Análisis predictivo y exploratorio de abandono en secundaria

Natalia da Silva & Juan José Goyeneche
IESTA, Universidad de la República

Estructura

- Motivación y objetivos
- Exploración de los datos
- Modelos predictivos, bosques aleatorios con datos desbalanceados
- Comentarios finales

Motivación

Educación formal en Uruguay:

- Educación preprimaria e inicial (entre 3 y 5 años)
- Educación primaria (entre 6 y 11 años)
- Educación secundaria (entre 12 y 17 años)
- Educación terciaria o superior (18 años o más)

Motivación

Si bien se ha logrado tener un nivel importante de universalización de la educación primaria, se encuentran serios problemas en la educación secundaria para lograr retener a los estudiantes en el sistema.

Objetivo

El primer año de secundaria es el que presenta mayores niveles de abandono.

- Analizar la población en riesgo de abandono es fundamental para entender el problema y obtener información relevante para el desarrollo de políticas públicas.
- Explorar y predecir el abandono de los alumnos pertenecientes a primer año de educación secundaria pública en Uruguay.

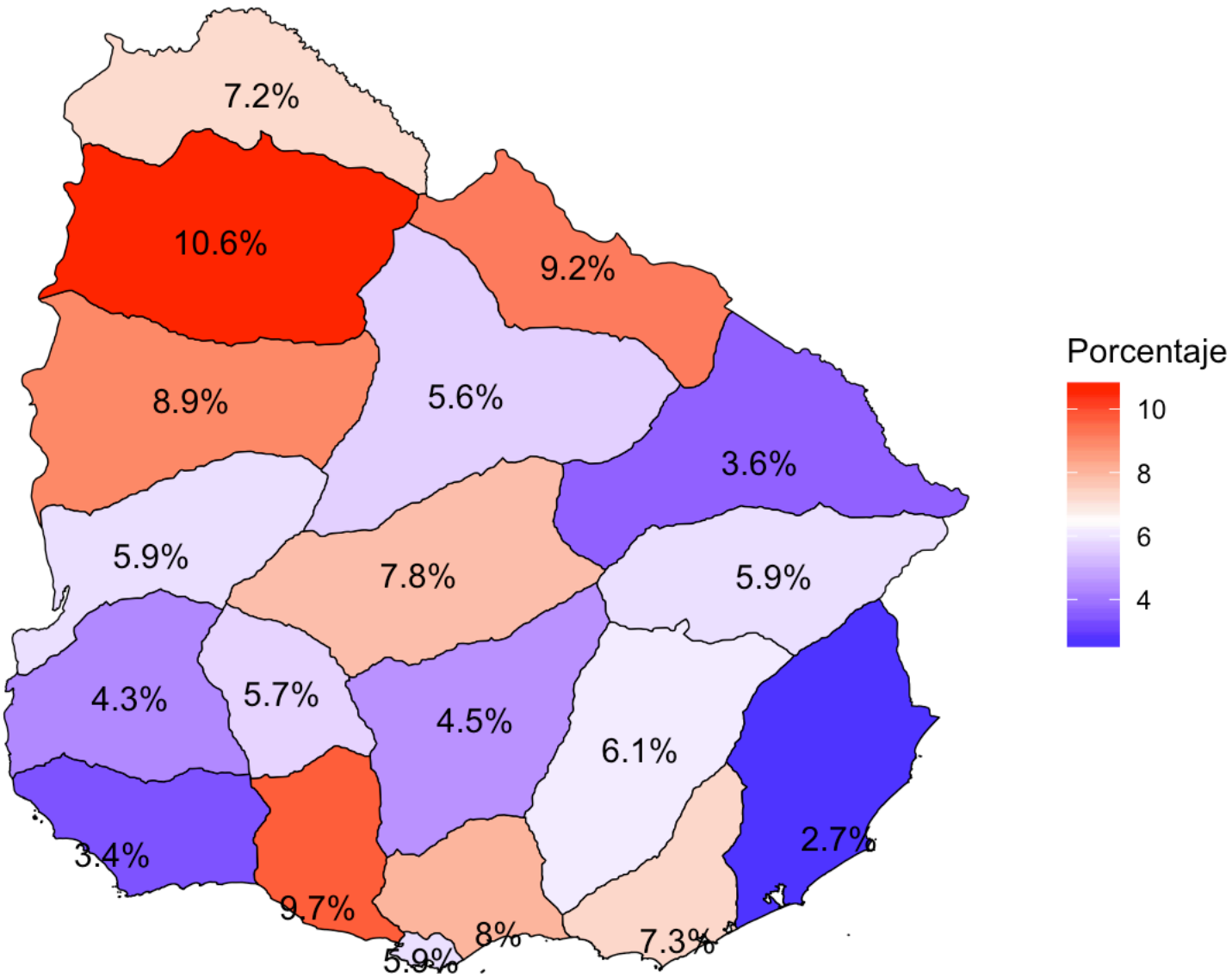
Población

- Población: estudiantes de primero de educación secundaria que cursan primero en el año 2016.
- Dicha población se compone de una matrícula de 40.233 alumnos, que pertenecen a 254 centros educativos.
- Se dispone de las transiciones de dichos alumnos entre los años 2016 y 2017.

Abandono

- Abandono: si cursó primero de CES en 2016 y en 2017 no se anotó en el sistema de educación pública.
- Limitante no hay datos de educación privada.
- Usando esta definición hay 7% de abandono en 2016-2017

