UNIVERSIDAD
DE LA REPUBLICA
URUGUAY

Modelización del desempeño educativo en la educación media mediante aprendizaje automático.

Emilio Aguirre

Federico Veneri

Licenciatura en Estadística
Facultad de Ciencias Económicas y Administración
Universidad de la República

Montevideo – Uruguay
Agosto de 2018

UNIVERSIDAD
DE LA REPUBLICA
URUGUAY

Modelización del desempeño educativo en la educación media mediante aprendizaje automático.

Emilio Aguirre

Federico Veneri

Monografía de grado presentada como parte de los requisitos necesarios para la obtención del título de Licenciado en Estadística , Licenciatura en Estadística, Facultad de Ciencias Económicas y Administración de la Universidad de la República,

Director de tesis:
Ph.D Mathias Bourel

Montevideo – Uruguay
Agosto de 2018

INTEGRANTES DEL TRIBUNAL DE DEFENSA DE TESIS

Prof. Mathias Bourel

Prof. Natalia da Silva

Prof. Laura Nalbarte

Montevideo – Uruguay
Agosto de 2018

Los estadísticos silenciosos han cambiado nuestro mundo; no descubriendo nuevos hechos o desarrollos técnicos, sino cambiando el formas en que razonamos, experimentamos y formamos nuestras opiniones

Ian Hacking
citado en Hastie et. al. (2017)

RESUMEN

En este documento se modeló la promoción educativa de estudiantes de cuarto año de educación media pública en Uruguay con datos provenientes del programa Compromiso Educativo (CE) del año 2012. Con tal fin se emplearon distintos modelos de clasificación: regresión logística, árboles de clasificación CART, árboles de clasificación de inferencia condicional (CTREE) y bosques aleatorios (con CART y CTREE); se exploraron distintas técnicas para seleccionar el conjunto de datos de aprendizaje: simple, Down y SMOTE. La estimación del poder predictivo se realizó mediante un proceso basado en la validación cruzada, utilizando herramientas de diseño de experimento para medir diferencias significativas entre distintas estrategias de modelización. Los resultados obtenidos indican que el bosque aleatorio basado en árboles CTREE y con muestreo simple del conjunto de aprendizaje, presenta en promedio mejores resultados al medir el área bajo la curva ROC (79.54 %) y su versión parcial (66.85 %), sin embargo, existen otras alternativas frente a las cuales no existen diferencias estadísticamente significativas.

Tomando este modelo como el mejor candidato, se estudió la importancia de las variables y sus efectos parciales. El deber exámenes, imaginarse estudiando en educación terciaria, haber repetido en educación media y si su familia lo imagina en educación terciaria son las variables que tienen mayor incidencia sobre el AUC. Los gráficos de dependencia parcial indicaron que no deber exámenes, no haber repetido, imaginarse en educación terciaría y que su familia lo imagine en educación terciaria tienen un efecto positivo sobre la probabilidad de aprobar. Finalmente, se sugiere un punto de corte de 0.78 para el problema de clasificación en base a este modelo y pruebas de entrenamiento y testeo.

Palabras claves:
Aprendizaje automático, Árboles de clasificación, Remuestreo, Desempeño educativo, Educación Media.

Tabla 1: Siglas y abreviaciones utilizadas.

AIC	Criterio de Información de Akaike (Akaike Information criterion)
ANEP	Administración Nacional de Educación Pública.
ANOVA	Análisis de Varianza (Analysis Of VAriance)
BAGGING	Bootstrap AGGreggtING
AUC	Área bajo la curva ROC (Area Under the Curve).
CART	Arboles de regresión y clasificación (Classificaction and Regressions Trees).
CES	Consejo de Educación Secundaria de la ANEP.
CETP	Consejo de Educación Técnico Profesional de la ANEP. También identificado por UTU [Universidad del Trabajo del Uruguay].
CTREE	Árbol de inferencia condicional (Conditional Inference Tree)
Down	Técnica de sobremuestreo que consiste en quitar observaciones de la clase mayoritaria aleatoriamente.
ECH	Encuesta Continua de Hogares.
EM	Educación Media.
EMB	Educación Media Básica. Corresponde a los primeros tres años de la Educación Media posteriores a la educación primaria.
EMS	Educación Media Superior. Corresponde a los últimos tres años de la Educación Media posteriores a la EMB.
ENAJ	Encuesta Nacional de Adolescencia y Juventud.
ICC	Índice de Carencias Críticas.
INE	Instituto Nacional de Estadística.
INEEd	Instituto Nacional de Evaluación Educativa.
MIDES	Ministerio de Desarrollo Social.
PAUC	Área parcial bajo la curva ROC.
PCE	Programa Compromiso Educativo.
RF	Bosque aleatorio (Random Forest).
ROC	Curva ROC (Receive Operating Curve).
SMOTE	Técnica de sobremuestreo de la clase minoritaria, generación de casos sineticos (Synthetic Minority Over-sampling TEchnique)
Up	Técnica de sobremuestreo que consiste en replicar observaciones de la clase minoritaria.

Tabla de contenidos

Lista de tablas

Lista de figuras

Introducción

Existe un amplio consenso sobre la **importancia de la educación** para el bienestar humano [53]. Por ejemplo, Sen sostiene que la educación tiene una relevancia directa para el bienestar y la libertad de las personas, así como un papel indirecto al influir en el cambio social y la producción económica [51]. Además del valor intrínseco de la educación en sí misma, la literatura sugiere relaciones positivas entre la educación, el crecimiento económico y la remuneración laboral. Esta relación se vuelve más pronunciada en los países más pobres. Una extensa literatura a proporcionado evidencia empírica de un vínculo entre mejores sistemas educativos y otros indicadores del desarrollo humano, incluido el estado de salud, la mortalidad materna e infantil, el menor crecimiento de la población y la reducción de la delincuencia. Las personas con altos niveles de educación tienen mayores probabilidades de tener un empleo, generar mayores ingresos, superar las conmociones económicas y mantener familias más saludables [6].

Pese a algunos avances en los últimos años, **Uruguay** presenta un problema estructural de bajas tasas de **finalización del nivel medio**. Este fenómeno posiciona a Uruguay como uno de los países de la región con menor tasa de egreso del nivel medio y mayor disparidad de logros educativos según estrato socioeconómico del hogar de origen del joven [2].

El **objetivo general** de este documento es contribuir al desarrollo de modelos predictivos sobre el desempeño educativo en la Educación Media. Los objetivos **específicos** son: (1) entrenar distintos modelos predictivos de promoción educativa; (2) estimar la capacidad predictiva de forma honesta (por fuera de la muestra de entrenamiento); (3) analizar el efecto de utilizar distintas estrategias de remuestro del conjunto de entrenamiento sobre la capacidad predictiva; (4) obtener un ordenamiento de la capacidad predictiva de los modelos, e identificar si existen diferencias estadísticamente significativas entre los mismos.

Este documento busca generar los siguientes **aportes**: (1) colaborar en el desarrollo de herramientas para detectar en forma temprana a estudiantes con alto riesgo de no aprobar; (2) emplear una fuente de información poco explorada en la literatura nacional; (3) utilizar técnicas de aprendizaje automático poco exploradas en la literatura educativa (arboles de inferencia condicional, técnicas de selección del conjunto de entrenamiento y herramientas de la literatura de diseño de experimentos para identificar si existen diferencias estadísticamente significativas en la capacidad predictiva de los distintos modelos analizados).

A partir de la **información** relevada por el Programa Compromiso Educativo (PCE) para los alumnos de cuarto año con los registros administrativos de educación media publica del año 2012, en este documento, se estimaron distintos modelos de aprendizaje automático supervisado para predecir la promoción, utilizando distintas técnicas de remuestreo sobre el conjunto de entrenamiento para intentar mejorar la capacidad predictiva.

Detectar tempranamente los jóvenes que están en riesgo de no aprobar el año lectivo permitiría establecer políticas focalizadas que permitieran abordar esta situación. En este documento se realizó una primera aproximación al desarrollo de este tipo herramientas.

Los problemas de clasificación de dos poblaciones, por ejemplo alumnos que aprueban o no, pueden presentar desbalance; es decir, existe una mayor proporción de estudiantes dentro de una categoría. Es esperable que los jóvenes que no aprueban sean una proporción menor. En este trabajo, distintas técnicas han sido propuestas para intentar mejorar el poder predictivo de distintos modelos.

Mediante una estrategia basada en la validación cruzada se obtuvieron estimaciones honestas del poder predictivo, medido por el área debajo de la curva ROC (AUC) y su versión parcial (PAUC), como principales métricas para evaluar la capacidad predictiva de la combinación de modelos y técnicas de remuestreo. Adicionalmente, se exploró cuáles son las variables más importantes y su efecto parcial para el modelo con mejor capacidad predictiva.

El presente documento se estructura en 5 capítulos. En el capítulo 1 se realiza una caracterización de la educación media en Uruguay con datos de encuestas de hogares, y en el 2 se pasa revista a la literatura previa, tanto a nivel teórico como empírico. En el capítulo 3 se describe brevemente las herramientas estadísticas empleadas en este documento, y en el 4 se aborda la

estrategia empírica para implementar los modelos. Por último en el capítulo 5 se discuten los resultados.

Capítulo 1

Breve caracterización de la educación media en Uruguay

La **educación media en Uruguay**[1] se estructura en dos ciclos, la educación media básica (EMB[2]) y la educación media superior (EMS[3]). La EMB corresponde a los tres años siguiente a la educación primaria y es obligatoria desde 1973. Su cometido es la profundización de las competencias obtenidas durante la educación primaria y adquirir competencias en diferentes disciplinas. La EMS constituye la continuación de este proceso teniendo un mayor grado de especialización respecto a la EMB, y es obligatoria desde el año 2008 en Uruguay [27]. La educación media (EM) es ofrecida en liceos (EM no vocacional, gestionado por el Consejo de Educación Superior) y en centros técnicos (o vocacionales, gestionados por el Consejo de Educación Técnico Profesional), en centros privados y públicos.

En este apartado se van a describir los logros educativos en EM en Uruguay, mediante los datos de las Encuesta Continua de Hogares (ECH) y la Encuesta Nacional de Adolescencia y Juventud (ENAJ) del Instituto Nacional de Estadísticas (INE).

De la ECH-INE se encuentra que entre el 2007 y 2015 la proporción de jóvenes entre 18 y 20 años que finalizó la EMB pasó del 69.9 % al 71.3 % y de la EMS pasó del 26.6 % al 29.5 % (Tabla 1.1).

A pesar de estas mejoras, **Uruguay** se encuentra en una situación desfavorable **en la región**. La proporción de jóvenes de 18 a 20 años que finalizaron la

[1] Este apartado este basado en [1, 2] .
[2] Grados del 7 al 9, con edades teóricas de 12 a 14.
[3] Grados del 10 al 12, con edades teóricas de 15 a 17.

EMS en el 2011 asciende a 76 % en Chile, 56 % en Bolivia, 48 % en Argentina, 47 % en Brasil, 43 % en Paraguay y 28 % en Uruguay [27].

Existen fuertes brechas en logros educativos de acuerdo al sexo del estudiante y el nivel socioeconómico de su familia. Las **mujeres**, finalizan la EMB y EMS 10 puntos porcentuales (pp) más que los hombres (Tabla 1.2). En el año 2015 el 47 % de los jóvenes de 18 a 20 años del primer quintil de **ingreso** del hogar, finalizó EMB mientras que este guarismo para los jóvenes del último quintil es del 87 %. En cuanto a la finalización de EMS, estos guarismos fueron respectivamente 12.4 % y 48.5 %. La brecha de finalización entre quintiles[4] para el periodo es significativamente distinta de cero, rondando 40 puntos porcentuales (pp) para la EMB y 36pp para EMS.

Al consultar a los jóvenes respecto al principal **motivo** por el cual **no finalizaron la EM** (Tabla 1.4), en el 2015 surge como principal razón la falta de interés o el interés por aprender otras cosas (45.2 %), seguido del inicio de la vida laboral (33.2 %).

Con el fin de describir las trayectorias educativas, Aguirre y Veneri [2] utilizan la ENAJ 2013 para ver los recorridos educativos de los jóvenes entre 23 y 29 años. La Figura 1.1 presenta un árbol descriptivo de potenciales **senderos educativos** que puede realizar un estudiante. Se definen como hitos importantes a lo largo de la trayectoria educativa, el inicio y la finalización de ciclos educativos, así como eventos de repetición en cada nivel. De la población entre 23 y 29 años en el 2013, únicamente el 46 % finalizó EM mientras que el 4 % continúa estudiando. El 41 % inicio el ciclo pero no lo finalizó, mientras que el 9 % no lo inició.

Se observa que los jóvenes que no repitieron en la primaria ni en la EM finalizaron el 74.8 %, mientras del grupo de jóvenes que no repitieron en primaria, pero si lo hicieron en EM terminaron sus estudios en el nivel medio el 35.2 %. De los estudiantes que sólo repitieron en primaria finalizaron la EM 21.2 %, a su vez, los que además repitieron en EM finalizaron 1.1 %.

En síntesis, Uruguay ha evidenciado una leve mejora en cuanto a los indicadores de finalización de EMB y EMS para los jóvenes de 18 a 20 años, sin embargo, una mirada comparativa con la región evidencia que aún se encuentra en una situación desfavorable. Se constató una brecha de finalización según ingresos y género significativa que permanece a lo largo del tiempo y se encontró una fuerte incidencia de la repetición en primaria y en EM sobre la

[4] La diferencia absoluta entre el quinto y el primer quintil.

trayectoria educativa posterior.

Tabla 1.1: Finalización EM, jóvenes 18 a 20 años en localidades de más de 5.000 habitantes, según años.

	EMB completa			EMS completa		
	% jovenes	IC inf. 95 %	IC sup. 95 %	% jóvenes	IC inf. 95 %	IC sup. 95 %
2015	71.30 %	69.88 %	72.71 %	29.51 %	28.13 %	30.90 %
2013	70.10 %	68.74 %	71.46 %	29.10 %	27.78 %	30.42 %
2011	69.40 %	67.93 %	70.87 %	29.01 %	27.59 %	30.43 %
2009	69.84 %	68.47 %	71.21 %	27.58 %	26.27 %	28.90 %
2007	69.88 %	68.59 %	71.17 %	26.54 %	25.30 %	27.78 %

Fuente: [2] en base a ECH-INE .

Tabla 1.2: Finalización EM según sexo y años, jóvenes de 18 a 20 años en localidades de 5.000 hab. o más

Año	Sexo	EMB			EMS		
		Estimación	IC inf. 95	IC sup. 95	Estimación	IC inf. 95	IC sup. 95
2015	Mujer	76.6 %	74.8 %	78.4 %	34.8 %	32.8 %	36.8 %
	Hombre	66.3 %	64.3 %	68.3 %	24.5 %	22.7 %	26.4 %
2013	Mujer	76.5 %	74.8 %	78.2 %	35.5 %	33.5 %	37.4 %
	Hombre	64.0 %	62.0 %	65.9 %	23.0 %	21.3 %	24.7 %
2011	Mujer	75.5 %	73.6 %	77.4 %	35.5 %	33.5 %	37.4 %
	Hombre	63.9 %	61.8 %	66.0 %	23.0 %	21.3 %	24.7 %
2009	Mujer	74.5 %	72.7 %	76.3 %	33.2 %	31.3 %	35.2 %
	Hombre	65.4 %	63.5 %	67.4 %	22.2 %	20.5 %	23.9 %
2007	Mujer	73.9 %	72.2 %	75.6 %	32.0 %	30.2 %	33.8 %
	Hombre	65.7 %	63.8 %	67.5 %	20.9 %	19.3 %	22.5 %

Fuente: [2] en base a ECH-INE.

Tabla 1.3: Finalización EM según quintil de ingresos y años, jóvenes de 18 a 20 años en localidades de 5.000 hab o más

Año	Quintil	EMB			EMS		
		Estimación	IC Inf. 95 %	IC sup. 95 %	Estimación	IC inf. 95 %	IC sup. 95 %
2015	1	47.8 %	43.5 %	52.2 %	12.4 %	9.8 %	15.1 %
	2	59.0 %	55.4 %	62.7 %	17.1 %	14.4 %	19.8 %
	3	68.5 %	65.4 %	71.7 %	23.0 %	20.1 %	25.8 %
	4	76.4 %	73.6 %	79.2 %	31.6 %	28.6 %	34.6 %
	5	87.5 %	85.4 %	89.5 %	48.5 %	45.5 %	51.5 %
2013	1	41.8 %	37.8 %	45.9 %	9.7 %	7.2 %	12.1 %
	2	62.6 %	59.0 %	66.2 %	15.4 %	12.7 %	18.0 %
	3	69.3 %	66.2 %	72.4 %	23.7 %	20.9 %	26.5 %
	4	72.2 %	69.5 %	74.9 %	30.4 %	27.6 %	33.1 %
	5	84.2 %	82.1 %	86.3 %	46.6 %	43.9 %	49.4 %
2011	1	48.3 %	43.3 %	53.3 %	11.4 %	8.2 %	14.6 %
	2	53.1 %	49.2 %	57.0 %	14.7 %	11.9 %	17.4 %
	3	67.2 %	63.9 %	70.5 %	24.5 %	21.6 %	27.5 %
	4	71.3 %	68.3 %	74.3 %	26.6 %	23.8 %	29.4 %
	5	84.5 %	82.4 %	86.7 %	47.2 %	44.3 %	50.1 %
2009	1	41.3 %	37.0 %	45.6 %	9.3 %	6.8 %	11.7 %
	2	56.5 %	52.8 %	60.1 %	14.9 %	12.2 %	17.6 %
	3	64.6 %	61.6 %	67.7 %	20.6 %	18.0 %	23.1 %
	4	72.1 %	69.2 %	75.0 %	26.6 %	23.9 %	29.2 %
	5	88.4 %	86.7 %	90.2 %	46.2 %	43.4 %	48.9 %
2007	1	47.6 %	43.5 %	51.7 %	10.1 %	7.5 %	12.6 %
	2	57.3 %	53.7 %	60.8 %	12.8 %	10.4 %	15.3 %
	3	64.7 %	61.8 %	67.7 %	19.5 %	17.1 %	22.0 %
	4	72.4 %	69.8 %	75.0 %	24.4 %	22.0 %	26.8 %
	5	85.7 %	83.8 %	87.5 %	46.1 %	43.5 %	48.6 %

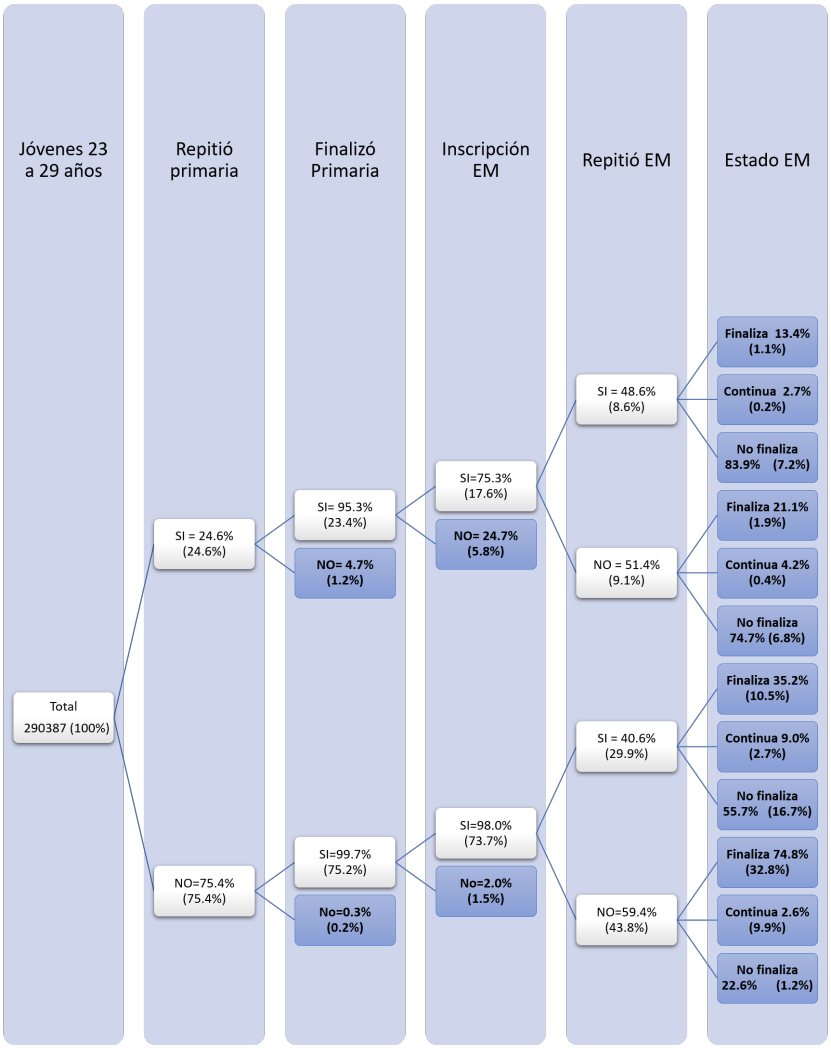
Fuente: [2] en base a ECH-INE.

Tabla 1.4: Principal motivo de no finalización, jóvenes 23 a 29 años localidades de 5.000 hab o más

	2011	2013	2015
Comenzó a trabajar	37.3 %	33.9 %	33.4 %
No tenía interés, le interesaba aprender otras cosas	38.5 %	44.3 %	45.2 %
Quedó usted o su pareja embarazada	7.0 %	7.9 %	6.8 %
Le resultaban difíciles las materias	4.4 %	3.0 %	3.8 %
Debió atender asuntos familiares	6.4 %	4.6 %	5.1 %
Otra razones y dificultades economicas	6.4 %	6.4 %	5.7 %
Total	100 %	100 %	100 %

Fuente: [2] en base a ECH-INE.

Figura 1.1: Trayectorias educativas jóvenes 23 a 29 años



Fuente: [2] en base a la ENAJ 2013.
Notas: En cada nodo se presenta información respecto a la proporción de la separación y entre paréntesis el tamaño del nodo en la población. Los nodos finales se encuentran resaltados en negrita, pudiendo sumar estos nodos para reproducir el 100 % de la población

Capítulo 2

Revisión teórica y antecedentes empíricos

En este capítulo, en la sección 2.1 se presentan los beneficios de la educación sobre la sociedad y los individuos, en la 2.2 se exponen algunos modelos teóricos sobre el desempeño educativo y en la 2.3, se sistematiza la evidencia empírica del impacto de distintos factores sobre el desempeño educativo en la Educación Media.

2.1. Beneficios de la educación

Siguiendo a **McMahon** [38], los **beneficios de la educación sobre el ciclo de vida** se puede esquematizar en una tabla de doble entrada. En las columnas se separa entre retornos privados y sociales de la educación, en tanto en las filas entre beneficios monetarios y no monetarios. En el cuadro 2.1 se presenta un listados no exhaustivo sobre los beneficios privados y sociales de la educación, donde se puede apreciar la importancia del fenómeno educativo para el individuo y la sociedad.

2.2. Modelo de desempeño educativo

Existen diversas teorías que explican el desempeño educativo, la mayoría de ellas enfatizan algún aspecto del fenómeno[1].

[1] Ver por ejemplo Fernández [15] para una revisión de la literatura.

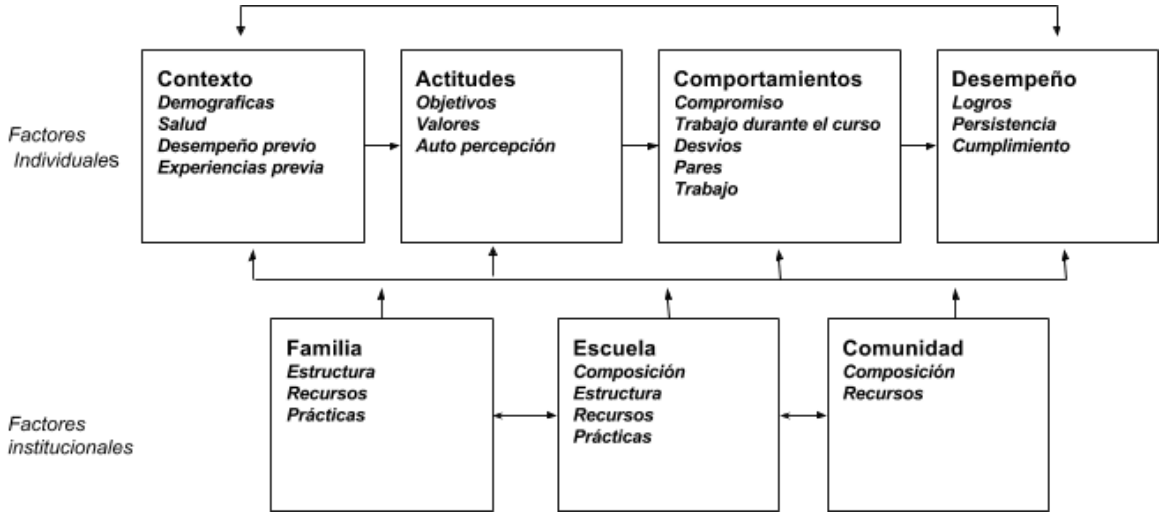
Tabla 2.1: Beneficios de la Educación

	Privados	Sociales
Monetarios	Mayores ingresos (skill premium)	Externalidades con efectos sobre el crecimiento
No Monetarios	1-Efectos sobres la salud 2-Capital humano producido en el hogar 3-Mas eficiente manejo de las finanzas del hogar 4-Mayor participación laboral 5-Mejor capacidad de adaptación y preparación para el re-entrenamiento 6-Efectos sobre la motivación 7-Satisfacción no monetaria del trabajo 8-Efecto consumo	1-Población y efectos sobre la salud 2-Democratización 3-Reducción de la pobreza y el crimen 4-Mejor cuidado del medio ambiente 5-Estructura familiar y retiro 6-Participación en servicios comunitarios 7-Distribución del ingreso

Fuente: Elaboración propia en base a McMahon. [38]

Rumberger y Lim [49] consideran que el desempeño educativo en la Educación Media (EM) se ve influenciado por dos grupos de **factores**. Por un lado factores **individuales** asociados al estudiante, y por otro factores **institucionales** asociados con tres grandes colectivos que afectan al estudiante: la familia, la institución educativa y la comunidad [ver Figura 2.1][2].

