Indicadores de riesgo en el rendimiento académico en grados de ingeniería

Primeras aproximaciones

Carlos de la Calle

22/10/2019

El problema

Analizar los datos de los estudiantes de la Escuela de Ingeniería Industrial y Aeroespacial de Toledo en la asignatura de estadística.

UCLM
UNIVERSIDAD DE CASTILLA-LA MANCHA
Escuela de Ingeniería Industrial de Toledo

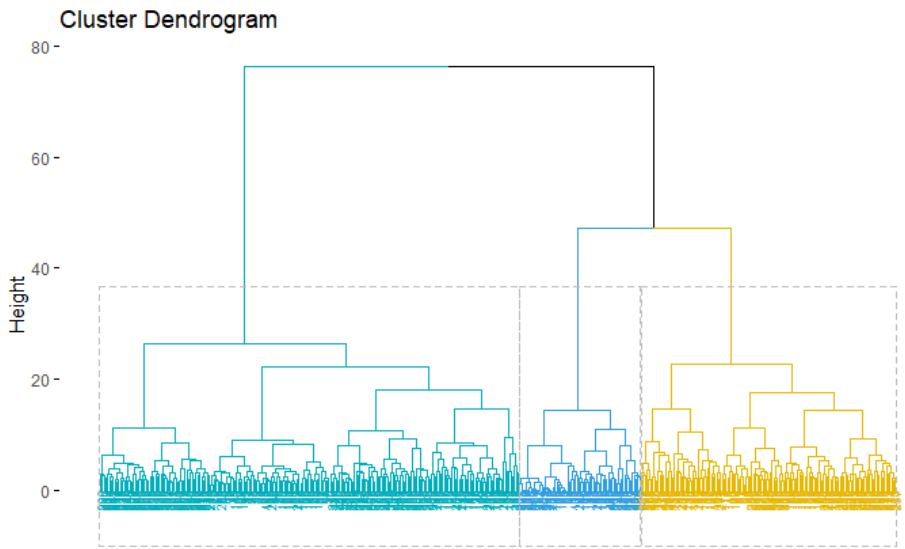
Los datos

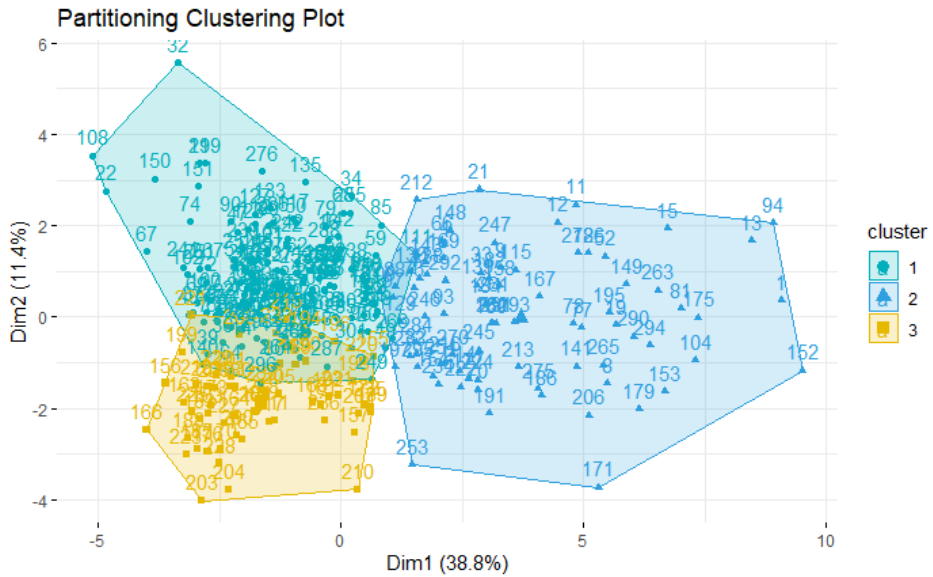
Los datos han sido recogido entre el curso 2015/16 y 2018/19, facilitados por los alumnos de forma voluntaria o no dependiendo de la variable, e incluye información de:

- Horas de estudio, asistencia y participación
- Resultados académicos: pruebas y exámenes, prácticas, preguntas, problemas...
- Cuestionario en escala likert

Análisis no supervisado

Se han probado distintos tipos de clustering jerárquico (aglomerativo: *hclust* y *agnes* y divisivo: *diana*), así como k-means, y se han valorado el número óptimo de clusters con *NbClust*.