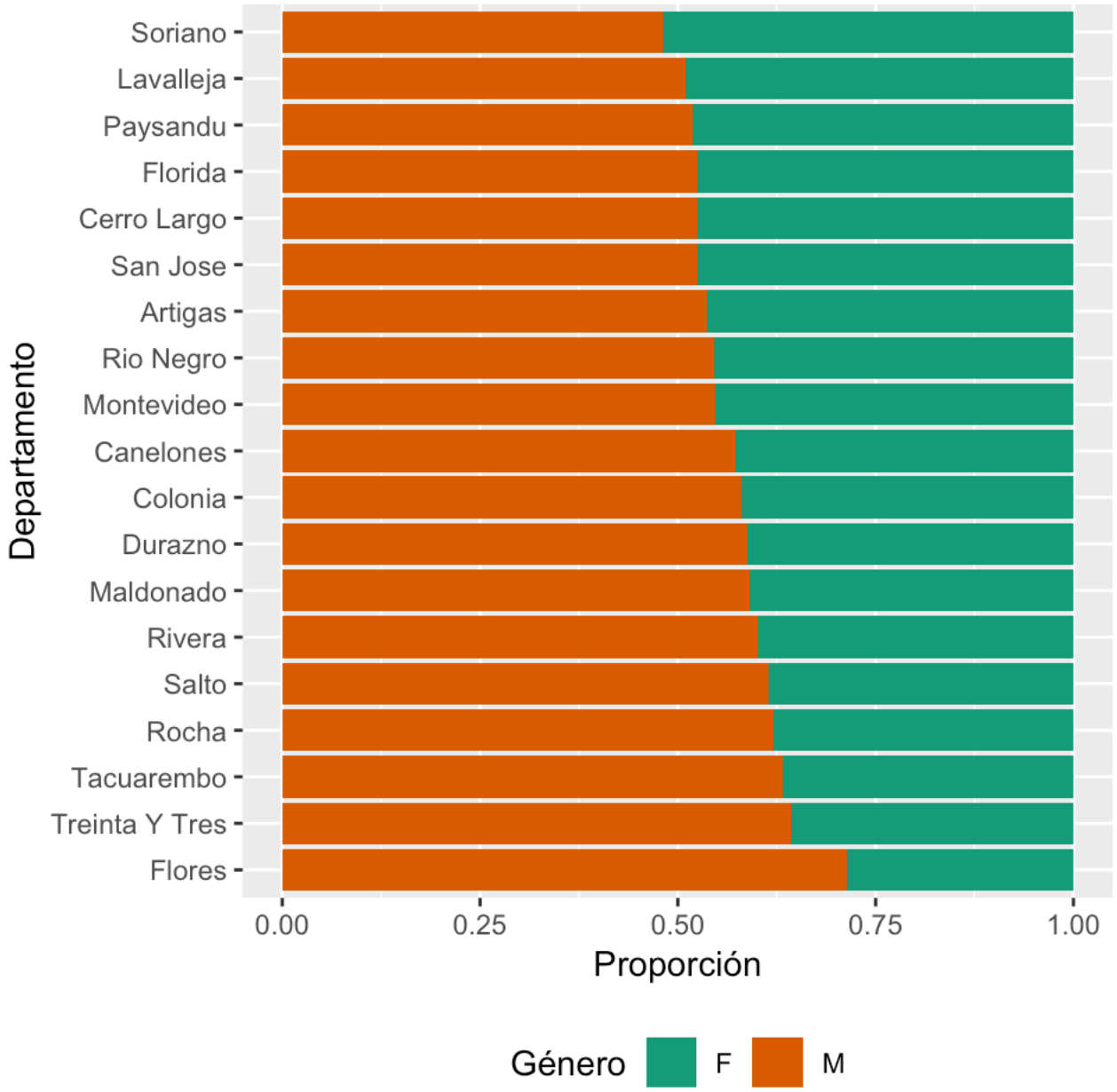
Abandono 1ero 2016-2017



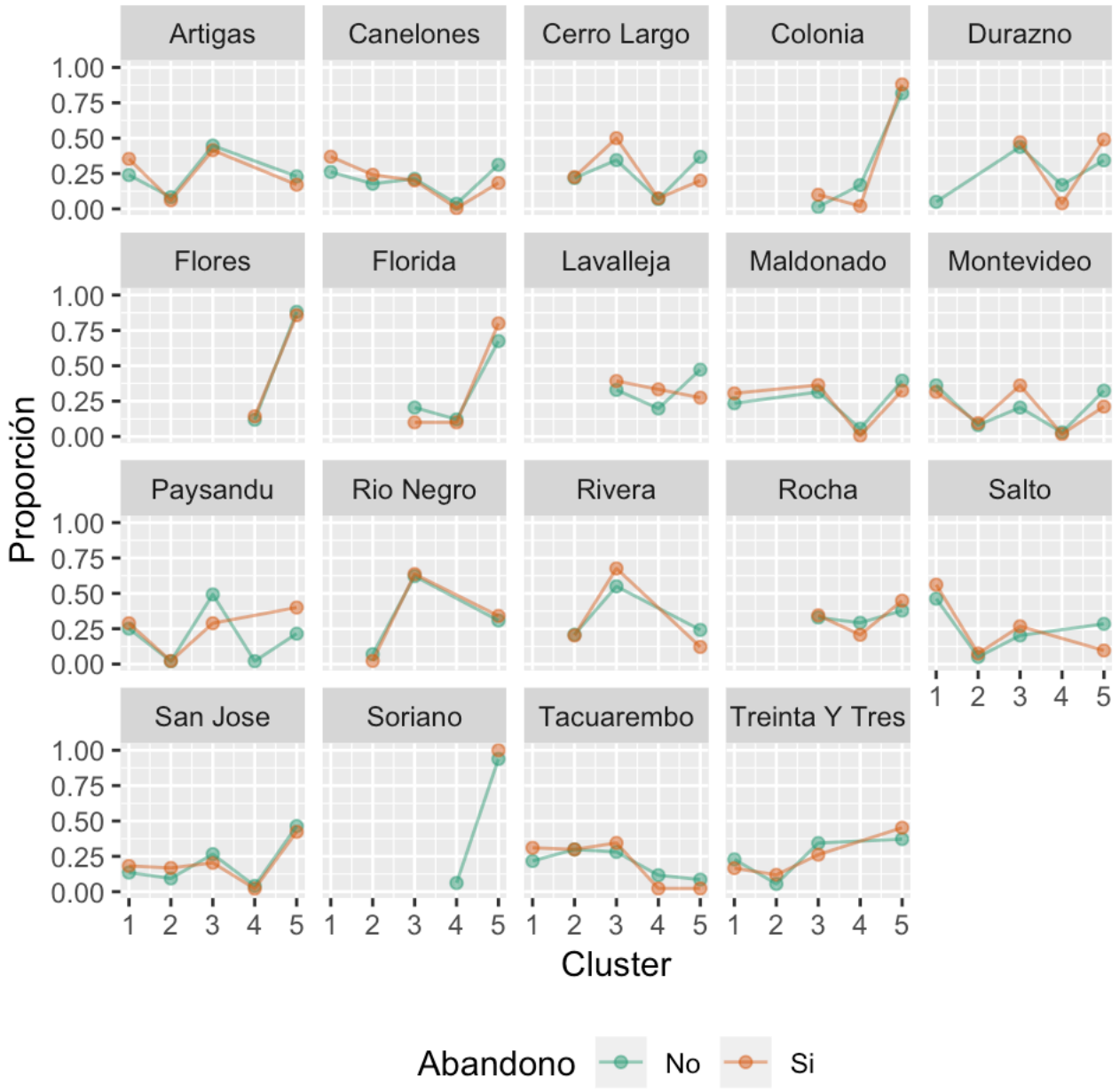
Abandono 1ero 2016-2017

- La distribución de los alumnos que no abandonan es similar para mujeres y