TFG – Educational Data Mining & Learning Analytics

Estudio de las Matriculaciones de A.D.E. en la UOC

Autor: Antonio Blanco Carpintero

Tutor: Ramón Caihuelas Quiles

Introducción

“Educational Data Mining”

- Soporte a profesores
- Recomendaciones para estudiantes
- Predicciones y modelaje de comportamientos
- Mediciones sobre rendimiento de los estudiantes
- Análisis de los grupos sociales
- Análisis, planificación y construcción de cursos y eventos

R como lenguaje para Data Mining

Aspectos principales de R

- Multitud de paquetes adicionales fácilmente instalables
- Posibilidad de comunicarse con bases de datos y ficheros
- Exportar resultados en diversos formatos
- Disponibilidad de un entorno gráfico
- Accesibilidad a grandes datos de internet como Google, Twitter y Facebook

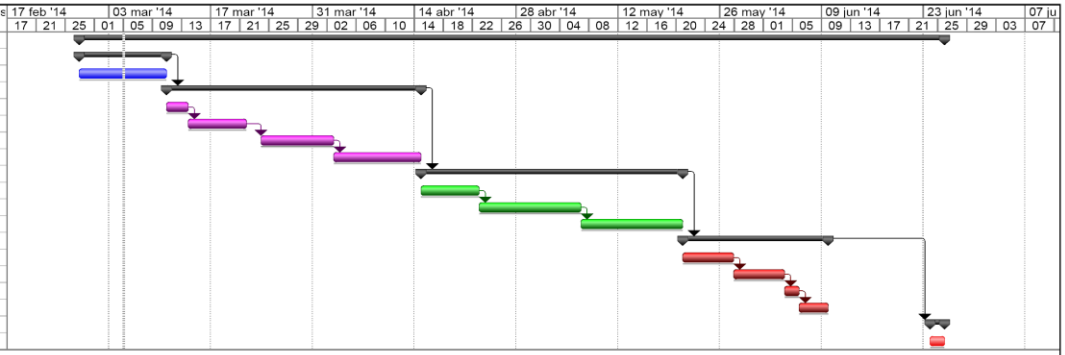
Enfoque y Método Seguido

Construcción de un Modelo

- Conocimientos de Minería de Datos
- Origen de los Datos
- Herramientas de Utilidad (lenguage R, Rattle, Rstudio, Weka)
- Preparación de los Datos (limpieza, organización, procesamiento, …)
- Construcción de Modelos (Agregación, Clasificación, …)

Planificación del Proyecto

Id	Nombre de tarea	Duración	Comienzo	Fin	Predeces
1	**TFG - Educational Data Mining & Learning Analytics**	**85 días**	**jue 27/02/14**	**mié 25/06/14**	
2	**Plan de Trabajo**	**8 días**	**jue 27/02/14**	**lun 10/03/14**	
3	Definición y Objetivo	8 días	jue 27/02/14	lun 10/03/14	
4	**Tratamiento de Datos**	**25 días**	**mar 11/03/14**	**lun 14/04/14**	2
5	Selección y Extracción de Datos	3 días	mar 11/03/14	jue 13/03/14	
6	Exploración de los Datos	6 días	vie 14/03/14	vie 21/03/14	5
7	Limpieza de Datos	8 días	lun 24/03/14	mié 02/04/14	6
8	Transformación de Datos	8 días	jue 03/04/14	lun 14/04/14	7
9	**Procesamiento de Datos**	**26 días**	**mar 15/04/14**	**mar 20/05/14**	4
10	Selección del Modelo	6 días	mar 15/04/14	mar 22/04/14	
11	Elaboración del Modelo	10 días	mié 23/04/14	mar 06/05/14	10
12	Verificación y Prueba de Contraste	10 días	mié 07/05/14	mar 20/05/14	11
13	**Lliurament Final y Presentación Virtual**	**14 días**	**mié 21/05/14**	**lun 09/06/14**	9
14	Preparación Entrega Final	5 días	mié 21/05/14	mar 27/05/14	
15	Preparación Presentación Virtual	5 días	mié 28/05/14	mar 03/06/14	14
16	Revisión	2 días	mié 04/06/14	jue 05/06/14	15
17	Referencias y Bibliografía	2 días	vie 06/06/14	lun 09/06/14	16
18	**Tribunal**	**2 días**	**mar 24/06/14**	**mié 25/06/14**	13
19	Tribunal	2 días	mar 24/06/14	mié 25/06/14	



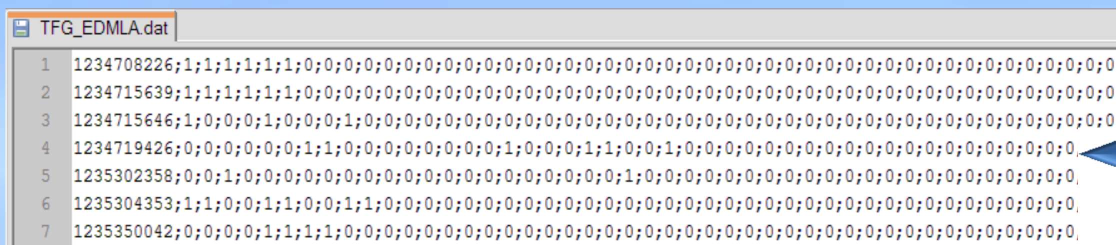
- **TFG - Educational Data Mining & Learning Analytics**
  - **Plan de Trabajo**
  - **Tratamiento de Datos**
  - **Procesamiento de Datos**
  - **Entrega Final y Presentación Virtual**
  - **Tribunal**