Figura 2.1: Modelo de desempeño educativo de Rumberger y Lim [49].



Fuente: Elaboración propia en base a [49].

Willms [62] considera como factores claves para el éxito estudiantil el compromiso y el aprendizaje, que se encuentra vinculados en una relación de causalidad reciproca. Desglosa el **compromiso** en tres aspectos. El primero

[2]Quedan por fuera de este enfoque factores biológicos, ver frith2011brain para un análisis de la educación desde la neurociencia.

es el social y se vincula al grado de involucramiento del estudiante con la vida social del centro. El segundo se vincula con lo institucional y analiza si el estudiante valora y se esfuerza por cumplir los requisitos formales para el éxito escolar. El tercero denominado intelectual, mide si el estudiante hace una inversión emocional y psicológica en el aprendizaje.

2.3. Evidencia empírica

2.3.1. Antecedentes sobre desempeño educativo

En un **meta análisis** sobre las causas de desafiliación educativa en la educación media para EEUU, **Rumberger y Lim** [49] analizan los resultados de 203 trabajos publicados sobre el tema.

Luego de revisar la literatura, los autores llegan a las siguientes **conclusiones**. La primera es que ningún factor por si solo es capaz de determinar la decisión de no continuar en el sistema educativo hasta la finalización de la educación media. La segunda es que la decisión de desafiliarse no depende exclusivamente de lo que pase en la institución educativa. La tercera, es que el fenómeno de desafiliación es más conveniente concebirlo como un proceso que como un evento. Los factores más influyentes para predecir dicho resultado, son los resultados académicos y el comportamiento social y académico. La cuarta es que el contexto importa, hay factores a nivel de las familias y comunidades que afectan la propensión a desafiliarse.

Los autores identifican dos tipos de factores, por un lado aquellos asociados a características individuales del sujeto y factores asociados a las instituciones o a las familias, escuelas o comunidades.

Respecto a los **factores individuales**, encuentran el desempeño educativo previo como un fuerte predictor de desafiliación. En tanto respecto al comportamiento se señala como importante el compromiso académico del estudiante y la integración del mismo al centro. A mayor ausentismo mayor riesgo de desafiliación.

Además se identifican varios factores que acentúan el riesgo de desafiliación, entre ellos: usos de drogas o alcohol, embarazo, responsabilidad en el cuidado de los niños, tener amigos que hayan cometidos delitos o que se hayan desafiliado; por último aquellos estudiantes que trabajen más de 20 horas semanales son más propensos a abandonar sus estudios.

Con respecto al rol de las creencias, valores y actitudes de los estudiantes, los autores señala que un mayor nivel de expectativas educativas se asocia a una menor propensión a desafiliarse. En relación a variables de contexto señalan que los hombres presentan peores resultados que las mujeres.

Por su parte con respecto a **predictores institucionales** los autores señalan que los estudiantes que viven con los dos padres poseen menores tasas de desafiliación. Además sostienen que aquellos cambios en la estructura familiar o otros eventos potencialmente estresantes aumentan el riesgo de desafiliación. Los autores encuentran que los estudiantes que pertenecen a aquellas familias con mayores recursos (medidos en educación de los padres, o ocupación o ingresos) presentan una menor probabilidad de desafiliación. Señalan algunas prácticas que reducen el riesgo de desafiliación, entre ellas: tener aspiraciones educativas altas para los hijos, monitorear el progreso escolar de los mismos, comunicarse con la escuela y conocer los padres de los amigos de sus hijos. Con respecto a la institución educativa señalan que cuatro características de las mismas son responsable del 20 % de la variabilidad en las tasas de desafiliación: la composición del estudiantado, recursos, aspectos estructurales, y políticas y prácticas. Señalan la presencia de un efecto composición, siendo menor la desafiliación cuanto menor sea la proporción de estudiantes de nivel socioeconómico bajo. Agregan que no se encuentra efecto en el tamaño de la clase, pero que si en cuanto a que la institución sea católica.

Los estudiantes son menos propensos a desafiliarse si asisten a instituciones con un fuerte clima académico, medido por más estudiantes tomando exámenes y haciendo las tareas domiciliarias. Adicionalmente señalan que obligar a los estudiantes a asistir hasta los 16 disminuye el riesgo de desafiliación. Por último, agregan que vivir en un barrio con un mayor poder adquisitivo se asocia con una menor probabilidad de desafiliación, esto podría estar operando mediante una mayor volumen de recursos de la comunidad y una positiva estructura de roles.

A modo de resumen se presenta el cuadro 2.2, que sintetiza lo presentado en este apartado. Esta revisión de la literatura, es utilizada para seleccionar el conjunto de variables a utilizar para la construcción de modelos predictivos.

Tabla 2.2: Factores que inciden sobre la desafiliación en la educación media.

Individuales	Performance educativa previa	[+] Resultados en pruebas
		[+] Logros Educativos previos
		[-] Repetición o abandono
	Comportamientos	[+] Involucramiento académico del estudiante
		[+] Integración social del estudiante
		[-] Uso de sustancias
		[-] Trabajar más de 20 horas
	Actitudes	[+] Expectativas educativas altas
	Contexto	[+] Mujer
Institucionales	Familia	[+] Vive con los dos padres
		[-] Cambios en la familia o eventos estresante
		[+] Recursos familiares
	Practicas familiares	[+] Altas aspiraciones educativas de los padres
		[+] Monitorear el progreso escolar de los hijos
		[+] Comunicarse con las instituciones educativa
		[+] Conocer los padres de los amigos de sus hijos
	Institución educativa	[+] Composición socioeconomica de la clase
		[?] Recursos
		[+] Aspectos estructurales (Católica)
		[+] Políticas y prácticas
	Comunidad	[+] Vivir en un barrio pudiente

Fuente: Elaboración propia en base a Rumberger y Lim [49]

2.3.2. Antecedentes nacionales

Patrón [43] analiza que la decisión de abandonar la EM en Uruguay puede ser resultado de una elección económica, totalmente racional. La autora estima la tasa interna de retorno de la inversión de continuar estudiando, contemplando la presencia de heterogeneidad de los estudiantes en probabilidad de repetición y en cuanto a la calidad de educación que reciben. Encuentra que para estudiantes de contexto desfavorecido continuar una año más en la EM, no les resulta una inversión económicamente conveniente.

Caballero y Jadra [10] estimaron mediante un modelo logit a partir de la ECH 2011, que elabora el INE, la probabilidad de asistencia de los jóvenes entre 14 y 17 años de edad. Encuentran que características como la edad (mayor), el ser activo, tener hijos a cargo, ser jefe de hogar y vivir en un hogar hacinado, disminuye la probabilidad de asistencia del joven. En tanto ser mujer, la cantidad de años de educación finalizados, el clima educativo del hogar y vivir en Montevideo, aumentan la probabilidad de asistencia.

Groso [20] analizó el Programa Aulas Comunitarias (PAC) que buscaba reinsertar a estudiantes entre 12 y 16 años que no ingresaron a primer año de EMB o que no lo aprobaron. En base a los datos administrativos del programa

edición 2008, construyen modelos probit para explicar la aprobación del curso y la permanencia. Los resultados muestran que en aprobación, incide de forma positiva: el nivel educativo de la madre, la actitud de la familia en cuanto al aprendizaje de su hijo, la actitud del adolescente frente al aprendizaje, su integración en el aula, así como la integración de la familia al proceso educativo; y de forma negativa el grado de privación del hogar; en tanto encuentra como significativa para explicar el abandono del curso: la cantidad de horas de trabajo del adolescente, si el estudiante ayuda en el hogar, el número de inasistencia, el grado de privación del hogar, si es que el joven fue derivado del liceo. En contraste, la actitud del adolescente y el grado de integración del mismo, operan potenciando positivamente su probabilidad de mantenerse en el curso.

Capítulo 3

Metodología

En este capítulo se aborda el herramental teórico usado en este documento. En la sección 3.1 se enuncia una definición del concepto de aprendizaje automático y en la 3.2 se expone el problema del aprendizaje supervisado. En la sección 3.3 se describen los modelos utilizados en este documento, y en la 3.4 se presentan las principales métricas para medir la capacidad predictiva de cada modelo. Por último, en la sección 3.5 se presentan distintas técnicas para remuestrar el conjunto de aprendizaje con el fin de incrementar el poder predictivo.

3.1. Aprendizaje automático

Se atribuye la primera definición del aprendizaje automático al trabajo seminal de **Samuel** [50], donde en lugar de programar una computadora para jugar a las damas se la programó para aprender a hacerlo. Este aprender refiere a que utilice la información de juegos anteriores para que, frente a escenarios similares, la computadora tome la mejor decisión posible.

El principio del aprendizaje automático es entonces la elaboración de algoritmos o reglas que aprendan de los datos de manera de que puedan tomar decisiones una vez que se enfrentan a nueva información. En el caso particular de los problemas de clasificación de estudiantes de acuerdo a su desempeño educativo, este aprendizaje refiere a utilizar la información de una muestra de casos donde ya se conoce el resultado y se han medido variables que permitan explicar el fenómeno, por ejemplo jóvenes de años anteriores, de manera que al año siguiente al contar con nuevos estudiantes sea posible predecir su

desempeño habiendo medido solamente estas variables explicativas.

En el resto de este capítulo se formalizaran estas ideas, se revisaran los algoritmos y el herramental teórico utilizados en este documento.

3.2. El problema del aprendizaje supervisado

Sea una **conjunto de aprendizaje** $\mathcal{L}_n = \{(\mathbf{x_1}, y_1), ..., (\mathbf{x_n}, y_n)\}$, donde cada observación $(\mathbf{x}_i, y_i)$ esta compuesta por un vector de variables explicativas o independientes $\mathbf{x}_i \in \mathcal{X}$, con $y_i \in \mathcal{Y}$ el valor de la variable dependiente o de respuesta. Se asume que las observaciones de $\mathcal{L}_n$ provienen de n realizaciones independientes de una distribución multidimensional desconocida. Siguiendo a **James et al.** [30] el **objetivo** es crear un modelo que relaciona la variable de respuesta con las variables explicativas, con el objetivo de predecir correctamente la variable de respuesta en nuevas observaciones (predicción) o ayudar a entender mejor la relación entre la variable de respuesta y las variables explicativas (inferencia).

En un problema de regresión la variable de respuesta es continua $\mathcal{Y} = \mathbb{R}$, en tanto en uno de clasificación es categórica o cualitativa $\mathcal{Y} = \{1, 2, ..., L\}$. En nuestro caso particular, se trata de un problema de clasificación donde la variable de respuesta es cualitativa y puede tomar únicamente dos niveles, ó que el joven apruebe el año que este cursando ($y_i = 1$) ó que no lo haga ($y_i = 0$).

3.3. Modelos

3.3.1. Modelo de Regresión Logística (logit)

Los modelos con variable dependiente dicotómica $y_i \in \{0, 1\}$ se distribuyen Bernoulli. Los modelos paramétricos más utilizados para modelar este tipo de variables de respuesta son el modelo de probabilidad lineal (MPL), el logit y el probit.

A estos modelos se les puede dar una interpretación de variable latente [13, 63]. Sea Y_i una variable observada e Y_i^* una variable latente continua e inobservable, tal que $Y_i^* = X_i'\beta + u_i$ e $y_i = \begin{cases} 1 \; si \; Y_i^* \geq 0 \\ 0 \; si \; Y_i^* < 0 \end{cases}$. Entonces, supo-

niendo que $u_i \sim F \ \forall \ i$ y que F es simétrica en cero[1]:

$$\mathbb{P}(y_i = 1 \mid \mathbf{X} = X_i) = \mathbb{P}(X_i'\beta + u_i \geq 0) = \mathbb{P}(u_i \geq -X_i'\beta) = 1 - F(-X_i'\beta) = F(X_i'\beta)$$

Los modelos mencionados pueden ser encontrado con una adecuada elección de F. Dado que F mide una probabilidad es deseable que cumpla que $0 \leq F(z) \leq 1, \ \forall z \in \mathbb{R}$, y esto se cumple para toda función de distribución acumulada.

El MPL supone implícitamente que el efecto parcial de cada regresor es constante y no garantiza que la probabilidad estimada este en el intervalo $[0, 1]$[2]. En el MPL los parámetros del modelo se estima por mínimos cuadrados ordinarios.

Los modelos logit y probit suponen una distribución logística y normal estándar respectivamente del término de error u_i y sus parámetros se estiman por máxima verosimilitud (MV)[3]. Estos modelos levantan el supuesto implícito de que el efecto parcial de una variable dicotómica o continua es constante, a diferencia del MPL.

Para el caso de datos de corte transversal, en la práctica no existe mucha diferencia entre las predicciones realizadas por los modelos probit y logit [63][4]. En este documento se optó por el modelo **logit** que presenta la siguiente forma funcional: $\mathbb{P}[y_i = 1 \mid \mathbf{X} = X_i] = \frac{e^{X_i'\beta}}{1+e^{X_i'\beta}}$

3.3.2. Modelo logit con selección de variables (GLM_m)

La decisión de qué variables explicativas incluir en un modelo es una etapa clave del proceso de entrenamiento. La inclusión de variables irrelevantes puede traducirse en un menor poder predictivo para observaciones por fuera de la muestra de entrenamiento; adicionalmente son preferibles modelos más parsimoniosos que favorecen la interpretación [30].

Por este motivo distintas técnicas han sido desarrolladas, las cuales pueden agruparse según su objetivo: reducción de dimensiones, regularización (shrin-

[1] F es simétrica en cero si se cumple $\forall \ x: \ F(-x) = 1 - F(x)$.

[2] Para el caso del MPL se asume implícitamente que F es la función identidad.

[3] Sea $P_i = P(Y = y_i \mid \mathbf{X} = \mathbf{x}_i)$ entonces la contribución de cada individuo a la verosimilitud será: $L_i = P_i^{y_i}(1 - P_i)^{1-y_i}$. Por lo que la función de verosimilitud de una muestra aleatoria $\{y_i, \mathbf{x}_i\}, i = \{1, 2, \ldots, n\}$ será: $L(\beta) = \prod_{i=1}^{n} L_i$. Los estimadores β_{MV} surgen de maximizar la verosimilitud con respecto a β: $\beta_{MV} = ArgMax_\beta \ L(\beta)$

[4] El logit posee una distribución con colas más pesadas que el probit, y posee una interpretación de odds ratio que es muy utilizada en bioestadística.

kage) o selección de variables. En este documento se trabajará con técnicas de selección de variables que buscan identificar el subgrupo de variables que generan el mejor modelo[5].

En un modelo con K variables explicativas (sin considerar interacciones), existen 2^K combinaciones posibles, por lo cual seleccionar el modelo con mejor ajuste se vuelve un problema computacionalmente complejo al aumentar K. Una estrategia posible es utilizar un método que paso a paso evalué la inclusión o exclusión de un variable explicativa en base a una regla práctica. El método hacia adelante (*forward* en inglés), comienza con un modelos sin variables y agrega sucesivamente una a una hasta que se cumpla una regla de parada; el método hacia atrás (*backward* en inglés) por el contrario, comienza con el modelo completo (con todas las variables) y va removiendo una a una las variables hasta llegar a un modelo donde se cumpla la regla de parada. Además existen métodos que combinan un paso hacia adelante (inclusión o no de una variable) con un paso hacia atrás (exclusión o no de una variable) decidiendo en cada etapa la dirección del paso (*step* en inglés). Como criterio para incorporar (remover) una variable es posible utilizar un test de hipótesis sobre la significación del parámetro asociado o la comparación de criterios de información [42, 21].

Los métodos secuenciales de búsqueda del mejor modelo resultan computacionalmente más eficientes ya que recorren un grupo reducido de modelos candidatos. Sin embargo, corresponde notar que estos métodos no asegura la obtención de un óptimo global. El procedimiento puede detenerse antes de alcanzar el mejor modelo ya que encontró un óptimo local.

El problema de selección de variables puede considerarse como un problema de optimización, donde buscamos una combinación de los predictores que devuelva el mejor modelo [32]. Los **algoritmos genéticos** de optimización numérica desarrollados por Holland [22] pueden utilizarse para resolver este problema siguiendo los principios de la evolución biológica donde los mejores cromosomas son pasados durante generaciones marcando un proceso evolutivo en la población.

En en el contexto de los modelos de predicción, la población son los posibles modelos candidatos. Sus cromosomas un vector de dimensión K donde cada coordenada es una indicatriz que vale 1 si el modelo incluye esa variable, y 0 en otro caso. Durante generaciones, los mejores cromosomas son pasados de

[5] Ver Hastie et al. [21] para una presentación exhaustiva de la técnica.

padre a hijos y sobreviven las mejores combinaciones posibles de variables.

El proceso comienza ($t = 0$) a partir de una población inicial seleccionada al azar, es decir, se estiman distintos modelos que incluyen o no aleatoriamente variables predictoras. En una siguiente etapa, los modelos son evaluados utilizando algún criterio de información (IC por su sigla en ingles), midiendo cual combinación de variables determinan el mejor modelo. Para cada modelo, se realiza una evaluación respecto al mejor modelo de la generación mediante el estadístico $w_i = exp(IC_i - IC_{mejor})$ que sirve como un ponderador muestral para determinar la probabilidad de utilizar su información genética, es decir esa combinación de variables predictoras, para la siguiente etapa.

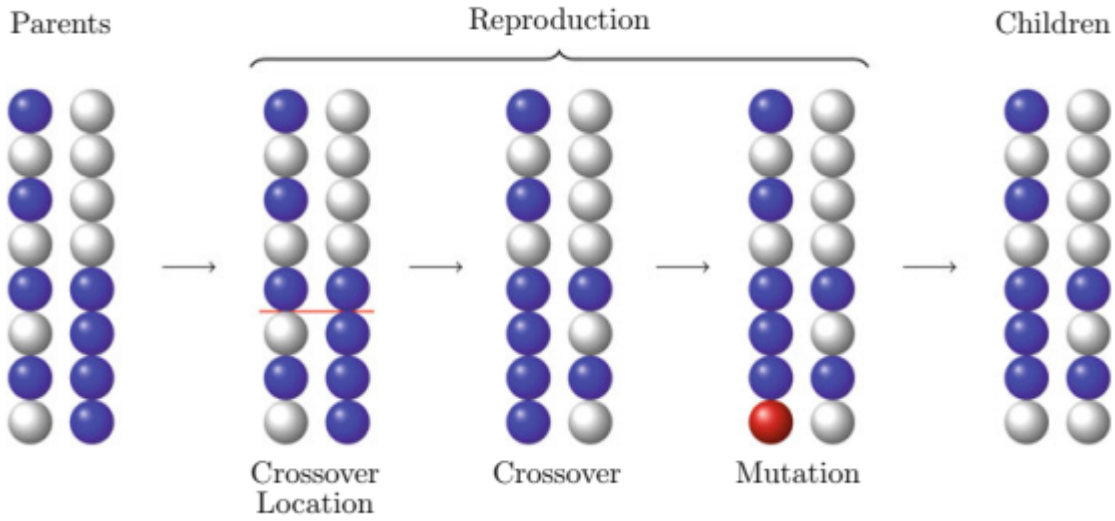
En cada generación (iteración t), el algoritmo crea G hijos o modelos nuevos durante una etapa denominada de reproducción. Para seleccionar a los padres, se realiza un muestreo con probabilidad proporcional al ajuste (w_i), es decir, los mejores modelos tienen más probabilidad de ser elegido y de trasmitir sus cromosomas o combinación de variables a futuras generaciones.

Habiendo seleccionado dos padres para una dupla de hijos cualquiera, se combina su información. Se selecciona un punto aleatorio donde dividir sus cromosoma (loci), es decir un punto donde cortar el vector de dimensión K que indica la presencia o no de las variables, la "cabeza" del cromosoma de uno de los padres es combinado con la "cola" del cromosoma del otro generando el cromosoma de los hijos. Durante esta etapa los hijos puede sufrir una mutación con probabilidad ($P_{mutación}$), es decir que la variable que indica presencia o no de la variable en el modelo pase de 0 a 1 o viceversa. Esta mutación permite explorar variantes que pueden haber sido descartadas en generaciones anteriores. Estos pasos descriptos previamente se presentan gráficamente en la Figura 3.1

Este proceso se repite en cada generación hasta que se cumpla una regla de convergencia establecida, esta puede ser por ejemplo que entre generaciones no se observen mejoras en los criterios de información. Si en una determinada iteración ($t = t'$) se cumple la regla, el algoritmo considera que se encuentra en un estado de convergencia y se selecciona el mejor modelo de la generación según el criterio de información como final. Un resumen del algoritmo es presentado en la Tabla 3.1 siguiendo a **Khun et al.** [32].

En este documento se implementó una aplicación de algoritmos genéticos para selección de las variables explicativas para modelos lineales generalizados (logit en nuestro caso), desarrollada por **Calcagno et al.** [12] donde la eva-

Figura 3.1: Reproducción algoritmo genético



Fuente: Presentado en [32]

luación del cromosoma se realiza mediante el criterio de información de Akaike (AIC) [3]. Como regla de parada los autores proponen seguir la evolución del AIC promedio y el mejor AIC en cada iteración, si en sucesivas generaciones no se observan cambios en estos indicadores el algoritmo declara convergencia.

Tabla 3.1: Algoritmo genético

- Definir tolerancia para el criterio de parada, el numero de hijos en cada generación (G) y la probabilidad de mutaciones ($P_{mutación}$).
- Seleccionar aleatoriamente G cromosomas con largo K
- Mientras no se cumpla regla de parada.
  1. **Fase de evaluación**: para cada cromosoma:
     a) Estimar un modelo y calcular AIC.
     b) calcular su ajuste relativo, $w_i = exp(IC_i - IC_{mejor})$
  2. **Reproducción**: para cada reproducción (G/2):
     a) Sortear dos progenitores con probabilidad proporcional a w_i
     b) **Cruzamiento:** dividir los cromosomas de los padres en una posición aleatoria (loci) y combinar las cadenas en los hijos.
     c) **Mutación:** Aleatoriamente cambiar valores en los hijos con probabilidad ($P_{mutación}$)

Fuente: Elaborado en base a capítulo 19 de Kuhn et al[32].

3.3.3. Arboles de Regresión y Clasificación (CART)

El algoritmo Classification and Regression Trees (CART por sus siglas en inglés) es una técnica de aprendizaje automático supervisada, desarrollada por Breiman et al (1984) [9][6]. Por construcción el método realiza particiones binarias sucesivas del espacio de variables explicativas $\mathcal{X}$. La salida final generada por el algoritmo CART es una partición del espacio de variables explicativas, que debido a que las particiones son recursivas y secuenciales, se puede representar mediante una estructura de árbol jerárquico. El nodo raíz o base, es aquel que contiene a toda la población agrupada y las ramas son los particiones a partir de las cuales se generan los nodos hijos. Aquellos nodos donde no se realizan particiones adicionales, se definen como nodos terminales o hojas.

En el presente apartado vamos a desarrollar el algoritmo CART para el problema de clasificación[7]; la **construcción** de un árbol **CART requiere** de la definición de **tres criterios**: de partición, de parada y de asignación.

El **criterio de partición** de un nodo, busca maximizar el incremento de una función de impureza. Sean $\pi_1, \ldots, \pi_L$, con $L \geq 2$, la proporción de cada clase de la variable categórica dependiente que se busca clasificar. Se define la función de impureza[8] del nodo τ como $i(\tau) = \phi(p_{1|\tau}, ..., p_{L|\tau})$, donde $p_{l|\tau}$ es una estimación de $\mathbb{P}(X \in \pi_l|\tau)$[9].

Las funciones de impureza más usadas en la práctica son el error de clasificación, el índice de Gini y el de Entropía (cuadro 3.2)[10].

Tabla 3.2: Criterios de impureza para árboles de clasificación.

Criterio de impureza	Forma funcional
Error de clasificación	$1 - max_l p_{(l\|\tau)}$
Índice de Gini	$\sum_{l \neq l'} p_{(l\|\tau)} p_{(l'\|\tau)} = 1 - \sum_{k=1}^{L} p_{l\|\tau}^2$
Entropía	$-\sum_{k=1}^{L} p_{l\|\tau} \log(p_{l\|\tau})$

Para realizar la partición de un nodo padre en dos nodos hijos (izquierdo y derecho), se busca en que variable y con que punto de corte la variación de

[6] Ver [37] para una revisión exhaustiva del algoritmo CART y otros modelos basados en arboles.