hombres (50.45% mujeres y 49.55 % hombres).
- Para los que abandonan, hay un porcentaje mayor de alumnos de sexo masculino (57 %) que femeninos (43 %).
- ¿Existen diferencias en la distribución del abandono según género a nivel departamental?

Abandono por departamento y género



Contextos socioculturales y abandono



Modelos predictivos

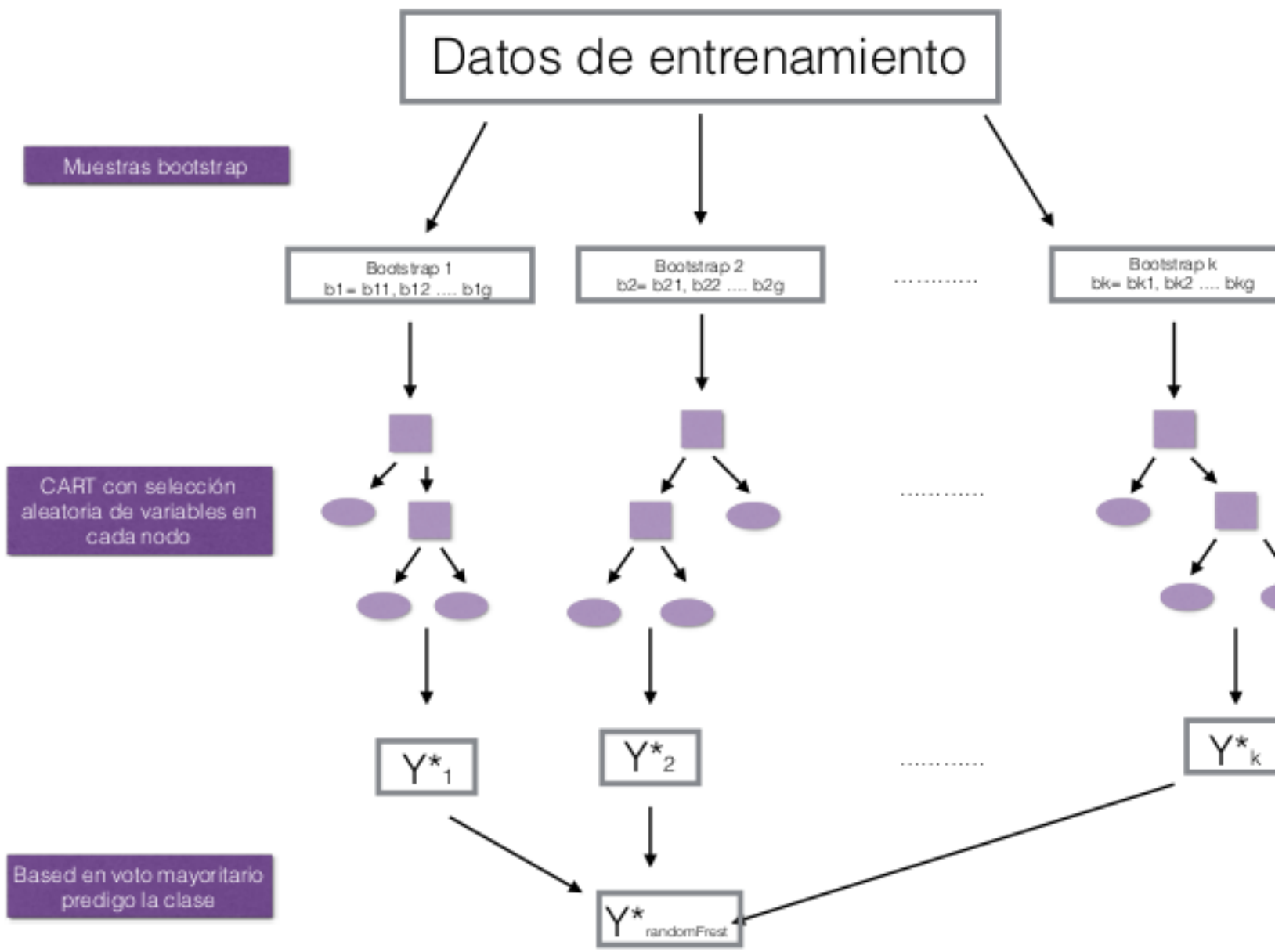
Modelos de clasificación para el abandono

