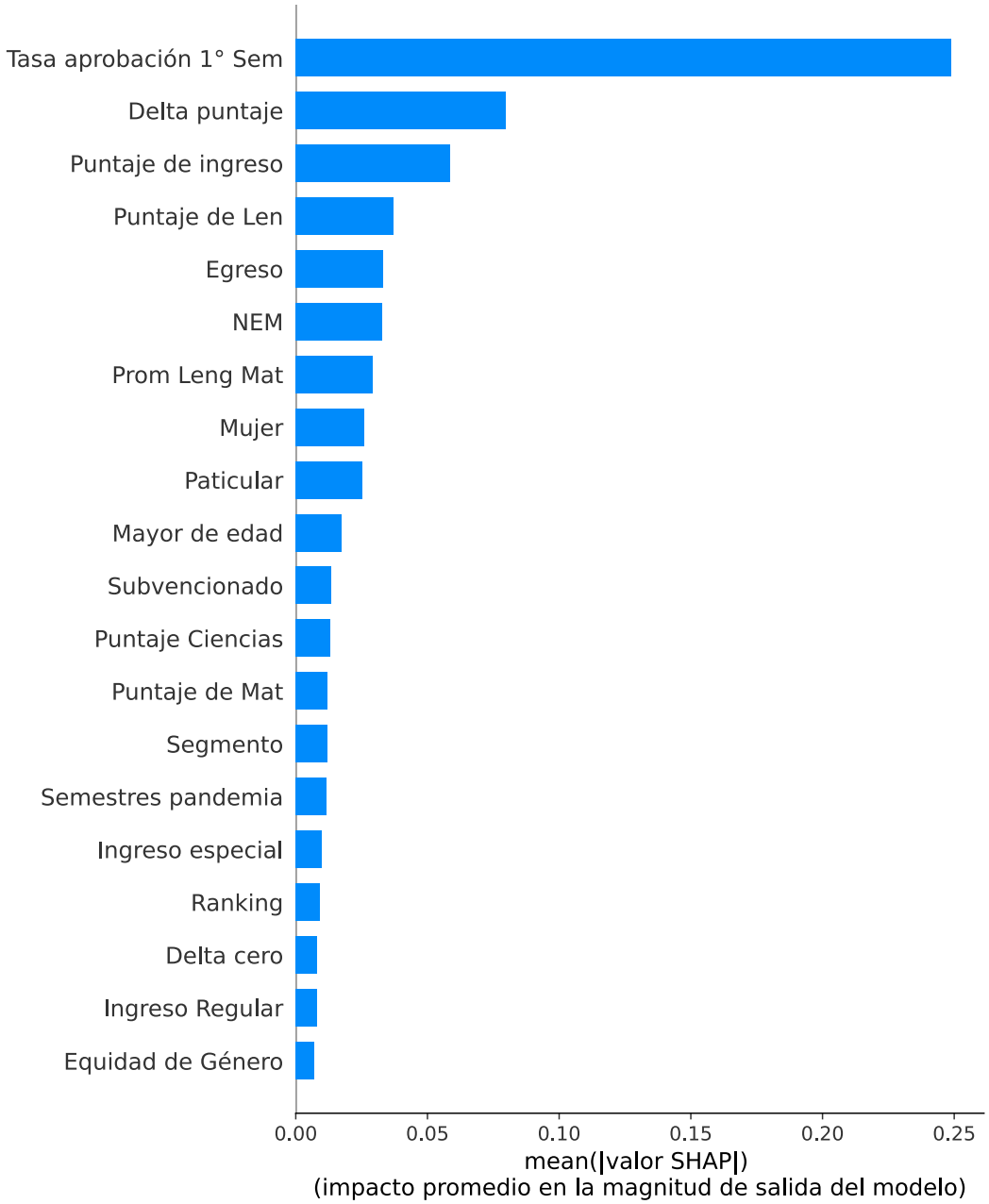


Figura 3.12: Impacto de las variables en RN, considerando la Tasa de Aprobación del 1º Semestre.

Se puede apreciar en la Figura 3.12, que la tasa de aprobación del 1º semestre aporta casi tres veces más que la segunda variable más importante, con lo cual se puede decir que esta variable puede marcar en gran parte las decisiones que tome el modelo al momento de

predecir. De igual manera, las variables relacionadas al colegio de egreso perdieron relevancia.

3.2.1.1.2. Análisis Logit

Parámentro	**Coeficiente**	**Desviación estándar**	**P-valor**	**Factor de inflación de la varianza**
Constante	-8.287	0.938	$< 10^{-6}$	-
NEM	0.002	0.001	0.016	1.086
Egreso	0.500	0.089	$< 10^{-6}$	1.051
Prom Leng Mat	0.003	0.001	0.001	1.075
Tasa aprobación 1º Sem	5.211	0.252	$< 10^{-6}$	1.060

Tabla 3.7: Detalle de las variables ocupadas en el modelo Logit, considerando la Tasa de Aprobación del 1º Semestre.

Todas las variables del modelo Logit están contenidas en el modelo RN. Es más, estas variables aportan en el mismo sentido a la probabilidad de predecir el valor y_i.

Parámentro	**Interpretación**	**Promedio de los efectos marginales**
NEM	Mayor NEM afecta positivamente la probabilidad de terminar Plan Común	0.0003
Egreso	Mientras menor sea la diferencia entre el año de egreso de la educación media y el ingreso a la universidad mayor será la probabilidad de terminar Plan Común	0.0856
Prom Leng Mat	Un punto más en el puntaje promedio de las pruebas PSU de matemáticas y lenguaje del colegio de egreso afecta positivamente la probabilidad de terminar Plan Común	0.0005
Tasa aprobación 1ºSem	Mientras mayor sea la cantidad de ramos aprobados el 1º semestre, mayor sera la probabilidad de terminar Plan Común	0.8908

Tabla 3.8: Interpretación del modelo Logit, considerando la Tasa de Aprobación del 1º Semestre.

Es importante considerar que en el modelo existe correlación moderada (Factor de inflación de la varianza). Por ello, no se rompe el supuesto de multicolinealidad, todas las variables del modelo son significativas, y su interpretación puede diferir con la realidad. Igualmente, se repiten 2 variables en comparación con el modelo Logit anterior y éstas afectan en el mismo

sentido la probabilidad. Además, se puede observar de la tabla 3.7 y 3.8 que la variable que refleja la tasa de aprobación del 1º semestre toma relevancia y refleja que a mayor tasa de aprobación, mayor será la probabilidad de terminar Plan Común correctamente, donde tener 100 % de aprobación marca en gran parte la predicción. Es importante recordar que los ramos del 1º semestre los designa la Facultad.

Las varibles consideradas en este modelos, contemplan los 3 mismos aspectos del modelo Logit pasado, pero ahora el rendimiento alumno antes de la universidad lo refleja la variable NEM y se agrega el aspecto del rendimiento en la universidad.

3.2.1.1.3. Análisis DT

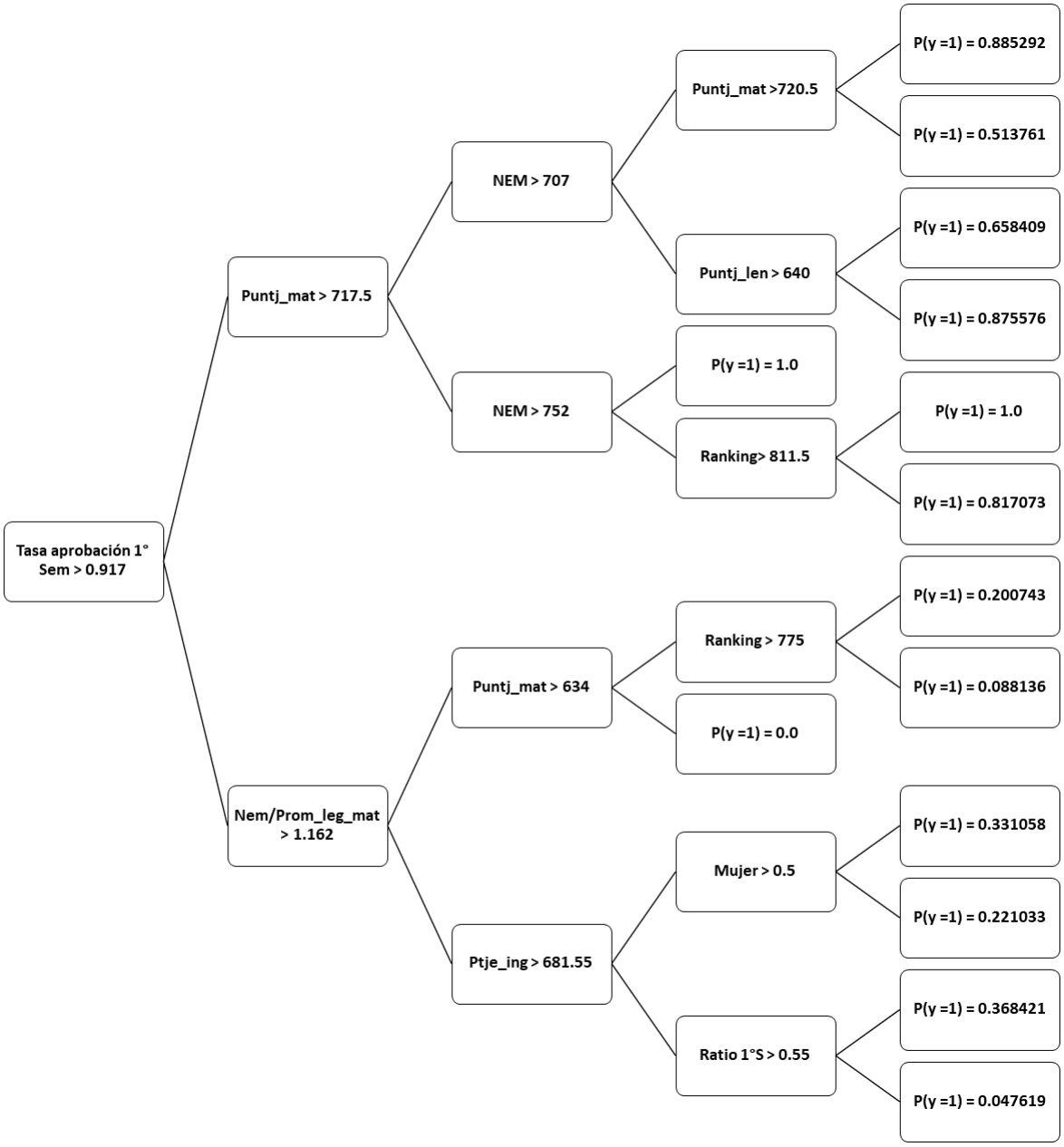


Figura 3.13: Árbol resultante del modelo DT, considerando la Tasa de Aprobación del 1º Semestre.

Del modelo DT, se puede apreciar en la Figura 3.13, que su decisión se basa totalmente en el valor de la tasa de aprobación del 1º semestre, donde el resto de las variables ocupadas sólo aportan en designar qué tan fuerte es la probabilidad de esta decisión. Es interesante mencionar que en base a solo una variable se puede captar alrededor del 80 % de cada categoría e intentar cambiar los parámetros configurados de este modelo causan un sobre ajuste

importante y hacen que se pierda una cantidad significativa de una categoría con respecto a lo que se gana en la otra.

Se refleja en los 3 modelos estudiados que la variable agregada es relevante para lo que se quiere predecir.

3.2.2. Modelos considerando tasa de aprobación 1º y 2º Semestre

Se obtuvieron 3 modelos por cada algoritmo de aprendizaje supervisado (detalles en el Anexo E), donde cada uno representa el resultado de la técnica utilizada para tratar el desequilibrio, de los cuales se escogió un modelo por algoritmo para ser analizado, basado en el criterio mencionado anteriormente.

Dado el momento en que se pueden calcular estos modelos, deberían ser las herramientas a ocupar desde el 3º semestre para buscar a los estudiantes que deben participar de las iniciáticas en contra de la deserción.

Modelo	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
KNN ***	0.97	0.87	0.92	0.39	0.73	0.51	0.86
SVM *	0.98	0.87	0.92	0.43	0.83	0.56	0.87
DT **	0.98	0.81	0.89	0.33	0.85	0.48	0.81
RF ***	0.96	0.92	0.94	0.51	0.69	0.58	0.90
RN **	0.98	0.86	0.92	0.40	0.82	0.54	0.86
Logit ***	0.98	0.88	0.93	0.43	0.85	0.57	0.87
** El Modelo escogido es con la técnica de peso de las clases equilibrado*							
*** El Modelo escogido es con la técnica de sobre-muestreo aleatorio*							
**** El Modelo escogido es con la técnica de sobre-muestreo sintético SMOTE*							

Tabla 3.9: Resumen del rendimiento de los diversos modelos en Test, considerando las Tasas de aprobación del 1º y 2º Semestre.

Se puede ver en la Tabla 3.9 una mejora en todos los modelos, ya que ahora, independiente del método ocupado, se puede captar sobre el 69 % de la categoría y_i=0 y sobre el 81 % de la categoría $y_i = 1$. Además, la precisión de cada modelo mejoró con respecto a su versión anterior, lo cual llevó a aumentar el f1-score y su acurracy.

Estas mejoras a pesar de perfeccionar la focalización de las iniciativas que busquen minimizar la cantidad de personas que no pueden aprobar Plan Común, permitirán comprobar en primera instancia qué tan importante son las intervenciones anteriores y si lograron su obje-

tivo, ya que, al momento de poder ocupar este modelo, se apreciarán los primeros estudiantes que deserten involuntariamente por razones académicas, con lo cual, se podrá visualizar la evolución de su probabilidad de desertar, a través del tiempo, considerando variables que miden el desempeño del estudiante a través de los semestres.