[7] Este apartado esta basado en [29].

[8] Una función de impureza $\phi : \{p \in \mathbb{R}^L; p_i \geq 0, \sum_{i=1}^{L} p_i = 1\} \to \mathbb{R}$, definida en el conjunta de L-tuplas $(p_1, ..., p_L)$ con suma unitaria, debe ser simétrica (es invariante a permutaciones de los argumentos de la función), tener mínimo en la base canónica y su único máximo en $(\frac{1}{L}, ..., \frac{1}{L})$.

[9] Como estimador se usa la proporción de la clase en el nodo $\frac{N_l(\tau)}{N(\tau)}$

[10] En este trabajo se emplea el índice de Gini como es usual en la literatura, en la práctica no hay muchas diferencias entre el índice de Gini y la Entropía.

impureza es maximiza. La variación entre la impureza del nodo τ y la impureza de sus dos nodos hijos izquierdo τ_{izq} y derecho τ_{der} al realizar la partición s es $\Delta i(\tau, s) = i(\tau) - p_{izq}(\tau)i(\tau_{izq}) - p_{der}(\tau)i(\tau_{der})$[11].

El algoritmo CART elige dentro del conjunto de todas las particiones posibles del nodo de τ (S_τ), aquella que maximiza el incremento en la función de impureza $s^*_\tau = \underset{s \in S_\tau}{\text{ArgMax}} \Delta i(\tau, s)$.

La impureza global del árbol T es $I(T) = \sum_{\tau \in \widetilde{T}} p(\tau)i(\tau)$, donde $\widetilde{T}$ es el conjunto de nodos terminales[12] u hojas de T, $p(\tau)$ es la probabilidad de pertenecer al nodo τ, e $i(\tau)$ es la impureza en τ.

El **criterio de parada** es definido por el usuario de antemano. Existen dos criterios principales: elegir un umbral de impureza a partir del cual decimos que un nodo es puro o definir una cantidad mínima de observaciones que deben tener los nodos terminales.

El **criterio de asignación** de la etiqueta en árboles de clasificación en cada nodo terminal, es el voto mayoritario[13].

Si bien a medida que crece el árbol por construcción el error de clasificación disminuye en el conjunto de entrenamiento, esto no es necesariamente cierto en la muestra de testeo. El algoritmo tendera a construir arboles grandes y por lo tanto más complejos, favoreciendo el sobreajuste[14] (overfitting en inglés).

A los efectos de evitar el sobreajuste Breiman et al. [9] plantearon un **algoritmo de poda**. La medida de costo complejidad de parámetro $\alpha \geq 0$ asociado al árbol T es: $C_\alpha(T) = R(T) + \alpha|\widetilde{T}|$, siendo $R(T)$ el error de clasificación[15] y $\widetilde{T}$ la complejidad del árbol T, medida como la cantidad de hojas del mismo.

El algoritmo de poda consta de dos pasos, partiendo del árbol maximal ($T_{maximal}$) [16]. El primero en suprimir los nodos hijos del mismo nodo padre con la misma etiqueta asignada, en este caso al suprimir estos nodos no disminuye el error de clasificación del árbol pero si disminuye la complejidad (cantidad

[11] Donde $p_{der}(\tau) = \frac{N(\tau_{der})}{N(\tau)}$, es la proporción de observaciones que pasan del nodo padre al nodo hijo por derecha, y p_{izq} se define de manera análoga.

[12] Un nodo terminal es aquel que no tiene nodos hijos.

[13] Sí el máximo es alcanzado por dos o más clases, la etiqueta se asigna por sorteo entre ellas.

[14] Se dice que un métodos sobreajusta cuando se ajusta al ruido de la muestra de entrenamiento, es decir cuando logra un ajuste en la muestra de entrenamiento que no es generalizable a la muestra de testeo.

[15] Proporción de observaciones mal clasificadas.

[16] El árbol maximal es aquel que minimiza el error de clasificación.

de nodos o hojas).

Sea T_1 el árbol resultado del paso anterior y $T_1 - T_\tau$ el subárbol que resulta de sacar la rama que yace del nodo τ al árbol T_1. Existe un valor $\alpha(\tau) = \frac{R(\tau)-R(T_\tau)}{|\widetilde{\mathrm{T}}|-1}$ mínimo, donde el costo complejidad de ambos árboles coinciden. El segundo paso de la poda es calcular $\alpha(\tau)$ para todos los nodos no terminales τ y seleccionar el o los nodos más débiles, es decir, todos los nodos internos t^* en T_1 tales que: $t^*_\alpha = \underset{\tau \in T_1}{\mathrm{ArgMin}}\ \alpha(\tau)$.

Iterando los pasos 1 y 2, se sigue con la poda un número finitos de veces desde el árbol maximal hasta llegar al nodo raíz (T_{raiz}). De esta forma se obtiene una secuencia de subárboles óptimos anidados $T_{raiz} \leq ... \leq T_2 \leq T_1 \leq T_{maximal}$ con respectivos parámetros de complejidad $\alpha_{raiz} \geq ... \geq \alpha_2 \geq \alpha_1 \geq \alpha_{maximal} = 0$. El árbol $T_{maximal}$ corresponde al valor de $\alpha = 0$, y es el árbol que minimiza el error de clasificación dentro de la muestra de entrenamiento.

El mecanismo de poda nos permite hallar un tamaño de árbol óptimo, que minimiza el error de clasificación sobre "datos frescos", es decir aquellos datos que no fueron utilizados para entrenar el modelo, mediante validación cruzada.

Las ventajas de los modelos CART son que poseen un interpretación gráfica y simple en un árbol de decisión. Respecto a las variables explicativas, incluye un mecanismo de selección de las mismas, permite detectar no linealidades e interacciones entre ellas, es invariante a sus transformaciones lineales y permite trabajar con valores faltantes [30]. Pero también posee sus desventajas, los CART tienden a no predecir bien en caso que la relación entre la variable dependiente y las explicativas es realmente lineal, y tienden a ser inestable a pequeñas variaciones del conjunto de aprendizaje (este fenómeno se acentúa en árboles sin podar). Además la selección de variables en cada partición es sesgada[17] a variables con mayor número de particiones posibles, es decir a variables con mayor número de valores diferentes y hacia variables categóricas[18] [37].

[17]Un algoritmo de partición recursiva es insesgado si bajo la hipótesis de independencia entre las Y y y las variables explicativas $X_1, X_2, .., X_K$, la probabilidad de seleccionar a la variable X_j es $\frac{1}{K}$ $\forall j = 1, .., K$, sin depender de la unidad de medida de las variables o de la cantidad de datos perdidos de la misma [24] .

[18]Sea X una variable con S valores distintos. La cantidad de particiones de una variable explicativa X es $S - 1$ si es ordinal, y $2^{S-1} - 1$ si X es categórica (no ordinal).

3.3.4. Arboles de Inferencia Condicional (CTREE)

Los métodos exhaustivos de búsqueda para la construcción de la mejor partición posible utilizando el conjunto de variables explicativas en cada nodo, por ejemplo los CART, poseen dos problemas: sobreajuste y sesgo en la selección en las variables.

El sobreajuste se debe, como se vera más adelante, a que el algoritmo no es capaz de determinar si la partición de un nodo es estadísticamente significativa. Si bien este problema puede ser contrarestado con una estrategia de poda para encontrar el tamaño óptimo del árbol, la interpretación del árbol esta afectada por el sesgo de selección en las variables [24].

Hothorn et al [24] muestran que el algoritmo CTREE (Conditional Inference Trees): (1) posee un mecanismo de selección de variables que es insesgado para la realización de la particiones; (2) no posee el problema del sobreajuste (como el CART sin podar); (3) el poder predictivo es equivalente al de un árbol podado.

El algoritmo CTREE describe la distribución condicional de una variable de respuesta $\mathbf{Y}$ categórica con K variables explicativas, por medio de particiones recursivas estructuradas en forma de árbol. La variable de respuesta $\mathbf{Y}$ pertenece al espacio muestral generado por $\mathcal{Y}$. El vector K-dimensional de variables explicativas $\mathbf{X} = (X_1, X_2, ..., X_K)$ pertenece al espacio generado por $\mathcal{X} = \mathcal{X}_1 * \cdots * \mathcal{X}_K$. Sea la distribución condicional $D(\mathbf{Y} \mid \mathbf{X})$, la respuesta de $\mathbf{Y}$ dado el estado de las K variables $X_1, X_2, \ldots, X_K$ que dependen de una función f de las mismas: $D(\mathbf{Y} \mid \mathbf{X}) = D(\mathbf{Y} \mid X_1, X_2, ..., X_K) = D(\mathbf{Y} \mid f(X_1, X_2, ..., X_K))$. Se restringe esta función f, a relaciones de partición del espacio de variables explicativas, por ejemplo, R celdas disjuntas $B_1, \ldots, B_R$ tal que $\cup_{r=1}^{R} B_r = \mathcal{X}$.

Un modelo de respuesta es construido con un conjunto de aprendizaje $\mathcal{L}_n$. A cada **nodo** del árbol se le puede asociar un vector de pesos de las observaciones $\mathbf{w} = (w_1, \ldots, w_n)$, donde cada observación tiene un valor 1 si la misma pertenece al nodo y 0 en otro caso.

En forma esquemática, en la Tabla 3.3 se presenta el algoritmo de CTREE.

El **paso 1** define la regla de detención y la selección de variables. Para cada nodo identificado con $\mathbf{w} = (w_1, \ldots, w_n)$, la **hipótesis de independencia** esta formulada en términos de K hipótesis parciales $H_0^j : D(\mathbf{Y} \mid X_j) = D(\mathbf{Y})$, $j = 1, ..., K$, siendo la hipótesis global de independencia $H_0 : \cap_{j=1}^{K} H_0^j$. Si no se

rechaza H_0 se detiene el algoritmo, en caso contrario se selecciona la variable $X_j, j = 1, .., K$ con mayor asociación con la $\mathbf{Y}$.

Pensando $\mathcal{L}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ como muestra aleatoria simple, se mide la asociación de $\mathbf{Y}$ con X_j $j = 1, .., K$ por el estadístico lineal de la forma

$$\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}) = vec(\sum_{i=1}^{n} w_i g_j(X_{ji}) h(Y_i, (Y_1, .., Y_n)^T) \in \mathbb{R}^{p_j q} \tag{3.1}$$

donde $g_j : \mathcal{X}_j \to \mathbb{R}^{p_j}$ es una transformación no aleatoria de X_j, y la función $h : \mathcal{Y} \times \mathcal{Y}^n \to \mathbb{R}^q$ se denomina función de influencia y depende de las respuestas $(Y_1, .., Y_n)$ en una permutación simétrica. Una matriz se convierte en un vector columna con el operador vec (vectorización).

La distribución de $\mathbf{T}_j(\mathcal{L}_n, \mathbf{w})$ bajo H_0^j depende de la distribución conjunta de $\mathbf{Y}$ y X_j, la cual es en general desconocida. Por lo menos bajo H_0^j es posible conocer la dependencia dejando las variables explicativas fijas y condicionando bajo todas las permutaciones posibles de la función de respuesta. Este principio conduce a procedimientos de prueba conocidos como pruebas de permutación. La esperanza condicional $\mu_j \in \mathbb{R}^{p_j q}$ y la matriz de coviaranzas $\Sigma_j \in \mathbb{R}^{p_j q \times p_j q}$ de $\mathbf{T}_j(\mathcal{L}_n, \mathbf{w})$ bajo H_0 dada todas las permutaciones de la variable de respuesta, que se notaran como $\sigma \in S(\mathcal{L}_n, \mathbf{w})$, fueron derivadas por Strasser y Weber [54].

$$\mu_j = \mathbb{E}(\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}) \mid S(\mathcal{L}_n, \mathbf{w})) = vec\big([\sum_{i=1}^{n} w_i g_j(X_{ji})]\mathbb{E}[h(Y_i, (Y_1, .., Y_n)]^T\big) \tag{3.2}$$

$$\begin{aligned} \Sigma_j = \mathbb{V}(\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}) \mid S(\mathcal{L}_n, \mathbf{w})) = \\ \frac{\mathbf{w}.}{\mathbf{w}. - 1}\mathbb{V}(h \mid S(\mathcal{L}_n, \mathbf{w})) \otimes \big(\sum_{i=1}^{n} w_i g_j(X_{ji}) \otimes w_i g_j(X_{ji})^T\big) - \\ \frac{1}{\mathbf{w}. - 1}\mathbb{V}(h \mid S(\mathcal{L}_n, \mathbf{w})) \otimes \big(\sum_{i=1}^{n} w_i g_j(X_{ji})\big) \otimes \big(\sum_{i=1}^{n} w_i g_j(X_{ji})\big)^T \end{aligned} \tag{3.3}$$

donde $\mathbf{w}. = \sum_{i=1}^{n} w_i$ denota la suma de los pesos individuales, $\otimes$ es el producto de Kronecker, y la esperanza condicional de la función de influencia es $\mathbb{E}(h \mid S(\mathcal{L}_n, \mathbf{w})) = \mathbf{w}.^{-1} \sum_{i=1}^{n} w_i h(Y_i, (Y_1, ..., Y_n)) \in \mathbb{R}^q$ con matriz $q \times q$ de coviaranzas

$$\mathbb{V}(h \mid S(\mathcal{L}_n, w)) = \mathbf{w}.^{-1} \sum_{i=1}^{n} w_i [h(Y_i, (Y_1, ..., Y_n)) - \mathbf{E}(h \mid S(\mathcal{L}_n, \mathbf{w}))][h(Y_i, (Y_1, ..., Y_n)) - \mathbf{E}(h \mid S(\mathcal{L}_n, \mathbf{w}))]^T$$

Una vez estimada la esperanza condicional y la matriz de covarianzas se puede estandarizar un estadístico lineal $\mathbf{T} \in \mathbb{R}^{pq}$ de la forma 3.1. Los test estadísticos univariado, que se notaran con la letra c, mapean un estadístico multivariado lineal a la recta real. Por ejemplo se puede elegir el máximo valor absoluto de los valores de los estadísticos **t** linearizados, utilizando la esperanza condicional μ y matriz de covarianzas Σ, $c_{max}(\mathbf{t}, \mu, \Sigma) = \underset{j=1,..,pq}{\text{Max}} \left| \frac{(t-\mu)_j}{\sqrt{(\Sigma)_{jj}}} \right|$[19].

Los test estadísticos $c(\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}), \mu_j, \Sigma_j), j = 1, .., K$ no pueden ser directamente comparados, a no ser que las variables explicativas estén medidas en la misma escala. Por tal razón para comparar estadísticos se utiliza la escala de los p-valores, porque los p-valores de la distribución condicional de los test estadístico $c(\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}), \mu_j, \Sigma_j)$ pueden ser directamente comparables entre variables explicativas medidas a diferentes escalas. En el paso del algoritmo se selecciona la variable X_{j*} con mínimo p-valor, $j = \underset{j=1,..,K}{ArgMin}\ P_j$, siendo:

$$P_j = \mathbb{P}_{H_0^j}(c(\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}), \mu_j, \Sigma_j) \geq c(\mathbf{t}_j, \mu_j, \Sigma_j) \mid S(\mathcal{L}_n, \mathbf{w}))$$

Se rechaza la hipótesis global de independencia (H_0) cuando el mínimo de los p-valores de los estadísticos obtenidos es menor que el nivel de significación corregido[20].

En caso de rechazar la hipótesis de independencia, se selecciona la variable X_j en el paso 1, y en el paso 2 el algoritmo CTREE define un criterio de partición óptimo de la variable X_j en el nodo en base a un criterio estadístico. Para todos los posibles subconjuntos A del espacio muestral $\mathcal{X}_j$ se define el estadístico lineal:

$$\mathbf{T}_j(\mathcal{L}_n, \mathbf{w}) = vec(\sum_{i=1}^{n} w_i I(X_{j^*i} \in A) h(Y_i, (Y_1, .., Y_n)^T) \in \mathbb{R}^q$$

que define un estadístico de dos muestras midiendo la discrepancia entre

$\{y_i \mid w_i > 0\ e\ X_{ji} \in A; i = 1, .., n\}$ y $\{y_i \mid w_i > 0\ e\ X_{ji} \notin A; i = 1, .., n\}$. La esperanza condicional μ_{j*}^A y coviaranza Σ_{j*}^A puede ser estimado de 3.2 y 3.3. La partición A^* se obtiene de maximizar un estadístico sobre todos los posibles

[19] Siendo $(t-\mu)_j$ el elemento en la coordenada j-esima del vector y el elemento $(\Sigma)_{jj}$ en la fila y columna j-esima de Σ.

[20] El nivel de significación es ajustado por el criterio de Bonferroni dado que se realizan comparaciones múltiples con el mismo conjunto de datos. El valor ajustado surge de dividir el nivel de significación escogido por el algoritmo CTREE dividido la cantidad de pruebas de hipótesis realizadas.

subconjuntos de A

$$A^* = \underset{A}{ArgMáx}\ c(\mathbf{t}_j^A, \mu_j^A, \Sigma_j^A) \tag{3.4}$$

En síntesis, los pasos 1 y 2 se realizan en cada nodo, para definir cual es la variable y el punto de corte adecuado. En caso de no poder rechazar la hipótesis de independencia con el nivel de significación bajo análisis, el algoritmo declara ese nodo como terminal. El árbol finaliza cuando declara todos los nodos como terminales.

Tabla 3.3: Mecanismo de partición del algoritmo CTREE

- Sea $\mathcal{L}_n$ un conjunto de entrenamiento con K variables explicativas, y $\mathbf{w} = (w_1, .., w_n)$ un vector de pesos con $w_i \in \mathbf{Z}^+$. Se define un nivel de significación α.

  1. Dado el vector de ponderadores $\mathbf{w}$ testear la hipótesis de independencia de la variable de respuesta con cada una de las variables explicativas. Detener si está hipótesis no puede ser rechazada ($pv > \frac{\alpha}{K}$). En otro caso seleccionar la variable j-esima X_{j*} con la mayor asociación con $\mathbf{Y}$.
  2. Elegir el conjunto $A^* \subset \mathcal{X}_{j*}$ para particionar $\mathcal{X}_{j*}$ en dos conjuntos disjuntos A^* y $\mathcal{X}_{j^*} \setminus A^*$. Los pesos $\mathbf{w}_{izquierda}$ y $\mathbf{w}_{derecha}$ determinan los dos subgrupos $w_{izquierda,i} = w_i I(\mathcal{X}_{j*i} \in A^*)$ y $w_{derecha,i} = w_i I(\mathcal{X}_{j*i} \notin A^*)$ para todo $i = 1, .., n$, I() denota la función indicatriz.
  3. Repetir recursivamente los pasos 1 y 2 con los pesos $\mathbf{w}_{izquierda}$ y $\mathbf{w}_{derecha}$ modificados.

Fuente: Elaboración propia en base a [24].

3.3.5. Ensamblaje

El algoritmo CART es sensible a pequeñas variaciones de la muestra de entrenamiento, es decir pequeños cambios en las observaciones pueden derivar en árboles muy distintos y por ende en predicciones distintas, por esta razón se dice que el algoritmo CART es inestable.

De acuerdo a Breiman [7] el poder predictivo de un algoritmo inestable, es mejorable mediante técnicas de **ensamblaje**. En términos generales estas técnicas pueden definirse como la combinación de modelos de base en una regla más general. A continuación se plantean algunas variantes y el método de construcción.

Bagging

El método de Bootstrap AGGRegatING (**Bagging**) fue propuesto por Breiman [7]. Sea un conjunto de aprendizaje $\mathcal{L}_n$, se sortean M muestras aleatorias

con reposición (bootstrap) de tamaño n de $\mathcal{L}_n$ y se construye un predictor h_m con cada muestra. En el caso de un problema de clasificación se ensamblan dichos predictores mediante una regla de voto mayoritario, para el caso de un problema de regresión se realiza un promedio simple de los valores predichos para la variable de respuesta.

Este método tiende a mejorar la capacidad predictiva de cualquier algoritmo con alta varianza (inestable). Sin embargo complejiza la interpretación, debido a que el clasificador final agrega un conjunto de modelos. Por ejemplo, en caso de utilizar CART como predictores base, si bien la lectura de cada CART es directa mediante un árbol, al tener un ensamblaje de arboles es compleja su interpretación, ya que no es posible realizar una representación gráfica sencilla en modo de árbol que sintetice todas las reglas de clasificación.

3.3.6. Bosques Aleatorios (RF y CRF)

El método **Random Forest** (RF) fue desarrollado por Breiman [8], es una generalización de Bagging que ensambla modelos CART siguiendo el algoritmo planteado en la Tabla 3.6. Si bien la formulación original utiliza CART como modelos de base, es posible utilizar arboles condicionales (CTREE) generando bosques aleatorios de arboles condicionales (CRF).

Tabla 3.4: Algoritmo RF

1. Sea $\mathcal{L}_n$ un conjunto de entrenamiento e $Y \in \mathcal{Y}$ una variable categórica con L clases.
2. Para $m = 1, 2, ..., M$
   - Se extrae una muestra bootstrap $\mathcal{L}^m$ del conjunto de aprendizaje $\mathcal{L}_n$.
   - Con $\mathcal{L}^m$ se construye un árbol de clasificación T^m. En cada nodo del árbol T^m se sortean aleatoriamente un subconjunto k^* de las K variables, y se encuentra la variable y el punto de corte que maximiza la impureza al igual que en CART. No se utiliza el algoritmo de poda.
   - A partir del árbol T^m se construye un clasificador que asigna la etiqueta para una nueva observación de la misma manera que en CART.
3. El conjunto de los M clasificadores de árbol T^m son denominados un bosque aleatorio (Random Forest o RF).
4. Una nueva observación es asignada a una clase mediante el voto mayoritario definido por los M arboles del bosque.

Fuente: Adaptado de Izenman [29] .

El método de RF puede mejorar el poder predictivo del algoritmo CART por dos motivos. En primer lugar al hacer variar las observaciones en cada árbol puede atenuar el problema de la inestabilidad. En segundo lugar al restringir

las variables explicativas en cada nodo de cada árbol evita que los árboles sean similares, esto se puede deber a que exista una variable explicativa fuerte que sea dominante, esta es la principal diferencia con el método bagging.

El algoritmo RF depende de dos parámetros: la cantidad de variables a sortear en cada nodo (k^*) y la cantidad de árboles. En este documento se utilizó el número por defecto de arboles en el paquete randomForest en R (500). El poder predictivo de RF puede ser sensible a la elección de la cantidad de variables a sortear en cada nodo del árbol (por defecto el paquete randomForest en R utiliza $\sqrt{K}$ para clasificación Para aplicaciones estándar con $n >> K$, Guener et al. [17] encuentran que los valores por defecto de los parámetros del algoritmo RF son adecuados[21].

Es posible testear la **importancia de las variables** en la capacidad predictiva en el modelo RF. Utilizando las observaciones no seleccionadas por la muestra boostrap ,observaciones fuera de la bolsa o Out of Bag (OOB). Para el m-ésimo árbol, la idea es comparar la diferencia en el error de clasificación entre testear el modelo con la variable X_j y una permutación aleatoria en la misma manteniendo el resto de las variables explicativas X_{-j} constantes, para luego realizar el promedio simple en cada conjunto OOB utilizados para los m arboles[22].

La intuición detrás permutar un predictor en una matriz de datos, radica en que si la variable no es relevante, no tendrá efecto sobre el poder predictivo del modelo.

Para este trabajo se utilizará la versión propuesta por Janitza et al. [31], que en vez de medir el efecto sobre el error propone medir el efecto sobre el área bajo de la curva ROC como:

$$VI_j^{(AUC)} = \frac{1}{\text{n arboles}} \sum_{t=1}^{\text{n arboles}} (AUC_{tj} - AUC_{tj*})$$

Donde AUC_{tj} indica el área bajo la curva ROC calculada paras las observaciones OOB del árbol t antes de permutar la variable j y AUC_{tj*}, el estadístico luego de realizar la permutación. De acuerdo a los autores, al utilizar el AUC

[21]Para esta aplicación se analizó la sensibilidad del RF al hacer variar la cantidad de arboles y la cantidad de variables a sortear en cada nodo. Como era esperable, con pocos variables en relación a la cantidad de obseraciones, no se observó grandes cambios al aumentar la cantidad de arboles o al aumentar la cantidad de variables.

[22]Esta medida no esta exenta de limitaciones, Izenman señala que el ranking de las variables depende de la cantidad de variables consideradas para construir los árboles, y del número de árboles en el RF [29].

se evita utilizar un punto de corte para calcular la tasa de error, lo que resulta más útil en casos de desbalance. Su interpretación es directa, cuanto mayor sea el valor del estadístico mayor es la importancia de la variable explicativa.

Para interpretar el efecto de cada variable explicativa se utilizarán los gráficos de dependencia parcial que permiten una aproximación a la dependencia entre las variable explicativa (X) y el valor predicho por el modelo [16, 21].

Sean $x = \{x_1, ..., x_p\}$ las variables explicativas, si se desea obtener el efecto de un subgrupo de variables z_s, se particiona x es el grupo z_s y su complemento $z_c = x \backslash z_s$, la función de dependencia parcial estara dada por:

$$f_s(z_s) = E_{Z_c}[\hat{f}(z_s, z_c)]$$

Lo cual puede ser estimada como $\bar{f}_z(z_s) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(z_s, z_{i,c})$, donde $z_{i,c}$ representa los valores de z_c para los individuos de la muestra de entrenamiento $(i = 1, .., n)$. Es decir, $f_s(z_s)$ representa el efecto de Z_s en la $f(x)$ luego de descontar el efecto de las otras variables z_c.