Productos Obtenidos

Entre los productos más relevantes obtenidos:

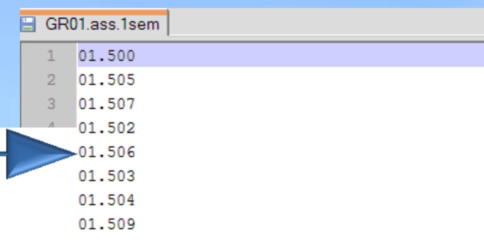
- Algoritmos en código R
- Diferentes estudios de los datos con técnicas complementarias
- Resultados de la ejecución de los algoritmos
- Gráficas explicativas de los resultados
- Ficheros de datos resultantes de la investigación
- Memoria del Proyecto
- Presentación Virtual

Extracción de los Datos



TFG_EDMLA.dat

```
1234708226;1;1;1;1;1;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;
1234715639;1;1;1;1;1;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;
1234715646;1;0;0;0;1;0;0;0;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;
1234719426;0;0;0;0;0;0;1;1;0;0;0;0;0;0;0;0;1;0;0;0;1;1;0;0;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
1235302358;0;0;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
1235304353;1;1;0;0;1;1;0;0;1;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
1235350042;0;0;0;0;1;1;1;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0
```



GR01.ass.1sem

```
01.500
01.505
01.507
01.502
01.506
01.503
01.504
01.509
01.501
```

Fichero TFG_EDMLA.dat. Datos de Matriculaciones ADE

Fichero GR01.ass.1sem. Códigos de Asignaturas ADE

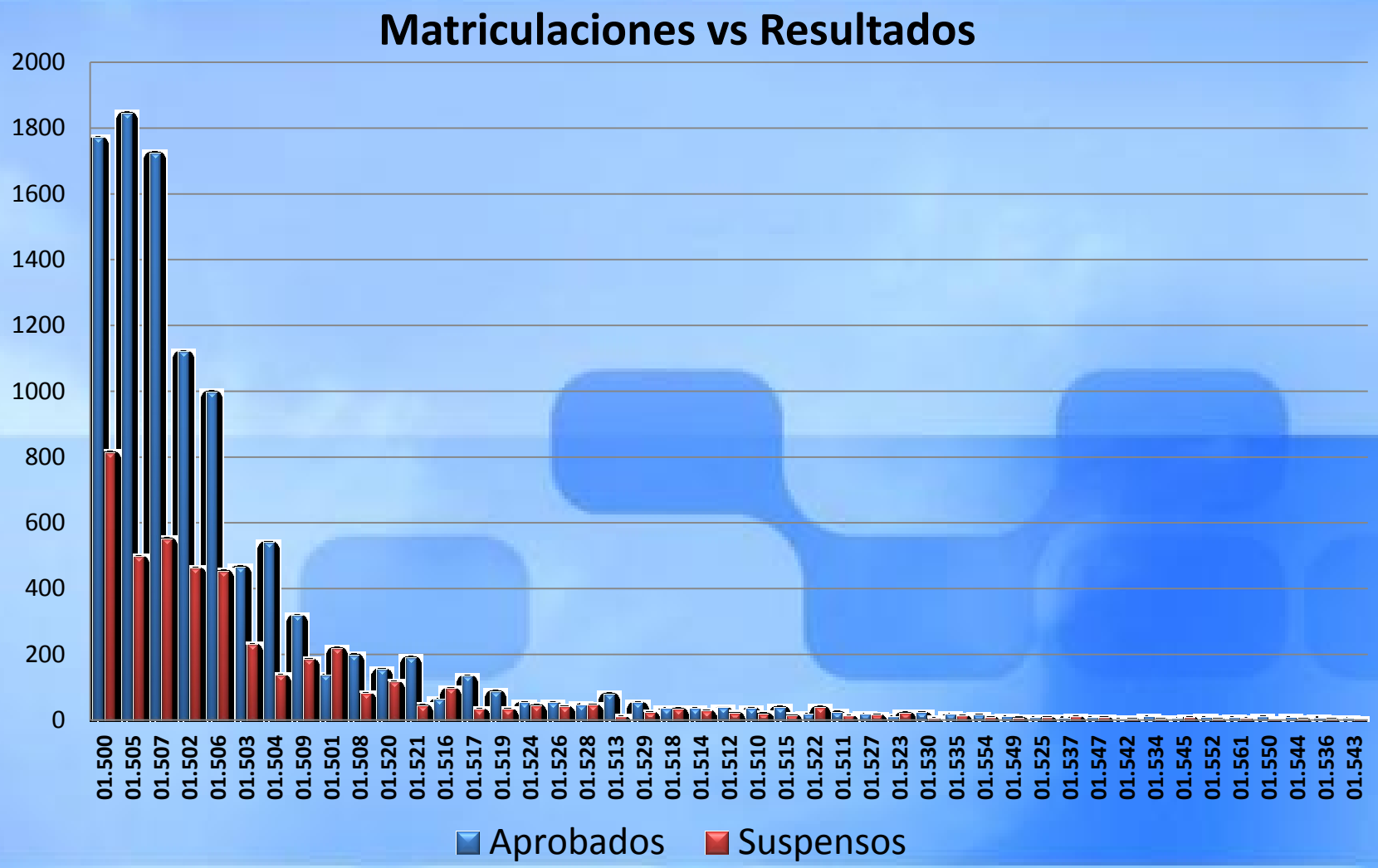
Campo de Datos	Descripción	Formato	Rango de Valores
Id_Alumno	Es el campo que identifica a un alumno	Alfanumérico 10 posiciones	0000000000 - 999999999
Fa_ass1	Indicador de la matriculación de la asignatura1	Alfanumérico 1 posición	0 – No matricula 1 – Sí matricula
Supera_ass1	Indicador de la superación de la asignatura1	Alfanumérico 1 posición	0 – No supera 1 – Sí supera
…	…	…	…
Fa_ass45	Indicador de la matriculación de la asignatura45	Alfanumérico 1 posición	0 – No matricula 1 – Sí matricula
Supera_ass45	Indicador de la superación de la asignatura45	Alfanumérico 1 posición	0 – No supera 1 – Sí supera