Algunos resultados

```
kmeans(VariablesExplicativas, 3, nstart = 25)
```

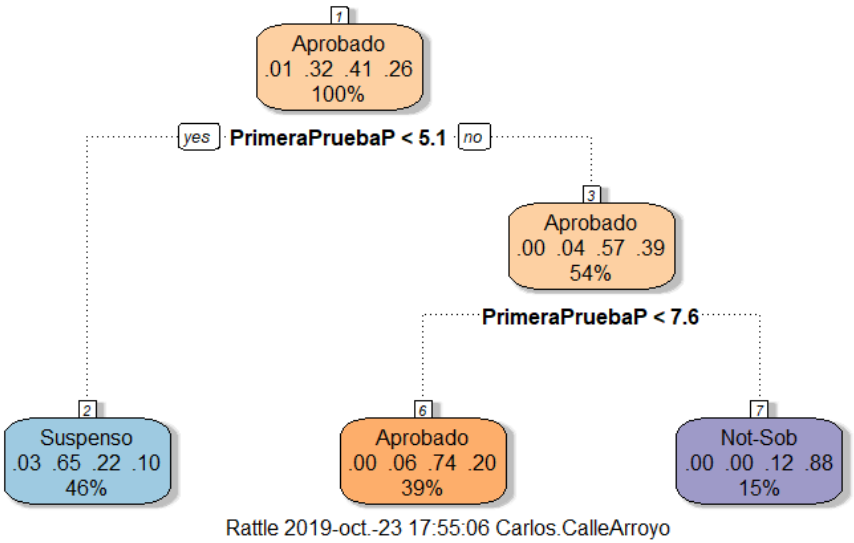
		No Presentado	Suspenso	Aprobado	Not-Sob
Cluster	1	2	33	83	45
	2	2	50	21	8
	3		13	20	25
	#Total cases	4	96	124	78

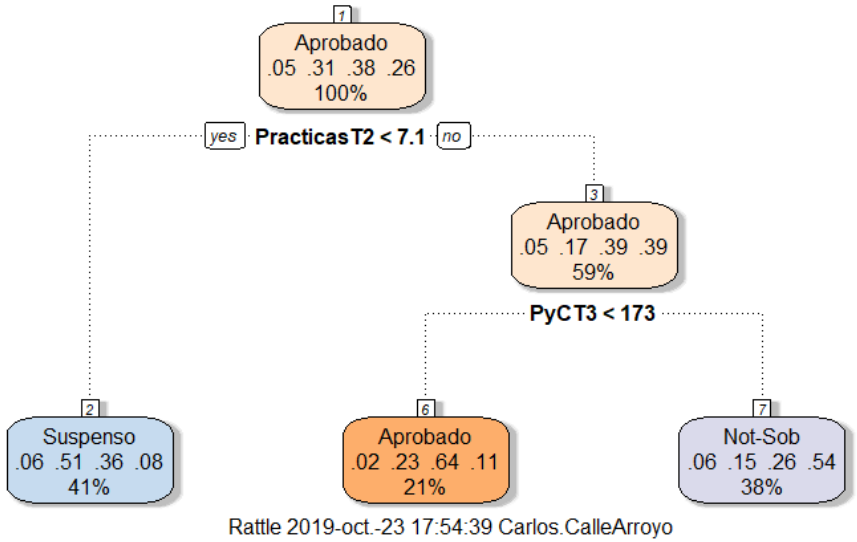
```
hclust(dist(VariablesExplicativas))
```

		No Presentado	Suspenso	Aprobado	Not-Sob
Corte	1	53	19	5	3
	2	25	54	40	14
	3	16	43	94	67
	#Total cases	94	116	139	84

Análisis supervisado

Se han probado varios métodos con la librería *caret*: árboles de decisión, bagged trees y random forest, y xgboost de la librería homónima.