La visualización de estos efectos esta limitada a bajas dimensiones, en este trabajo nos concentraremos en los efectos parciales de cada cada variable sobre la probabilidad estimada. Greenwell [19] presenta la implementación del algoritmo utilizado en nuestro trabajo.

En un problema de clasificación de k clases, la función $f(x)$ esta dada por

$$f_k(x) = \log p_k(X) - \frac{1}{K}\sum_{l=1}^{K} \log p_l(X)$$

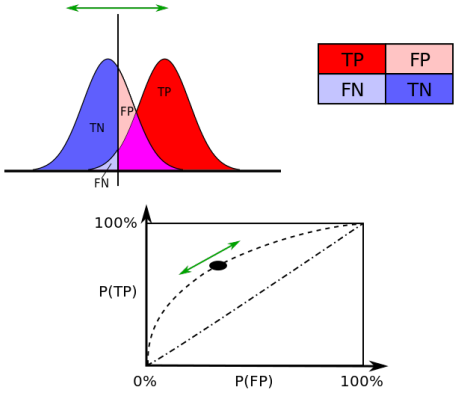
Donde $p_k(x)$ es la probabilidad predicha para la k-esima clase. Un valor más grandes de $f_k(x)$ indica que existe una probabilidad más alta de observar la clase k en el valor x.

3.4. Evaluación de la capacidad predictiva de un clasificador

Sea Y una variable dependiente binaria de interés, donde interesa separar los casos positivos $(y = 1)$ de los negativos $(y = 0)$, siendo G_1 (G_0) el grupo de individuos en la población con $y = 1$ $(y = 0)$[23]. Sea Pr la probabilidad

[23]Este apartado esta basado en [28].

Figura 3.2: Curva ROC



Fuente: [23]

a posteriori de pertenecer a G_1 estimada por cualquiera de los modelos de la sección anterior, la cuál es usada para clasificar los casos entre G_0 y G_1. Se define punto de corte c $\in [0,1]$, de manera que una observación es clasificada como G_1 si $\{Pr \geq c\}$, y 0 en caso contrario.

El objetivo es evaluar la capacidad de separar las dos subpoblaciones G_1 y G_0. Sea D una variable que indica la pertenencia a una de las dos subpoblaciones $D = \begin{cases} 1 \; si \; el \; caso \; pertenece \; a \; G_1 \\ 0 \; si \; el \; caso \; pertenece \; a \; G_0 \end{cases}$ y τ una variable que indica la clase atribuida por el clasificador $\tau = \begin{cases} 1 \; si \; \{Pr \geq c\} \\ 0 \; si \{Pr \leq c\} \end{cases}$, es posible construir la matriz de confusión de un clasificador binario como se ilustra en la Tabla 3.5.

Tabla 3.5: Resultado del test (Matriz de confusión)

		Observado		
		$D=1$	$D=0$	Total
Predicción	$\tau = 1$	TP	FP	$n_{1.}$
	$\tau = 0$	FN	TN	$n_{2.}$
	$Total$	$n_{.1}$	$n_{.2}$	n

Fuente: [61]

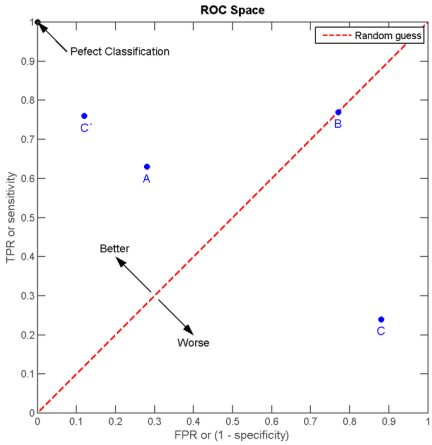
Hay dos tipos de error: (1) FN (falso negativo) G_1 es clasificado como G_0 [Error de tipo 1 (ET1)] ; (2) FP (falso positivo) G_0 es clasificado como G_1 [Error de tipo 2 (ET2)].

Sea n= $TP+FN+FP+TN$, la proporción de casos correctamente clasificados o $Acc = \frac{TP+TN}{n}$. La tasa de falsos negativos se define como: $FPR = \frac{FP}{TN+FP}$

la sensibilidad (o TPR) es la capacidad del clasificador de detectar los casos

Figura 3.3: Espacio de la curva ROC

(A): TPR = 0.63, FPR = 0.28, ACC = 0.68		
$TP = 63$	$FN = 37$	100
$FP = 28$	$TN = 37$	100
91	109	200
(B): TPR = 0.77, FPR = 0.77, ACC = 0.5		
$TP = 77$	$FN = 23$	100
$FP = 77$	$TN = 23$	100
154	46	200
(C): TPR = 0.24, FPR = 0.88, ACC = 0.18		
$TP = 24$	$FN = 76$	100
$FP = 88$	$TN = 12$	100
112	88	200
(C'): TPR = 0.76, FPR = 0.12, ACC = 0.82		
$TP = 76$	$FN = 24$	100
$FP = 12$	$TN = 88$	100
88	112	200



Nota: 4 puntos sobre el espacio de la curva ROC con matriz de confusión asociada. Fuente: [61]

positivos, la cual se calcula mediante $TPR = \frac{TP}{TP+FN}$. En tanto la especificidad (o TNR) es la capacidad del clasificador de detectar los casos negativos, $TNR = \frac{TN}{TN+FP}$

La **curva ROC** (Receiver Operating Characteristic) es el gráfico que vincula la falsa clasificación de los casos G_0 (1-TNR), y la correcta clasificación de los casos G_1 (TPR), sobre todos los puntos de corte c.

En la Figura 3.2 se observa como se construye la curva ROC. Para cada punto de corte es posible construir una nueva matriz de confusión, que define una relación entre la TPR y la FPR de un clasificador.

En la Figura 3.3 se ilustra un ejemplo sobre el plano de la curva ROC (FPR,TPR). Un clasificador perfecto logra una sensibilidad y especificidad del 100 %, esto define el punto de clasificación óptimo (0,1).

Se dice que un clasificador clasifica al azar, es decir no posee capacidad predictiva, si para todo punto de corte la tasa de verdaderos positivos es igual a la de falsos negativos (TPR=FPR), esto define la diagonal roja en la Figura 3.3. Cuanto más cerca del punto (0,1) se encuentre la curva ROC mejor es el clasificador.

El área bajo la curva ROC permite evaluar la bondad de un clasificador para todos los puntos de corte, y se notará como **AUC** (Area Under the ROC Curve).

Sin embargo hay zonas de la curva ROC donde no seriá conveniente operar[24]. Por ejemplo, en un test biológico para diagnosticar una enfermedad es

[24]Ver [47] para una presentación del tema.

importante que el test tenga alta especificidad (o baja tasa de falso negativo o error de tipo 1). Es decir la proporción de pacientes con la enfermedad diagnosticados como sanos, debe ser muy baja.

Un refinamiento del AUC es considerar el área bajo la curva en un segmento en particular de la curva ROC (**PAUC**, por partial AUC)[25]. En este documento se va analizar la capacidad de separación de las clases medidos por la curva ROC, en la región donde el error de tipo 1 (TFN) sea menor al 10 %.

3.5. Métodos de selección de la muestra de aprendizaje

En aplicaciones empíricas en modelos de clasificación, es usual que la proporción de clases presenten frecuencias no equidistribuidas, por ejemplo en la predicción de una patología poco frecuente, es esperable que la muestra tenga una baja proporción de personas con la enfermedad. En los casos extremos donde la frecuencia relativa de las clases es muy dispar, la literatura define esta situación como desbalance. Kuhn [32] establece que el desbalance tiene efectos importantes durante el proceso de entrenamiento sobre la capacidad predictiva de la clase minoritaria, ya que los modelos tenderán a predecir correctamente la clase mayoritaria logrando un mejor clasificador global.

Para afrontar este problema, la literatura plantea distintas estrategias, como por ejemplo, la selección del puntos de corte de la probabilidad a posteriori predichas por los modelos, o la asignación de costos diferenciales por clasificar erróneamente los casos. Por ejemplo, en el contexto de un modelo de riesgo crediticio, sería posible utilizar los costos financieros para las empresas, sin embargo en general no es común saber a priori cual es el costo y el beneficio de clasificación para el contexto educativo. Una posibilidad es modificar el puntos de corte de asignación de la clase, por defecto los algoritmos establecen un punto de corte en 0.5, si la probabilidad predicha por el modelo es mayor a este punto de corte, se asigna 1 como clase predicha. Sería posible establecer otro valor, una posibilidad es seleccionar ese valor en la curva ROC, por ejemplo el valor del punto de corte que devuelva una mayor proporción correcta, o mayor sensibilidad y especificidad.

[25] Al considerar el área sobre un segmento del cuadrante unitario, el valor máximo del área potencial deja de ser uno. A los efectos de simplificar la interpretación de se normaliza el AUC entre 0 y 1.

En este trabajo, se optó por utilizar técnicas de remuestreo que tienen como objetivo obtener una muestra de entrenamiento donde las clases de la variable de respuesta presenten una frecuencia equidistribuida [40], lo que puede mejorar la capacidad predictiva.

En la aplicación de este documento, la clase minoritaria es la correspondiente a los jóvenes que no aprobaron. La preferencia por utilizar una estrategia de remuestreo se debe a que no se conoce un costo diferencial por clasificar erróneamente un joven. En segundo lugar, la principal herramienta de comparación de modelos (AUC y PAUC) utilizadas en este trabajo, son independientes del punto de corte. En una siguiente etapa, una vez que se determinó el mejor modelo, se estimará un punto de corte adecuado para la probabilidad a posteriori estimada por los modelos.

Las estrategias de remuestreo pueden agruparse de acuerdo a la forma en que se selecciona la muestra de entrenamiento.

1. **Down-Sampling**: Consiste en equiparar las clases de la variable de respuesta, removiendo casos de la etiqueta mayoritaria aleatoriamente. Empíricamente se ha encontrado que el método puede ser efectivo, sin embargo, tiene el inconveniente que puede descartar información importante, ya que no incluye todas las observaciones de la clase mayoritaria.

2. **Up-Sampling**: Consiste en incrementar el tamaño del grupo de observaciones de la clase minoritaria seleccionando aleatoriamente casos a ser replicados. Esta técnica tiene como inconveniente de realizar copias exactas lo que puede generar un sobre ajuste dentro de la muestra de entrenamiento.

Entre estos dos grandes métodos algunos autores plantean variantes que permiten superar algunas deficiencias antes mencionadas. Chawla et al. [14] proponen una variante denominada Synthetic Minority Over-sampling Technique (**SMOTE**), que combina el Upsampling en clases minoritarias y el Downsampling para las mayoritarias. Para evitar la replica exacta de casos, los autores proponen seleccionar aleatoriamente casos del grupo minoritario y sus k vecinos más cercanos, generando nuevas observaciones, construidas a partir de la interpolación de las variables explicativas de los casos seleccionados y sus k vecinos [32].

Originalmente, el SMOTE fue propuesto para variables contínuas, sin embargo, puede adaptarse a los casos donde se cuenten con variable categóricas o binarias [26]. Este es el caso del presente documento, donde la información

Tabla 3.6: Algoritmo SMOTE simplificado

Parámetros: Datos minoritarios (N casos por K atributos), cantidad de casos sintéticos a generar (ns), números de vecinos (k)

1. Selección al azar de ns candidatos y sus k vecinos:
   - Seleccionar aleatoriamente ns candidatos de las N observaciones pertenecientes a la clase minoritaria.
   - Calcular los k vecinos más cercanos y guardar su información.
2. Generación de observaciones sintéticas. Para cada caso seleccionado (ns):
   - Para cada atributo K del caso seleccionado, distinguir entre las variables continuas y categóricas:
     - Para el grupo de variables continuas:
       - Seleccionar aleatoriamente entre los k vecinos el caso "donante".
       - Calcular la distancia entre la observación y el caso seleccionado (dist)
       - Seleccionar un número aleatorio entre 0 y 1 (λ)
       - Imputar en el caso sintético, coordenadas de las variables continuas como: $ns + dist * \lambda$
     - Si K es una variable binaria o categórica, imputar la moda de sus k-vecinos más cercanos.

relevada sobre los estudiantes en general toma la forma de variables categórica, por ejemplo si tuvo un evento de repetición en el pasado o no.

La Tabla 3.6 presenta una versión simplificado del pseudocódigo presentado en [14] donde en la base de datos se encuentran variables numéricas y categóricas.

El algoritmo tiene dos pasos fundamentales: (1) la selección aleatoria de observaciones y sus k vecinos que servirán como insumos (2) la extrapolación de casos. La selección de los k vecinos descansa en el cálculo de una distancia adecuada. Para ver como funcionan estas técnicas en un caso práctico se presentan ejemplos simulados en el anexo 5.3.

Para este trabajo se utilizaran tres estrategias para construir la muestra de aprendizaje, y se comparan su efecto sobre la capacidad predictiva de los modelos: estrategia Simple (sin realizar remuestreo), Down-Sampling y SMOTE-Sampling.

Capítulo 4

Estrategia empírica

En el presente capítulo, en la sección 4.1 se describen los datos empleados en este documento, en la sección 4.2 se presenta la estrategia empírica empleada para estimar los modelos y para determinar si existen diferencias significativas en su poder predictivo.

4.1. Fuentes de información

El Programa Compromiso Educativo (**PCE**[1]) es un programa interinstitucional de inclusión educativa, que tiene por objetivo general apoyar a los adolescentes y jóvenes para que permanezcan y puedan potenciar sus trayectorias en el sistema educativo público, completando la educación media superior (EMS).

Para este trabajo se utilizó la encuesta del PCE 2012, realizada a centros y niveles educativos donde el programa se encontraba vigente en el 2012[2]. Esta encuesta fue utilizada previamente para evaluar el impacto del programa [2, 5]. Si bien la encuesta es representativa de centros donde se encontraba vigente el PCE en el 2012, no necesariamente lo es del sistema educativo en su conjunto. Potencialmente esto puede generar un sesgo de selección de la

[1] Ver Ambrosi et al. [4] para una descripción más detallada del programa. Ver Mides [39] para un relevamiento de programas sociales en Uruguay.

[2] En esta encuesta se empleo la técnica de conglomerados, donde primero se seleccionaron centros para luego censar los grados para los cuales el programa estaba vigente. Para centros que empezaron con el programa en el 2011 se encuestaron a primeros y segundo de EMS; en tanto para centros que ingresaron en el 2012 al programa, se analizan solo cuartos. La encuesta se realizó en formato panel, se realizan una encuesta a los mismos centros y estudiantes, en julio y en noviembre. Los centros encuestados son 41 y los individuos relevados son 3330 en julio y 2772 en noviembre.

muestra, ya que se desconoce si los resultados obtenidos son extrapolables al universo bajo consideración. A pesar de esta limitación, esta fuente de datos permite incorporar variables no relevadas en otras encuestas, como el vínculo del alumno con el centro educativo y sus expectativas sobre los estudios.

La variable de respuesta a predecir es la promoción del grado y se mide a partir de los registros administrativos de la educación media mediante el fallo final. Dado que para el computo de la variable dependiente se utilizan registros administrativo, no se posee un problema de desgaste de la muestra (attrition[3] en inglés) como es usual al trabajar con encuestas.

Es importante señalar que los registros administrativos de fallos educativos para este estudio, solamente poseen los registros para cuarto año de EMS general (CES)[4] y cuarto y quinto de EMS técnica (CETP).

En este documento a los efectos de simplificar y seleccionar una muestra más homogénea, se analizó a los estudiantes de cuarto año de EM. Adicionalmente, se excluyen del alcance de este documento a aquellos estudiantes becados por el PCE en el 2011 y 2012, ya que potencialmente podrían presentar un comportamiento diferencial [1, 2]. Además se excluyen del estudio las observaciones de la muestra con datos faltantes en la primera encuesta[5]. Con este procedimiento, se generó una base de datos con 1529 observaciones, donde se tomó el dato de aprobación[6] o no del estudiante de los registros administrativo de educación media. La base de datos consta de 18 variables explicativas que fueron construidas a partir del formulario de la encuesta del PCE, de la cuales solamente ICC es continua y las restantes son binarias.

4.1.1. Estudio descriptivo

Se define como variable de interés la aprobación del año lectivo, es decir, sí es que logró pasar al siguiente nivel educativo. Para entrenar los modelos, se plantea el evento de aprobación o no como una variable dicotómica, que toma

[3]Se dice que hay un problema de attrition cuando la probabilidad de permanecer en la muestra es no aleatoria.

[4]En el año que se construyeron los datos (2013), el CES no computaba el fallo final del curso para estudiantes de segundo y tercero de EMS.

[5]A los estudiantes seleccionados se le aplicaron dos formularios uno cerca de junio y otro en noviembre, los estudiantes sin el formulario llenado en junio faltaron el día que se realizó la encuesta en el centro educativo.

[6]Un estudiante que cursa un año puede promover o no el año lectivo. Un estudiante promueve el año lectivo, sí al último período de examen, se queda con la cantidad mínima de materias que lo habilita a cursar el año siguiente.

valores 1 si aprueba y 0 sino. En este caso la probabilidad predicha por los modelos puede considerarse como la probabilidad de pasar al siguiente grado, siendo los jóvenes con mayor riesgo los estudiantes con una baja probabilidad de promoción.

En la muestra utilizada, solo el 24.4 % de los jóvenes **no fueron promovidos**, presentando esta subpoblación algunas **características particulares**.

En la Tabla 4.1, se computa el promedio de cada variable explicativa según el valor de la variable de respuesta (promueve o no promueve), y el p-valor asociado al test de diferencias en proporciones[7] y un test t de diferencia de medias para el caso del Índice de Carencias Criticas (ICC).

Estos dos grupos presentan varias diferencias significativas. El grupo de alumnos que no aprobaron se caracterizan por tener una mayor proporción de jóvenes con algún evento de abandono (11.8pp) y repetición (26.7pp) en su trayectoria previa, así como una mayor proporción de jóvenes que deben exámenes (32.3pp). Adicionalmente los que no aprobaron presentan una mayor proporción de jóvenes que se encontraban trabajando al momento de la encuesta (4.8pp).

En cuanto a los vínculos, los jóvenes que no aprobaron presentan una menor proporción que respondió contar con mucho apoyo familiar (-9.2pp), así como una menor proporción que respondió tener vínculos muy buenos con compañeros (8.7pp) y docentes (5pp). Así mismo, cuentan con una mayor proporción de jóvenes que respondieron que muchos de sus amigos dejaron de estudiar (17.8pp).

En cuanto a las expectativas sobre la educación terciaria, los jóvenes que no aprobaron, presentan una menor proporción que se imagina estudiando en educación terciaria (-30.6pp), lo mismo sucede con las expectativas familiares (-29.6pp).

Finalmente, los jóvenes que no aprobaron presenta una mayor proporción de hombres (9.8pp), de Montevideo (7.1pp) y una mayor vulnerabilidad económica evidenciada por un nivel más alto del ICC (0.041).

[7] Con correcciones de Yates para el caso de las variable dicotómicas.

Tabla 4.1: Estadísticas descriptivas

	Aprobó		Dif. en medias	p-valor
	No	Si	$\mu_{y=1} - \mu_{y=0}$	
Algún abandono previo en EMS	21.4 %	9.6 %	-11.8 pp	0.00
Repitió en EM	46.6 %	20.0 %	-26.7 pp	0.00
Repitió en la escuela	16.4 %	8.0 %	-8.3 pp	0.00
Debe Exámenes	51.2 %	18.9 %	-32.3 pp	0.00
Trabaja Actualmente	13.1 %	8.3 %	-4.8 pp	0.01
Ayuda Padres en el Trabajo	15.5 %	14.4 %	-1.1 pp	0.66
Cuida Familia	29.0 %	26.3 %	-2.7 pp	0.35
Ayuda en las tareas del Hogar	79.1 %	81.3 %	2.2 pp	0.38
Apoyo familiar	66.8 %	76.0 %	9.2 pp	0.00
Vínculo muy bueno con compañeros	40.8 %	49.5 %	8.7 pp	0.00
Vínculo muy bueno docentes	12.6 %	17.6 %	5.0 pp	0.03
SECLI (secundaria)	59.5 %	62.5 %	2.9 pp	0.34
Montevideo	36.2 %	29.1 %	-7.1 pp	0.01
Hombre	54.7 %	44.9 %	-9.8 pp	0.00
Se imagina en educación terciaria	19.8 %	50.4 %	30.6 pp	0.00
Su familia lo imagina en educación terciaria	21.4 %	51.0 %	29.6 pp	0.00
Muchos amigos dejaron de estudiar	44.2 %	26.5 %	-17.8 pp	0.00
Índice de Carencias Criticas (ICC)	0.167	0.126	-0.041	0.00

Nota: pp (puntos porcentuales), p-valor de estadísticos t de diferencia de medias en variable aprobó. N=1529

4.2. Estrategia de modelización

Para poder comparar los modelos se implementó una estrategia basada en la **validación cruzada**. En cada iteración se realiza la partición en muestra de entrenamiento (70 %) y testeo (30 %) de la base original, siguiendo un **muestreo estratificado** según la variable de respuesta, siendo los estratos los grupos de jóvenes que aprueban y los que no. Esto asegura que se mantenga la proporción de clases que se observan en la muestra completa.

Sobre la **muestra de entrenamiento** obtenida, se aplicaron técnicas de remuestreo, generando tres submuestras sobre las cuales se estimará los modelos; una muestra "Simple" que consiste en la muestra de entrenamiento original, una muestra "Down" resultante de aplicar down sampling sobre la original y una muestra "Smote" que se obtuvo a partir de la aplicación de la variante sintética.

Sobre cada muestra se entrenan los modelos descritos en la sección 3.3, y se predice la probabilidad a posteriori de promoción de las observaciones perteneciente a la muestra de testeo. Sobre estos casos se evalúan la capacidad predictiva de los modelos utilizando el AUC y PAUC descriptos en la sección 3.4. En la tabla 8 del Apéndice se señalan los paquetes de R [46] utilizados.

Los pasos anteriores tienen como resultado para cada iteración, con 100

réplicas, el computo de estos estadísticos para cada combinación de estrategias de muestreo y modelos. En la tabla 4.2 se resumen los pasos seguidos.

Tabla 4.2: Estrategia de modelización inspirada en validación cruzada

- Parámetros de entrada: Número de repeticiones (rep)
  1. Para de 1 hasta rep
     - I Se particiona la muestra en un conjunto de entrenamiento (70 %) y otro de testeo (30 %) siguiendo un muestreo estratificado en la variable de respuesta, y se generan tres muestras de entrenamiento:
       1) Simple
       2) Down
       3) Smote
     - II Se estiman los modelos con las muestras de entrenamiento.
     - III Se predice sobre la base de testeo.
     - IV Se calculan métricas de resultado para cada combinación de modelo y de remuestreo.
  2. Salida: Base de datos con las métricas para cada combinación de modelo y estrategia de muestreo.

El resultado de este proceso es una base de datos, donde para iteración, se obtiene una estimación de AUC y PAUC para cada combinación de muestreo (Up, down, SMOTE) y modelo (GLM, GLM_m, CART, CTREE, RF, CRF), lo que permite comparar su capacidad predictiva. Luego se calculan promedios y desvío estándar de ambos estadísticos.

Corresponde notar que la utilización de esta estrategia inspirada en la validación cruzada repetida, asegura obtener una medida honesta del desempeño predictivo que no esté condicionada a una muestra de entrenamiento y testeo en particular.

Como estadísticos principales se computaron el área debajo de la curva ROC (AUC) y el área parcial (PAUC). Para la estimación del PAUC integramos el área bajo la curva ROC con errores de tipo I menor al 10 %, en nuestro ejemplo este error es predecir que el estudiante va a aprobar cuando no aprueba. Esto sería el peor error porque no permite realizar un acompañamiento, o un tratamiento especial a un estudiante con alto riesgo de no promoción.

4.2.1. Comparación de modelos

Para identificar la mejor combinación de modelo y estrategia de muestreo, es necesario hacer inferencia sobre la capacidad predictiva de los modelos. Para

esto es posible utilizar herramientas de análisis de experimentos [8].

Los pasos establecidos en la sección anterior generan una serie de observaciones que pueden considerarse realizaciones de un experimento de aprendizaje automático, donde se desea concluir si existen factores como el tipo de modelo y estrategias de remuestro que impactan sobre el estadístico de resultado (AUC o PAUC). Bajo el enfoque del diseño de experimentos se trata de un diseño cruzado, modelos por estrategias de muestreo, donde se puede aplicar el siguiente modelo:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}$$

- y_{ijk} la variable de resultado para la k-esima replica asociada a la familia de modelos i y estrategia de muestreo j
- μ es la media global
- α_i efecto principal del factor modelos
- β_j efecto principal del factor muestreo
- $(\alpha\beta)_{ij}$ efecto interacción
- e_{ijk} restantes causas de variabilidad del experimento.

Donde se supone que los errores e_{ijk} satisfacen las hipótesis de media cero, ser independientes y presentan igual varianza.