Complementando lo anterior, de la Figura 3.14 se puede decir que el modelo DT es el modelo que mejor clasifica los y_i= 0 seguido por Logit, con una diferencia mínima, mientras que RF es el modelo que mejor clasifica los y_i= 1 seguido por Logit, que se diferencia por muy poco de SVM y KNN. Además, se puede apreciar que la curva del SVM, crece menos que el resto cuando sobrepasa el umbral del eje y igual 0.5, lo que significa que este modelo predice una probabilidad que difiere por menos con la probabilidad correcta que permitiría clasificar bien a los y_i= 1 mal clasificados.

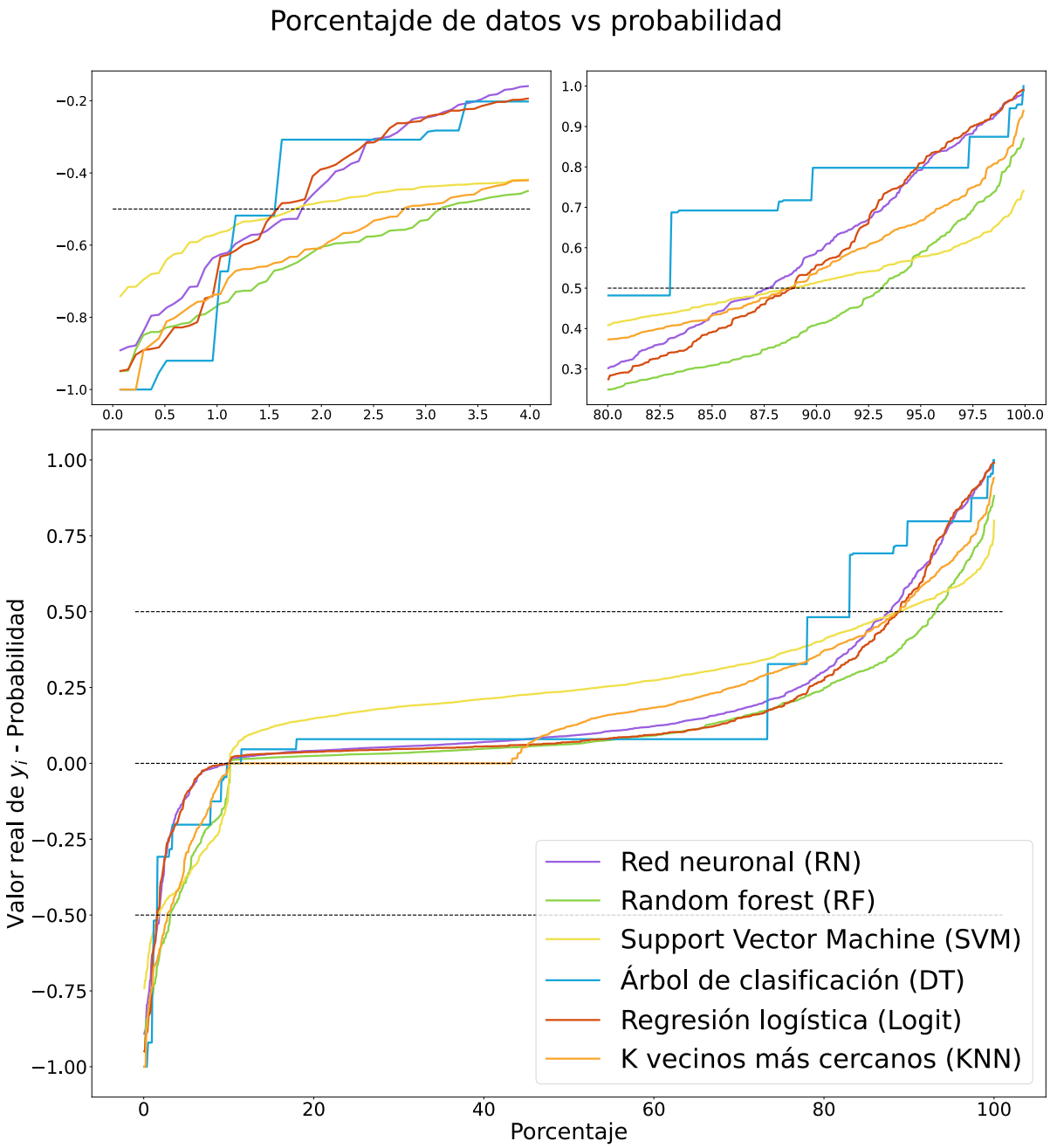


Figura 3.14: Porcentaje de datos vs probabilidad, considerando las Tasas de aprobación del 1º y 2º Semestre

Modelo	Máximo error $y_i = 1$	Máximo error $y_i = 0$	Mínimo error $y_i = 1$	Mínimo error $y_i = 0$	Promedio error $y_i = 1$	Promedio error $y_i = 0$	Promedio de todos los errores
KNN ***	0.440	0.500	0.001	0.021	0.168	0.195	0.173
SVM *	0.300	0.242	0.000	0.007	0.080	0.099	0.082
DT **	0.500	0.500	0.188	0.018	0.274	0.302	0.277
RF ***	0.381	0.449	0.007	0.008	0.163	0.192	0.172
RN **	0.493	0.391	0.001	0.027	0.242	0.181	0.235
Logit ***	0.491	0.449	0.000	0.011	0.260	0.252	0.259

* *El Modelo escogido es con la técnica de peso de las clases equilibrado*

** *El Modelo escogido es con la técnica de sobre-muestreo aleatorio*

*** *El Modelo escogido es con la técnica de sobre-muestreo sintético SMOTE*

Tabla 3.10: Detalle del error en lo datos mal clasificados en Test por modelo, considerando las Tasas de aprobación del 1º y 2º Semestre.

De la Tabla 3.10 se observa que SVM es el modelo que en promedio se equivocan por menos y posee los errores máximos más bajos independiente de la clase. Además, en el resto de modelos se mantuvo o aumentó su error promedio en comparación con los modelos anteriores.

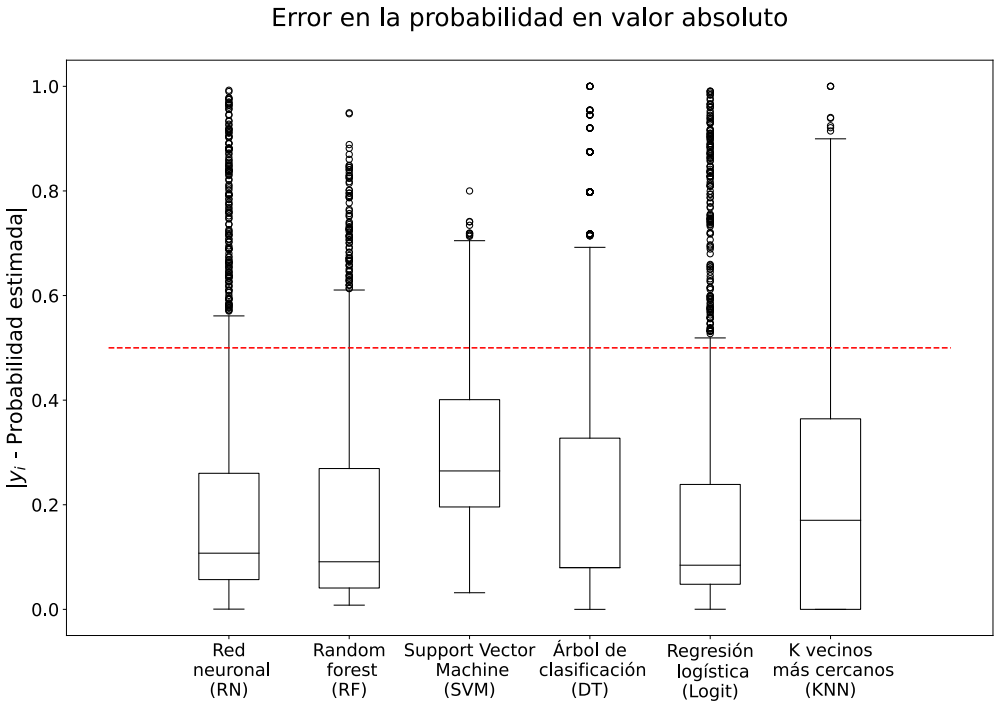


Figura 3.15: Error en la probabilidad en valor absoluto, considerando las Tasas de aprobación del 1º y 2º Semestre.

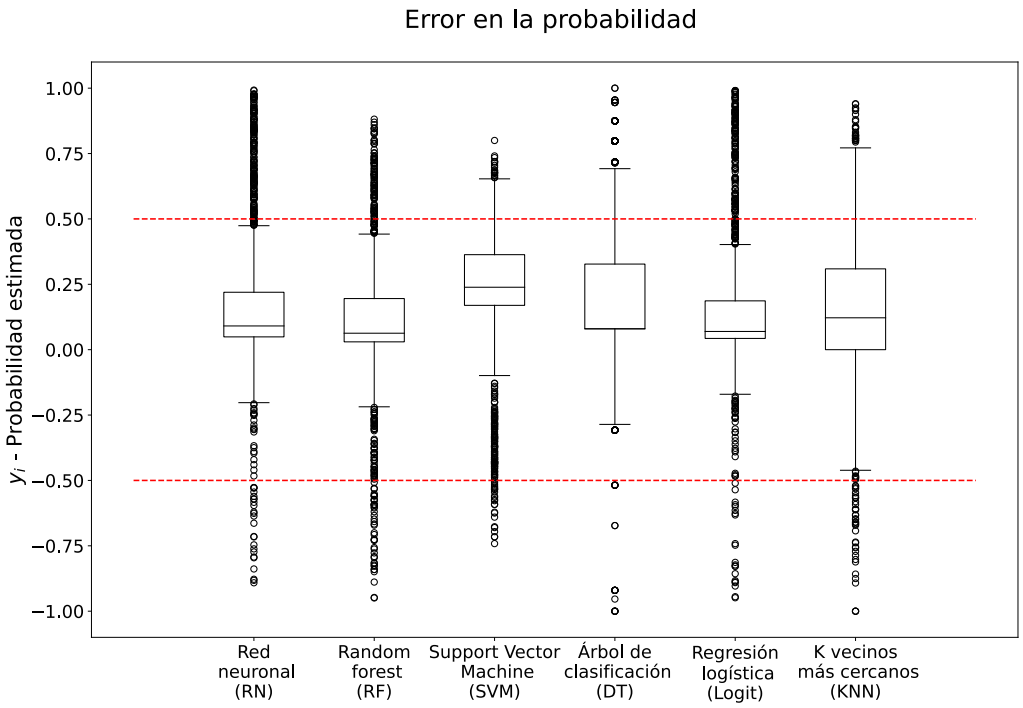


Figura 3.16: Error en la probabilidad, considerando las Tasas de aprobación del 1º y 2º Semestre.

En las Figura 3.15 y 3.16 se aprecia que la mediana y la caja de los distintos modelos se acercan más a 0, en comparación con los modelos anteriores. Sumado a ello, se refleja que una mayor cantidad de datos queda dentro de los márgenes de los resultados correctamente clasificados, lo cual demuestra que la variable agregada aporta en la detección de las personas que terminan o no Plan Común y otorga fuerza en la predicción, ya que, ahora la probabilidad correcta se diferencia por menos con el valor real de y_i. Sin embargo, aún existe una gran presencia de outliers o datos difíciles de clasificar. Destacan RF y Logit como los modelos que más datos tienen dentro de los márgenes y son de los modelos con la mediana más cercana a 0.

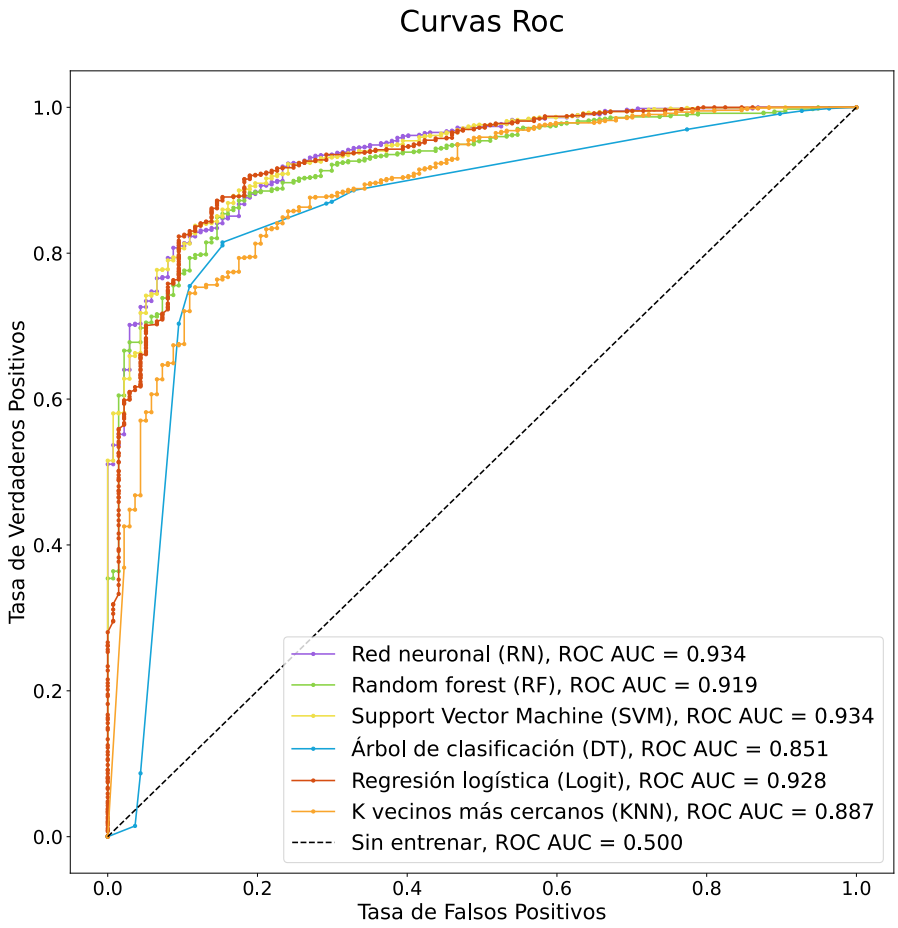


Figura 3.17: Curvas ROC de los diversos modelos, considerando las Tasas de aprobación del 1º y 2º Semestre.