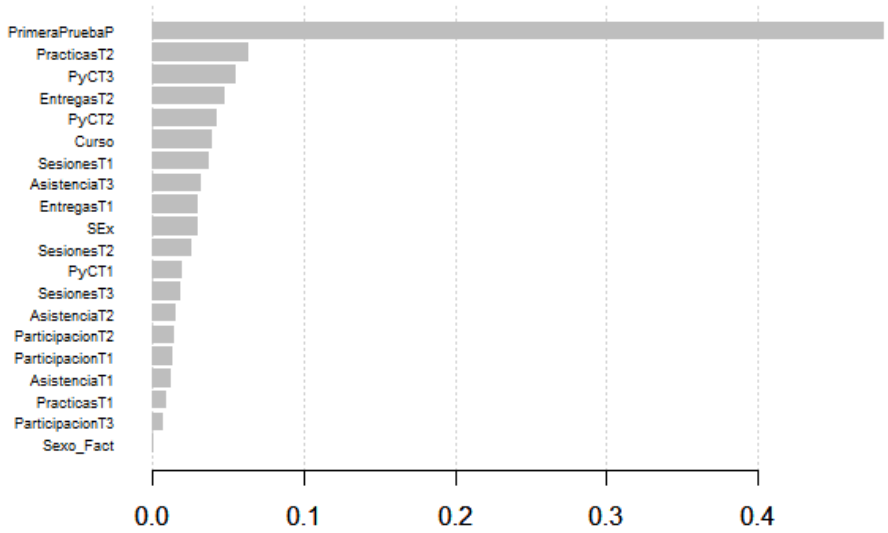


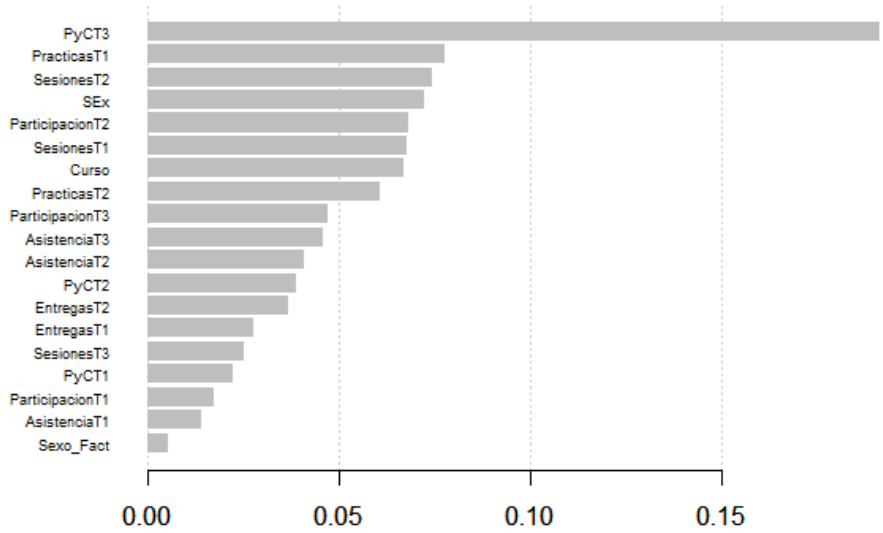


Algunos resultados

Hemos podido ver la importancia de las variables, así como los resultados de precisión de los diferentes modelos (hay diferente importancia de los errores):

Árbol de decisión	Random Forest	XGBoost	Bagged Trees
50 %	53 %	53 %	57 %





Análisis de Resultados y futuros pasos

De los resultados hasta ahora se obtiene que:

- La primera prueba tiene un peso muy alto
- Se puede clasificar bastante bien (no supervisado) aún sin ella
- Si quitamos la primera prueba, dominar la última parte de la asignatura y el estudio al principio/medio aparecen como variables más importantes, así como la nota de prácticas.

Desafíos y futuros pasos:

- Crear perfil de estudiantes para poder hacerles recomendaciones
- Definir y aplicar uno (o varios) tratamientos a los NAs de las diferentes variables
- Escoger un modelo para identificar perfiles de riesgo y llevarlo a la práctica

Comentarios y preguntas

Comentarios, dudas, quejas y sugerencias:

carlos.callearroyo@uclm.es