Realizando la descomposición de la varianza para este modelo se obtiene[9]:

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{...})^2}_{SC_{total}} = \underbrace{bn\sum_{i=1}^{a}\hat{\alpha}_i^2}_{SC_\alpha} + \underbrace{an\sum_{i=1}^{a}\hat{\beta}_j^2}_{SC_\beta} + \underbrace{n\sum_{i=1}^{a}\sum_{j=1}^{b}\widehat{\alpha\beta}_{ij}^2}_{SC_{\alpha\beta}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{ij.})^2}_{SC_\epsilon}$$

Donde a es igual a 6 la cantidad de tipos de modelos estimados (por ejemplo: CART, GLM, etc), b es igual a 3 e indica la cantidad de estrategias de muestreo ensayadas y n[10] el número de iteraciones o pruebas realizadas. El

[8] Ver Montgomerry [41] o Tahane [55], para una introducción al análisis de experimentos y comparaciones múltiples.

[9] Se notara con $\bar{y}_{ij.} = \frac{\sum_{k=1}^{n} y_{ijk}}{n}$ e $\bar{y}_{...} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}}{abn}$.

[10] Observar que en este caso, con la letra n se nota la cantidad de réplicas del experimento realizadas y no con las observaciones de la base de datos.

primer elemento (SC_{total}) presenta la variabilidad total que se calcula como la suma de cuadrados entre las k observaciones pertenecientes a las 100 iteraciones, la estrategia de muestreo j y l familia de modelos i. Es posible contrastar la significación de las distintas fuentes de variabilidad, es decir si existe variabilidad explicada por los modelo (SC_{α}), muestreo (SC_{β}) o la combinación de ambos ($SC_{\alpha\beta}$) mediante una prueba F.

En nuestro caso, el supuesto de errores independientes se estaría violando ya que en cada iteración se utiliza la misma muestra de entrenamiento y testeo. Por este motivo se utiliza una variante del modelo anterior donde se introduce un efecto aleatorio por iteración. En otras palabras, consideramos a cada iteración del experimento como un individuo sobre el cual se realiza un experimento midiendo distintas combinaciones de factores (modelo y estrategia de muestreo).

Habiendo contrastado la significación de los factores, resulta relevante testear si entre los distintos niveles de los mismos, las diferencias resultan estadísticamente significativas. Para eso se utilizara un test de diferencias de medias para comparaciones multiples corregido por Tukey [41].

Capítulo 5

Resultados

En este capítulo, en la sección 5.1 se mide y se compara el poder predictivo de los modelos utilizados, analizando si existen diferencias estadísticamente significativas en su desempeño según tipo de modelo y estrategia de muestreo. Habiendo seleccionado el mejor modelo, en la sección 5.2 se interpreta la importancia de las variables y el efecto parcial de las mismas. Finalmente, la sección 5.1 se plantea un método para seleccionar un punto de corte para el problema de clasificación.

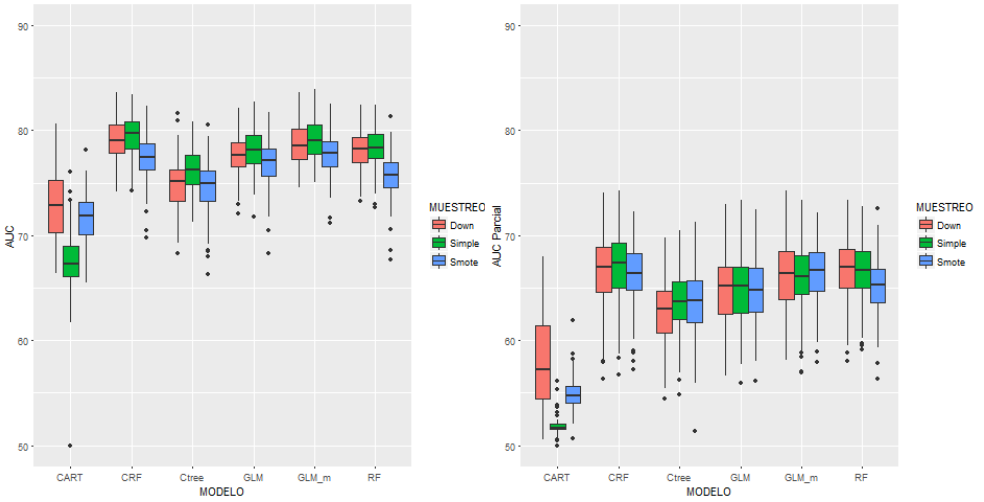
5.1. Estimación de la capacidad predictiva

Aplicando la estrategia de validación cruzada de la Tabla 4.2, se computan los estadísticos descriptivos de las métricas de desempeño predictivo (AUC y PAUC) para cada estrategia de muestreo y modelo, en la Tabla 5.1. La Figura 5.1 presenta los diagramas de caja para cada combinación de modelo y estrategia re escalando la variable de respuesta entre 0 y 100 para el PAUC.

En promedio, el CRF simple presentó los mejores resultados en PAUC (66.85) seguidos muy de cerca de RF-Down (66.62), RF-simple (66.56) y los restantes CRF, down (66.56) y smote (66.28). En cuanto al AUC, el CRF simple continua siendo el que presenta el mejor resultado con AUC de 79.54 seguido del GLM_m simple (79.12) y el CRF con estrategia de down sampling (79.04).

En síntesis, los métodos basados en agregaciones de árboles ya sea condicionados o no, parecen los mejores resultados en cuanto al AUC parcial. Al observar el AUC se observa que el CRF simple presenta los mejores resultados.

Figura 5.1: Diagramas de caja del AUC y PAUC



(a) AUC **(b)** PAUC

Nota: AUC (área bajo la curva ROC), PAUC (área parcial bajo la curva ROC) con error de tipo 1 menor a 0.1.. Algoritmos: RF (Random Forest) , CART , CTREE, CRF (RF de CTREE), GLM (logit), GLM_m (logit con selección genética de variables explicativas). Estrategias de selección de la muestra de entre: simple, down and smote.

Para verificar si existen diferencias significativas en el AUC ó PAUC atribuibles a los modelos o estrategias de muestreo, se utilizan herramientas de análisis de experimentos. Las tablas Anova para AUC y PAUC (Tabla 5.2) muestran que los factores así como su interacción son significativos para explicar las diferencias observadas en los indicadores de resultado.

El cuadro 5.3 presenta las medias corregidas y el resultado del test de diferencias ajustado por Tukey indicando la pertenencia a un grupo mediante letras. Se observa que CRF simple es el mejor modelo en cuanto a AUC aunque no resulta significativamente distinta de GLM_m y los restantes modelos, exceptuando los modelos CTREE con muestreo down y SMOTE y todos los CART los cuales presentan los peores resultados.

En cuanto al PAUC, el cuadro 5.4 muestra que CRF, RF (down y simple) y GLM_m down (down y simple) conforman un único grupo aunque no es significativamente distinto de GLM_m y las restante estrategias exceptuando los modelos basados en CART los cuales presentan los peores resultados.

En síntesis, si bien las diferencias no resultan significativas en todos los casos, los mejores resultados medidos como el área debajo de la curva ROC (AUC) y su versión parcial (PAUC), indican que el CRF sin estrategia de muestreo puede considerarse como la mejor estrategia de modelado.

Tabla 5.1: Estadísticos AUC y PAUC.

Modelo	Estr	AUC		AUC parcial	
		Promedio	Desv est.	Promedio	Desv est.
RF	Simple	78.35	1.89	66.56	2.94
RF	Smote	75.70	2.23	65.10	2.79
RF	down	78.10	1.98	66.62	3.18
CART	Simple	67.21	3.49	51.83	0.82
CART	Smote	71.56	2.37	54.89	1.59
CART	down	72.74	3.00	58.11	4.26
Ctree	Simple	76.28	1.96	63.46	2.86
Ctree	Smote	74.64	2.55	63.87	3.42
Ctree	down	74.79	2.36	62.76	3.19
CRF	Simple	79.54	2.03	66.85	3.62
CRF	Smote	77.35	2.15	66.28	3.07
CRF	down	79.04	1.99	66.56	3.67
GLM	Simple	78.02	2.04	64.87	3.43
GLM	Smote	76.83	2.19	64.75	3.00
GLM	down	77.70	2.01	64.79	3.43
GLM_m	Simple	79.12	2.02	65.98	3.49
GLM_m	Smote	77.76	2.03	66.25	2.94
GLM_m	down	78.76	2.01	66.17	3.25

Nota: RF (Random Forest) , CART , CTREE, CRF (RF de CTREE), GLM (logit), GLM_m (logit con selección genética de variables explicativas).

Tabla 5.2: Prueba Anova

	AUC		PAUC	
	Estadistico F	p valor	Estadístico F	p valor
Intercepto	185189.6	<.0001	68193.91	<.0001
Estrategia	105.77	<.0001	32.42	<.0001
Modelo	1337.3	<.0001	1408.86	<.0001
Estrategia:Modelo	105.73	<.0001	46.53	<.0001

Tabla 5.3: Prueba de diferencias en medias: AUC

Estrategia	Modelo	Promedio	IC inf	IC sup	Grupo
Simple	CART	67.21	66.51	67.91	a
Smote	CART	71.56	70.86	72.25	b
down	CART	72.74	72.05	73.44	c
Smote	Ctree	74.64	73.95	75.34	d
down	Ctree	74.79	74.09	75.48	d
Smote	RF	75.70	75.00	76.39	e
Simple	Ctree	76.28	75.58	76.98	ef
Smote	GLM	76.83	76.14	77.53	fg
Smote	CRF	77.35	76.66	78.05	gh
down	GLM	77.70	77.00	78.40	hi
Smote	GLM_m	77.76	77.07	78.46	hi
Simple	GLM	78.02	77.33	78.72	hi
down	RF	78.10	77.40	78.80	ij
Simple	RF	78.35	77.66	79.05	ijk
down	GLM_m	78.76	78.06	79.45	jkl
down	CRF	79.04	78.35	79.74	klm
Simple	GLM_m	79.12	78.43	79.82	lm
Simple	CRF	79.54	78.84	80.23	m

Tabla 5.4: Prueba de diferencias en medias: PAUC

Estrategia	Modelo	Promedio	IC inf	IC sup	Grupo
Simple	CART	51.83	50.87	52.80	a
Smote	CART	54.89	53.93	55.85	b
down	CART	58.12	57.15	59.08	c
down	Ctree	62.76	61.80	63.72	d
Simple	Ctree	63.46	62.50	64.42	de
Smote	Ctree	63.87	62.91	64.83	ef
Smote	GLM	64.75	63.78	65.71	fg
down	GLM	64.79	63.83	65.75	fg
Simple	GLM	64.87	63.91	65.83	fg
Smote	RF	65.10	64.14	66.06	gh
Simple	GLM_m	65.98	65.02	66.94	hi
down	GLM_m	66.17	65.21	67.13	i
Smote	GLM_m	66.25	65.29	67.21	i
Smote	CRF	66.28	65.32	67.24	i
down	CRF	66.56	65.60	67.52	i
Simple	RF	66.56	65.60	67.52	i
down	RF	66.62	65.66	67.58	i
Simple	CRF	66.85	65.89	67.81	i

5.2. Interpretación del modelo CRF

Utilizando la muestra completa, re estimamos el modelo CRF calculando la importancia de las variables y los gráficos de dependencia parciales.

La Figura 5.2 muestra que deber exámenes, el hecho de imaginarse estudiando educación terciaria, si repitió en Educación media y si su familia lo imagina estudiando educación terciaria son las cuatro factores que resultan más importantes en cuanto a su impacto sobre el AUC.

La Figura 5.3 muestra los gráficos de dependencia parcial utilizando una escala de probabilidad para la clase aprobado. Valores más altos indican que es más probable observar jóvenes que aprobaron para esos valores de la variable explicativa.

Es más probable observar casos de aprobación para jóvenes que no deben exámenes y no experimentaron un evento de repetición en educación media. En cuanto a las expectativas, es más probable observar aprobación cuando los estudiantes y su familia lo imaginan estudiando en educación terciaria.

Figura 5.2: Importancia de variables: Modelo CRF

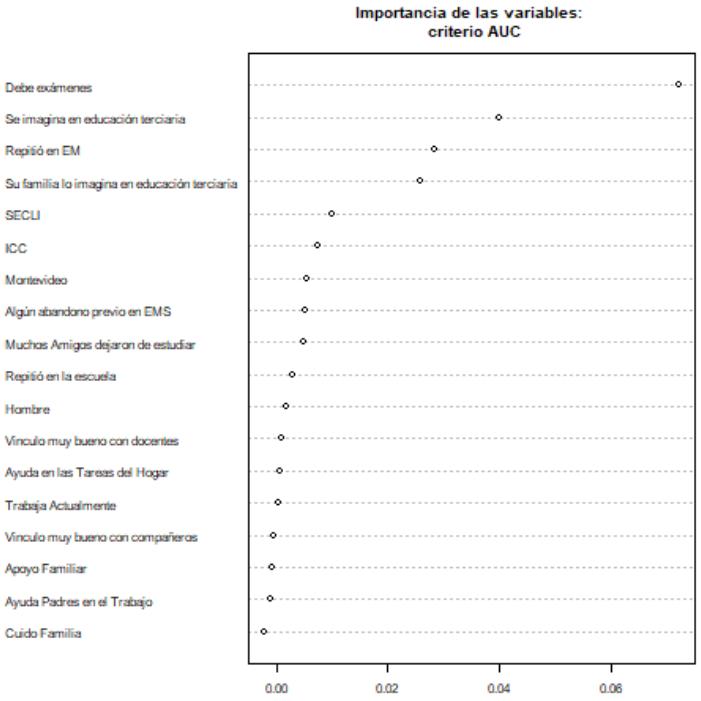
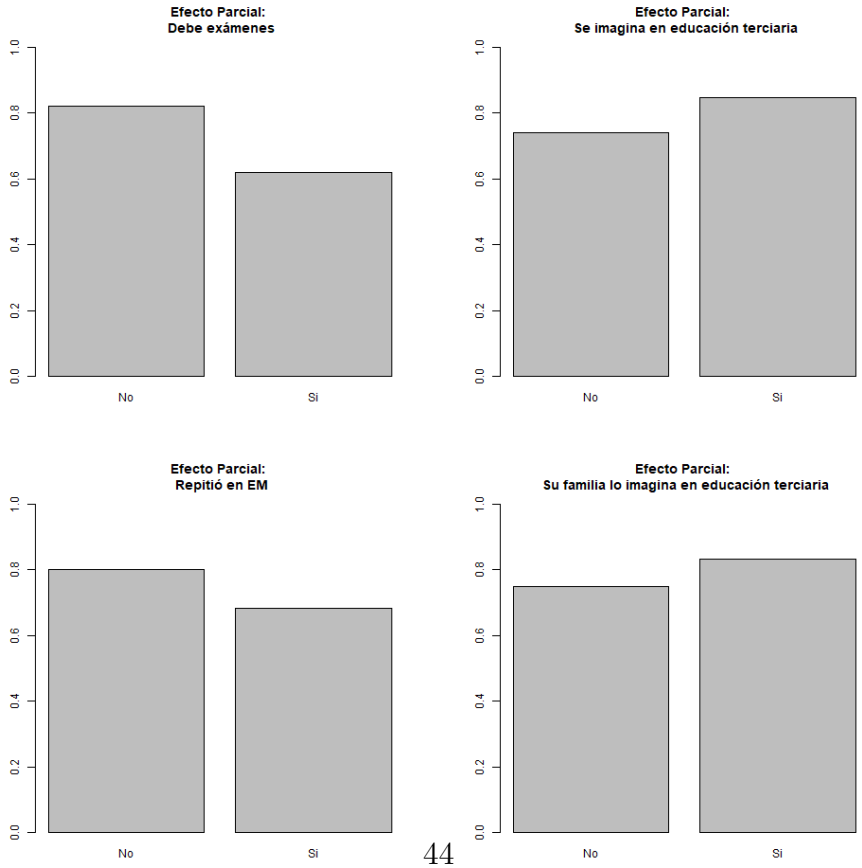
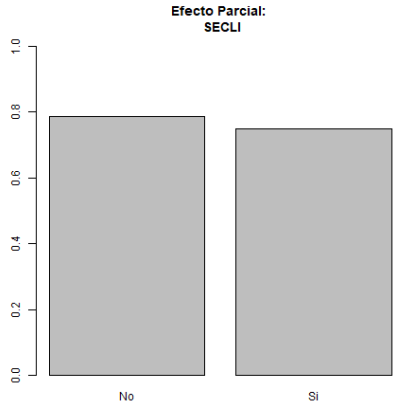
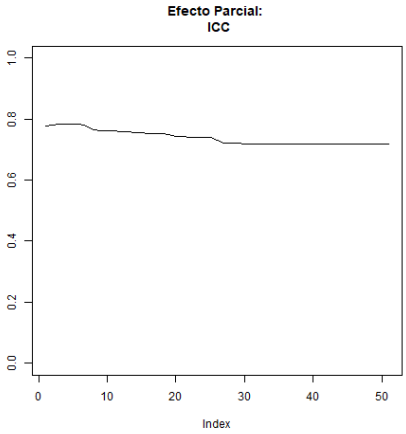
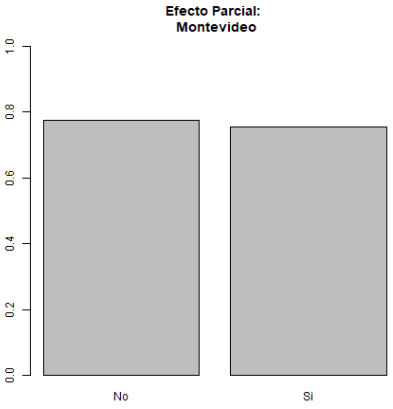


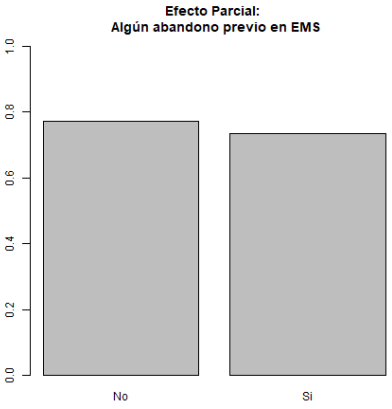
Figura 5.3: Gráficos de efectos parciales (CRF).

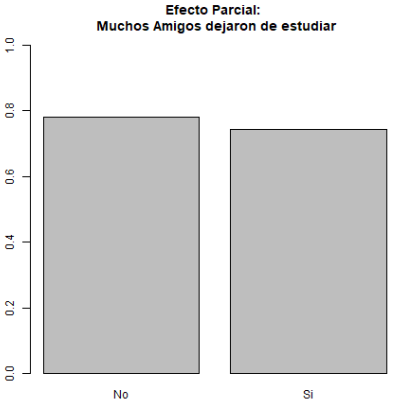


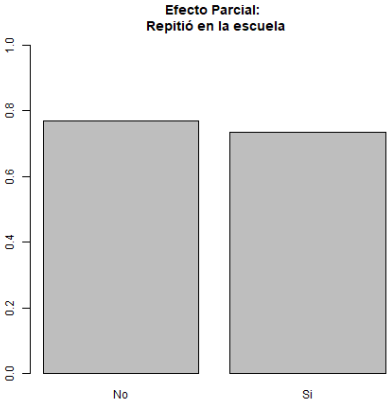

Efecto Parcial:
SECLI
0.0
0.2
0.4
0.6
0.8
1.0
No
Si

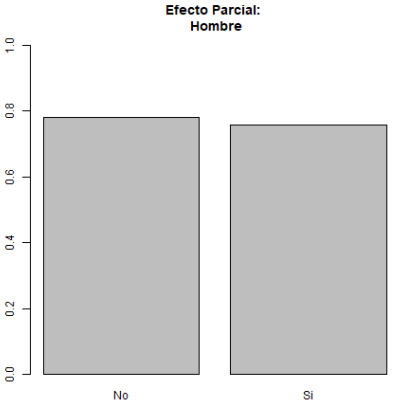

Efecto Parcial:
ICC
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
Index

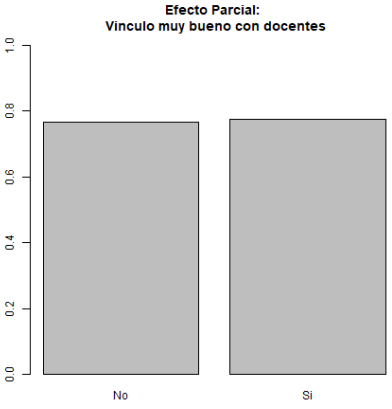

Efecto Parcial:
Montevideo
0.0
0.2
0.4
0.6
0.8
1.0
No
Si

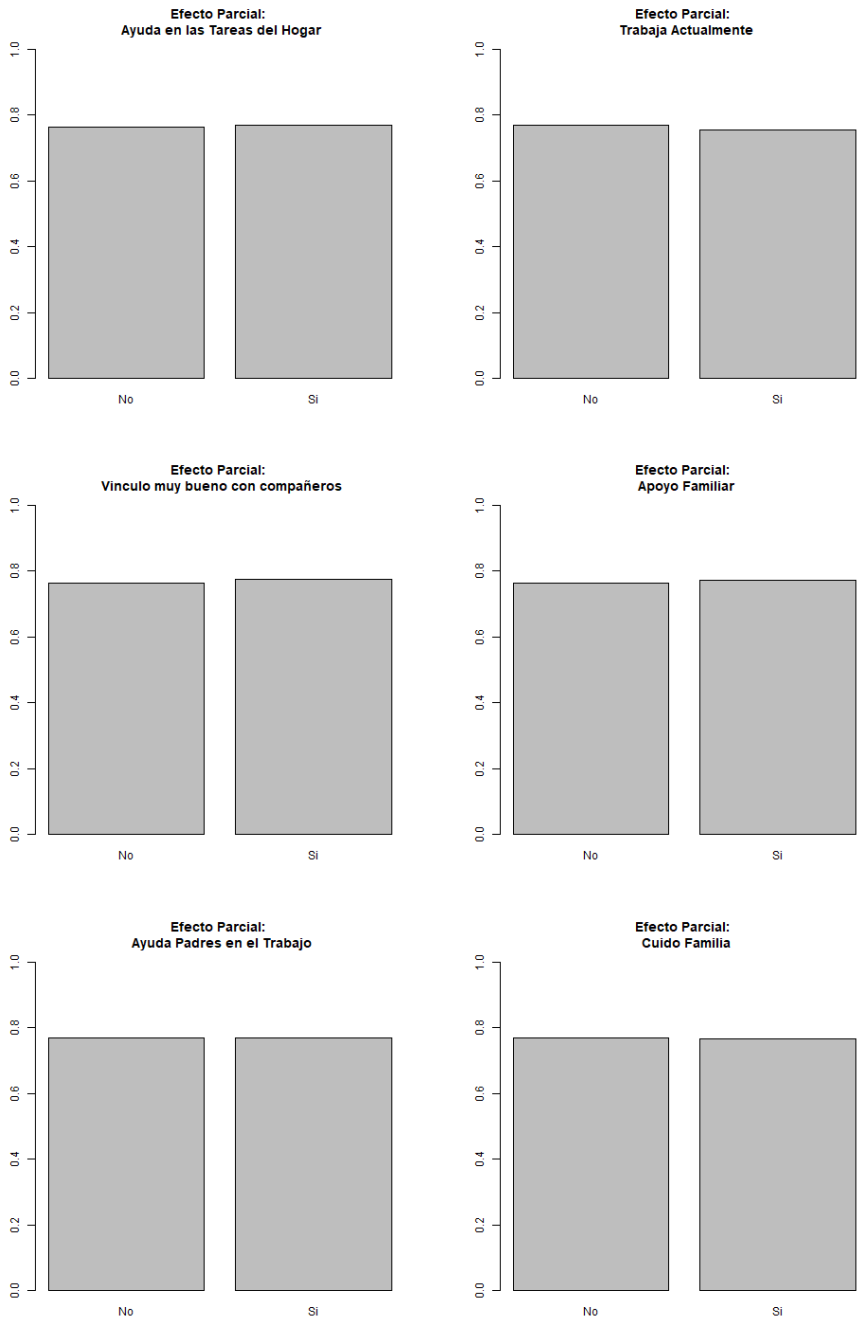

Efecto Parcial:
Algún abandono previo en EMS
0.0
0.2
0.4
0.6
0.8
1.0
No
Si


Efecto Parcial:
Muchos Amigos dejaron de estudiar
0.0
0.2
0.4
0.6
0.8
1.0
No
Si


Efecto Parcial:
Repitió en la escuela
0.0
0.2
0.4
0.6
0.8
1.0
No
Si


Efecto Parcial:
Hombre
0.0
0.2
0.4
0.6
0.8
1.0
No
Si


Efecto Parcial:
Vinculo muy bueno con docentes
0.0
0.2
0.4
0.6
0.8
1.0
No
Si



Nota: Variables explicativas ordenadas de acuerdo a su importancia, criterio AUC

5.3. Selección de un punto de corte

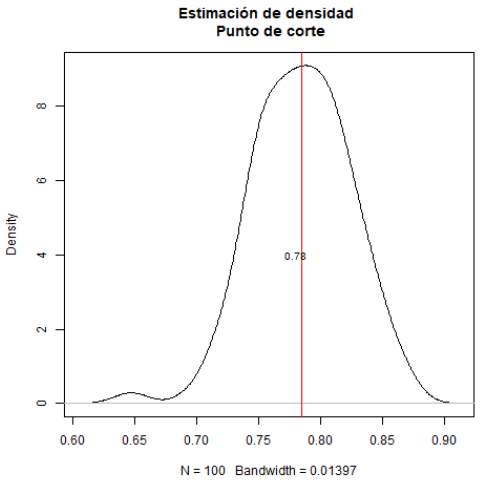
En la sección anterior se identificó al CRF simple como el mejor modelo de acuerdo al PAUC y AUC. Estas medidas son independientes del punto de corte seleccionado. Sin embargo, en la práctica, el objetivo es poder clasificar un

nuevo caso a partir de la probabilidad a posteriori predicha por el modelo. Para esto es necesario sugerir un punto de corte adecuado para esta probabilidad.