Extracción de los Datos

Estudio de los Datos

	Característica	Asignatura	Valor
Frecuencias Absolutas	Total de Matriculaciones		14666
	Total de Aprobados		10348
	Total de Suspensos		4318
	Asignatura con mayor número de matriculaciones	01.500	2581
	Asignatura con menor número de matriculaciones	01.542	4
		01.543	4
	Asignatura con mayor número de aprobados	01.505	1844
	Asignatura con menor número de aprobados	01.542	2
	Asignatura con mayor número de suspensos	01.500	812
	Asignatura con menor número de suspensos	01.543	1
		01.550	1
Tasas	Asignatura con mayor tasa de aprobados/matriculados	01.513	0,9070
	Asignatura con menor tasa de aprobados/matriculados	01.523	0,2857
	Asignatura con mayor tasa aprobados/suspensos	01.513	9,7500
	Asignatura con menor tasa aprobados/suspensos	01.523	0,4000

Descripción de los Datos

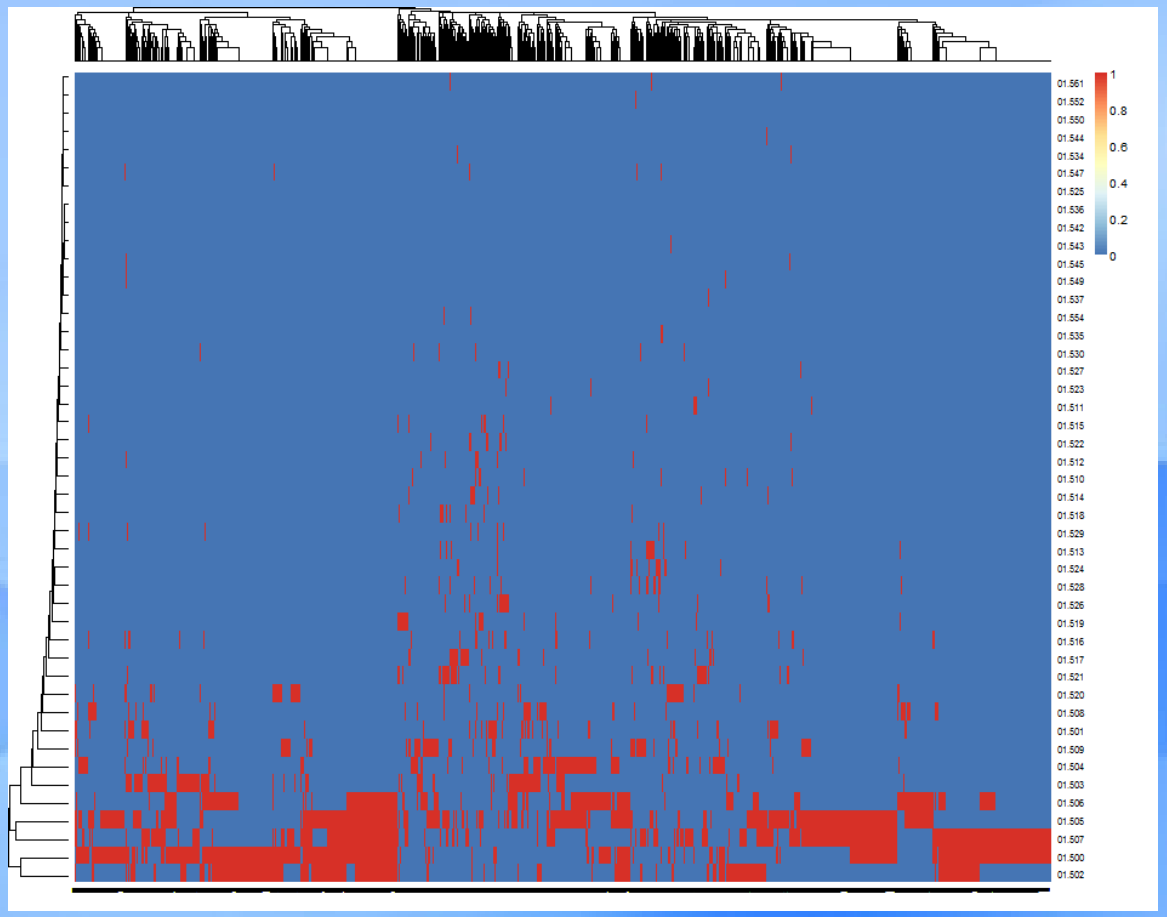


NNFM – Matriz de Factores No Negativos

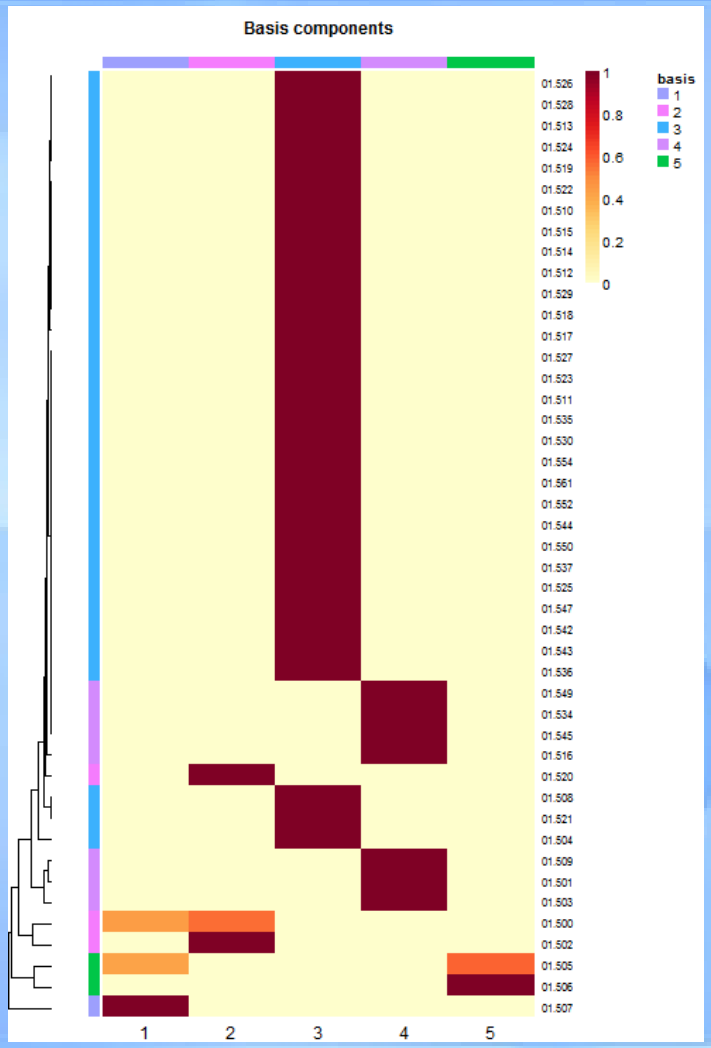
Descomposición de datos binarios

$X \sim W \cdot H$

Uso en clustering



Matriz de Bases



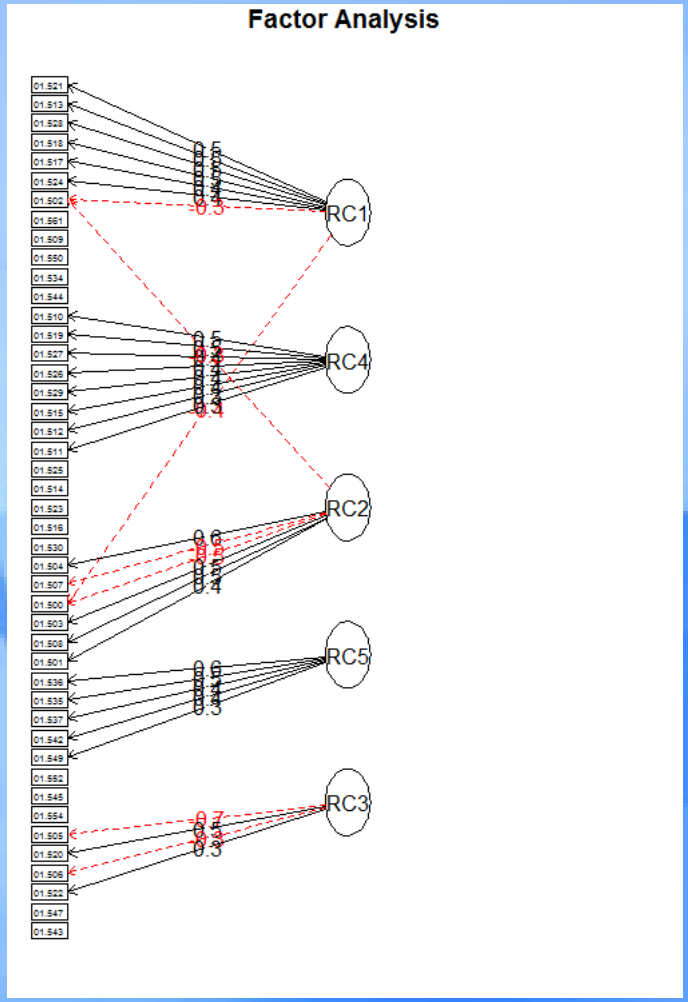
- Clústers de Asignaturas
  - Clúster 1: 01.500, 01.505, 01.507
  - Clúster 2: 01.500, 01.502, 01,520
  - Clúster 3: 01.504, 01.508, 01.521 y otras
  - Clúster 4: 01.501, 01.503, 01.509 y otras
  - Clúster 5: 01.505, 01.506
- Obtenemos las mayores similitudes entre las asignaturas entre grupos reducidos