De la figura 3.17 se puede observar que los modelos con mayor AUC son RN y SVM, los cuales coinciden con los modelos anteriores. No obstante, ahora tienen un valor de 0.934, lo que se traduce en que es el modelo con mejor poder discriminatorio, lo que significa que puede distinguir mejor entre un alumno que pasará Plan Común y uno que no. Además, se aprecia que si se quiere modificar el umbral con el objetivo de aumentar la cantidad de y_i= 0 (verdaderos negativos) se puede llegar a identificar correctamente cerca del 90 % de esta población, manteniendo un recall de y_i= 1 superior al 80 %. Igualmente, se aprecia que el Logit posee un curva ROC similar.

Todo lo anterior refleja una mejora considerable con respecto a los modelos anteriores; sin embargo, reafirma que existen datos difíciles de clasificar correctamente, sumado a que nos

entrega que el modelo más equilibrado es Logit, ya que abarca un área bajo la curva ROC similar a los mejores modelos en esta categoría, difiere muy poco con el mejor clasificador de $y_i = 0$ y es el segundo mejor clasificando de $y_i = 1$. A pesar de lo anterior, si se quiere identificar más $y_i = 1$ sin cambiar el punto de corte en los modelos, es recomendable ocupar RF, pero este llevará una cantidad importante de falsos negativos, lo cual provoca poca identificación de la clase relevante $y_i = 0$.

3.2.2.1. Variables importantes para los modelos considerando la tasa de aprobación del 1º y 2º Semestre.

Con el objetivo de comprender cómo afectan las variables considerando la tasa de aprobación del 1º y 2º Semestre en la probabilidad de predecir si una persona terminará Plan Común o no, se estudiaron los efectos en los modelos Logit y RF, dado lo mencionado anteriormente.

3.2.2.1.1. Análisis Logit

Parámentro	**Coeficiente**	**Desviación estándar**	**P-valor**	**Factor de inflación de la varianza**
Constante	-3.075	0.667	$< 10^{-5}$	-
Puntaje Lenguaje	-0.007	0.001	$< 10^{-6}$	1.009
Tasa aprobación 1º Sem	5.487	0.269	$< 10^{-6}$	1.156
Tasa aprobación 2º Sem	5.402	0.258	$< 10^{-6}$	1.154

Tabla 3.11: Detalle de las variables ocupadas en el modelo Logit, considerando las Tasas de aprobación del 1º y 2º Semestre.

Es importante considerar que en el modelo existe correlación moderada (Factor de inflación de la varianza). Por ello, no se rompe el supuesto de multicolinealidad, todas las variables del modelo son significativas y su interpretación puede diferir con la realidad.

Parámentro	**Interpretación**	**Promedio de los efectos marginales**
Puntaje Lenguaje	Mayor NEM afecta positivamente la probabilidad de terminar Plan Común	-0.0007
Tasa aprobación 1º Sem	Mientras mayor sea la cantidad de ramos aprobados el 1º semestre, mayor sera la probabilidad de terminar Plan Común	0.5565
Tasa aprobación 2º Sem	Mientras mayor sea la cantidad de ramos aprobados el 2º semestre, mayor sera la probabilidad de terminar Plan Común	0.5478

Tabla 3.12: Interpretación del modelo Logit, considerando las Tasas de aprobación del 1º y 2º Semestre.

Es interesante mencionar que se repiten variables en comparación con el modelo Logit anterior y estas afectan en el mismo sentido la probabilidad, la cual es la tasa de aprobación del 1º semestre, y que afecta más que la tasa de aprobación del 2º semestre en la probabilidad de terminar Plan Común correctamente, lo cual se puede observar de la tabla 3.11 y 3.12. Ahora sólo está el factor del rendimiento en el colegio y en la universidad presentes en este modelo Logit.

3.2.2.1.2. Análisis RF

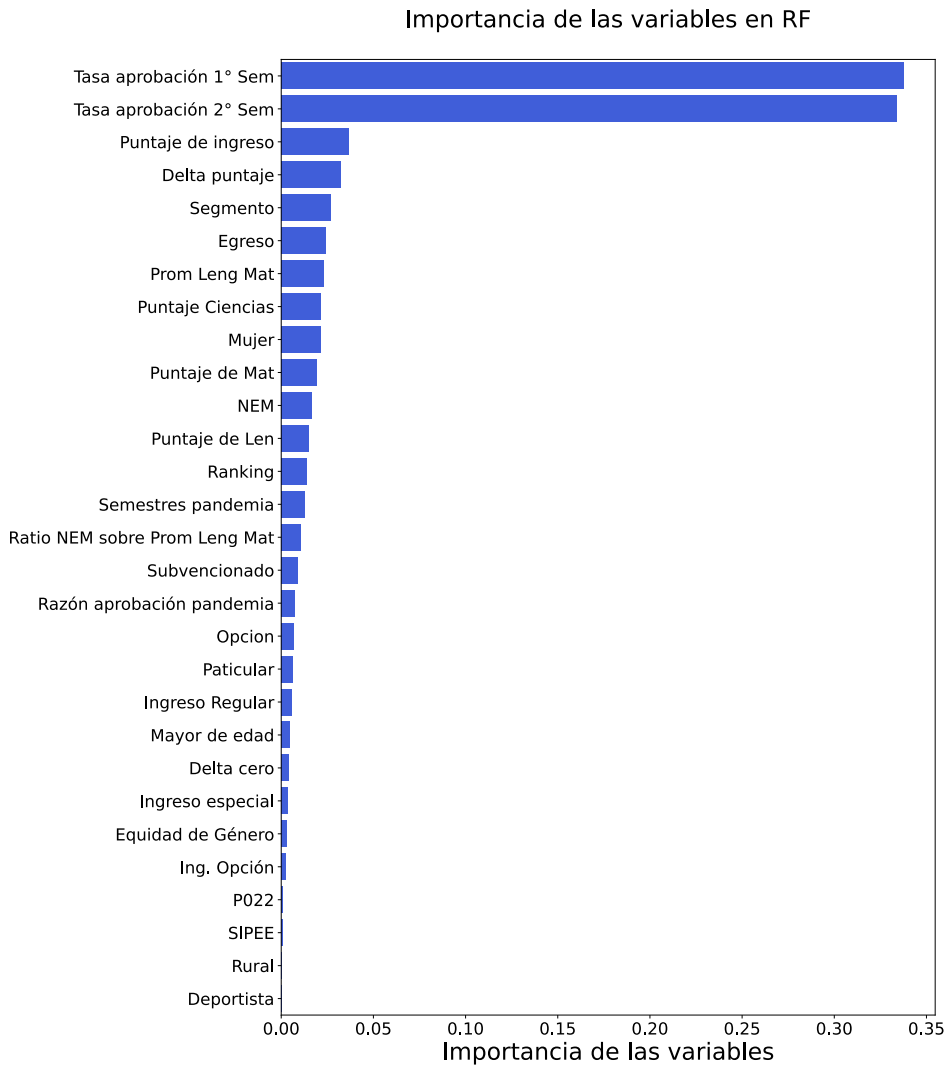


Figura 3.18: Importancia de las variables en RF, considerando las Tasas de aprobación del 1º y 2º Semestre.

En la Figura 2.20, se puede observar que las tasas de aprobación de los diferentes semestres afectan casi 10 veces más la probabilidad de terminar Plan Común que el resto de las variables, entre las cuales destaca el Puntaje de ingreso y Delta puntaje, que son variables altamente correlacionadas y concentran el rendimiento previo a la Facultad y tipo de ingreso .

Considerando los resultados del modelo RF y Logit, se observa que la Tasa de aprobación

del 1º semestre afecta más en la predicción que la Tasa del 2º semestre, y que al cabo del 2º semestre la probabilidad de terminar Plan Común dependerá mayoritariamente del rendimiento en el programa que de variables previas. A pesar de lo anterior, este punto representa la mitad de la duración ideal del programa.

3.2.3. Evolución de los modelos

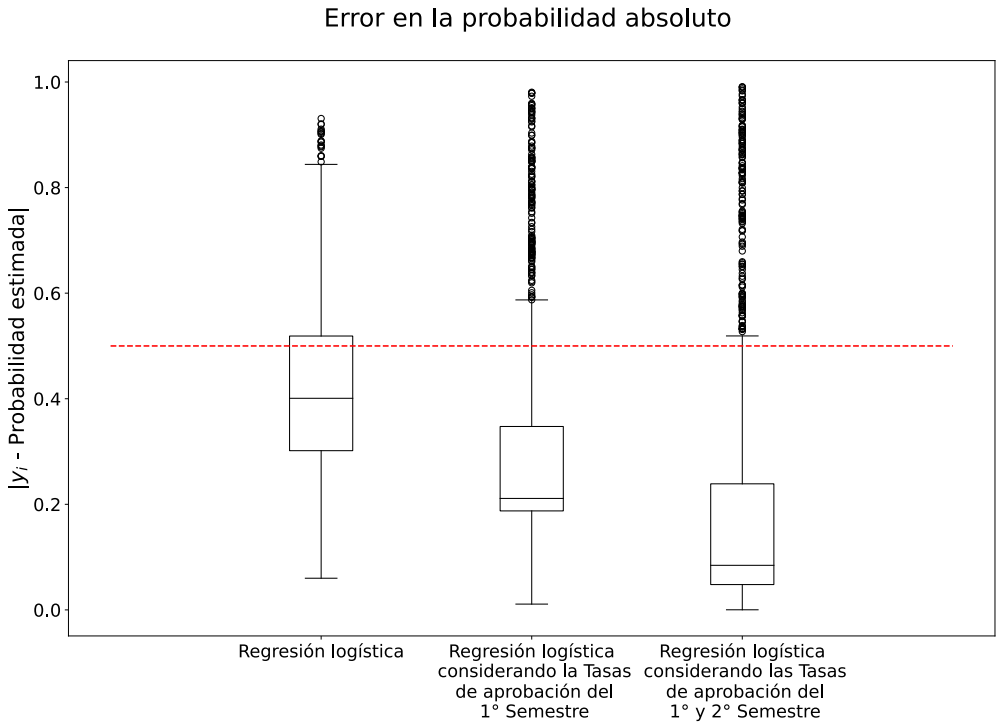


Figura 3.19: Error en la probabilidad en valor absoluto, de los diferentes modelos Logit.

Los modelos mejoran en cada iteración, lo cual se puede apreciar en la figura 3.19, que generaliza el comportamiento a través del modelo Logit, ya que, este destaca en las diferentes iteraciones.

En primera instancia, es posible identificar de la base de datos de testeo sobre el 66 % de personas que no terminaron Plan Común y el 70 % que sí. En base a sólo variables previas al ingreso de los estudiantes a la Facultad, a través del modelo más "equilibrado".

Considerando la tasa de aprobación del primer semestre, es posible mejorar los modelos logrando captar sobre el 72 % de personas que no terminan Plan Común y 80 % que sí terminan, existentes en la base de datos de testeo. En este punto las variables previas al

ingreso pierden relevancia ante la Tasa de aprobación del primer semestre, ya que esta variable captura información que las variables previas no pueden. Además, se aprecia que existen casos difíciles de predecir, lo cual lleva a que aumente el porcentaje por el cual se equivocan los modelos.

Sumando al análisis la tasa de aprobación del segundo semestre, es posible captar sobre el 85 % de personas que no terminan Plan Común y 88 % que sí existentes en la base de datos de testeo con el modelo más "equilibrado", en este punto las variables previas al ingreso pierden casi toda su importancia, y se ven suplantadas por las tasas de aprobación, donde la más influyente es la del primer semestre. Además, se observa un aumento en el error de casi todos los modelos, lo cual está marcado por las personas que son difíciles de clasificar.

Es importante mencionar que, en cada iteración se pueden ocupar los modelos para focalizar las iniciativas de la FCFM, en contra de la deserción involuntaria. A pesar de lo anterior, los modelos que ocupan la tasa de aprobación del 2º semestre pueden llegar a pronosticar tardíamente, ya que, en este punto existirán las primeras personas que deserten, a causa del problema estudiado.