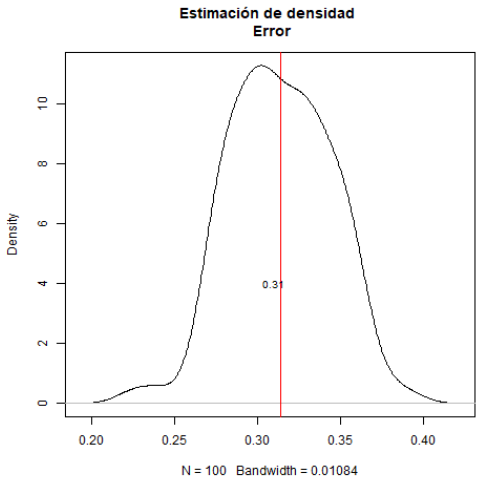
Buscando elegir el punto de corte, se realizaron 100 iteraciones donde se dividió la muestra en entrenamiento y testeo. Se estimó el algoritmo CRF con la muestra entrenamiento original y se seleccionó como valor de corte el punto sobre la curva ROC con mayor valor de sensibilidad y especificidad en la muestra de testeo. Los resultados de las iteraciones son presentados mediante la estimación kernel de su densidad y promedio (Figura 5.4).

Los resultados sugieren la utilización del valor 0.78 como punto de corte, adicionalmente las iteraciones mostraron que al utilizar como criterio de punto de corte sensibilidad y especificidad, se encontró una tasa de error de aproximadamente un 30 % así como un valor de sensibilidad de 0.63 y especificidad de 0.86.

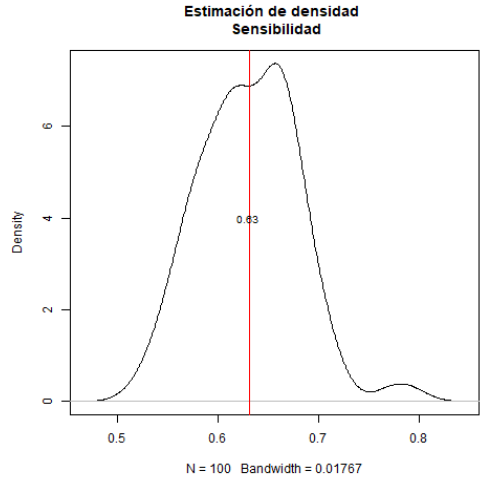
Figura 5.4: Resultados de la estimación del punto de corte



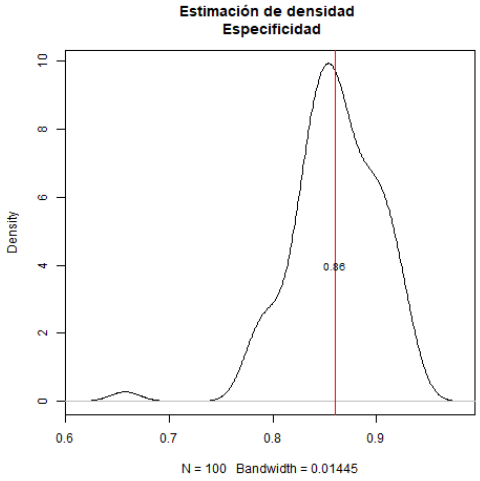
(a) Punto de corte



(b) Proporción correcta



(c) Sensibilidad



(d) Especificidad

Conclusiones

Muchos problemas prácticos de política pública pueden concebirse como problemas predictivos, para lo cual es importante ensayar con distintas estrategias de modelización para medir su poder predictivo. Si bien los distintos modelos nos permiten obtener distintas miradas de los datos, muchas veces en el trabajo aplicado resulta necesario seleccionar un modelo final adecuado. Para dicho fin ordenar los modelos según su capacidad predictiva, resulta ser un criterio útil para seleccionar un modelo.

En el contexto educativo, dada la importancia de contar con técnicas de focalización, los modelos de aprendizaje automático pueden resultar una herramienta útil para identificar a los estudiantes con mayor riesgo de no promoción.

En este documento se midió la capacidad predictiva de distintos modelos para predecir la aprobación por parte de los estudiantes en cuarto años de educación media superior pública, utilizando los datos de la encuesta del Programa Compromiso Educativo para la cohorte de estudiantes del año 2012. Además se indagó en la utilidad de usar técnicas de remuestreo para mejorar el desempeño predictivo de estos modelos.

En términos generales, las técnicas de remuestreo propuestas en la literatura no evidenciaron una mejora significativa de forma uniforme. Esto puede deberse a que en esta aplicación, la gran mayoría de las variables explicativas son categóricas.

Los resultados muestran, que el CRF sin utilizar técnicas de remuestreo presentó los mejores resultados medidos como el área debajo de la curva ROC (AUC) y su versión parcial (PAUC), aunque en algunos casos las diferencias con otras estrategias de modelización no resultaron estadísticamente significativa distinta de cero.

Tomando este modelo como base se estudio la importancia de las variables y sus efectos parciales. El deber exámenes, imaginarse estudiando en educación terciaria, haber repetido en educación media y si su familia lo imagina en

educación terciaria son las variables que tienen mayor incidencia sobre el AUC. Los gráficos de dependencia parcial indicaron que no deber exámenes, no haber repetido, imaginarse en educación terciaría y que su familia lo imagine en educación terciaria tienen un efecto positivo sobre la probabilidad de aprobar.

Finalmente se estimó cual sería el punto de corte adecuado mediante una partición entre muestra de testo y entrenamiento repetidas, eligiendo en cada iteración el punto de corte que devolviera mayor sensibilidad y especificidad en la muestra de testeo. Los resultados sugieren utilizar .78 como punto de corte.

Si bien los resultados parecen alentadores, futuras investigaciones se beneficiarían de un mayor acceso al historial académico del estudiante para lograr mejorar la capacidad predictiva de nuevos algoritmos, que permitirán realizar una focalización más certera de las políticas públicas.

Referencias bibliográficas

[1] E. Aguirre. Impacto de ser becado del programa compromiso educativo. *Documento de Trabajo DECON UdelaR; 16/16*, 2016.

[2] E. Aguirre and F. Veneri. El impacto de un programa de inclusión educativa. evidencia para uruguay del programa compromiso educativo. In *Una mirada joven a la juventud: aportes para las políticas públicas de Uruguay.*, pages 221–262. Banco Interamericano de Desarrollo, 2017.

[3] H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[4] N. Ambrosi, C. Conteri, and L. Cousillas. *Miradas. A cuatro años de Compromiso Educativo.* 2015.

[5] ANEP and MIDES. Informe de evaluación del programa compromiso educativo (ediciones 2011-2012). Technical report, 2014. unpublished.

[6] Banco Mundial. Learning for all: investing in people's knowledge and skills to promote development. 2011.

[7] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[8] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[9] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees.* CRC press, 1984.

[10] N. Caballero and G. Jadra. Caracterización de los jóvenes uruguayos que no asisten al sistema educativo. 2013.

[11] V. Calcagno. *glmulti: Model selection and multimodel inference made easy*, 2013. R package version 1.0.7.

[12] V. Calcagno, C. de Mazancourt, et al. glmulti: an r package for easy automated model selection with (generalized) linear models. *Journal of Statistical Software*, 34(12):1–29, 2010.

[13] A. C. Cameron and P. K. Trivedi. *Microeconometrics using stata*, volume 2. Stata press College Station, TX, 2010.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[15] T. Fernández. Enfoques para explicar la desafiliación. In *La desafiliación en la Educación Media y Superior de Uruguay: conceptos, estudios y políticas*. 2010.

[16] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[17] R. Genuer, J.-M. Poggi, and C. Tuleau. Random forests: some methodological insights. *arXiv preprint arXiv:0811.3619*, 2008.

[18] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871, 1971.

[19] B. M. Greenwell. pdp: An r package for constructing partial dependence plots. *R Journal*, 9(1), 2017.

[20] S. Grosso. Factores promotores o bloqueadores del éxito educativo en poblaciones vulnerables. resultados y reflexiones a partir del programa de aulas comunitarias. Master's thesis, Documento de Trabajo / FCS-DE, 2010.

[21] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning: Data mining, inference, and prediction. 2017.

[22] J. H. Holland. Adaptation in natural and artificial systems. an introductory analysis with application to biology, control, and artificial intelligence. *Ann Arbor, MI: University of Michigan Press*, pages 439–444, 1975.

[23] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[24] T. Hothorn, K. Hornik, and A. Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674, 2006.

[25] T. Hothorn and A. Zeileis. partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, 16:3905–3909, 2015.

[26] P. Huarhuachi and J. Clauss. Clasificación de fuga de clientes en una entidad financiera utilizando el algoritmo smote para datos desbalanceados en una regresión logística. 2017.

[27] INEEd. Informe sobre el estado de la educación en uruguay 2014. Technical report, INEEd, 2014.

[28] H. Ivanka, K. Jan, and Z. Jiri. *Kernel Smoothing in MATLAB: theory and practice of kernel smoothing*. World scientific, 2012.

[29] A. J. Izenman. Modern multivariate statistical techniques. *Regression, classification and manifold learning*, 2008.

[30] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 6. Springer, 2013.

[31] S. Janitza, C. Strobl, and A.-L. Boulesteix. An auc-based permutation variable importance measure for random forests. *BMC bioinformatics*, 14(1):119, 2013.

[32] M. Kuhn and K. Johnson. *Applied predictive modeling*. Springer Science & Business Media, 2013.

[33] M. Kuhn, J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, the R Core Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan, and T. Hunt. *caret: Classification and Regression Training*, 2017. R package version 6.0-78.

[34] R. Kunert. Smote explained for noobs - synthetic minority over-sampling technique line by line, 2017.

[35] R. V. Lenth. Least-squares means: The R package lsmeans. *Journal of Statistical Software*, 69(1):1–33, 2016.

[36] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[37] W.-Y. Loh. Fifty years of classification and regression trees. *International Statistical Review*, 82(3):329–348, 2014.

[38] W. McMahon. Conceptual framework for the analysis of the social benefits of lifelong learnings. *Education Economics*, 6(3):309–346, 1998.

[39] MIDES. *Sistematización básica de resultados de relevamiento de programas y proyectos sociales nacionales - 2014/2015.* 2016.

[40] R. Mollineda, R. Alejo, and J. Sotoca. The class imbalance problem in pattern classification and learning. In *II Congreso Español de Informática (CEDI 2007). ISBN*, pages 978–84, 2007.

[41] D. C. Montgomery. *Design and analysis of experiments.* John wiley & sons, 2017.

[42] R. H. Myers. Classical and modern regression with applications. Technical report, 1990.

[43] R. Patrón. When more schooling is not worth the effort: another look at the dropout decisions of disadvantaged students in uruguay. 2011.

[44] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2017. R package version 3.1-131.

[45] R Core Team. *foreign: Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ...*, 2017. R package version 0.8-69.

[46] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017.

[47] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12(1):77, 2011.

[48] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77, 2011.

[49] R. W. Rumberger and S. A. Lim. Why students drop out of school:a review of 25 years of research. 2008.

[50] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

[51] A. Sen. *Development as freedom.* Oxford Paperbacks, 2001.

[52] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):7881, 2005.

[53] B. Snilstveit, E. Gallagher, D. Phillips, M. Vojtkova, J. Eyers, D. Skaldiou, J. Stevenson, A. Bhavsar, and P. Davies. Education interventions for improving the access to, and quality of, education in low and middle income countries: A systematic review. Technical report, The Campbell Collaboration, 2014.

[54] H. Strasser and C. Weber. On the asymptotic theory of permutation statistics. 1999.

[55] A. C. Tamhane. *Statistical analysis of designed experiments: theory and applications*, volume 609. John Wiley & Sons, 2009.

[56] T. Therneau, B. Atkinson, and B. Ripley. *rpart: Recursive Partitioning and Regression Trees*, 2017. R package version 4.1-11.

[57] L. Torgo. *Data Mining with R, learning with case studies*, 2010.

[58] J. VanDerWal, L. Falconi, S. Januchowski, L. Shoo, and C. Storlie. *SDMTools: Species Distribution Modelling Tools: Tools for processing data associated with species distribution modelling exercises*, 2014. R package version 1.1-221.

[59] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

[60] H. Wickham, R. Francois, L. Henry, and K. Müller. *dplyr: A Grammar of Data Manipulation*, 2017. R package version 0.7.4.

[61] Wikipedia contributors. Receiver operating characteristic — Wikipedia, the free encyclopedia, 2018. [Online; accessed 8-July-2018].

[62] D. Willms. *PISA Student Engagement at School A Sense of Belonging and Participation: Results from PISA 2000: A Sense of Belonging and Participation: Results from PISA 2000*. PISA. OECD Publishing, 2003.

[63] J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

Técnicas de muestreo

En este apartado se presenta con mayor grado de detalle la técnica SMOTE para seleccionar una muestra. Además se ilustra el efecto de la técnicas de muestro utilizando una simulación con dos variables continuas.

En primer lugar, se utilizaran los primeros casos del dataset Iris y dos variables de interés, largo del sépalo (sepal.length) y el ancho del mismo (sepal.width), utilizando como parámetro de entrada dos vecinos (k=2)[1].

Las figura 5a presenta los datos originales, donde se representa en rojo las 4 observaciones de la clase minoritaria y en verde la mayoritaria. Los casos de la clase minoritaria están conectados por rectas que representan el espacio donde es posible generar casos sintéticos como un punto entre dos observaciones.

El primer paso del algoritmo consiste en seleccionar un caso entre las 4 observaciones originales. La figura 5b presenta este primer paso donde se selecciona el caso marcado como un rectángulo rojo y sus dos vecinos más cercanos.

Entre estos dos vecinos se selecciona aleatoriamente un caso con el cual se combinara la información del caso original. Habiendo seleccionada uno de los casos (figura 5c), es posible generar un caso sintético entre el caso original y el donante. En el paso final se selecciona mediante una variable aleatoria uniforme entre 0 y 1 el parámetro λ, que determina las coordenadas del nuevo caso en el segmento. En caso de ser cercano a 1, el nuevo caso estará más próximo al caso donante mientras que si es cercano a 0, estará próximo al caso original. La tabla 5 muestra los cálculos realizados en este paso. Este proceso se repiten hasta generar todos los casos sintéticos.

Lo anterior es valido en la medida de que la base contenga únicamente variables continuas. En muchas casos y en particular en educación, en los individuos se miden variables categóricas, como el sexo, en este contexto no es valido usar la distancia euclídea para seleccionar los k vecinos más cercanos

[1]Este ejemplo fue presentado por Richard Kunert [34]

Tabla 5: Generación de una nueva observación sintética con SMOTE

Considere el punto inicial (4.9,3) y el vecino más cercano seleccionado (4.4,2.9) siendo las coordenadas *Sepal.Length* y *Sepal.Width*. El vector de distancia entre los dos puntos esta dado por la columna dist, y la columna λ indica el valor sorteado utilizando una distribución uniforme[0,1].

heightVariables	Punto inicial	Vecino	dist	λ	Punto nuevo
Sepal.Length	4.9	4.4	-0.5	0.8	4.50
Sepal.Width	3	2.9	-0.1		2.92

Las coordenadas del nuevo punto son calculadas como:

$$\begin{pmatrix} \text{Sepal.Length}_{nuevo} \\ \text{Sepal.Width}_{nuevo} \end{pmatrix} = \begin{pmatrix} 4.9 \\ 3 \end{pmatrix} + 0.8 \begin{pmatrix} -0.5 \\ -0.1 \end{pmatrix} = \begin{pmatrix} 4.50 \\ 2.92 \end{pmatrix}$$

y generar casos sintéticos, pudiendo optar por una distancias más adecuada. Si se cuentan con variables continuas y binarias es posible usa la distancia de **Gower** [18]. Se define la distancia entre dos individuos i y j como $d_{ij} = 1 - s_{ij}$ donde s_{ij} representa la similitud entre estos dos individuos medido como:

$$s(i,j) = \frac{\sum_{k=1}^{K} w_k(i,j) S_k(i,j)}{\sum_{k=1}^{K} w_k(i,j)}$$

donde $w_k(i,j)$ indica el peso de la variable en la comparación. Esta puede valer 1, si es utilizada en la comparación, 0 en otro caso. $S_k(i,j)$ establece la similitud entre dos observaciones i y j que toman valores en las variables k igual a $X_{k,i}$ y $X_{k,j}$ respectivamente. Si la variable k es numérica se considera $S_k(i,j) = 1 - \frac{|X_{k,i} - X_{k,j}|}{rango(X_k)}$, en caso de ser binarias $S_k(i,j) = 1$ si $X_{k,i} = X_{k,j}$ y 0 en otro caso.

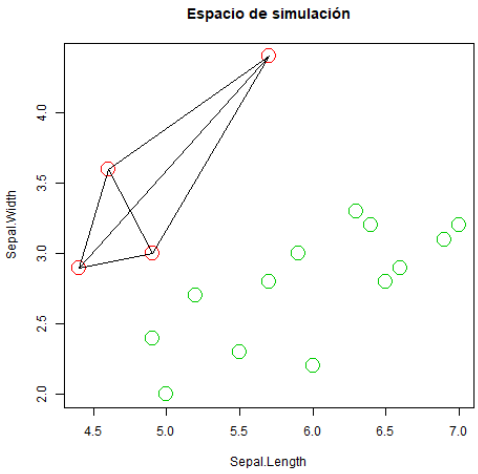
Habiendo seleccionado los k-vecinos más cercanos, el algoritmo procederá de la misma manera para imputar las variables continuas mientras que para las variables binarias considerara la moda de sus k vecinos.

Se modifica el ejemplo anterior basado en los datos iris para mostrar esta variante. Se agrega un caso adicional respecto al ejemplo anterior y se transformó el largo del sépalo en una variable binaria que vale 1 si el largo se encuentra por encima de 3 y 0 en otro caso (tabla 6).

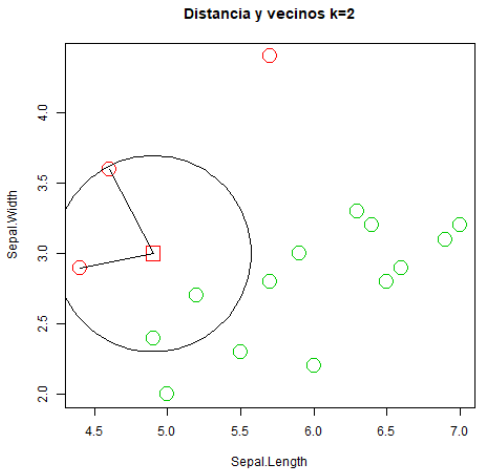
Tomando los casos de especie 1, se calcula la matriz de distancia de Gower 7. Si durante el primer paso del algoritmo hubiera sido seleccionado el individuo 1, sus tres vecinos más cercanos son las observaciones: 2, 16 y 23.

Para construir el nuevo caso sintético, el algoritmo procedería de la misma manera que para las variables continuas, es decir, se calcula la distancia euclídea y se sortea un valor λ para asignar el nuevo valor de Sepal.Length. En

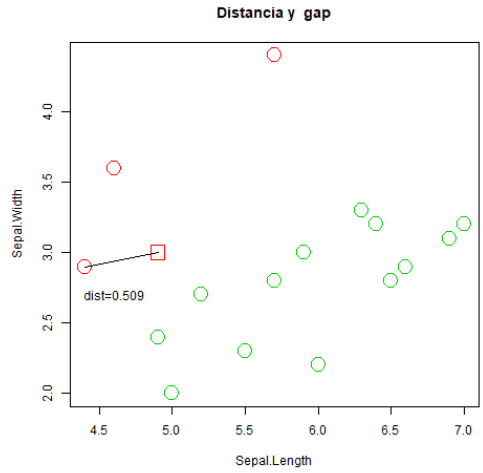
Figura 5: Ejemplo SMOTE variables continuas



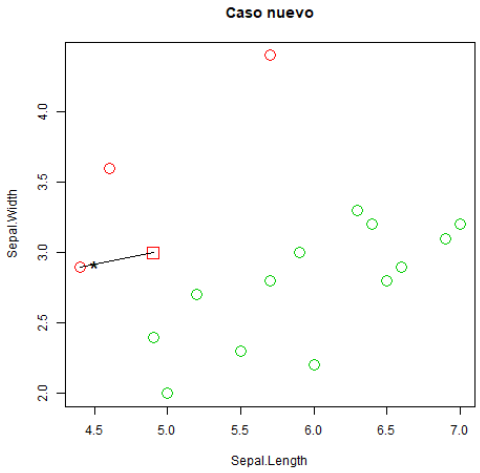
(a) Datos Setosa



(b) Distancia



(c) Distancia vecino seleccionado



(d) Caso nuevo

el caso de las variables categóricas, calculara la moda y asignara este valor a la nueva observación. En nuestro ejemplo, la moda de los vecinos del individuo 2 es 0 por lo que se asignara este valor como $Sepal.Width_{bin}$ al caso sintético.

Para mostrar el efecto que tiene las técnicas de remuestreo sobre la muestra de entrenamiento, se muestra un ejemplo adicional donde se generaran valores de dos normales bivariadas con distintas medias, por lo cual puede apreciarse dos sub poblaciones distintas (ver figura 6a) siendo el grupo de 0 la clase minoritaria. Para el grupo 0, se generaron 50 observaciones a par-

Tabla 6: Datos binarios Iris

ID	Sepal.Length	Species	Sepal.Width$_{bin}$
1	5.10	1	0
2	4.90	1	1
9	4.40	1	1
16	5.70	1	0
23	4.60	1	0
51	7.00	2	0
52	6.40	2	0
53	6.90	2	0
54	5.50	2	1
55	6.50	2	1
56	5.70	2	1
57	6.30	2	0
58	4.90	2	1
59	6.60	2	1
60	5.20	2	1
61	5.00	2	1
62	5.90	2	1
63	6.00	2	1

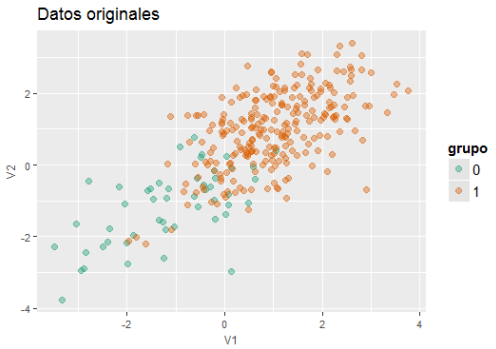
Tabla 7: Matriz de distancia de Gower

ID	1	2	9	16	23
1	0				
2	0.38	0			
9	0.51	0.13	0		
16	0.15	0.54	0.67	0	
23	0.13	0.41	0.38	0.28	0

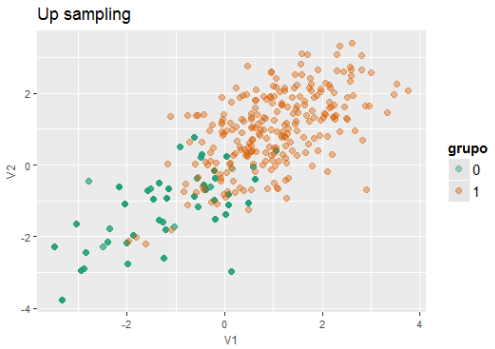
tir de la siguiente distribución: $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} -1 \\ -1 \end{pmatrix}\begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}\right)$. Para el grupo 1, se generaron 250 observaciones a partir de la siguiente distribución: $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}\begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}\right)$.

El resultado de aplicar Upsampling, figura 6b, es la replica de los casos minoritarios, se observa que en la figura aumentan la densidad de puntos de la clase minoritaria. En el caso de Downsampling, figura 6c, el efecto del remuestreo es la eliminación de casos mayoritarios equiparando la proporción de clases. Finalmente con la figura 6d, se presentan los resultados de aplicar el SMOTE. Se observa que se eliminaron casos de la clase mayoritaria, así como la generación de nuevos casos del grupo minoritario obtenidos como la combinación de los casos sorteados y sus vecinos.