PCA – Análisis de Componentes Principales

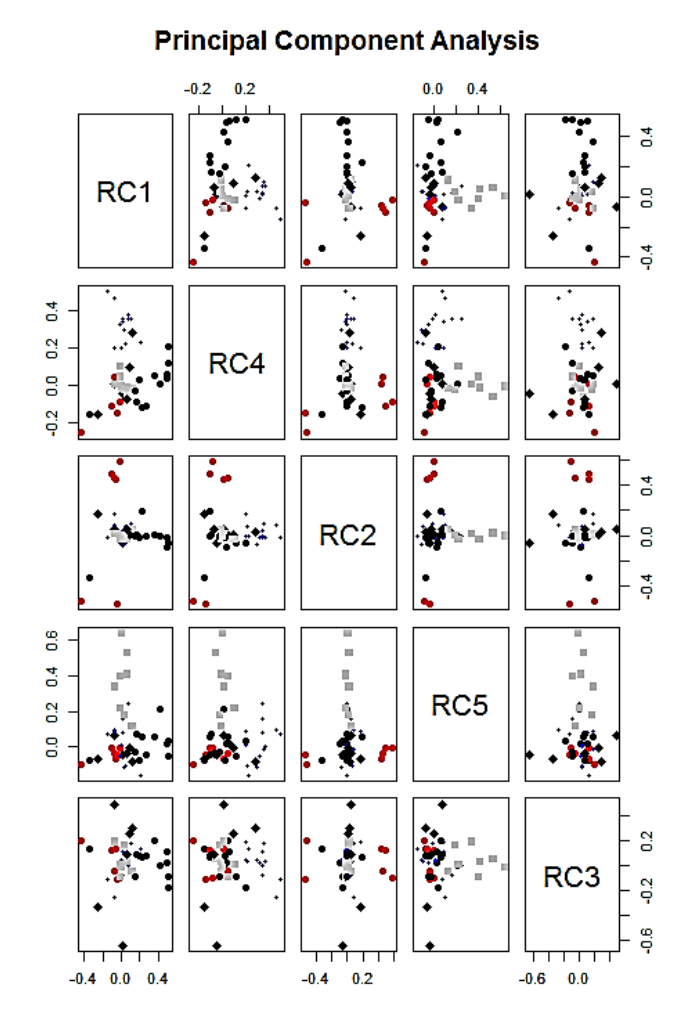
- Componente Principal ‘i’

$$PC_i = \sum_{j=1}^{n} a_{i,j} \cdot X_{i,j}$$

- Diferentes tipos de rotaciones practicadas
  - “varimax”
  - “oblimin”
  - “promax”



PCA – Análisis de Componentes Principales



RC1. **01.500, 01.502 (en valor negativo)** junto con 01.513, 01.518, 01.517, 01.521, 01.524, 01.528 (en valor positivo)

RC2. **01.500, 01.502, 01.507 (en valor negativo)** junto con 01.501, 01.503, 01.504 y 01.508 (en valor positivo)

RC3. **01.505, 01.506 (en valor negativo)** junto con 01.520 y 01.522 (en valor positivo)

RC4. 01.510, 01.511, 01.512, 01.515, 01.519, 01.526, 01.527, 01.529 (todas ellas positivas)

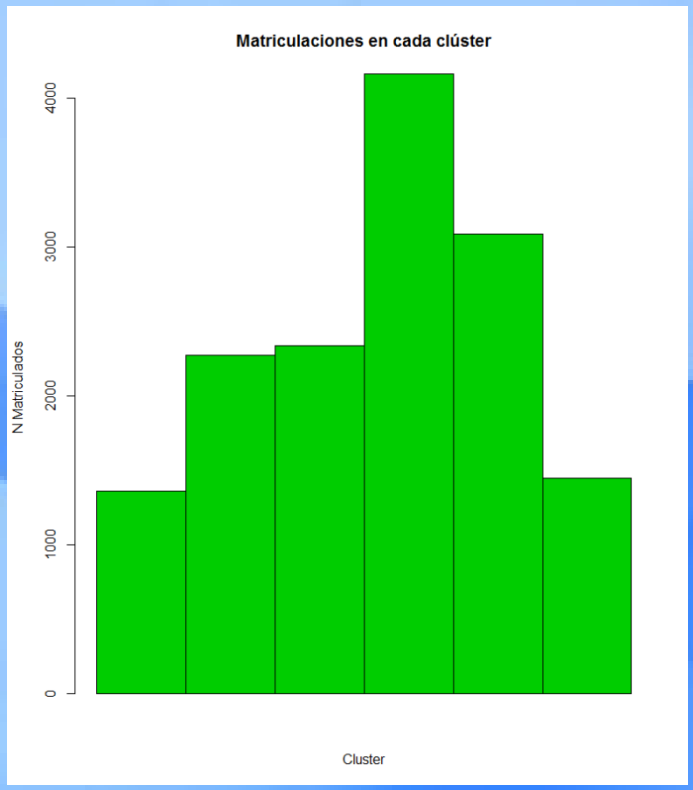
RC5. 01.535, 01.536, 01.537, 01.542, 01.549 (todas ellas en valores positivos)