3.2.4. Discusión

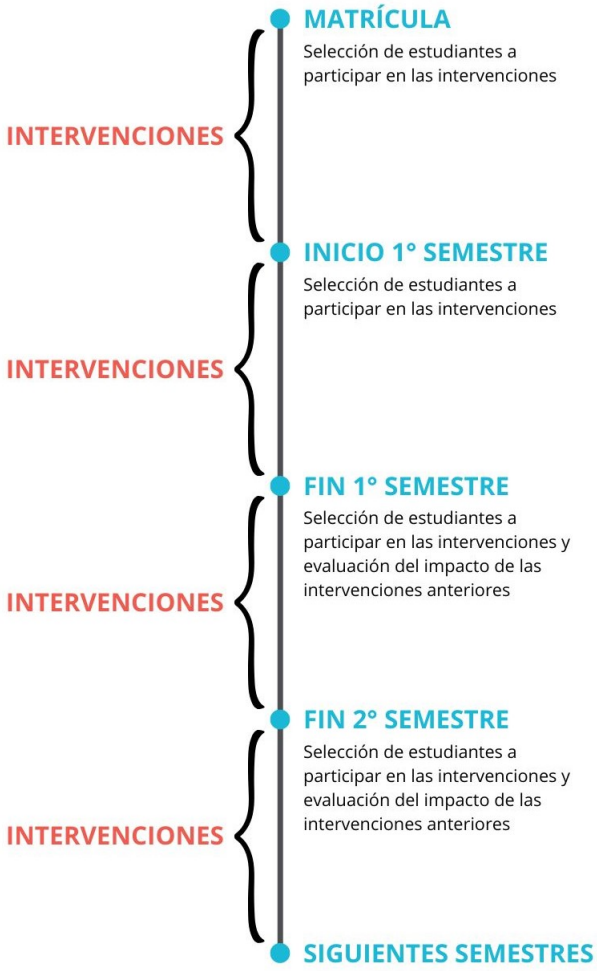


Figura 3.20: Diagrama del uso de los modelos

Considerando los resultados obtenidos, sumado a que los modelos buscan ser herramientas para distintos períodos de tiempo que permitan comprender tempranamente qué factores afectan en tener éxito en el programa e identificar a las personas que corren mayor riesgo de no tenerlo, se reconocen cuatro períodos donde estos modelos pueden lograr su objetivo y ayudar a aplicar intervenciones en base a sus resultados, lo cual se puede apreciar en la figura 3.20. En detalle los períodos son:

1. **Entre el período de matrícula y el inicio del 1º semestre**: En este período se deben aplicar los modelos que sólo consideran las variables previas, para encontrar a los estudiantes que deberían ser parte de las intervenciones contra la deserción donde, si se modifica el umbral, se puede identificar al 80 % de estas personas y sobre el 50 % de

las que no desertan. En este período se recomienda principalmente crear intervenciones orientadas a nivelar los conocimientos, ya que, en los modelos analizados se reflejó que los estudiantes con peor rendimiento previo y de colegios con baja preparación para las pruebas PSU, corren mayor riegos de no terminar Plan Común. Alternativamente, se recomienda aplicar intervenciones enfocadas en enseñar los conocimientos mínimos para comprender los primeros cursos de Plan Común, sustentado en la causa de la recomendación anterior y en que la tasa de aprobación del 1º semestre es importante en los otros modelos, por lo cual formar una buena base, teórica y práctica, para entender estos ramos, antes de ser cursados, debería afectar positivamente la tasa de aprobación del 1º semestre y así la probabilidad de terminar Plan Común.

2. **Durante el 1º semestre**: En este período al igual que en el anterior se deben aplicar los modelos que sólo consideran las variables previas, para encontrar a los estudiantes que deberían ser parte de las intervenciones contra la deserción. Se recomienda realizar intervenciones del tipo tutorías académicas, enfocadas a los cursos con mayor tasa de reprobación, ya que la tasa de aprobación del 1º semestre de un estudiante juega un papel importante en la probabilidad de terminar correctamente Plan Común o no, según los modelos.

3. **Durante el 2º semestre**: En este período se deben aplicar los modelos que consideran las variables previas y la tasa de aprobación del 1º semestre, para encontrar a los estudiantes que deberían ser parte de las intervenciones contra la deserción, se debe destacar que la calidad de las predicciones mejora, logrando identificar correctamente cerca del 90 % de las personas que deserta y sobre 50 % de las que no, si se modifica el umbral. Se recomienda realizar intervenciones del tipo tutorías académicas, enfocadas en primera instancia a los cursos que corresponden al 1º semestre, ya que, en esta instancia si un estudiante reprueba un curso correspondiente al 1º semestre, indica que entra en causa de ser eliminado, sumado con que una mayor tasa de aprobación en el 2º semestre, afecta positivamente. Además, se puede evaluar a priori las intervenciones realizadas en el 1º semestre, ya que en esta instancia la probabilidad de terminar Plan Común correctamente, para quienes participaron de alguna intervención, debería cambiar positivamente en comparación con la pronosticada anteriormente.

4. **Durante los siguientes semestres**: En este período se deben aplicar los modelos que

consideran las variables previas y las tasas de aprobación del 1º y 2º semestre, para encontrar a los estudiantes que deberían ser parte de las intervenciones contra la deserción. No se puede recomendar ninguna intervención, ya que las variables importantes de este modelo no se pueden alterar finalizado el 2º semestre, pero sí se puede evaluar a priori las intervenciones realizadas en el 2º semestre, ya que se espera que la probabilidad de terminar Plan Común aumente en las personas que participaron en las intervenciones, en comparacion con los modelos ocupados en el resto de períodos.

Cabe mencionar que, independiente del período y modelo, siempre se considerada una gran cantidad de estudiantes que terminaran Plan Común correctamente ($y_i = 1$) en las intervenciones, este problema irá perdiendo relevancia al pasar los distintos períodos de predicción y focalización, sin embargo, a pesar de lo anterior, ser partícipe de estas intervenciones debería generar una externalidad positiva, ya que, disminuirían aún más su riego de no terminar Plan Común correctamente.

Por último, es importante mencionar que cada intervención tiene requerimientos y limitantes distintos, por lo que se debe evaluar cuál de todos los modelos y umbrales obtenidos para el período de intervención, se ajusta de mejor manera a estos.

Capítulo 4

Conclusiones

Predecir si un alumno terminará Plan Común o desertará involuntariamente por razones académicas, es una tarea compleja de estudiar, debido a diversos factores, tales como el desbalance de datos, heterogeneidad en el desempeño de los estudiantes, e información relevante que no se posee en los datos, como por ejemplo características psicosociales. A pesar de lo anterior, los modelos predictivos que usaron métodos de balance de clases entregaron buenos resultados para predecir la aprobación o abandono de Plan Común por razones académicas.

Estos modelos no sólo permiten predecir si un alumno aprobará o no Plan Común, sino también permiten comprender este fenómeno, a través de las relaciones e impactos que existen entre las variables explicativas consideradas y la variable a explicar.

Acompañado de lo anterior, la predicción de los modelos va mejorando considerablemente cuando, no sólo se consideran variables previas, sino también factores que se van creando una vez que el estudiante ingresa a la Facultad, lo cual se ve reflejado en los diversos modelos que mejoran su predicción gracias al uso de variables que van considerando el desempeño de los estudiantes en los primeros semestres del programa, donde destaca la tasa de aprobación del 1º semestre, ya que, es una variable que ayudo considerablemente en mejorar la predicción y además, afecta positivamente la probabilidad de terminar Plan Común.

La utilización de estos modelos permite detectar a los alumnos que corren mayor riesgo de desertar por motivos académicos, antes de que ingresen al Plan Común como también cuando ya lo están cursando. Lo cual permite que las iniciativas de la FCFM que vayan en contra de esto, puedan seleccionar a sus participantes de forma justificada. Cabe mencionar que cada iniciativa tiene requerimientos y limitaciones distintas, por lo cual la FCFM, puede

escoger el modelo y umbral de corte que más se ajuste a sus requerimientos.

Se destacan cuatro períodos de intervención en donde los modelos pueden ser ocupados, los cuales son: (1) Entre el período de matrícula y el inicio del 1º semestre, (2) Durante el 1º semestre, (3) Durante el 2º semestre y (4) Durante los siguientes semestres.

Al avanzar los semestres por cohorte de ingreso, los grupos abarcados por las intervenciones van mejorando cada vez más, donde se incluirían al 66 % de las personas que desertan con una precisión del 20 % en las intervenciones de los dos primeros períodos, pasando a abarcar al 85 % de las personas que desertan con una precisión del 43 % en las intervenciones del último período, esto considerando los modelos más equilibrados para focalizar.

Además, se recomiendan diversos ejes para las intervenciones en los diferentes períodos, debido a los insight que entregan estos modelos, donde se pasa de recomendar nivelaciones de conocimiento en (1) a recomendar apoyos y tutorías en los cursos más reprobados en el (2) y (3) . Para (4) no se puede recomendar ejes de intervención debido a que las variables consideradas no se pueden cambiar finalizado el 2º semestre.

Sin embargo, se debe tener en consideración que independiente del modelo y el período, existirán estudiantes que se clasificarán erróneamente, donde los estudiantes que terminarán Plan Común (y_i=1) y que sean mal clasificados ($\widehat{y}_i$=0), obtendrá una externalidad positiva, ya que, disminuirían aún más su riesgo de no terminar Plan Común correctamente, mientras que los estudiantes que no terminarán Plan Común (y_i=0) y que sean mal clasificados ($\widehat{y}_i$=1), no alcanzarán a participar de las intervenciones para contrarrestar que deserten.

Por último, es importante destacar que, a modo general los modelos Logit lograron buenos resultados, aún cuando son de las herramientas más sencillas, lo cual es relevante, ya que, este tipo de modelo es sencillo de interpretar. Además, se espera que esta clase de modelos mejore en el futuro, debido, principalmente, al reemplazo de la PSU por la PAES, ya que se pasa de utilizar datos de una prueba estandarizada, que normaliza sus puntajes a través de una escala que tiene promedio 500 puntos y desviación estándar de 150, con máximo 850 puntos y mínimo 150 (DEMRE, sf), a una que no aplica ninguna transformación en sus resultados, por lo cual hace que estos sean comparables en los diferentes años, además de que PAES tiene como objetivo evaluar competencias, en otras palabras quiere medir tanto “el saber” como “el saber hacer”, integrando los conocimientos y habilidades requeridos para el éxito universitario (DEMRE, sf).

Bibliografía

Acuña, C. (2012). *Acceso y deserción en la educación superior: caso aplicado a Chile*. Universidad de Chile. Santiago, Chile. Recuperado de: https://repositorio.uchile.cl/handle/2250/112062

Aguirre, N. (2012). *Factores que predicen el rendimiento académico en la escuela de ingeniería de la Universidad de Chile*. Universidad de Chile. Santiago, Chile. Recuperado de: https://repositorio.uchile.cl/bitstream/handle/2250/112299/cf-aguirre_ng.pdf?sequence=1&isAllowed=y

Amazon Web Services (s.f). *¿Qué es la regresión logística?*. Auckland, Nueva Zelanda. Recuperado de: https://aws.amazon.com/es/what-is/logistic-regression/#:~:text=La%20regresi%C3%B3n%20log%C3%ADstica%20es%20una,factores%20bas%C3%A1ndose%20en%20el%20otro.

Amazon Web Services (s.f.). *¿Qué es una red neuronal?*. Auckland, Nueva Zelanda. Recuperar de: https://aws.amazon.com/es/what-is/neural-network/#:~:text=Una%20red%20neuronal%20es%20un,lo%20hace%20el%20cerebro%20humano.

Amestica, L., King, A., Gutiérrez, D. y Ramírez, V. (2021). *Efectos económicos de la deserción en la gestión universitaria: el caso de una universidad pública chilena*. Universidad del Bio-Bio. Concepción, Chile. Recuperado de: https://www.researchgate.net/publication/346703020_Efectos_economicos_de_la_desercion_en_la_gestion_universitaria_el_caso_de_una_universidad_publica_chilena

Barrios, A. (2011). *Deserción universitaria en Chile: incidencia del financiamiento y otros factores asociados*. Revista CIS. Vol. 9, Nº14, p. 59-72. Recuperado de: https://dialnet.unirioja.es/servlet/articulo?codigo=6310278

Bean, J. (1981). *Student Attrition, Intentions, and Confidence: Interaction Effects in a Path Model. Part I, The 23 Variable Model*. Universidad de Nebraska-Lincoln, Nebraska,

Estados Unidos. Recuperado de: https://eric.ed.gov/?id=ED202443

Campbell, C. y Fuqua, D. (2008). *Factor predictive of student completion in a collegiate honors program*. Journal of college student retention. Vol. 10. Estados Unidos. Recuperado de: https://journals.sagepub.com/doi/10.2190/CS.10.2.b

Canales, A. y De los Ríos, D. (2007). *Factores explicativos de la deserción universitaria*. Universidad de Santiago de Chile. Santiago, Chile. Recuperado de: https://www.calidadenlaeducacion.cl/index.php/rce/article/view/239

Celis, S., Moreno, L., Poblete, P., Villanueva, J. y Weber, R. (2015). *Un modelo analítico para la predicción del rendimiento académico de estudiantes de ingeniería*. Revista Ingeniería de Sistemas. Vol. XXIX. Santiago, Chile. Recuperado de: https://www.dii.uchile.cl/~ris/RIS2015/rendimientoac.pdf

Centro de Estudios MINEDUC (s.f.). *Datos abiertos*. Ministerio de Educación. Santiago, Chile. Recuperado de: https://datosabiertos.mineduc.cl/

Centro de Microdatos (2008). *Estudio sobre causas de la deserción universitaria*. Universidad de Chile. Santiago, Chile. Recuperado de: https://bibliotecadigital.mineduc.cl/bitstream/handle/20.500.12365/17988/E08-0041.pdf?sequence=1&isAllowed=y

Clark Labs (s.f.). *Classification Tree Analysis*. Worcester, Massachusetts. Recuperado de: https://clarklabs.org/classification-treeanalysis/#:~:text=A%20classification%20tree%20is%20a,that%20leads%20to%20categorical%20decisions.