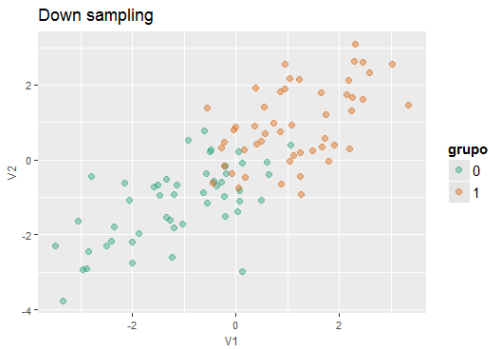
Figura 6: Efecto de estrategias de selección de la muestra de entrenamiento



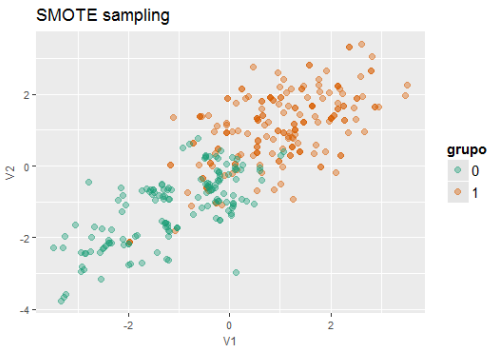
(a) Datos originales



(b) upsampling



(c) Downsampling



(d) Smote

Listado de paquetes utilizados

Tabla 8: Listado de paquetes utilizados en R.

Modelo	Paquete en R	Observaciones
GLM	glm [46]	Modelo logit sin selección de variables explicativas.
GLM_m	glmulti [11]	Modelo logit con selección genética de variables. Se utilizó AIC como criterio de información, una población de 100 modelos y tolerancia para el cambio en AIC promedio y mejor AIC de 0.05.
CART	rpart [56]	Árbol podado con la regla de 1SD de Breiman[1]
CTREE	rparty [25]	Árbol construido con un nivel de significación (α) del 5 %
RF	randomForest [36]	Se usan 500 árboles y en cada nodo se sortean $\sqrt{k}$ variables.
CRF	rparty [25]	Se utilizaron 500 arboles CTREE con $\alpha = 5\,\%$.
Muestreo	**Paquete en R**	**Observaciones**
Simple	caret [33]	Muestreo estratificado, entrenamiento 70 % y testeo 30 %
Down	caret [33]	Remuestreo con down sampling del conjunto de entrenamiento.
SMOTE	DMwR [57]	Remuestreo con SMOTE del conjunto de entrenamiento. Se utilizaron 5 vecinos más cercanos.
Métricas	**Paquete en R**	**Observaciones**
AUC	pROC [48]	Estimación por defecto con la regla del trapezoide.
PAUC	pROC [48]	Área bajo la curva ROC con error tipo 1 menor al 10 %.
Sensibilidad, Especificidad, Error	SDMTools [58] , ROCR [52]	
Diseño experimentos	**Paquete en R**	**Observaciones**
Modelo ANOVA	nlme [44]	Efecto aleatorio por iteración.
Comparación múltiple	lsmeans [35]	Comparación multiple ajustada por el criterio de Tukey, nivel de significación del 5 %
Otros	**Paquete en R**	**Observaciones**
Gráficos	ggplot [59], gráficos base [46]	
Lectura base de datos	foreing [45]	Base de datos original en formato .dta (STATA).
Estadísticos	dplyr [60]	

Nota: RF (Random Forest), CART, CTREE, CRF (RF de CTREE), GLM (logit), GLM_m (logit con selección genética de variables explicativas).

[1] Se selecciona el árbol más parsimonioso con error menor a un desvío estándar del mejor modelo.

Script en R

Script Principal

Datos y descriptivos

```
library(foreign)
library(ggplot2)
library(pROC)
library(caret)
library(dplyr)

orig <- read.dta("Encuesta2012ConTray12_14v3.dta")

## Recorto la muestra.
orig <- subset(orig,orig$t0==1); A2=dim(orig)[1]
#voy a sacar a los becados de la base.
orig <- subset(orig,orig$solicitantebeca2012asig=="SinBeca"); A3=dim(orig)[1]
orig <- subset(orig,orig$solicitantebeca2011asig==0); A4=dim(orig)[1]
## Me quedo co  cuarto
orig <- subset(orig,orig$Quinto==0); A5=dim(orig)[1]
##Ultimo , cedulas de identidad incorrectas
orig <- subset(orig,orig$cirecuperadas!="."); A6=dim(orig)[1]

paste("Nos quedamos con la primera ola, perdemos", A1-A2, "observaciones, representan"
 ,       round((A1-A2)/A1,4) ,"% de las observaciones.")

paste("Excluimos a estudiantes becados por el PCE 2012",A2-A3,"observaciones,
 representan",         round((A2-A3)/A2,4) ,"% de las observaciones.")

paste("Excluimos a estudiantes becados por el PCE 2011",A3-A4,"observaciones,
 representan",         round((A3-A4)/A3,4) ,"% de las observaciones.")

paste("Excluimos a estudiantes de quinto ano", A4-A5,   "observaciones, representan",
          round((A3-A5)/A5,4) ,"% de las observaciones.")
rm("A1","A2","A3","A4","A5")


###RECODIFICAR ABANDONO EN UNA SOLA
orig$algun_abandonoEMS=ifelse(orig$AbondonoEM==1|
orig$AbandonoAnAnt==1,1,0)
##Le agrego un NA si tiene missing en alguno de los dos
orig[(is.na(orig$AbondonoEM)|
is.na(orig$AbandonoAnAnt)),]$algun_abandonoEMS<-NA


##Selecciono el data frame
data_frame_admin=subset(orig,select=
c("Aprobo", "algun_abandonoEMS","DebeExamenes"   ,"RepitioEM","RepitioEscuela","
 TrabajActual","AyudaPadresTrabajo","CuidoFlia","AyudTarHog","TuFlApoyMuch","
 VincMBCompan","VincMBDocent",     "secli","Mvdeo","hombre","icc2008_ce2012t0","
 ImTuFliaTerc", "ImEdTerc","MuchAmiDejEst"))


```

```
#Casos completos###
data_complete=na.omit(data_frame_admin)
###Defino data.
data=data_complete
str(data)


###Factoreo las variables######
table(data$Aprobo)/nrow(data)
data$Aprobo <- factor(data$Aprobo, labels=c("No", "Si"), ordered=FALSE)

names <- c("algun_abandonoEMS","DebeExamenes","RepitioEM","RepitioEscuela","
 TrabajActual","AyudaPadresTrabajo","CuidoFlia","AyudTarHog", "TuFlApoyMuch","
 VincMBCompan","VincMBDocent",
"secli","Mvdeo","hombre",#,"ConHijos","ImTuFliaTerc","ImEdTerc","MuchAmiDejEst")

data[,names] <- lapply(data[,names] , factor)
str(data)

#Descriptivos
library('dplyr')

Salida<-as.data.frame(data) %>% group_by(Aprobo) %>% summarize(
algun_abandonoEMS=mean(algun_abandonoEMS==1),
RepitioEM=mean(RepitioEM==1),
RepitioEscuela=mean(RepitioEscuela==1),
DebeExamenes=mean(DebeExamenes==1),
TrabajActual=mean(TrabajActual==1),
AyudaPadresTrabajo=mean(AyudaPadresTrabajo==1),
CuidoFlia=mean(CuidoFlia==1),
AyudTarHog=mean(AyudTarHog==1),
TuFlApoyMuch=mean(TuFlApoyMuch==1),
VincMBCompan=mean(VincMBCompan==1),
VincMBDocent=mean(VincMBDocent==1),
secli=mean(secli==1),
Mvdeo=mean(Mvdeo==1),
hombre=mean(hombre==1),
ImEdTerc=mean(ImEdTerc==1),
ImTuFliaTerc=mean(ImTuFliaTerc==1),
MuchAmiDejEst=mean(MuchAmiDejEst==1),
ICC=mean(icc2008_ce2012t0))

t(as.data.frame(Salida))

table(data$Aprobo,data$MuchAmiDejEst)

table(data$Aprobo,data$algun_abandonoEMS)
TABLITA  <-rbind(
cbind(prop.test(t(table(data$algun_abandonoEMS,data$Aprobo)),
correct = TRUE)$estimate[1],
prop.test(t(table(data$algun_abandonoEMS,data$Aprobo)), correct = TRUE)$estimate[2],
prop.test(t(table(data$algun_abandonoEMS,data$Aprobo)), correct = TRUE)$statistic,
prop.test(t(table(data$algun_abandonoEMS,data$Aprobo)), correct = TRUE)$p.value) ,

cbind(
prop.test(t(table(data$RepitioEM,data$Aprobo)),
correct = TRUE)$estimate[1],
prop.test(t(table(data$RepitioEM,data$Aprobo)),
correct = TRUE)$estimate[2],
prop.test(t(table(data$RepitioEM,data$Aprobo)),
correct = TRUE)$statistic,
prop.test(t(table(data$RepitioEM,data$Aprobo)),
correct = TRUE)$p.value) ,

cbind(
prop.test(t(table(data$RepitioEscuela,data$Aprobo)),
correct = TRUE)$estimate[1],
prop.test(t(table(data$RepitioEscuela,data$Aprobo)),
correct = TRUE)$estimate[2],
prop.test(t(table(data$RepitioEscuela,data$Aprobo)),
correct = TRUE)$statistic,
prop.test(t(table(data$RepitioEscuela,data$Aprobo)),
correct = TRUE)$p.value) ,
```

```
114
115     cbind(
116     prop.test(t(table(data$DebeExamenes,data$Aprobo)),
117     correct = TRUE)$estimate[1],
118     prop.test(t(table(data$DebeExamenes,data$Aprobo)),
119     correct = TRUE)$estimate[2],
120     prop.test(t(table(data$DebeExamenes,data$Aprobo)),
121     correct = TRUE)$statistic,
122     prop.test(t(table(data$DebeExamenes,data$Aprobo)),
123     correct = TRUE)$p.value) ,
124
125     cbind(
126     prop.test(t(table(data$TrabajActual,data$Aprobo)),
127     correct = TRUE)$estimate[1],
128     prop.test(t(table(data$TrabajActual,data$Aprobo)),
129     correct = TRUE)$estimate[2],
130     prop.test(t(table(data$TrabajActual,data$Aprobo)),
131     correct = TRUE)$statistic,
132     prop.test(t(table(data$TrabajActual,data$Aprobo)),
133     correct = TRUE)$p.value) ,
134
135     cbind(
136     prop.test(t(table(data$AyudaPadresTrabajo,data$Aprobo)),
137     correct = TRUE)$estimate[1],
138     prop.test(t(table(data$AyudaPadresTrabajo,data$Aprobo)),
139     correct = TRUE)$estimate[2],
140     prop.test(t(table(data$AyudaPadresTrabajo,data$Aprobo)),
141     correct = TRUE)$statistic,
142     prop.test(t(table(data$AyudaPadresTrabajo,data$Aprobo)),
143     correct = TRUE)$p.value) ,
144
145
146     cbind(
147     prop.test(t(table(data$CuidoFlia,data$Aprobo)),
148     correct = TRUE)$estimate[1],
149     prop.test(t(table(data$CuidoFlia,data$Aprobo)),
150     correct = TRUE)$estimate[2],
151     prop.test(t(table(data$CuidoFlia,data$Aprobo)),
152     correct = TRUE)$statistic,
153     prop.test(t(table(data$CuidoFlia,data$Aprobo)),
154     correct = TRUE)$p.value) ,
155
156     cbind(
157     prop.test(t(table(data$AyudTarHog,data$Aprobo)),
158     correct = TRUE)$estimate[1],
159     prop.test(t(table(data$AyudTarHog,data$Aprobo)),
160     correct = TRUE)$estimate[2],
161     prop.test(t(table(data$AyudTarHog,data$Aprobo)),
162     correct = TRUE)$statistic,
163     prop.test(t(table(data$AyudTarHog,data$Aprobo)),
164     correct = TRUE)$p.value) ,
165
166
167     cbind(
168     prop.test(t(table(data$TuFlApoyMuch,data$Aprobo)),
169     correct = TRUE)$estimate[1],
170     prop.test(t(table(data$TuFlApoyMuch,data$Aprobo)),
171     correct = TRUE)$estimate[2],
172     prop.test(t(table(data$TuFlApoyMuch,data$Aprobo)),
173     correct = TRUE)$statistic,
174     prop.test(t(table(data$TuFlApoyMuch,data$Aprobo)),
175     correct = TRUE)$p.value) ,
176
177
178     cbind(
179     prop.test(t(table(data$VincMBCompan,data$Aprobo)),
180     correct = TRUE)$estimate[1],
181     prop.test(t(table(data$VincMBCompan,data$Aprobo)),
182     correct = TRUE)$estimate[2],
183     prop.test(t(table(data$VincMBCompan,data$Aprobo)),
184     correct = TRUE)$statistic,
185     prop.test(t(table(data$VincMBCompan,data$Aprobo)),
186     correct = TRUE)$p.value) ,
```

```
187
188     cbind(
189     prop.test(t(table(data$VincMBDocent,data$Aprobo)),
190     correct = TRUE)$estimate[1],
191     prop.test(t(table(data$VincMBDocent,data$Aprobo)),
192     correct = TRUE)$estimate[2],
193     prop.test(t(table(data$VincMBDocent,data$Aprobo)),
194     correct = TRUE)$statistic,
195     prop.test(t(table(data$VincMBDocent,data$Aprobo)),
196     correct = TRUE)$p.value) ,
197
198
199     cbind(
200     prop.test(t(table(data$secli,data$Aprobo)),
201     correct = TRUE)$estimate[1],
202     prop.test(t(table(data$secli,data$Aprobo)),
203     correct = TRUE)$estimate[2],
204     prop.test(t(table(data$secli,data$Aprobo)),
205     correct = TRUE)$statistic,
206     prop.test(t(table(data$secli,data$Aprobo)),
207     correct = TRUE)$p.value) ,
208
209     cbind(
210     prop.test(t(table(data$Mvdeo,data$Aprobo)),
211     correct = TRUE)$estimate[1],
212     prop.test(t(table(data$Mvdeo,data$Aprobo)),
213     correct = TRUE)$estimate[2],
214     prop.test(t(table(data$Mvdeo,data$Aprobo)),
215     correct = TRUE)$statistic,
216     prop.test(t(table(data$Mvdeo,data$Aprobo)),
217     correct = TRUE)$p.value) ,
218
219     cbind(
220     prop.test(t(table(data$hombre,data$Aprobo)),
221     correct = TRUE)$estimate[1],
222     prop.test(t(table(data$hombre,data$Aprobo)),
223     correct = TRUE)$estimate[2],
224     prop.test(t(table(data$hombre,data$Aprobo)),
225     correct = TRUE)$statistic,
226     prop.test(t(table(data$hombre,data$Aprobo)),
227     correct = TRUE)$p.value) ,
228     cbind(
229     prop.test(t(table(data$ImEdTerc,data$Aprobo)),
230     correct = TRUE)$estimate[1],
231     prop.test(t(table(data$ImEdTerc,data$Aprobo)),
232     correct = TRUE)$estimate[2],
233     prop.test(t(table(data$ImEdTerc,data$Aprobo)),
234     correct = TRUE)$statistic,
235     prop.test(t(table(data$ImEdTerc,data$Aprobo)),
236     correct = TRUE)$p.value) ,
237
238     cbind(
239     prop.test(t(table(data$ImTuFliaTerc,data$Aprobo)),
240     correct = TRUE)$estimate[1],
241     prop.test(t(table(data$ImTuFliaTerc,data$Aprobo)),
242     correct = TRUE)$estimate[2],
243     prop.test(t(table(data$ImTuFliaTerc,data$Aprobo)),
244     correct = TRUE)$statistic,
245     prop.test(t(table(data$ImTuFliaTerc,data$Aprobo)),
246     correct = TRUE)$p.value) ,
247
248     cbind(
249     prop.test(t(table(data$MuchAmiDejEst,data$Aprobo)),
250     correct = TRUE)$estimate[1],
251     prop.test(t(table(data$MuchAmiDejEst,data$Aprobo)),
252     correct = TRUE)$estimate[2],
253     prop.test(t(table(data$MuchAmiDejEst,data$Aprobo)),
254     correct = TRUE)$statistic,
255     prop.test(t(table(data$MuchAmiDejEst,data$Aprobo)),
256     correct = TRUE)$p.value),
257
258     cbind(
259     t.test(data$icc2008_ce2012t0~data$Aprobo)$estimate[1],
```

```
    t.test(data$icc2008_ce2012t0~data$Aprobo)$estimate[2],
    t.test(data$icc2008_ce2012t0~data$Aprobo)$statistic,
    t.test(data$icc2008_ce2012t0~data$Aprobo)$p.value)
    )

```

Comparación de modelos

```

inicio=Sys.time()
DF_S_1<-Compara_models_f(1,data,prop_train_test=0.7)
DF_S_1$error=ifelse(is.na(DF_S_1$AUC),1,0)
fin=Sys.time()
fin-inicio

inicio=Sys.time()
DF_S_2<-Compara_models_f(99,data,0.7)
DF_S_2$error=ifelse(is.na(DF_S_2$AUC),1,0)
fin=Sys.time()
fin-inicio

DF_SALIDA_FINAL<-rbind(DF_S_1,DF_S_2)
write.csv(as.data.frame(DF_SALIDA_FINAL),"DF_SALIDA_FINALcsv")


##Estadisticos
DF_SALIDA_FINAL$AUC_true<-as.numeric(DF_SALIDA_FINAL$AUC_true)
DF_SALIDA_FINAL$Partial_auc<-as.numeric(DF_SALIDA_FINAL$Partial_auc)

ESTADISTICOS<-  DF_SALIDA_FINAL %>%
  group_by(Modelo,Estr) %>%
  summarise (mean_AUC_true=mean(AUC_true,na.rm = T),
             sd_AUC_true= sd(AUC_true,na.rm = T),
             mean_Partial_auc=mean(Partial_auc,na.rm = T),
             sd_Partial_auc= sd(Partial_auc,na.rm = T)

  )


library(xtable)
write.csv(as.data.frame(ESTADISTICOS),"ESTADISTICO.csv")
xtable(ESTADISTICOS, digits = 3,caption = "Estadisticos" )

##Graficos
DF_SALIDA_FINAL$INTER <- interaction(DF_SALIDA_FINAL$Modelo,
DF_SALIDA_FINAL$Estr)

ggplot(aes(y = AUC_true*100, x = Modelo, fill = Estr),
data = DF_SALIDA_FINAL) + geom_boxplot()+ylab("AUC")+
ylim(50,90)

ggplot(aes(y = Partial_auc, x = Modelo, fill = Estr),
data = DF_SALIDA_FINAL) + geom_boxplot()+ylab("AUC Parcial")+
ylim(50,90)

###EXPERIMENTO
library(nlme)
DF_SALIDA_FINAL$IDD<-rep(1:100,each=18)
library(emmeans)

##AUC
results.lme_AUC <- lme(AUC_true*100~Estr*Modelo,
                       random=~1|IDD, data=DF_SALIDA_FINAL)
anova(results.lme_AUC)

marginal_AUC = lsmeans(results.lme_AUC,
                       ~ Estr:Modelo)
cld(marginal_AUC,alpha   = 0.05,
Letters = letters, adjust  = "tukey")

##PARCIAL
CASO_COMP<-DF_SALIDA_FINAL[complete.cases(DF_SALIDA_FINAL),]
```

```
results.lme_PAUC <- lme(Partial_auc~Estr*Modelo,
                        random=~1|IDD, data=CASO_COMP)

anova(results.lme_PAUC)

marginal_PAUC = lsmeans(results.lme_PAUC,
                        ~ Estr:Modelo)

cld(marginal_PAUC,alpha=0.05,
Letters=letters,adjust="tukey")
```

Importancia de variables y dependencia parcial

```

library(party)
c_Rf_ALL <- party::cforest(formula = formula_sint, data = data)

VAR_IMP2<-varimpAUC(c_Rf_ALL)

VAR_IMP<-as.data.frame(cbind(names(VAR_IMP2),as.numeric(VAR_IMP2)))
VAR_IMP$var

VAR_IMP$labels<-c("Alg[U+FFFD]n abandono previo en EMS",
"Repiti[U+FFFD] en EM",
"Repiti[U+FFFD] en la escuela",
"Debe ex[U+FFFD]menes",
"Trabaja Actualmente",
"Ayuda Padres en el Trabajo","Cuido Familia",
"Ayuda en las Tareas del Hogar",
"Apoyo Familiar",
"Vinculo muy bueno con compa[U+FFFD]eros",
"Vinculo muy bueno con docentes",
"SECLI",
"Montevideo",
"Hombre",
"ICC",
"Su familia lo imagina en educaci[U+FFFD]n terciaria",
"Se imagina en educaci[U+FFFD]n terciaria",
"Muchos Amigos dejaron de estudiar")


VAR_IMP <- VAR_IMP[order(VAR_IMP$IMP),]

png(paste("C:/Users/Usuario/Dropbox/Tesis IESTA/Estimaci[U+FFFD]n en R/SALIDAS PDP/",
          "VIM",".png"))
dotchart(VAR_IMP$IMP,
         labels=VAR_IMP$labels,cex=.7,
main="Importancia de las variables: \n criterio AUC")
dev.off()


Lista<-c("algun_abandonoEMS",
  "RepitioEM","RepitioEscuela","DebeExamenes","TrabajActual",
  "AyudaPadresTrabajo","CuidoFlia","AyudTarHog",
  "TuFlApoyMuch","VincMBCompa[U+FFFD]","VincMBDocent",
  "secli","Mvdeo","hombre","ImTuFliaTerc",
  "ImEdTerc","MuchAmiDejEst","icc2008_ce2012t0")

Lista_etiquetas<-c("Alg[U+FFFD]n abandono previo en EMS",
          "Repiti[U+FFFD] en EM","Repiti[U+FFFD] en la escuela","Debe ex[U+FFFD]menes",
          "Trabaja Actualmente",
          "Ayuda Padres en el Trabajo","Cuido Familia","Ayuda en las Tareas del Hogar",
          "Apoyo Familiar","Vinculo muy bueno con compa[U+FFFD]eros",
          "Vinculo muy bueno con docentes",
          "SECLI","Montevideo","Hombre","Su familia lo imagina en educaci[U+FFFD]n terciaria",
          "Se imagina en educaci[U+FFFD]n terciaria","Muchos Amigos dejaron de estudiar","ICC")



library(pdp)
for (i in 1:18){
  auxi_pp<-partial(c_Rf_ALL,paste(Lista[i]),prob=TRUE,which.class=2)
```

```
60
61 if (i!=18) {
62   png(paste("C:/Users/Usuario/Dropbox/Tesis IESTA/Estimaci[U+FFFD]n en R/SALIDAS PDP/",
63               as.character(Lista[i]),".png"))
64     barplot(auxi_pp$yhat,ylim=c(0,1),main=paste("Efecto Parcial: \n",as.character(
       Lista_etiquetas[i])),
65               names.arg = c("No","Si"))
66   dev.off()
67
68 } else {
69   png(paste("C:/Users/Usuario/Dropbox/Tesis IESTA/Estimaci[U+FFFD]n en R/SALIDAS PDP/",
70               as.character(Lista[i]),".png"))
71   plot(auxi_pp$yhat,ylim=c(0,1),main=paste("Efecto Parcial: \n",as.character(
       Lista_etiquetas[i])),type="l",ylab="")
72   dev.off()
73 }}
```

Punto de corte

```
1  library(partykit)
2  require(pROC)
3  require(SDMTools)
4  library(party)
5
6  for (i in 1:100){
7  listado <- createDataPartition(data$Aprobo,
8  p = 0.7, list = FALSE)
9  data_train <- data[listado, ]
10 data_test <- data[-listado, ]
11
12
13 Y_TESTEO=as.data.frame(data_test$Aprobo)
14 colnames(Y_TESTEO)<-"Aprobo"
15
16 Y_TRAINEO=as.data.frame(data_train$Aprobo)
17 colnames(Y_TRAINEO)<-"Aprobo"
18
19 c_Rf_ALL <- party::cforest(formula = formula_sint,
20 data = data_train)
21
22 ##Agrego  hat
23 Y_TESTEO$HAT<-unlist(predict(c_Rf_ALL,
24 newdata=data_test,type ="prob"))[seq(2,dim(data_test)[1]*2,2)]
25 Y_TRAINEO$HAT<-unlist(predict(c_Rf_ALL,
26 newdata=data_train,type ="prob"))[seq(2,dim(data_train)[1]*2,2)]
27
28 ##ARMO  ROC
29 Y_TRAINEO$aprobo_num=ifelse(Y_TRAINEO$Aprobo=="Si",1,0)
30 pred <-prediction(Y_TESTEO$HAT,Y_TESTEO$Aprobo)
31
32
33 #mayor sensibilidad y especificadad
34 ss <-performance(pred, "sens","spec")
35 cut_prob_ss <-ss@alpha.values[[1]][which.max(ss@x.values[[1]]+
36 ss@y.values[[1]])]
37
38 Y_TESTEO$aprobo_num=ifelse(Y_TESTEO$Aprobo=="Si",1,0)
39 una_iter<-accuracy(Y_TESTEO$aprobo_num,Y_TESTEO$HAT,cut_prob_ss)
40
41 if (i==1) {
42   Salida<-una_iter
43 } else {
44   Salida=rbind(Salida,una_iter)
45 }
46
47 }
48
49
50 plot(density(Salida$threshold),
51 main="Estimacion de densidad \n Punto de corte")
52 abline(v=mean(Salida$threshold),col="red")
53 text(round(mean(Salida$threshold),2),
```

```
4,paste(round(mean(Salida$threshold),2)),cex=.8)

plot(density(1-Salida$prop.correct),
main="Estimacion de densidad \n Error")
abline(v=mean(1-Salida$prop.correct),col="red")
text(round(mean(1-Salida$prop.correct),2),
4,paste(round(mean(1-Salida$prop.correct),2)),cex=.8)

plot(density(Salida$sensitivity)
,main="Estimacion de densidad \n Sensibilidad")
abline(v=mean(Salida$sensitivity),col="red")
text(round(mean(Salida$sensitivity),2),
4,paste(round(mean(Salida$sensitivity),2)),cex=.8)

plot(density(Salida$specificity),
main="Estimacion de densidad \n Especificidad")
abline(v=mean(Salida$specificity),col="red")
text(round(mean(Salida$specificity),2),
4,paste(round(mean(Salida$specificity),2)),cex=.8)
```

