

# Ceibal en inglés: Modelos Predictivos para el resultado del pruebas adaptativas de inglés

## Ciencia de datos en plataformas educativas

Estudiar relación entre uso de plataformas educativas y desempeño educativo.

En el marco de Ceibal en Inglés.

- ▶ Prueba Adaptativa de Inglés
- ▶ Plataforma Little Bridge

**Modelos predictivos** del desempeño brinda información sobre determinantes de buen (mal) desempeño, puede disparar intervenciones tempranas.

Ciencia de datos en plataformas educativas

Fuentes de información

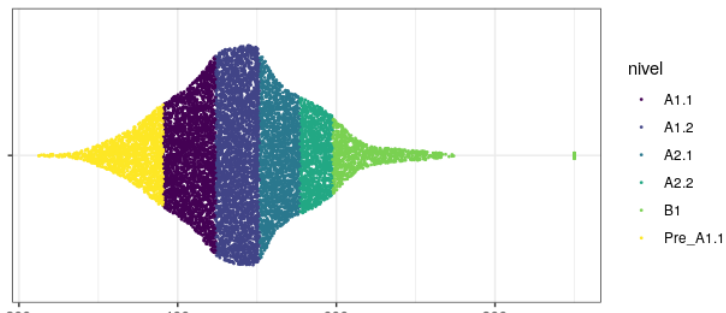
Modelos predictivos

## Datos de resultados

### Prueba adaptativa de Inglés

- ▶ Componentes evaluados: Vocabulario-gramática (VG) y Lectura (R)
- ▶ entre 15 Noviembre y 15 Diciembre (aprox.)
- ▶  $\approx$  35 mil estudiantes

## Datos de resultados



Hay 12% de estudiantes que no alcanzan el nivel A1.1

## Datos de uso

### Plataforma Little Bridge

- ▶ plataforma educativa e interactiva de inglés
- ▶ sistema de evaluación automático
- ▶ para 4º, 5º y 6º año de Primaria

### Datos para 2021

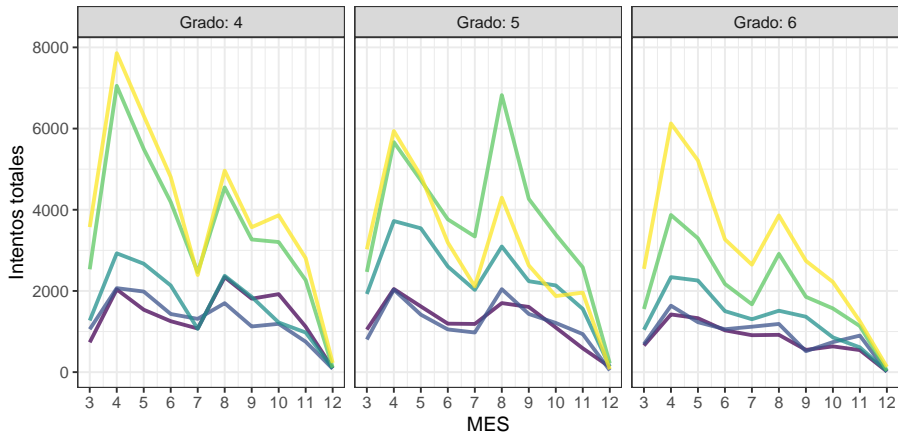
- ▶  $\approx$  70 mil estudiantes
- ▶ Actividad en LB por día
- ▶ Información de efecto docente

## Variables LB

##	Act	min.pts	max.pts	ActTot	Preguntas	Correctas
## 1	act_32	0.50	0.50	1	10	5
## 2	act_32	0.50	0.50	1	10	5
## 3	act_33	1.00	1.00	1	2	2
## 4	act_402	1.00	1.00	1	1	1
## 5		NA	NA	NA	NA	NA
## 6	act_16	0.30	0.60	2	20	9
## 7	act_18	1.00	1.00	1	12	12
## 8	act_19	1.00	1.00	1	5	5
## 9	act_20	0.88	0.88	1	8	7
## 10	act_21	1.00	1.00	1	5	5

Otras variables: estudiante, centro, contexto, fecha, ....

# Intentos por mes



— Quintil 1 — Quintil 2 — Quintil 3 — Quintil 4 — Quintil 5



**Ciencia de datos en plataformas educativas**

**Fuentes de información**

**Modelos predictivos**

## Modelos predictivos

¿Es posible predecir el resultado de la prueba en base a la información de uso de plataformas ?