Contreras, C. (2021). *Determinación de variables predictivas de deserción inicial para generar un sistema de alerta temprana. Análisis sobre una muestra de estudiantes beneficiarios de la beca de nivelación académica en una universidad pública en Chile*. Calidad en la Educación Nº 54. Universidad de Valparaíso. Valparaíso, Chile. Recuperado de: https://www.scielo.cl/scielo.php?pid=S0718-45652021000100012&script=sci_arttext

Čukušić, M., Garača, Z. y Jadrić, M. (2010). *Student Dropout Analysis with Application of Data Mining Methods*. Management: Journal of Contemporary Management Issues. Vol. 15, Nº 1, p. 31-46. Recuperado de : https://hrcak.srce.hr/en/file/81744

Daza, A. (2016). *Un modelo basado en árboles de decisión para predecir la deserción estudiantil en la Educación Superior Privada*. UCV-Scientia. Vol. 8, Nº1, p. 59–73. Recuperado de : https://doi.org/10.18050/RevUcv-Scientia.v8n1a7

Departamento de Evaluación, Medición y Registro Educacional (s.f). *¿Cómo se calcula el puntaje?*, Santiago, Chile. Recuperado de: https://demre.cl/psu/la-prueba/que-es-la-psu/c

alculo-puntaje-psu

Departamento de Evaluación, Medición y Registro Educacional (s.f). *Preguntas frecuentes - PAES 2022*, Santiago, Chile. Recuperado de: https://demre.cl/mesa-de-ayuda/preguntas-frecuentes-PAES

Díaz, C. (2008). *Modelo conceptual para la deserción estudiantil universitaria chilena.* Estudios Pedagógicos XXXIV, Nº 2. Concepción, Chile. Recuperado de: https://www.scielo.cl/scielo.php?pid=S071807052008000200004&script=sci_arttext&tlng=pt

Díaz, C. (2009). *Factores de Deserción Estudiantil en Ingeniería: Una Aplicación de Modelos de Duración.* Información tecnológica. Vol. 20, Nº 5, p.129-145. Recuperado de : https://dx.doi.org/10.4067/S0718-07642009000500016

Dombrovskaia, L., del Rio, J. y Rodríguez, P. (2020). *Prediction of student's retention in first year of engineering program at a technological chilean university.* 39th International Conference of the Chilean Computer Science Society (SCCC), Coquimbo, Chile. Recuperado de: https://www.researchgate.net/publication/347544547_Prediction_of_student's_retention_in_first_year_of_engineering_program_at_a_technological_chilean_university

Fazio, M. (2004). *Incidencia de las Horas Trabajadas en el Rendimiento Académico de Estudiantes Universitarios Argentinos.* Universidad Nacional de La Plata, La Plata, Argentina. Recuperado de: http://sedici.unlp.edu.ar/handle/10915/3543

González, Á. (2018). *Transición con equidad hacia universidades selectivas: El caso de las vías de acceso inclusivo en la Universidad de Santiago.* Editorial Universidad Santiago de Chile, Colección Divulgación. Santiago, Chile. Recuperado de: https://www.paiep.usach.cl/sites/paiep/files/documentos/transicion_con_equidad_web_0.pdf

Himmel, E. (2002). *Modelo de análisis de la deserción estudiantil en la educación superior.* Calidad en la Educación. Nº 17, Santiago, Chile. Recuperado de: https://doi.org/10.31619/caledu.n17.409

International Business Machines Corporation (2021). *El modelo de redes neuronales.* Nueva York, Estados Unidos. Recuperado de: https://www.ibm.com/docs/es/spss-modeler/saas?topic=networks-neural-model

International Business Machines Corporation (2021). *Funcionamiento de SVM.* Nueva York, Estados Unidos. Recuperado de: https://www.ibm.com/docs/es/spss-modeler/saas?topic=models-how-svm-works

International Business Machines Corporation (s.f.). *¿Qué es el aprendizaje supervisado?.*

Nueva York, Estados Unidos. Recuperado de: https://www.ibm.com/mx-es/topics/supervised-learning

International Business Machines Corporation (s.f.). *¿Qué es KNN?*. Nueva York, Estados Unidos. Recuperado de: https://www.ibm.com/mx-es/topics/knn?mhsrc=ibmsearch_a&mhq=knn

International Business Machines Corporation (s.f.). *¿Qué es un árbol de decision?*. Nueva York, Estados Unidos. Recuperado de: https://www.ibm.com/es-es/topics/decision-trees

Inostroza, C. (2020). *Análisis de desempeño de estudiantes del programa de ingreso especial de equidad de género de la Facultad de Ciencias Físicas y Matemáticas*. Universidad de Chile. Santiago, Chile. Recuperado de: https://repositorio.uchile.cl/bitstream/handle/2250/176869/An%c3%a1lisis-de-desempe%c3%b1o-de-estudiantes-del-programa-de-ingreso-especial-de-equidad-de-g%c3%a9nero-de-la-Facultad-de-Ciencias-F%c3%adsicas-y-Matem%c3%a1ticas.pdf?sequence=1&isAllowed=y

Kokkelenberg, E. y Sinha, E. (2010). *Who succeeds in STEM studies? An analysis of Binghamton University undergraduate students*. Economics od Education Review. Vol. 29. Washington, Estados Unidos. Recuperado de: https://www.sciencedirect.com/science/article/abs/pii/S0272775710000816

Kuna, H., García, R. y Villatoro, F. (2010). *Identificación de causales de abandono de estudios universitarios. Uso de procesos de explotación de información*. Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología. Nº 5, p. 39-44. Recuperado de: http://sedici.unlp.edu.ar/bitstream/handle/10915/18991/Documento_completo.pdf?sequence=1&isAllowed=y

Larroucau, T. (2013). *Estudio de los factores determinantes de la deserción en el sistema universitario chileno*. Universidad de Chile. Santiago, Chile. Recuperado de: https://repositorio.uchile.cl/bitstream/handle/2250/114843/cf-larroucau_td.pdf?sequence=1&isAllowed=y

Merkle (2020). *El algoritmo K-NN y su importancia en el modelado de datos*. Aubing, Alemania. Recuperado de: https://www.merkle.com/es/es/blog/algoritmo-knn-modelado-datos

Ministerio de Educación (2022), *Mineduc presenta la nueva Prueba de Acceso a la Educación Superior (PAES) que reemplaza definitivamente a la PSU*. Santiago, Chile. Recuperado de: https://www.mineduc.cl/prueba-de-acceso-a-la-educacion-superior-paes/

Moreno, J., Rodríguez, D., Sicilia, M., Riquelme, J. y Ruiz, R. (2009). *SMOTE-I: mejora*

del algoritmo SMOTE para balanceo de clases minoritarias. Actas de los talleres de las jornadas de ingeniería del software y bases de datos. Vol. 3, Nº 1. Sevilla, España. Recuperado de: http://www.cc.uah.es/drg/adis2009/articles/adis-09-Moreno-ISMOTE.pdf

Rilling, C. (2022). *10 años de políticas de equidad de acceso en la Universidad de Chile*. Universidad de Chile. Santiago, Chile. Recuperado de: https://repositorio.uchile.cl/handle/2250/184717

Rodríguez, M. y Zamora, J. (2014). *Análisis de la deserción en la Universidad Nacional desde una perspectiva longitudinal*. Universidad Nacional de Costa Rica. Herida, Costa Rica. Recuperado de: https://doi.org/10.13140/RG.2.2.30416.66569

Scikit-learn (s.f.). *sklearn.tree.DecisionTreeClassifier*. Recuperado de: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier

Soria, K. y Zúñiga, S. (2014). *Aspectos Determinantes Del Éxito Académico De Estudiantes Universitarios*. Formación universitaria. Vol. 7, Nº 5. La Serena, Chile. Recuperado de: https://dx.doi.org/10.4067/S0718-50062014000500006

Tinto, V. (1975), *Dropout in higher education: A theoretical synthesis of recent research*, Review of Educational Research. Vol 45, Nº1, p. 89-125.

Universidad de Chile (s.f.). *Departamento de Evaluación, Medición y Registro Educacional*. Vicerrectoría de Asuntos Académicos. Santiago, Chile. Recuperado de: https://demre.cl/

Universidad de Chile (s.f.). *Ingresos especiales. Facultad de Ciencias Físicas y Matemáticas*. Santiago, Chile. Recuperado de: https://ingenieria.uchile.cl/admision/ingresos-especiales.html

Universidad de Chile (s.f.). *¿Qué se estudia en la FCFM?*. Facultad de Ciencias Físicas y Matemáticas. Santiago, Chile. Recuperado de: https://ingenieria.uchile.cl/carreras/que-se-estudia-en-la-fcfm

Julca, J., Larios, A., Navarro, Á y Valero, J. (2022). *Deserción universitaria: Evaluación de diferentes algoritmos de Machine Learning para su predicción*. Revista de Ciencias Sociales (Ve). Vol. 28, Nº3. Recuperado de: https://www.redalyc.org/articulo.oa?id=28071865024

Veloso, C. (2020). *Herramienta para la simulación de trayectorias estudiantiles: aplicación al cálculo de los ingresos por gratuidad de la Universidad de Chile*. Universidad de Chile. Santiago, Chile. Recuperado de: https://repositorio.uchile.cl/bitstream/handle/2250/176184/Herramienta-para-la-simulacion-de-trayectorias-estudiantiles-aplicacion-al-calculo-de-l

os.pdf?sequence=1&isAllowed=y

Yiu, T. (2019). *Understanding Random Forest*. Medium. Recuperado de: https://towardsdatascience.com/understanding-random-forest-58381e0602d2#:~:text=The%20random%20forest%20is%20a,that%20of%20any%20individual%20tree.

Anexos

Anexo A. Información descriptiva

Variable	Valor variable	y_i	Cantidad
Opción	1	1	3608
		0	382
	2	1	424
		0	55
	3	1	32
		0	4
	4	1	15
		0	2
Rural	Urbano	1	4064
		0	438
	Rural	1	15
		0	5
Tipo de colegio	Particular	1	1774
		0	171
	Subvencionado	1	1235
		0	152
	Municipal	1	1070
		0	120
Mayor edad	1	1	3021
		0	351
	0	1	1058
		0	92
Mujer	Si	1	1063
		0	113
	No	1	3016
		0	330

Variable	Valor variable	y_i	Cantidad
Segmento	0	1	548
		0	84
	1	1	1177
		0	143
	2	1	2354
		0	216
Ing. Opción	1	1	3535
		0	377
	0	1	544
		0	66
Semestres pandemia	0	1	2222
		0	221
	1	1	630
		0	79
	2	1	0
		0	1
	3	1	623
		0	80
	4	1	604
		0	62
Ingreso especial	1	1	362
		0	98
	0	1	3717
		0	345
P076	Humanista Científico Diurno	1	4034
		0	428
	Humanista Científico Nocturno	1	10
		0	3
	Técnico Profesional Comercial	1	9
		0	6
	Técnico Profesional Industrial	1	26
		0	6
Egreso	0	1	122
		0	49
	1	1	380
		0	57
	2	1	3577
		0	337

Tabla A.1: Información descriptiva variables no continuas

Variable	y_i	Mediana	Media	Desviación estándar
Puntaje de ingreso	0.0	728.5	727.79	26.95
	1.0	740.3	747.04	27.99
NEM	0.0	711.0	706.90	53.23
	1.0	723.0	723.69	48.17
Ranking	0.0	753.0	680.26	245.32
	1.0	771.0	690.04	250.01
P022	0.0	0.0	0.57	4.30
	1.0	0.0	0.15	2.17
Puntaje de Mat	0.0	738.0	740.92	51.43
	1.0	762.0	764.87	44.14
Puntaje de Len	0.0	681.0	685.06	64.38
	1.0	685.0	687.34	59.49
Puntaje Ciencias	0.0	694.0	691.60	52.55
	1.0	716.0	718.98	48.75
Delta puntaje	0.0	8.00	7.13	27.09
	1.0	19.65	26.41	27.92
Delta cero	0.0	0.0	6.40	16.84
	1.0	0.0	1.18	6.27
Prom Leng Mat	0.0	553.96	546.73	49.74
	1.0	570.14	566.36	42.05
Ratio NEM sobre Prom Leng Mat	0.0	1.07	1.08	0.09
	1.0	1.05	1.05	0.08
Razón aprobación pandemia	0.0	0.0	0.01	0.04
	1.0	0.0	0.00	0.02