Funciones auxiliares

Testeo cruzado

```

Compara_models<-function(repite,data,prop_train_test){
  require(caret)
  require(DMwR)
  require(ROSE)
  require(e1071)
  require(randomForest)
  require(ROCR)
  require(pROC)
  require(party)
  require(glmulti)
  require(rpart)

##Defino una barra de avance.
pb = txtProgressBar(min = 0, max = repite, initial = 0,style=3)
for (i in 1:repite){ ##Loop principal
  setTxtProgressBar(pb,i)
##Generamos train y test.
listado <- createDataPartition(data$Aprobo,
p = prop_train_test, list = FALSE)
data_train <- data[listado, ]
data_test <- data[-listado, ]
Y_test=as.data.frame(data_test$Aprobo)
colnames(Y_test)<-"Aprobo"

Realizamos el paso de re muestreo
smote_train <- SMOTE(Aprobo ~ ., data  = data_train)
down_train <- downSample(x = data_train[,2:length(data_train)],
y = data_train$Aprobo)
down_train$Aprobo=down_train$Class

#Guardo los Y  train
Y_train_simple=as.data.frame(data_train$Aprobo)
colnames(Y_train_simple)<-"Aprobo"
Y_train_down=as.data.frame(down_train$Aprobo)
colnames(Y_train_down)<-"Aprobo"
Y_train_smote=as.data.frame(smote_train$Aprobo)
colnames(Y_train_smote)<-"Aprobo"

###Defino la formula para predecir.
formula_sint=Aprobo~ algun_abandonoEMS +
RepitioEM+RepitioEscuela+DebeExamenes+TrabajActual+
AyudaPadresTrabajo+CuidoFlia+AyudTarHog+
TuFlApoyMuch+VincMBCompan+VincMBDocent+
secli+Mvdeo+hombre+icc2008_ce2012t0+ImTuFliaTerc+
```

```
ImEdTerc+MuchAmiDejEst

###LOGIT###
logit_0 <- tryCatch(glm(formula_sint, data = data_train,  family =  "binomial"),error=
    function(e)e,finally="BAD")

logit_s <- tryCatch(glm(formula_sint, data = smote_train, family =  "binomial"),error=
    function(e)e,finally="BAD")

logit_d <- tryCatch(glm(formula_sint, data = down_train,  family =  "binomial"),error=
    function(e)e,finally="BAD")

##GLMmulti, usamos un algoritmo genetico
o<-capture.output(glmulti_aux_0 <-   tryCatch(glmulti(formula_sint,
data = data_train,
level = 1,method = "g",crit = "aicc",confsetsize = 5,plotty = F,
report = F,fitfunction = "glm",family=binomial),
error=function(e)e,finally="BAD"))

s<-capture.output(glmulti_aux_s <-   tryCatch(glmulti(formula_sint,
data = smote_train,
level = 1,method = "g",crit = "aicc",confsetsize = 5,plotty = F,
report = F,fitfunction = "glm",family=binomial),
error=function(e)e,finally="BAD"))

d<-capture.output(glmulti_aux_d <-   tryCatch(glmulti(formula_sint,
data = down_train,
level = 1,method = "g",crit = "aicc",confsetsize = 5,plotty = F,
report = F,fitfunction = "glm",family=binomial),
error=function(e)e,finally="BAD"))

#Defino  modelos
glmulti_0<-glmulti_aux_0@objects[[1]]
glmulti_d<-glmulti_aux_d@objects[[1]]
glmulti_s<-glmulti_aux_s@objects[[1]]


###CART- RPART###
tree_0_aux <- rpart(formula_sint, data_train)
tree_0 <- prune(tree_0_aux,cp =  tree_0_aux$cptable[which.min(
rowSums(tree_0_aux$cptable[, 4:5])), 1])

tree_s_aux <- rpart(formula_sint, smote_train)
tree_s <- prune(tree_s_aux,cp =  tree_s_aux$cptable[which.min(
rowSums(tree_s_aux$cptable[, 4:5])), 1])

tree_d_aux <- rpart(formula_sint, down_train)
tree_d <- prune(tree_d_aux,
cp =  tree_d_aux$cptable[which.min(
rowSums(tree_d_aux$cptable[, 4:5])), 1])



###RF###
Rf_0 <- tryCatch(randomForest(formula = formula_sint,
data = data_train,importance=F),error=function(e)e,finally="BAD")
Rf_s <- tryCatch(randomForest(formula = formula_sint,
data = smote_train,importance=F),error=function(e)e,finally="BAD")
Rf_d <- tryCatch(randomForest(formula = formula_sint,
data = down_train,importance=F),error=function(e)e,finally="BAD")

##Ctree ###
ctre_0<-tryCatch(ctree(formula = formula_sint,
data = data_train),error=function(e)e,finally="BAD")
ctre_s<-tryCatch(ctree(formula = formula_sint,
data = smote_train),error=function(e)e,finally="BAD")
ctre_d<-tryCatch(ctree(formula = formula_sint,
data = down_train),error=function(e)e,finally="BAD")

## C Random Forest###
c_Rf_0 <- tryCatch(cforest(formula = formula_sint,
data = data_train),error=function(e)e,finally="BAD")
c_Rf_s <- tryCatch(cforest(formula = formula_sint,
```

```
data = smote_train),error=function(e)e,finally="BAD")
c_Rf_d <- tryCatch(cforest(formula = formula_sint,
data = down_train),error=function(e)e,finally="BAD")


######Genero predicciones para cada modelo
##Logit##
if(any(class(logit_0)=="error"))
{Y_test$pred_logit_0=rep(NA, nrow(data_test)) }
else { Y_test$pred_logit_0=predict(logit_0,data_test,
type ="response")}
if(any(class(logit_0)=="error"))
{Y_train_simple$pred_logit_0=rep(NA, nrow(data_train)) }
else {Y_train_simple$pred_logit_0=predict(logit_0,data_train,
type ="response")}

if(any(class(logit_s)=="error"))
{Y_test$pred_logit_s=rep(NA, nrow(data_test)) }
else {Y_test$pred_logit_s=predict(logit_s,data_test,type ="response")}
if(any(class(logit_s)=="error"))
{Y_train_smote$pred_logit_s=rep(NA, nrow(smote_train)) }
else {Y_train_smote$pred_logit_s=predict(logit_s,smote_train,
type ="response")}

if(any(class(logit_d)=="error"))
{Y_test$pred_logit_d=rep(NA, nrow(data_test)) }
else {Y_test$pred_logit_d=predict(logit_d,data_test,type ="response")}
if(any(class(logit_d)=="error"))
{Y_train_down$pred_logit_d=rep(NA, nrow(down_train)) }
else {Y_train_down$pred_logit_d=predict(logit_d,down_train,
type ="response")}

##GLMmulti
if(any(class(glmulti_0)=="error")){Y_test$pred_glmmutli_0=rep(NA, nrow(data_test)) }
else {Y_test$pred_glmmutli_0=predict(glmulti_0,data_test,
type ="response")}
if(any(class(glmulti_0)=="error"))
{Y_train_simple$pred_glmmutli_0=rep(NA, nrow(data_train)) }
else {Y_train_simple$pred_glmmutli_0=predict(glmulti_0,data_train,
type ="response")}

if(any(class(glmulti_s)=="error"))
{Y_test$pred_glmmutli_s=rep(NA, nrow(data_test)) }
else { Y_test$pred_glmmutli_s=predict(glmulti_s,data_test,
type ="response")}
if(any(class(glmulti_s)=="error"))
{Y_train_smote$pred_glmmutli_s=rep(NA, nrow(smote_train)) }
else {Y_train_smote$pred_glmmutli_s=predict(glmulti_s,smote_train,
type ="response")}

if(any(class(glmulti_d)=="error"))
{Y_test$pred_glmmutli_d=rep(NA, nrow(data_test)) }
else {Y_test$pred_glmmutli_d=predict(glmulti_d,data_test,
type ="response")}
if(any(class(glmulti_d)=="error"))
{Y_train_down$pred_glmmutli_d=rep(NA, nrow(down_train)) }
else {Y_train_down$pred_glmmutli_d=predict(glmulti_d,down_train,
type ="response")}

##RPART
if(any(class(tree_0)=="error"))
{Y_test$pred_tree_0=rep(NA, nrow(data_test)) }
else {Y_test$pred_tree_0=predict(tree_0,data_test,type ="prob")[,2]}
if(any(class(tree_0)=="error"))
{Y_train_simple$pred_tree_0=rep(NA, nrow(data_train)) }
else {Y_train_simple$pred_tree_0=
predict(tree_0,data_train,type ="prob")[,2]}

if(any(class(tree_s)=="error"))
{Y_test$pred_tree_s=rep(NA, nrow(data_test)) }
else {Y_test$pred_tree_s=predict(tree_s,data_test,type ="prob")[,2]}
if(any(class(tree_s)=="error"))
{Y_train_smote$pred_tree_s=rep(NA, nrow(smote_train)) }
```

```
else {Y_train_smote$pred_tree_s=
predict(tree_s,smote_train,type ="prob")[,2]}

if(any(class(tree_d)=="error"))
{Y_test$pred_tree_d=rep(NA, nrow(data_test)) }
else {Y_test$pred_tree_d=predict(tree_d,data_test,type ="prob")[,2]}
if(any(class(tree_d)=="error"))
{Y_train_down$pred_tree_d=rep(NA, nrow(down_train)) }
else  {Y_train_down$pred_tree_d=
predict(tree_d,down_train,type ="prob")[,2]}


##RF
if(any(class(Rf_0)=="error"))
{Y_test$pred_RF_0=rep(NA, nrow(data_test)) }
else {Y_test$pred_RF_0=predict(Rf_0,data_test,type ="prob")[,2]}
if(any(class(Rf_0)=="error"))
{Y_train_simple$pred_RF_0=rep(NA, nrow(data_train)) }
else {Y_train_simple$pred_RF_0=predict(Rf_0,data_train,
type ="prob")[,2]}

if(any(class(Rf_s)=="error"))
{Y_test$pred_RF_s=rep(NA, nrow(data_test)) }
else {Y_test$pred_RF_s=predict(Rf_s,data_test,type ="prob")[,2]}
if(any(class(Rf_s)=="error"))
{Y_train_smote$pred_RF_s=rep(NA, nrow(smote_train)) }
else { Y_train_smote$pred_RF_s=predict(Rf_s,smote_train,
type ="prob")[,2]}

if(any(class(Rf_d)=="error"))
{Y_test$pred_RF_d=rep(NA, nrow(data_test)) }
else {Y_test$pred_RF_d=predict(Rf_d,data_test,type ="prob")[,2]}
if(any(class(Rf_d)=="error"))
{Y_train_down$pred_RF_d=rep(NA, nrow(down_train)) }
else {Y_train_down$pred_RF_d=predict(Rf_d,down_train,
type ="prob")[,2]}


####ctree
if(any(class(ctre_0)=="error"))
{Y_test$pred_ctree_0=rep(NA, nrow(data_test)) }
else {Y_test$pred_ctree_0=unlist(predict(ctre_0,data_test,
type ="prob",simplify=T))[seq(2,dim(data_test)[1]*2,2)]}
if(any(class(ctre_0)=="error"))
{Y_train_simple$pred_ctree_0=rep(NA, nrow(data_train)) }
else {Y_train_simple$pred_ctree_0=unlist(predict(ctre_0,data_train,
type ="prob",simplify=T))[seq(2,dim(data_train)[1]*2,2)]}

if(any(class(ctre_s)=="error"))
{Y_test$pred_ctree_s=rep(NA, nrow(data_test)) }
else {Y_test$pred_ctree_s=unlist(predict(ctre_s,data_test,
type ="prob",simplify=T))[seq(2,dim(data_test)[1]*2,2)]}
if(any(class(ctre_s)=="error"))
{Y_train_smote$pred_ctree_s=rep(NA, nrow(smote_train)) }
else {Y_train_smote$pred_ctree_s=unlist(predict(ctre_s,smote_train,
type ="prob",simplify=T))[seq(2,dim(smote_train)[1]*2,2)]}

if(any(class(ctre_d)=="error"))
{Y_test$pred_ctree_d=rep(NA, nrow(data_test)) }
else {Y_test$pred_ctree_d=unlist(predict(ctre_d,data_test,
type ="prob",simplify=T))[seq(2,dim(data_test)[1]*2,2)]}
if(any(class(ctre_d)=="error"))
{Y_train_down$pred_ctree_d=rep(NA, nrow(down_train)) }
else { Y_train_down$pred_ctree_d=unlist(predict(ctre_d,down_train,
type ="prob",simplify=T))[seq(2,dim(down_train)[1]*2,2)]}


##Random forest ctree
if(any(class(c_Rf_0)=="error"))
{Y_test$pred_c_RF_0=rep(NA, nrow(data_test)) }
else {Y_test$pred_c_RF_0=unlist(predict(c_Rf_0,
newdata=data_test,type ="prob"))[seq(2,dim(data_test)[1]*2,2)]}
if(any(class(c_Rf_0)=="error"))
```

```
{Y_train_simple$pred_c_RF_0=rep(NA, nrow(data_train)) }
else {Y_train_simple$pred_c_RF_0=unlist(predict(c_Rf_0,
newdata=data_train,type ="prob"))[seq(2,dim(data_train)[1]*2,2)]}

if(any(class(c_Rf_s)=="error"))
{Y_test$pred_c_RF_s=rep(NA, nrow(data_test)) }
else { Y_test$pred_c_RF_s=unlist(predict(c_Rf_s,newdata=data_test,
type ="prob"))[seq(2,dim(data_test)[1]*2,2)]}
if(any(class(c_Rf_s)=="error"))
{Y_train_smote$pred_c_RF_s=rep(NA, nrow(smote_train)) }
else {Y_train_smote$pred_c_RF_s=unlist(predict(c_Rf_s,
newdata=smote_train,type ="prob"))[seq(2,dim(smote_train)[1]*2,2)]}

if(any(class(c_Rf_d)=="error"))
{Y_test$pred_c_RF_d=rep(NA, nrow(data_test)) }
else {Y_test$pred_c_RF_d=unlist(predict(c_Rf_d,newdata=data_test,type ="prob"))[seq(2,dim(
    data_test)[1]*2,2)]}
if(any(class(c_Rf_d)=="error")){Y_train_down$pred_c_RF_d=rep(NA, nrow(down_train))} else {
    Y_train_down$pred_c_RF_d=unlist(predict(c_Rf_d,
newdata=down_train,type ="prob"))[seq(2,dim(down_train)[1]*2,2)]}

###CALCULO METRICAS##
corte_05=0.5

## MAtrix con los resultados, utilizo la fun Performa_R
una_iter<-rbind(

##Random Forest
cbind(performa_R(Y_test$pred_RF_0,Y_test$Aprobo,
Y_train_simple$pred_RF_0,Y_train_simple$Aprobo,
corte_05),Modelo="RF",Estr="Simple"),
cbind(performa_R(Y_test$pred_RF_s,Y_test$Aprobo,
Y_train_smote$pred_RF_s,Y_train_smote$Aprobo,
corte_05),Modelo="RF",Estr="Smote"),
cbind(performa_R(Y_test$pred_RF_d,Y_test$Aprobo,
Y_train_down$pred_RF_d,Y_train_down$Aprobo,
corte_05),Modelo="RF",Estr="down"),

##TREE
cbind(performa_R(Y_test$pred_tree_0,Y_test$Aprobo,
Y_train_simple$pred_tree_0,Y_train_simple$Aprobo,
corte_05), Modelo="rpart",Estr="Simple"),
cbind(performa_R(Y_test$pred_tree_s,Y_test$Aprobo,
Y_train_smote$pred_tree_s,Y_train_smote$Aprobo,
corte_05),Modelo="rpart",Estr="Smote"),
cbind(performa_R(Y_test$pred_tree_d,Y_test$Aprobo,
Y_train_down$pred_tree_d,Y_train_down$Aprobo,
corte_05),Modelo="rpart",Estr="down"),

#Conditional tree
cbind(performa_R(Y_test$pred_ctree_0,Y_test$Aprobo,
Y_train_simple$pred_ctree_0,Y_train_simple$Aprobo,
corte_05), Modelo="Ctree",Estr="Simple"),
cbind(performa_R(Y_test$pred_ctree_s,Y_test$Aprobo,
Y_train_smote$pred_ctree_s,Y_train_smote$Aprobo,
corte_05),Modelo="Ctree",Estr="Smote"),
cbind(performa_R(Y_test$pred_ctree_d,Y_test$Aprobo,
Y_train_down$pred_ctree_d,Y_train_down$Aprobo,
corte_05),Modelo="Ctree",Estr="down"),

#Conditional Random  Forest
cbind(performa_R(Y_test$pred_c_RF_0,Y_test$Aprobo,
Y_train_simple$pred_c_RF_0,Y_train_simple$Aprobo,
corte_05), Modelo="CRF",Estr="Simple"),
cbind(performa_R(Y_test$pred_c_RF_s,Y_test$Aprobo,
Y_train_smote$pred_c_RF_s,Y_train_smote$Aprobo,
corte_05),Modelo="CRF",Estr="Smote"),
cbind(performa_R(Y_test$pred_c_RF_d,Y_test$Aprobo,
Y_train_down$pred_c_RF_d,Y_train_down$Aprobo,
corte_05), Modelo="CRF",Estr="down"),

#Logit
cbind(performa_R(Y_test$pred_logit_0,Y_test$Aprobo,
```

```
Y_train_simple$pred_logit_0 , Y_train_simple$Aprobo ,
corte_05 ), Modelo="GLM" , Estr="Simple" ) ,
cbind( performa_R( Y_test$pred_logit_s , Y_test$Aprobo ,
Y_train_smote$pred_logit_s , Y_train_smote$Aprobo ,
corte_05 ), Modelo="GLM" , Estr="Smote" ) ,
cbind( performa_R( Y_test$pred_logit_d , Y_test$Aprobo ,
Y_train_down$pred_logit_d , Y_train_down$Aprobo ,
corte_05 ), Modelo="GLM" , Estr="down" ) ,

#Glm multi
cbind( performa_R( Y_test$pred_glmmutli_0 , Y_test$Aprobo ,
Y_train_simple$pred_glmmutli_0 , Y_train_simple$Aprobo ,
corte_05 ) , Modelo="GLM_m" , Estr="Simple" ) ,
cbind( performa_R( Y_test$pred_glmmutli_s , Y_test$Aprobo ,
Y_train_smote$pred_glmmutli_s , Y_train_smote$Aprobo ,
corte_05 ), Modelo="GLM_m" , Estr="Smote" ) ,
cbind( performa_R( Y_test$pred_glmmutli_d , Y_test$Aprobo ,
Y_train_down$pred_glmmutli_d , Y_train_down$Aprobo ,
corte_05 ), Modelo="GLM_m" , Estr="down" ))

if (i==1) {
Salida<-una_iter
} else {
Salida=rbind( Salida , una_iter )}
}
  ##Cierro el loop
    return( Salida )
}
```

Estadísticos de resultado

```

performa_R<-function( hat_test , obs_test , hat_train , obs_train , corte_05 ){
   require( pROC )
   require( SDMTools )
   obs_test_2=ifelse( obs_test=="No" ,0 ,1 )
   obs_train_2=ifelse( obs_train=="No" ,0 ,1 )

acc_05<-accuracy( obs_test_2 , hat_test , corte_05 )
AUC_P_05<- roc( obs_test_2 , hat_test )$auc
P_AUC_100a90<-roc( obs_test_2 , hat_test , partial.auc = c(100, 90),
partial.auc.correct = TRUE, percent = TRUE)$auc

     salida<-rbind( cbind( TIPO="0.5" , acc_05 , AUC_true=AUC_P_05 , Partial_auc=P_AUC_100a90 ))

   return( salida )
}
```

Muestreo

SMOTE en variables continuas

```
dat = iris[c(2, 9, 16, 23, 51:63),
           c(1, 2, 5)]

library(DMwR)
library("plotrix")
library(smotefamily)

dat$Species<-as.factor(as.numeric(dat$Species))
SMOTE_DATA<-SMOTE( Species ~ ., data = dat , k=2)

plot(dat[,-3], col=as.numeric(dat[,3])+1)
plot(SMOTE_DATA[,-3], col=as.numeric(SMOTE_DATA[,3])+1)


dat_plot = smotefamily::SMOTE(dat[,-3],  # feature values
                  as.numeric(dat[,3]),  # class labels
```

```
                    K = 2, dup_size = 0)  # function parameters

setosa<-dat[1:4,]

plot(dat[,-3],col=as.numeric(dat[,3])+1, main="Datos originales" )

plot(dat[,-3],col=as.numeric(dat[,3])+1,main="Espacio de simulaci[U+FFFD]n")
s <- seq(length(setosa))  # one shorter than data
segments(setosa[s,1],setosa[s,2],
         setosa[s+1,1],setosa[s+1,2])
segments(setosa[3,1],setosa[3,2],
         setosa[1,1],setosa[1,2])
segments(setosa[2,1],setosa[2,2],
         setosa[4,1],setosa[4,2])
segments(setosa[2,1],setosa[2,2],
         setosa[4,1],setosa[4,2])
segments(setosa[2,1],setosa[2,2],
         setosa[1,1],setosa[1,2])
segments(setosa[1,1], setosa[1,2],
         setosa[4,1], setosa[4,2])
dev.off()
AUXI<-rep(1,nrow(dat))
AUXI[1]<-0

plot(dat[,-3],col=as.numeric(dat[,3])+1,main="Distancia y vecinos k=2",
     pch=AUXI)
draw.circle(setosa[1,1], setosa[1,2], dist(setosa[c(1,4),-3])[1])
segments(setosa[1,1], setosa[1,2],
         setosa[4,1], setosa[4,2])
segments(setosa[1,1], setosa[1,2],
         setosa[2,1], setosa[2,2])

plot(dat[,-3],col=as.numeric(dat[,3])+1,main="Distancia y  gap",
     pch=AUXI)
segments(setosa[1,1], setosa[1,2],
         setosa[2,1], setosa[2,2])
text(setosa[1,-3]-.3,"dist=0.509")

plot(dat[,-3],col=as.numeric(dat[,3])+1,main="Caso nuevo",
     pch=AUXI)
segments(setosa[1,1], setosa[1,2],
         setosa[2,1], setosa[2,2])
points(SMOTE_DATA[24,-3],pch="*",cex=2)

plot(SMOTE_DATA[,-3])
```

SMOTE binario y distancia Gower

```

library(xtable)
library(cluster)

dat = iris[c(1,2, 9, 16, 23, 51:63),
           c(1, 2, 5)]

dat$Species<-as.factor(as.numeric(dat$Species))

dat$Sepal.Width_bin <-ifelse(dat$Sepal.Width<=3,1,0)
dat<-subset(dat,select=-Sepal.Width)
dat$Sepal.Width_bin<-as.factor(dat$Sepal.Width_bin)
xtable(dat)

dat_1<-subset(dat,Species==1)
matr<-daisy(dat_1,metric = "gower")
matr<-as.matrix(matr)
xtable(matr)
```

Efectos de Muestreo

```
library(MASS)
```

```
library(caret)
library(DMwR)
library(sjPlot)
library(sjmisc)

mu <- c(1,1)
Sigma <- matrix(c(1, .5, .5, 1), 2)
bivn_11 <- mvrnorm(250, mu = mu, Sigma = Sigma )

mu <- c(-1,-1)
bivn_00 <- mvrnorm(50, mu = mu, Sigma = Sigma )

MATRIX<-as.data.frame(rbind(cbind(bivn_00,grupo=0),
                cbind(bivn_11,grupo=1)))

set_theme(theme = "scatter",
              geom.label.size = 3.5,
              axis.textsize = .85,
              axis.title.size = .85,
              geom.alpha = .4)

sjp.scatter(MATRIX$V1, MATRIX$V2,MATRIX$grupo)


MATRIX$grupo<-as.factor(MATRIX$grupo)
smote_train <- SMOTE(grupo ~ ., data  = MATRIX)
down_train <- downSample(x = MATRIX[, 1:2],y = MATRIX$grupo)
down_train$grupo=down_train$Class
up_train<-upSample(x = MATRIX[, 1:2],y = MATRIX$grupo)
up_train$grupo=up_train$Class

table(up_train$grupo)
table(down_train$grupo)
table(smote_train$grupo)

sjp.scatter(MATRIX$V1, MATRIX$V2,MATRIX$grupo,
title="Datos originales")
sjp.scatter(up_train$V1, up_train$V2,up_train$grupo,
title="Up sampling")
sjp.scatter(down_train$V1, down_train$V2,down_train$grupo,
title="Down sampling")
sjp.scatter(smote_train$V1, smote_train$V2,smote_train$grupo,
title="SMOTE sampling")
```