K-means

Clúster de Asignaturas

Clúster	Matriculaciones	Matr_%	Asignaturas
1	1363	9,29	01.503, 01.504
2	2271	15,48	01.507
3	2339	15,95	01.505
4	4159	28,36	01.500, 01.502
5	3087	21,05	otras
6	1447	9,87	01.506

6 clústers



Preparación de un Clasificador

El clasificador de Éxito o Fracaso se basa en la diferencia entre asignaturas matriculadas y las asignaturas aprobadas de tal manera:

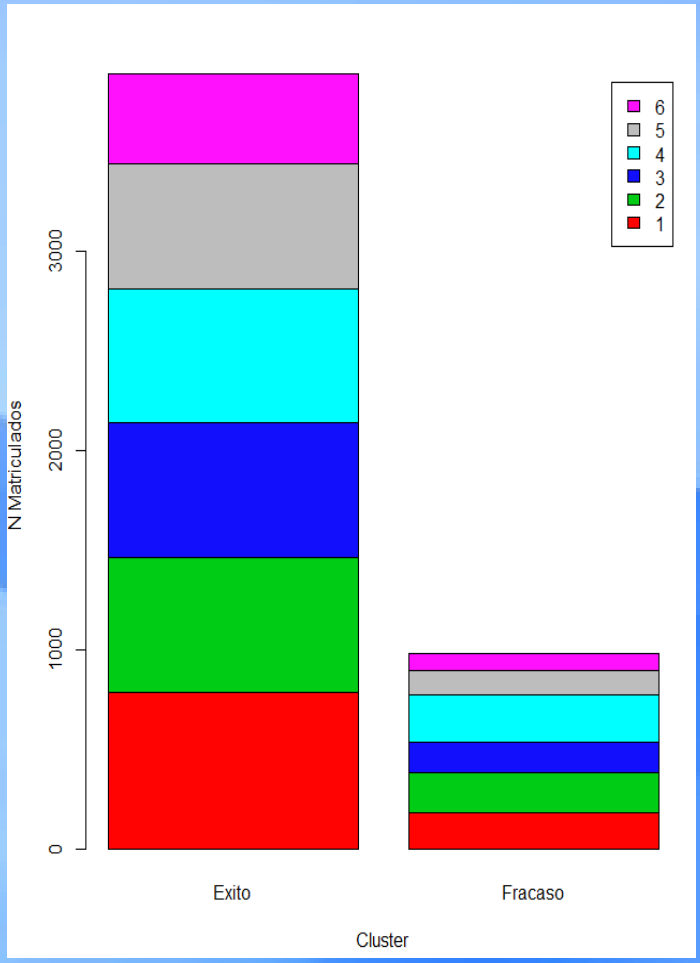
$$\boldsymbol{Asignaturas\ Matriculadas\ -\ Asignaturas\ Aprobadas} > \mathbf{0} \Rightarrow \text{"Éxito"}$$
$$\boldsymbol{en\ caso\ contrario}\ \text{"}\boldsymbol{Fracaso}\text{"}$$

N_Alumno	1	2	3	…	45	Clasificador Éxito/Fracaso
Al_1	0/1	0/1	0/1	…	0/1	"Éxito" o "Fracaso"
…	…	…	…	…	…	…
Al_n	0/1	0/1	0/1	…	0/1	"Éxito" o "Fracaso"