Tabla A.2: Información descriptiva variables continuas

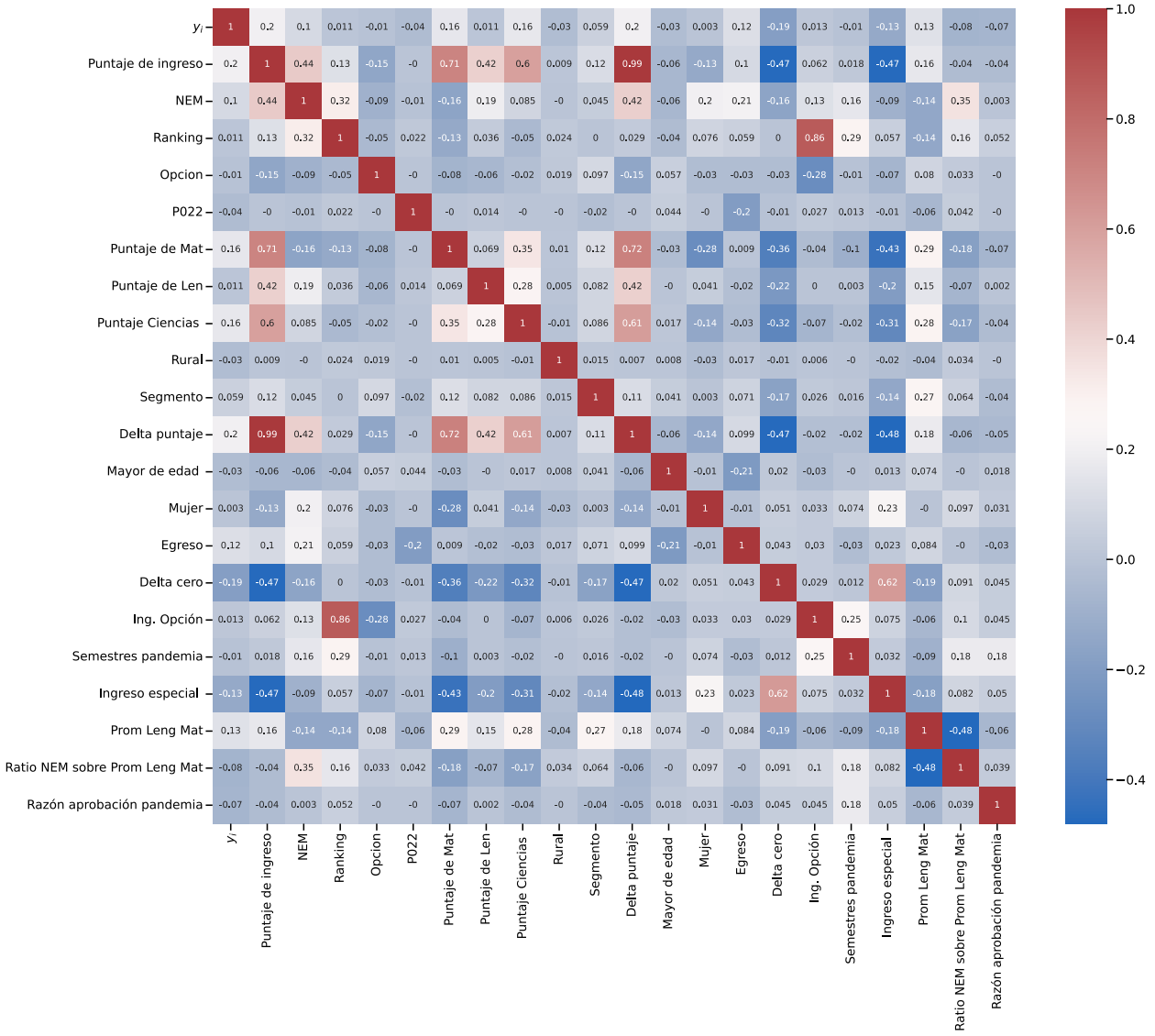


Figura A.1: Matriz de correlación variables no categóricas

Anexo B. Información descriptiva de las Tasas de aprobación

Variable	y_i	Mediana	Media	Desviación estándar
Tasas de aprobación 1º Sem	0.0	0.6	0.56	0.30
	1.0	1.0	0.92	0.18
Tasas de aprobación 2º Sem	0.0	0.5	0.47	0.34
	1.0	1.0	0.90	0.19

Tabla B.1: Información descriptiva Tasas de aprobación

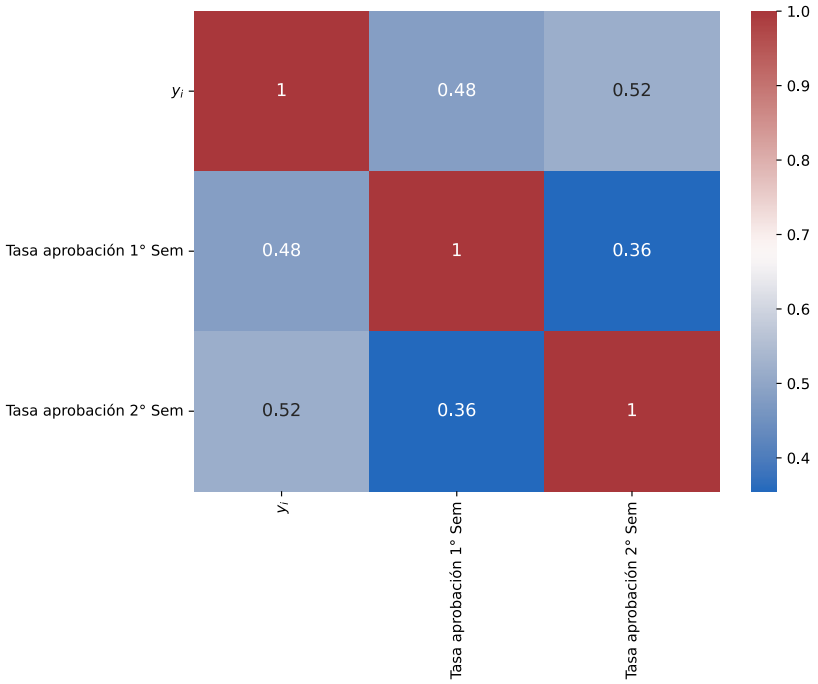


Figura B.1: Matriz de correlación Tasas de aprobación

Anexo C. Resultados de los modelos

C.1. KNN

	Train							Test						
	Precisión $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
KNN *	1	1	1	1	1	1	1	0.91	0.92	0.91	0.16	0.15	0.15	0.84
KNN **	0.96	0.70	0.81	0.21	0.75	0.33	0.70	0.93	0.68	0.79	0.16	0.55	0.25	0.67
KNN ***	0.96	0.68	0.80	0.20	0.74	0.31	0.68	0.93	0.65	0.77	0.15	0.56	0.24	0.64
* Sin ajustes en el balanceo, ** Sobre-muestreo aleatorio, *** Sobre muestro sintético SMOTE (Synthetic Minority Oversampling Technique)														

Tabla C.1: Resultados de los modelos KNN

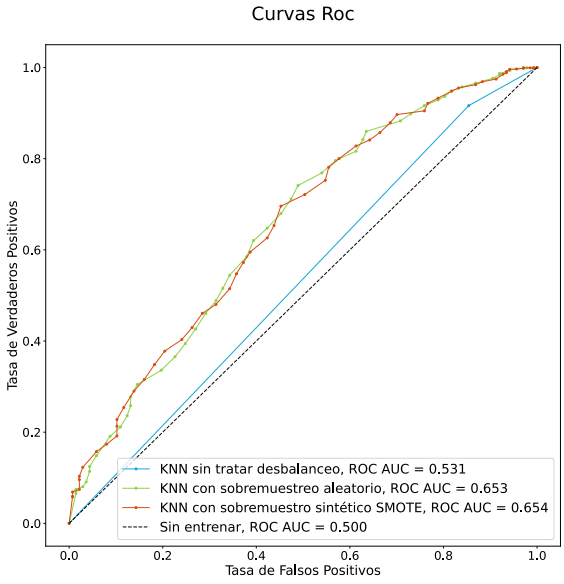


Figura C.1: Curvas ROC modelos KNN

C.2. SVM

	Train							Test						
	Precisión $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
SVM *	0.95	0.68	0.79	0.19	0.69	0.29	0.68	0.95	0.67	0.78	0.19	0.69	0.29	0.67
SVM **	0.96	0.68	0.79	0.19	0.71	0.30	0.68	0.95	0.66	0.78	0.18	0.69	0.29	0.66
SVM ***	0.95	0.64	0.77	0.17	0.72	0.28	0.65	0.95	0.62	0.75	0.17	0.69	0.27	0.63
* Peso de las clases equilibrado, ** Sobre-muestreo aleatorio, *** Sobre muestro sintético SMOTE (Synthetic Minority Oversampling Technique)														

Tabla C.2: Resultados de los modelos SVM

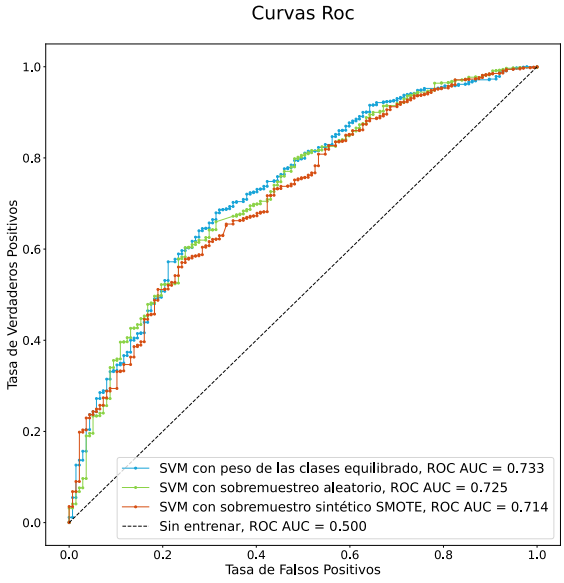


Figura C.2: Curvas ROC modelos SVM

C.3. DT

	Train							Test						
	Precisión $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
DT *	0.95	0.60	0.73	0.16	0.72	0.26	0.61	0.93	0.59	0.72	0.14	0.60	0.23	0.59
DT **	0.95	0.70	0.80	0.19	0.66	0.30	0.69	0.93	0.69	0.79	0.16	0.53	0.24	0.67
DT ***	0.96	0.69	0.80	0.20	0.71	0.31	0.69	0.92	0.66	0.77	0.15	0.52	0.23	0.65
* Peso de las clases equilibrado, ** Sobre-muestreo aleatorio, *** Sobre muestro sintético SMOTE (Synthetic Minority Oversampling Technique)														

Tabla C.3: Resultados de los modelos DT

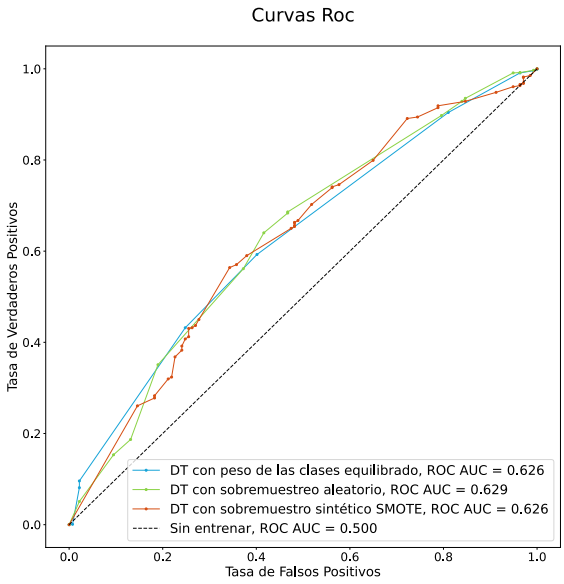


Figura C.3: Curvas ROC modelos DT

C.4. RF

	Train							Test						
	Precisión $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
RF *	0.96	0.63	0.76	0.18	0.73	0.28	0.64	0.94	0.60	0.73	0.15	0.64	0.25	0.61
RF **	0.96	0.66	0.79	0.19	0.74	0.30	0.67	0.93	0.65	0.77	0.16	0.58	0.25	0.64
RF ***	0.94	0.69	0.79	0.17	0.62	0.27	0.68	0.93	0.67	0.78	0.15	0.53	0.24	0.66
* Peso de las clases equilibrado, ** Sobre-muestreo aleatorio, *** Sobre muestro sintético SMOTE (Synthetic Minority Oversampling Technique)														

Tabla C.4: Resultados de los modelos RF

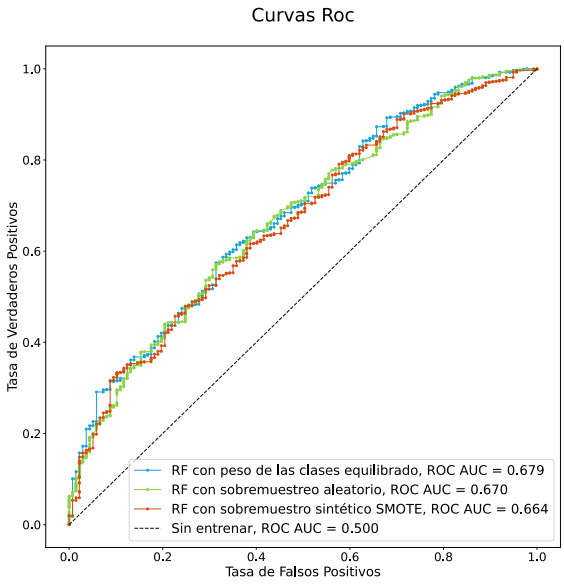


Figura C.4: Curvas ROC modelos RF

C.5. RN

	Train							Test						
	Precisión $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
RN *	0.91	0.99	0.95	0.51	0.10	0.16	0.90	0.91	0.99	0.94	0.45	0.11	0.18	0.90
RN **	0.95	0.70	0.81	0.20	0.67	0.30	0.70	0.95	0.70	0.80	0.20	0.66	0.30	0.69
RN ***	0.95	0.67	0.79	0.18	0.68	0.29	0.68	0.94	0.66	0.78	0.18	0.64	0.28	0.66
* Sin ajustes en el balanceo, ** Sobre-muestreo aleatorio, *** Sobre muestro sintético SMOTE (Synthetic Minority Oversampling Technique)														