- Variable de respuesta categórica con dos niveles (abandona, no abandona).
- Variables explicativas: Contexto sociocultural, sexo, extra edad fuerte, extra edad leve, inasistencias relativas, centro educativo, departamento.
- Problema: los datos están muy desbalanceados sólo un 7% abandonan
- Distintas estrategias usando bosques aleatorios (RF)

Bosque aleatorio, RF

Bosque aleatorio es un método de agregación supervisado basado en combinar modelos individuales de tipo árbol. Dos fuentes de aleatoriedad son introducidas, agregación bootstrap y selección aleatoria de variables en la partición del nodo.

Diagrama bosque aleatorio clásico



Datos desbalanceados

- Los algoritmos de ML no tienen mucha información de la clase minoritaria para tener buenas predicciones en la clase minoritaria.
- Los algoritmos son guiados por la precisión, minimizan el error global donde la clase minoritaria tiene poco peso.
- Los algoritmos de ML asumen que las clases están balanceadas.
- También asumen que los errores obtenidos de diferentes clases tienen el mimo costo.

Random Forest, clásico

```
##
## Call:
##  randomForest(formula = Abandono ~ nro_doc_centro_educ
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 5.69%
## Confusion matrix:
##       0   1 class.error
## 0 26503  63 0.002371452
## 1  1543 123 0.926170468
```

Random Forest, datos desbalanceados

- Podemos incorporar pesos a las clases y penalizar el error de clasificación en la clase minoritaria (WRF)
- Combinar técnicas de muestreo y métodos de agregación. Sub muestrea la clase mayoritaria y crece los árboles con muestras más balanceadas (BRF).

Random Forest ponderado, WRF

- Aprendizaje sensible al costo
- Debemos incorporar mayor penalidad al error de clasificación en la clase minoritaria.
- En WRF los pesos se traducen en dos lugares: usa pesos para encontrar las particiones de cada árbol. En los nodos terminales de cada árbol se ponderan las clases y se usa voto mayoritario ponderado.

Random Forest ponderado (WRF)

```
##
## Call:
##  randomForest(formula = Abandono ~ nro_doc_centro_educ
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 38.33%
## Confusion matrix:
##       0     1 class.error
## 0 15909 10657  0.40115185
## 1   163  1503  0.09783914
```

Random Forest, balanceado (BRF)

- Saco muestras bootstrap para la clase minoritaria y aleatoriamente con reposición, selecciono el mismo número de casos para la clase mayoritaria.
- Ajusto CART para cada muestra bootstrap y selecciono aleatoriamente un subconjunto de variables para la partición de cada nodo
- Basado en voto mayoritario obtengo la predicción del bosque