Clústers Alumnos y Clasificación Éxito - Fracaso

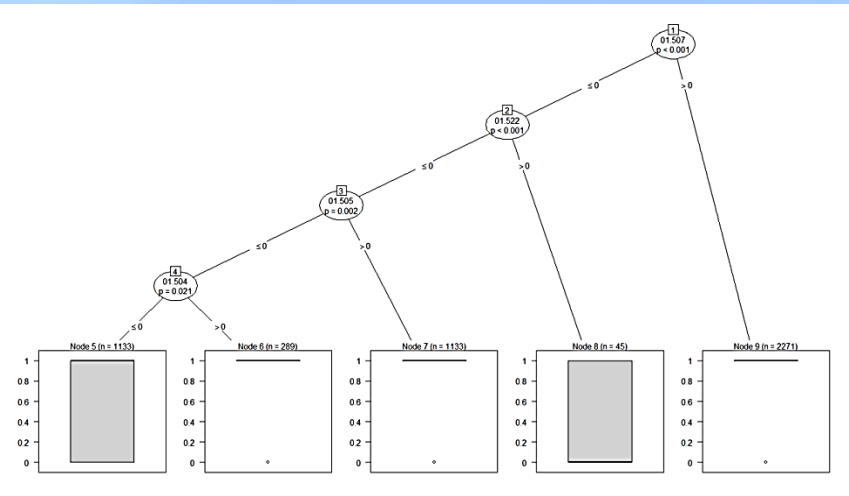
Relación de Clústers

Clúster	Éxito	Fracaso
1	787	183
2	676	205
3	678	149
4	673	236
5	628	122
6	448	86



Árboles de Decisión

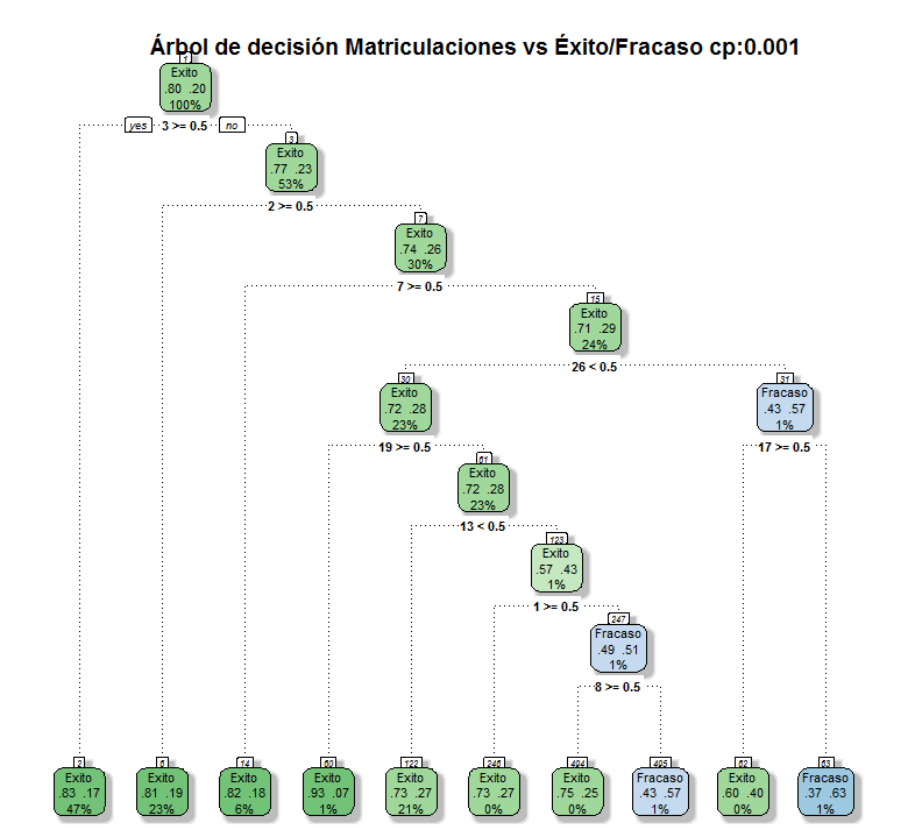
Modelo ‘ctree’



- Asignaturas más relevantes:
  - 01.507
    - 01.522
      - 01.505
        - » 01.504

Árboles de Decisión

Modelo ‘rpart’



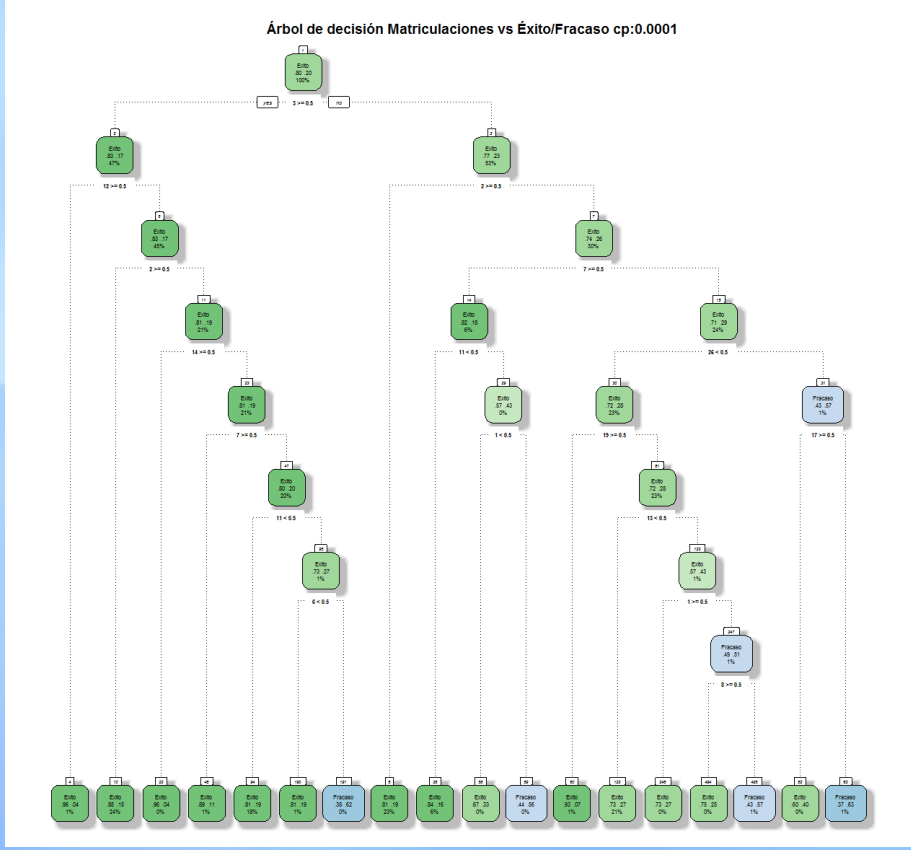
Complejidad 0,001

Casos significativos de Fracaso:

- EF=Fracaso cover=27 (1%) prob=0.63
  3<0.5 && 2<0.5 && 7<0.5 && **26>=0.5** && 17<0.5
- EF=Fracaso cover=35 (1%) prob=0.57
  3<0.5 && 2<0.5 && 7<0.5 && 26<0.5 && 19<0.5 && **13>=0.5** && 1< 0.5 && 8<0.5

Árboles de Decisión

Modelo ‘rpart’