Tabla C.5: Resultados de los modelos RN

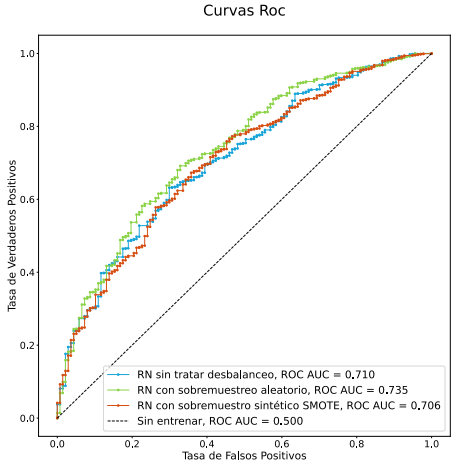


Figura C.5: Curvas ROC modelos RN

C.6. Logit

	Train							Test						
	Precisión $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
Logit *	0.94	0.69	0.80	0.17	0.60	0.27	0.68	0.93	0.69	0.79	0.16	0.53	0.25	0.67
Logit **	0.94	0.69	0.80	0.17	0.61	0.27	0.68	0.93	0.69	0.70	0.17	0.55	0.25	0.67
Logit ***	0.94	0.73	0.82	0.17	0.53	0.26	0.71	0.94	0.73	0.82	0.19	0.57	0.29	0.72
* Peso de las clases equilibrado, ** Sobre-muestreo aleatorio, *** Sobre muestro sintético SMOTE (Synthetic Minority Oversampling Technique)														

Tabla C.6: Resultados de los modelos Logit

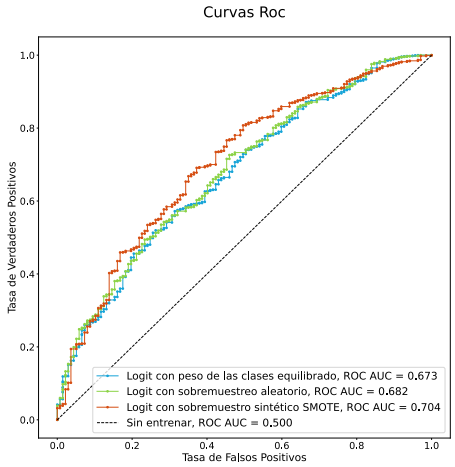


Figura C.6: Curvas ROC modelos Logit

Anexo D. Resultados de los modelos, considerando la Tasa de Aprobación del 1º Semestre

D.1. KNN, considerando la Tasa de Aprobación del 1º Semestre

	Train							Test						
	Precisión $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
KNN *	0.90	1	0.95	0.5	0.01	0.03	0.90	0.90	1	0.95	0.67	0.01	0.03	0.9
KNN **	1	1	1	1	1	1	1	0.96	0.82	0.88	0.29	0.69	0.41	0.80
KNN ***	1	1	1	1	1	1	1	0.95	0.82	0.88	0.29	0.65	0.40	0.80
* Sin ajustes en el balanceo, ** Sobre-muestreo aleatorio, *** Sobre muestro sintético SMOTE (Synthetic Minority Oversampling Technique)														

Tabla D.1: Resultados de los modelos KNN, considerando la Tasa de Aprobación del 1º Semestre

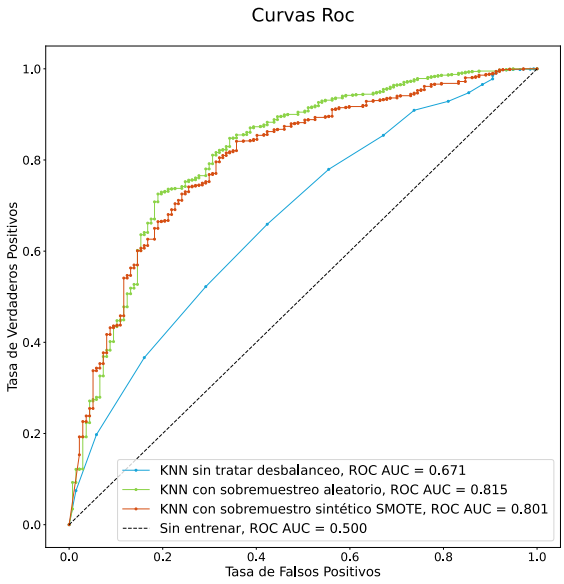


Figura D.1: Curvas ROC modelos KNN, considerando la Tasas de aprobación del 1º Semestre

D.2. SVM, considerando la Tasa de Aprobación del 1º Semestre

	Train							Test						
	Precisión $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
SVM *	0.97	0.84	0.90	0.34	0.77	0.47	0.83	0.97	0.84	0.90	0.34	0.73	0.47	0.83
SVM **	0.98	0.80	0.88	0.30	0.81	0.44	0.80	0.97	0.80	0.88	0.30	0.76	0.43	0.80
SVM ***	0.96	0.88	0.92	0.35	0.62	0-45	0.85	0.95	0.89	0.92	0.36	0.55	0.43	0.85
* Peso de las clases equilibrado, ** Sobre-muestreo aleatorio, *** Sobre muestro sintético SMOTE (Synthetic Minority Oversampling Technique)														

Tabla D.2: Resultados de los modelos SVM, considerando la Tasa de Aprobación del 1º Semestre

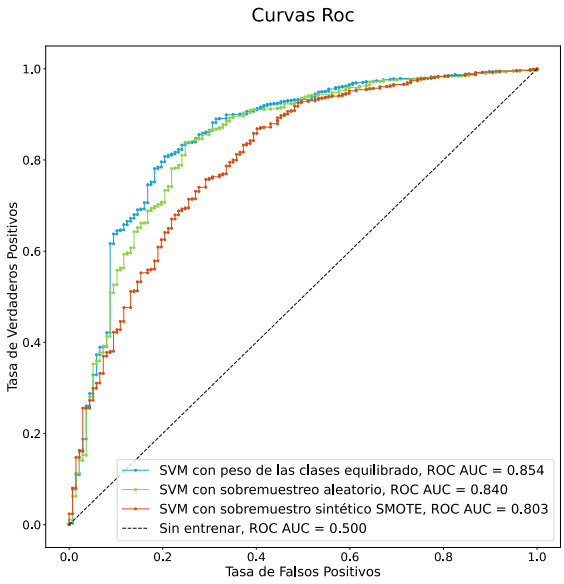


Figura D.2: Curvas ROC modelos SVM, considerando la Tasas de aprobación del 1º Semestre

D.3. DT, considerando la Tasa de Aprobación del 1º Semestre

	Train							Test						
	Precisión $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
DT *	0.97	0.82	0.89	0.32	0.79	0.46	0.82	0.96	0.82	0.89	0.31	0.69	0.42	0.81
DT **	0.98	0.78	0.86	0.28	0.82	0.42	0.78	0.97	0.80	0.88	0.30	0.80	0.44	0.80
DT ***	0.97	0.82	0.89	0.32	0.79	0.45	0.81	0.96	0.82	0.88	0.30	0.70	0.42	0.81
* Peso de las clases equilibrado, ** Sobre-muestreo aleatorio, *** Sobre muestro sintético SMOTE (Synthetic Minority Oversampling Technique)														

Tabla D.3: Resultados de los modelos DT, considerando la Tasa de Aprobación del 1º Semestre

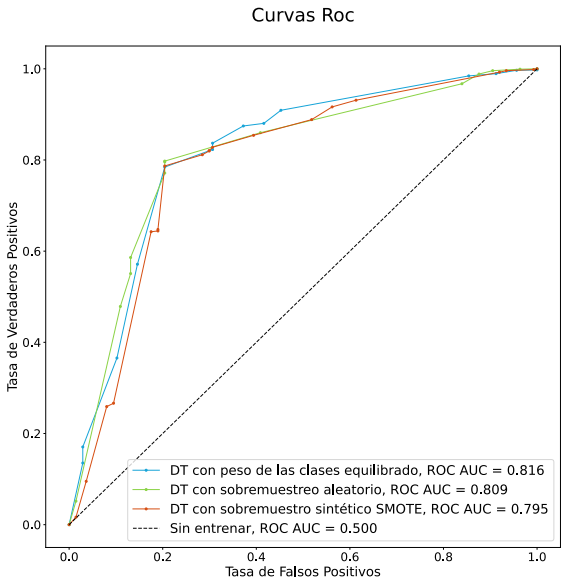


Figura D.3: Curvas ROC modelos DT, considerando la Tasas de aprobación del 1º Semestre

D.4. RF, considerando la Tasa de Aprobación del 1º Semestre

	Train							Test						
	Precisión $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
RF *	0.98	0.90	0.94	0.47	0.87	0.61	0.89	0.95	0.88	0.92	0.38	0.63	0.47	0.86
RF **	0.99	0.91	0.95	0.53	0.95	0.68	0.91	0.95	0.88	0.91	0.36	0.61	0.45	0.85
RF ***	0.98	0.91	0.94	0.50	0.82	0.62	0.90	0.94	0.89	0.92	0.35	0.52	0.42	0.86
* Peso de las clases equilibrado, ** Sobre-muestreo aleatorio, *** Sobre muestro sintético SMOTE (Synthetic Minority Oversampling Technique)														

Tabla D.4: Resultados de los modelos RF, considerando la Tasa de Aprobación del 1º Semestre

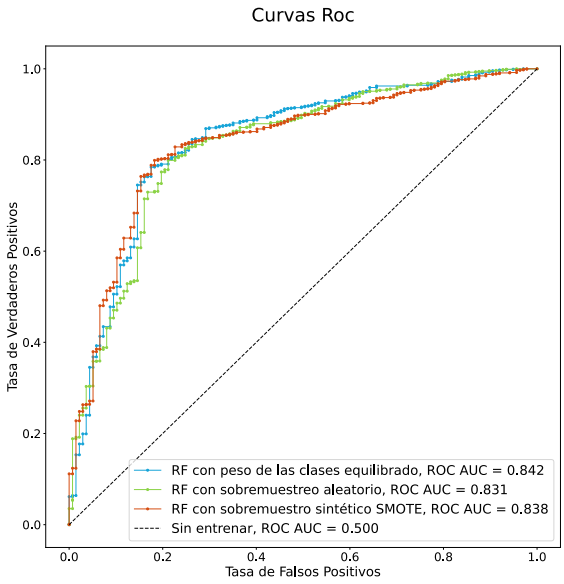


Figura D.4: Curvas ROC modelos RF, considerando la Tasas de aprobación del 1º Semestre

D.5. RN, considerando la Tasa de Aprobación del 1º Semestre

	Train							Test						
	Precisión $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
RN *	0.93	0.98	0.95	0.56	0.27	0.37	0.91	0.92	0.97	0.95	0.53	0.26	0.34	0.90
RN **	0.97	0.84	0.90	0.34	0.77	0.47	0.83	0.97	0.84	0.90	0.35	0.74	0.48	0.83
RN ***	0.96	0.88	0.92	0.36	0.65	0.46	0.85	0.95	0.89	0.92	0.37	0.58	0.45	0.86
* Sin ajustes en el balanceo, ** Sobre-muestreo aleatorio, *** Sobre muestro sintético SMOTE (Synthetic Minority Oversampling Technique)														

Tabla D.5: Resultados de los modelos RN, considerando la Tasa de Aprobación del 1º Semestre

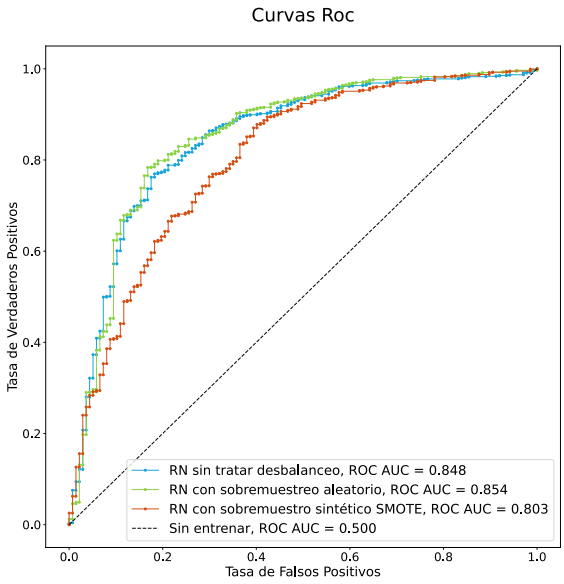


Figura D.5: Curvas ROC modelos RN, considerando la Tasas de aprobación del 1º Semestre

D.6. Logit, considerando la Tasa de Aprobación del 1º Semestre

	Train							Test						
	Precisión $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
Logit *	0.96	0.86	0.91	0.35	0.70	0.46	0.84	0.96	0.87	0.91	0.37	0.68	0.48	0.85
Logit **	0.97	0.86	0.91	0.35	0.72	0.47	0.84	0.96	0.87	0.92	0.39	0.72	0.50	0.86
Logit ***	0.96	0.87	0.92	0.37	0.70	0.49	0.86	0.96	0.88	0.92	0.39	0.68	0.50	0.86
* Peso de las clases equilibrado, ** Sobre-muestreo aleatorio, *** Sobre muestro sintético SMOTE (Synthetic Minority Oversampling Technique)														

Tabla D.6: Resultados de los modelos Logit, considerando la Tasa de Aprobación del 1º Semestre