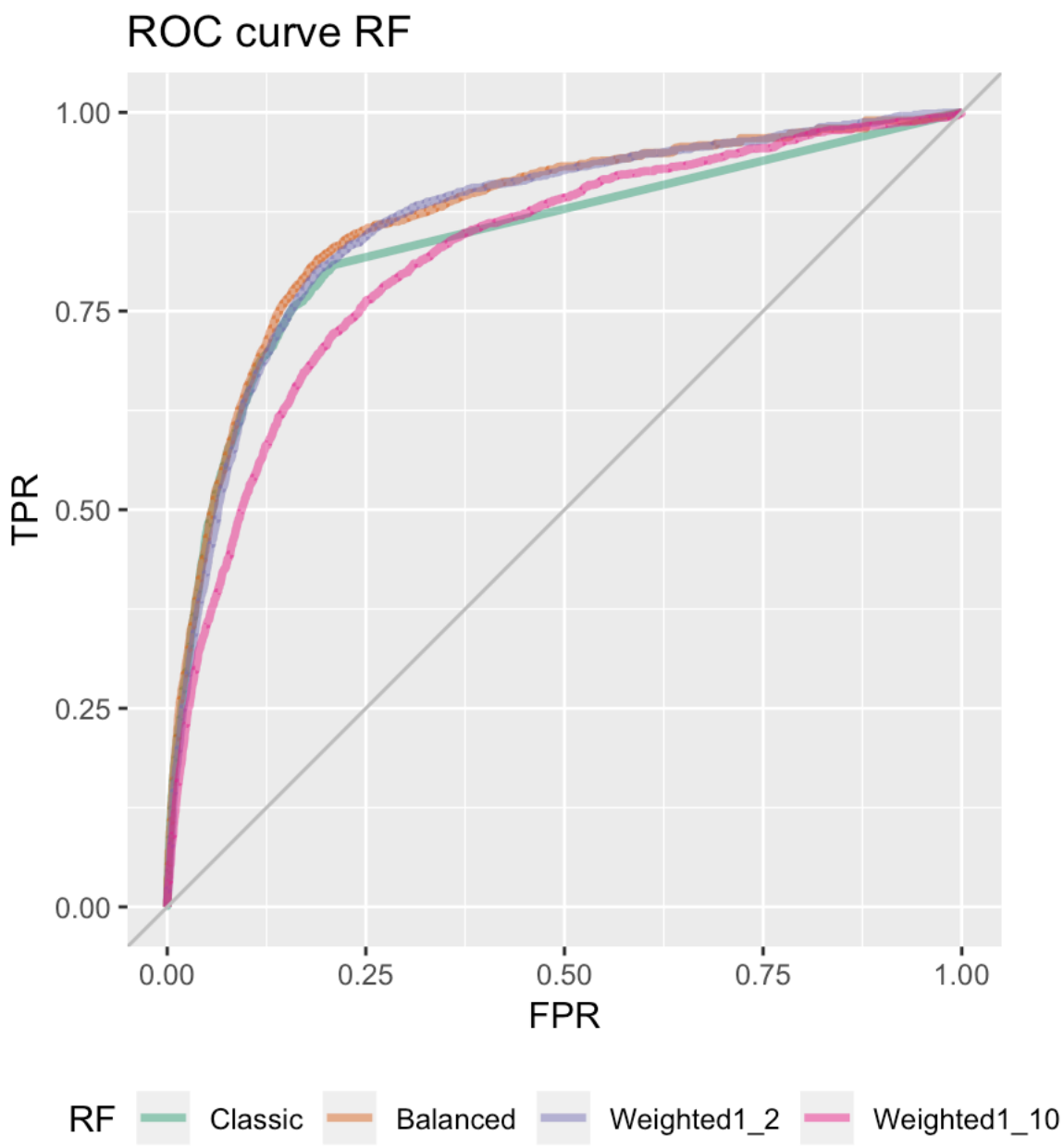
Random Forest, balanceado (BRF)

```
##
## Call:
##  randomForest(formula = Abandono ~ nro_doc_centro_educ
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 17.76%
## Confusion matrix:
##       0    1 class.error
## 0 21893 4673   0.1759015
## 1   340 1326   0.2040816
```

Random Forest, ROC

- Comparar la performance, uso ROC (Receiver operating characteristic curve)
- ROC es una representación gráfica que muestra el trade off entre los TPR y FPR para todos los puntos de corte

Random Forest, ROC



Comentarios Finales

Datos desbalanceados:

- Importante usar modelos que tomen en cuenta esta característica
- Seleccionar medidas de performance apropiadas para evaluar los modelos
- Seleccionar modelos en base a nuestro interés en el problema particular
- En nuestro caso reducir los Falsos negativos (que abandone y clasificarlo erróneamente)

Información e Invitación

1. Slides: https://github.com/natydasilva/riocuarto
2. email: natalia@iesta.edu.uy, twitter:@pacocuak, webpage: http://natydasilva.com
3. LatinR: Conferencia sobre el uso y desarrollo de R, 3 al 5 de Setiembre en Buenos Aires
4. Web: https://latinr.github.io, twitter: @LatinR2018
5. Querés saber qué es R-Ladies, preguntame