### Variable de respuesta

- ▶ Puntos en la prueba, Nivel obtenido, Obtiene nivel suficiente

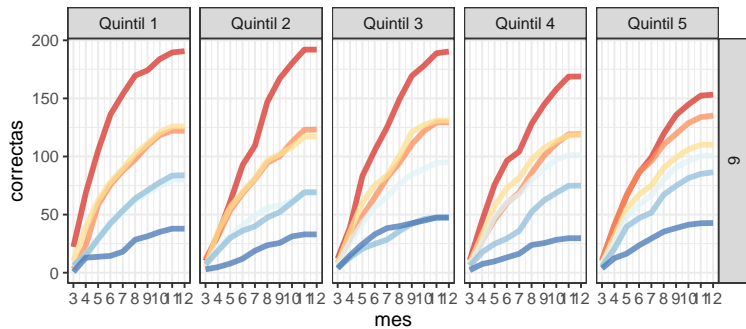
### Variable explicativas

- ▶ Uso en LB por mes, Uso de CREA (en inglés), Otras

### Muestra

- ▶ Quedan  $\approx$  9000 estudiantes con información de LB y la prueba

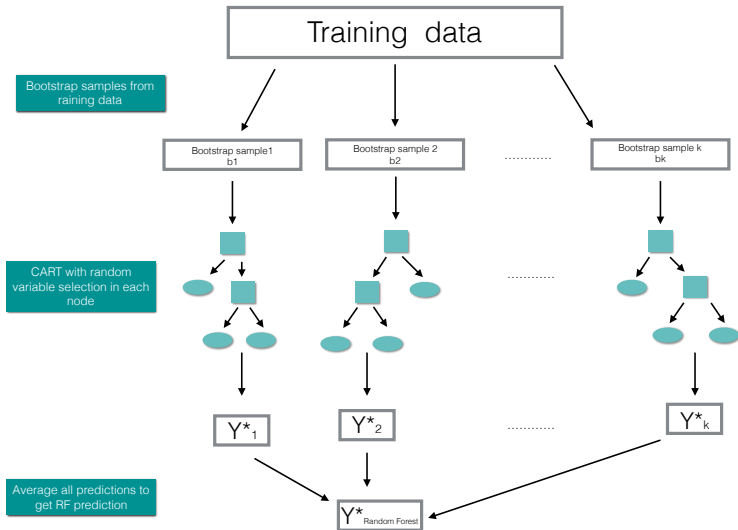
# Preguntas correctas en LB y nivel obtenido



— B1    — A2.1    — A1.1  
— A2.2    — A1.2    — Pre\_A1.1

9

# Random Forest

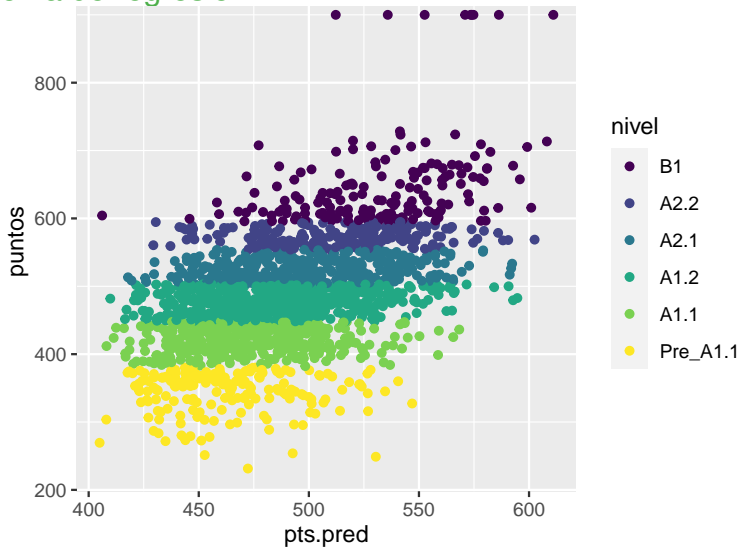


## Problema de regresión

- ▶ biblioteca `ranger`
- ▶ usamos importancia de variables

	rmse	mape	met
1	74.01	12.26	rf
2	74.17	12.31	rf.imp
3	73.43	12.20	rf.tun

## Problema de regresión



## Clasificación binaria

- ▶ Para complementar el modelo continuo en la cola inferior.
- ▶ seguimos usando `ranger`
- ▶ clases no balanceadas

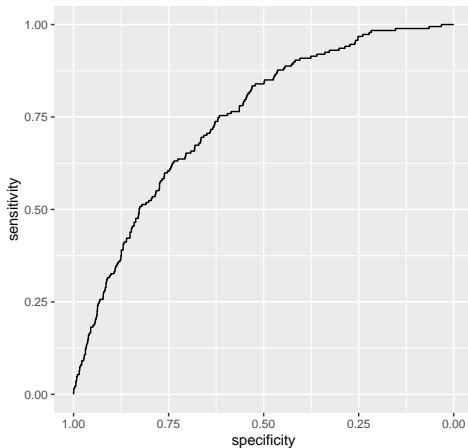
Bosque Balanceado: sub-muestras ponderadas en cada árbol

# Clasificación binaria

Matriz de  
confusión

	Insuf	Suf
Suf	0.30	0.70
Insuf	0.65	0.35

Curva ROC con AUC= 0.75





## Siguientes pasos

Continuar mejorando performance predictiva!

- ▶ Incorporar otras fuentes de información
- ▶ Mejorar variables de uso
- ▶ Trabajar con modelos a medida tomando en cuenta la estructura de los datos.

## Efecto de clase

El proceso de aprendizaje sucede en clases, ¿cómo lo reflejamos en los modelos?

Para un modelo lineal:  $y_{il} = x_i\beta + u_l + \epsilon_{il}$ , tenemos que

$$\text{Cov}(y_{il}, y_{jl}) = \sigma_u \quad \text{donde } u_l \sim N(0, \sigma_u^2)$$

¿De que manera se incorpora este tipo de efectos en algoritmos como *Random Forest*?