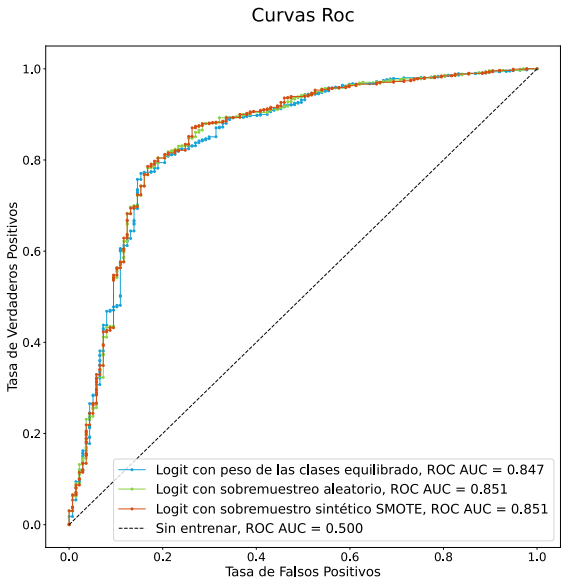


Figura D.6: Curvas ROC modelos Logit, considerando la Tasas de aprobación del 1º Semestre

Anexo E. Resultados de los modelos, considerando las Tasas de aprobación del 1º y 2º Semestre

E.1. KNN, considerando las Tasas de aprobación del 1º y 2º Semestre

	Train							Test						
	Precisión $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
KNN *	1	1	1	1	1	1	1	0.91	1	0.95	0.88	0.11	0.19	0.91
KNN **	1	1	1	1	1	1	1	0.97	0.85	0.90	0.35	0.74	0.48	0.84
KNN ***	1	1	1	1	1	1	1	0.97	0.87	0.92	0.39	0.73	0.51	0.86
* Sin ajustes en el balanceo, ** Sobre-muestreo aleatorio, *** Sobre muestro sintético SMOTE (Synthetic Minority Oversampling Technique)														

Tabla E.1: Resultados de los modelos KNN, considerando las Tasas de aprobación del 1º y 2º Semestre

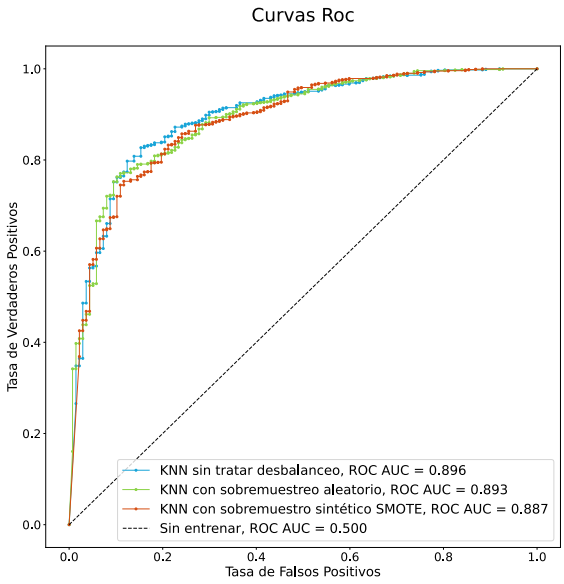


Figura E.1: Curvas ROC modelos KNN, considerando las Tasas de aprobación del 1º y 2º Semestre

E.2. SVM, considerando las Tasas de aprobación del 1º y 2º Semestre

	Train							Test						
	Precisión $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
SVM *	0.98	0.86	0.92	0.40	0.83	0.54	0.86	0.98	0.87	0.92	0.43	0.83	0.56	0.87
SVM **	0.98	0.84	0.90	0.36	0.86	0.51	0.84	0.98	0.84	0.90	0.38	0.87	0.53	0.84
SVM ***	0.97	0.91	0.94	0.46	0.76	0.58	0.89	0.96	0.90	0.93	0.44	0.70	0.54	0.88
* Peso de las clases equilibrado, ** Sobre-muestreo aleatorio, *** Sobre muestro sintético SMOTE (Synthetic Minority Oversampling Technique)														

Tabla E.2: Resultados de los modelos SVM, considerando las Tasas de aprobación del 1º y 2º Semestre

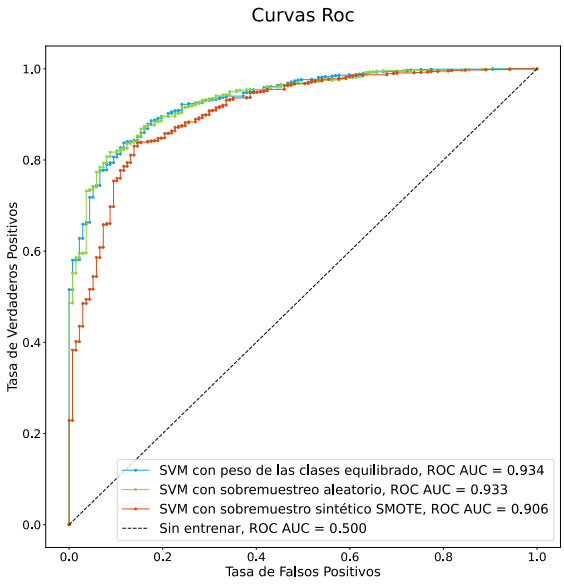


Figura E.2: Curvas ROC modelos SVM, considerando las Tasas de aprobación del 1º y 2º Semestre

E.3. DT, considerando las Tasas de aprobación del 1º y 2º Semestre

	Train							Test						
	Precisión $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
DT *	0.97	0.85	0.91	0.35	0.74	0.48	0.84	0.96	0.86	0.91	0.36	0.71	0.48	0.85
DT **	0.98	0.80	0.88	0.32	0.87	0.47	0.81	0.98	0.81	0.89	0.33	0.85	0.48	0.81
DT ***	0.96	0.82	0.89	0.30	0.72	0.43	0.81	0.96	0.84	0.89	0.31	0.65	0.42	0.82
* Peso de las clases equilibrado, ** Sobre-muestreo aleatorio, *** Sobre muestro sintético SMOTE (Synthetic Minority Oversampling Technique)														

Tabla E.3: Resultados de los modelos DT, considerando las Tasas de aprobación del 1º y 2º Semestre

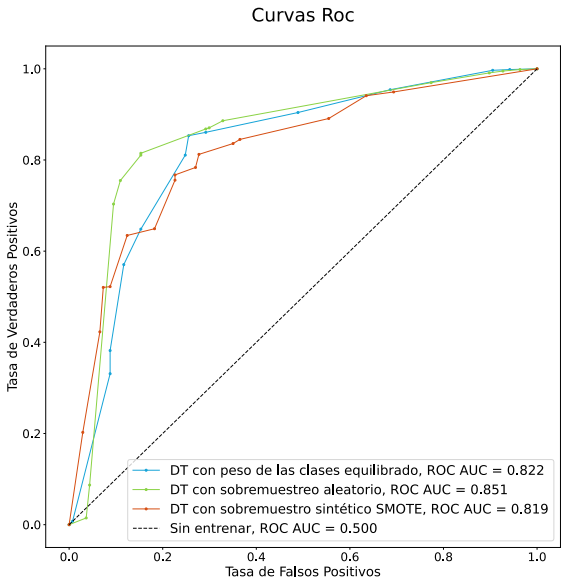


Figura E.3: Curvas ROC modelos DT, considerando las Tasas de aprobación del 1º y 2º Semestre

E.4. RF, considerando las Tasas de aprobación del 1º y 2º Semestre

	Train							Test						
	Precisión $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
RF *	0.99	0.94	0.96	0.62	0.91	0.74	0.94	0.96	0.92	0.94	0.51	0.69	0.58	0.90
RF **	1	0.97	0.98	0.77	0.99	0.87	0.97	0.96	0.93	0.94	0.51	0.68	0.58	0.90
RF ***	0.98	0.93	0.95	0.55	0.79	0.65	0.92	0.96	0.92	0.94	0.51	0.69	0.58	0.90
* Peso de las clases equilibrado, ** Sobre-muestreo aleatorio, *** Sobre muestro sintético SMOTE (Synthetic Minority Oversampling Technique)														

Tabla E.4: Resultados de los modelos RF, considerando las Tasas de aprobación del 1º y 2º Semestre

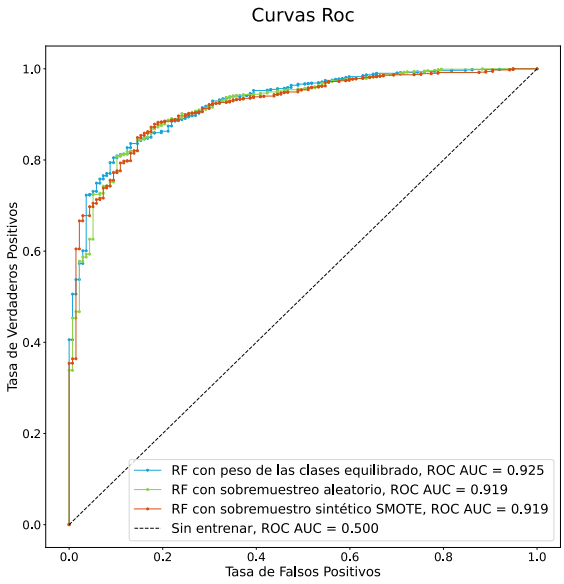


Figura E.4: Curvas ROC modelos RF, considerando las Tasas de aprobación del 1º y 2º Semestre

E.5. RN, considerando las Tasas de aprobación del 1º y 2º Semestre

	Train							Test						
	Precisión $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
RN *	0.95	0.98	0.97	0.78	0.51	0.62	0.94	0.94	0.98	0.96	0.73	0.43	0.58	0.93
RN **	0.98	0.86	0.92	0.40	0.84	0.54	0.86	0.98	0.86	0.92	0.40	0.82	0.54	0.86
RN ***	0.97	0.90	0.93	0.43	0.70	0.53	0.88	0.96	0.90	0.93	0.43	0.68	0.53	0.88
* Sin ajustes en el balanceo, ** Sobre-muestreo aleatorio, *** Sobre muestro sintético SMOTE (Synthetic Minority Oversampling Technique)														

Tabla E.5: Resultados de los modelos RN, considerando las Tasas de aprobación del 1º y 2º Semestre

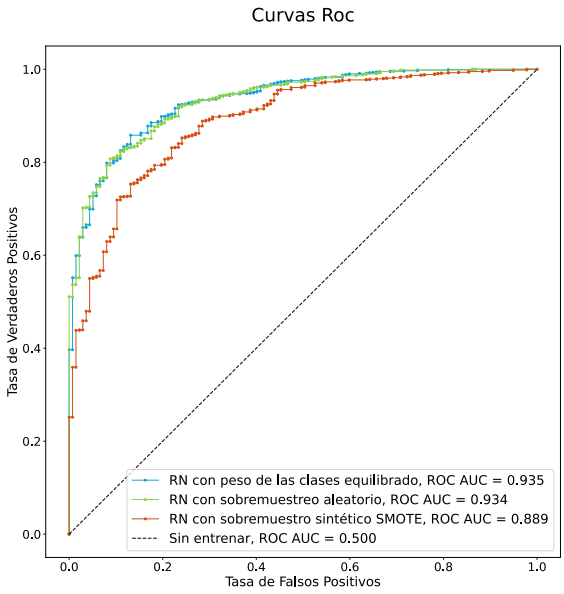


Figura E.5: Curvas ROC modelos RN, considerando las Tasas de aprobación del 1º y 2º Semestre

E.6. Logit, considerando las Tasas de aprobación del 1º y 2º Semestre

	Train							Test						
	Precisión $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**	**Precisión** $y_i = 1$	**Recall** $y_i = 1$	**F1-score** $y_i = 1$	**Precisión** $y_i = 0$	**Recall** $y_i = 0$	**F1-score** $y_i = 0$	**Accuracy**
Logit *	0.98	0.87	0.92	0.40	0.80	0.53	0.86	0.98	0.87	0.92	0.43	0.85	0.57	0.87
Logit **	0.98	0.86	0.91	0.38	0.82	0.52	0.85	0.98	0.86	0.92	0.41	0.85	0.56	0.86
Logit ***	0.98	0.87	0.92	0.40	0.80	0.53	0.86	0.98	0.88	0.93	0.43	0.85	0.57	0.87
* Peso de las clases equilibrado, ** Sobre-muestreo aleatorio, *** Sobre muestro sintético SMOTE (Synthetic Minority Oversampling Technique)														

Tabla E.6: Resultados de los modelos Logit, considerando las Tasas de aprobación del 1º y 2º Semestre

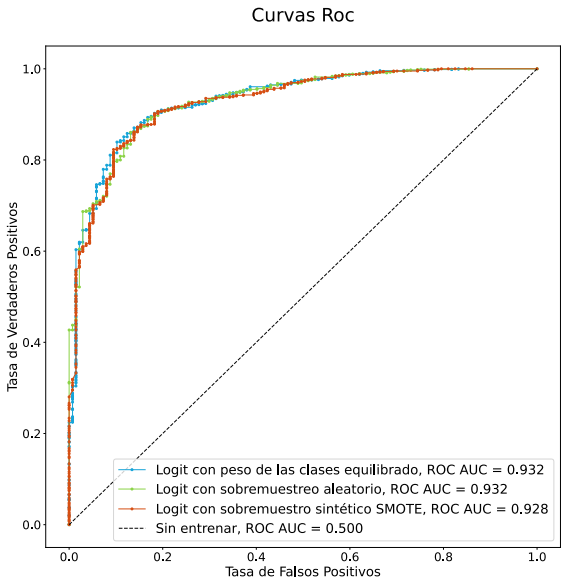


Figura E.6: Curvas ROC modelos Logit, considerando las Tasas de aprobación del 1º y 2º Semestre