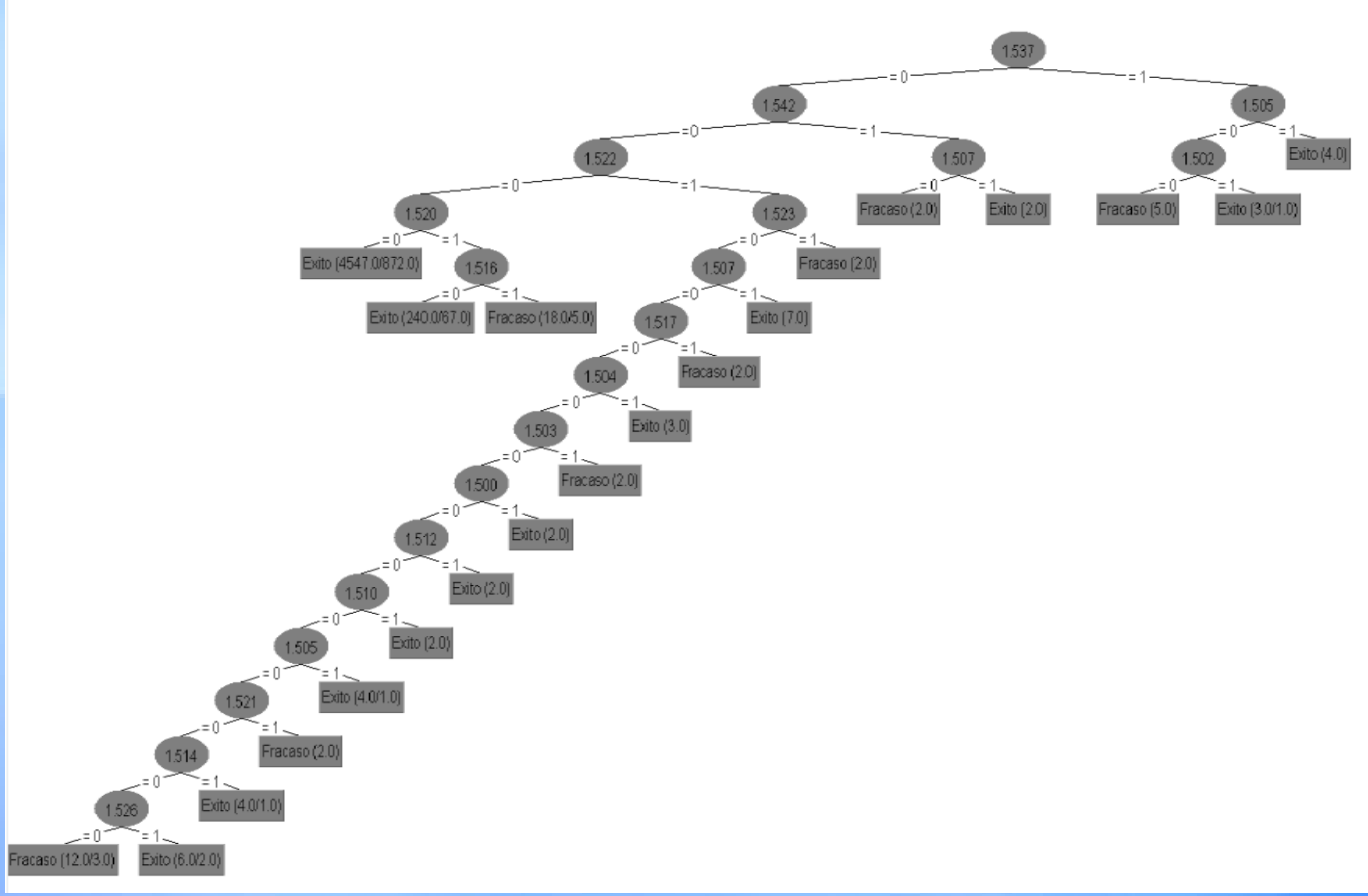
Complejidad 0,0001

Casos significativos de Fracaso:

- EF=Fracaso cover=27 (1%) prob=0.63
  3<0.5 && 2<0.5 && 7<0.5 &&
  26>=0.5 && 17<0.5
- EF=Fracaso cover=13 (0%) prob=0.62
  3>=0.5 && 12<0.5 && 2<0.5 &&
  14<0.5 && 7<0.5 && **11>=0.5** &&
  6>=0.5
- EF=Fracaso cover=35 (1%) prob=0.57
  3<0.5 && 2<0.5 && 7<0.5 &&
  26<0.5 && 19<0.5 && **13>=0.5** &&
  1<0.5 && 8<0.5
- EF=Fracaso cover=9 (0%) prob=0.56
  3<0.5 && 2<0.5 && **7>=0.5** &&
  11>=0.5 && **1>=0.5**

Árboles de Decisión

Árbol J48 con Weka



Leyenda:

'=0' → "No matricula"

'=1' → "Sí matricula"

Nodos relevantes Asociados al Fracaso:

- 01.537
- 01.520 && 01.516
- 01.522 && 01.523
  - 01.517
  - 01.503
  - 01.521
  - 01.526

Conclusiones

- Trabajo Evolutivo de Proceso de Aprendizaje y crecimiento Personal
- Obtención de una visión global de lo que es un proyecto, su planteamiento, puesta en marcha y finalización
- Evaluación de riesgos y enfrentamiento a inconvenientes
- Añade Conocimiento y Experiencia