TFC-Minería de datos

Ingeniería Técnica Informática

Estudio sobre el abandono de los estudiantes de la carrera de Ciencias Empresariales de la UOC

Autor: Víctor Aguilera Arranz
Tutor: Ramón Caihuelas Quiles

Temas

- Herramientas
- Objetivo
- Los datos
- Descripción de los datos
- Preparación de los datos
- Modelos
  - Modelo 1: Cluster
  - Evaluación del modelo 1
  - Modelo 2: Cluster + árboles
  - Evaluación del modelo 2
  - Modelo 3: Árbol
  - Evaluación del modelo 3
- Conclusiones
- Posibles mejoras

Herramientas

- En el mercado existen numerosas aplicaciones útiles para la minería de datos algunas de las mas utilizadas de código abierto son:
  - RapidMiner
  - R
  - Orange
  - Weka
  - JhepWork
  - Knime

- Se decide realizar el proyecto con la aplicación RapidMiner por su potencia y facilidad de uso.

Objetivo

Planteamiento del problema:

La UOC ha detectado que todos los años alumnos de la carrera de Ciencias Empresariales abandonan sus estudios dejando la carrera sin terminar.

Objetivo:

- Llegar a conocer cuales son las causas que producen el abandono de los estudiantes
- Clasificar a los alumnos en función de los datos de matriculación
- Predecir el comportamiento de los alumnos

Los datos

La UOC proporciona dos ficheros con información sobre los alumnos entre los años 1998 y 2008.

- Un fichero "TFC_MD.txt" con la descripción de los datos
- Un fichero "TFC_MD.dat" con los datos de los alumnos

Esta información es tratada y añadida a un repositorio de RapidMiner

Descripción de los datos

- Dos campos imprescindibles:
  - ID: que identifica a cada alumno
  - ABANDONA: Que indica si el estudiante abandona o no.

- Otros campos que aparecen son:
  - Personales, como sexo, edad, vía de acceso.
  - Relativos a las asignaturas matriculadas, superadas, presentadas, etc.
  - Relativos a los créditos matriculados, superados
  - Otros (como el número de semestres o algunos campos calculados a partir de otros ya nombrados)

Preparación de los datos

- Los datos pasan por varias transformaciones para que puedan ser utilizados en los modelos de la forma mas eficiente.

- Se sustituyen algunos valores numéricos por su nominal correspondiente para que la información obtenida sea mas clara:

  - 0, 1 NO, SI
  - 0, 1 FALSO, VERDADERO
  - 0, 1 MUJER, HOMBRE
  - 0, 1 no_matricula, si_matricula
  - 1,2,3,4 no_cou, cou, est_inacabados, titulado
  - etc.

- Se transforman campos como en el caso de SEMESTRE que originalmente indicaba año y semestre, para quedarse únicamente con la información sobre semestre.

Modelos

Desarrollo de tres modelos basados en las asignaturas matriculadas por los alumnos:

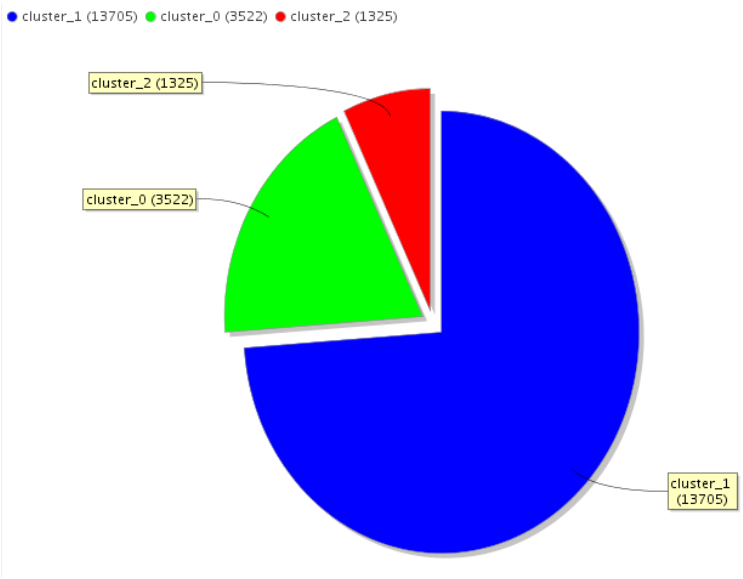
- Modelo 1: modelo de clusters
- Modelo 2: modelo de árboles a partir de los clusters del modelo 1
- Modelo 3: modelo de árbol a partir de los datos del repositorio original

Modelo 1: Clusters

- Se agrupan los datos en 3 clusters utilizando el algoritmo k-means y en base a las asignaturas de las que se matriculan los alumnos.

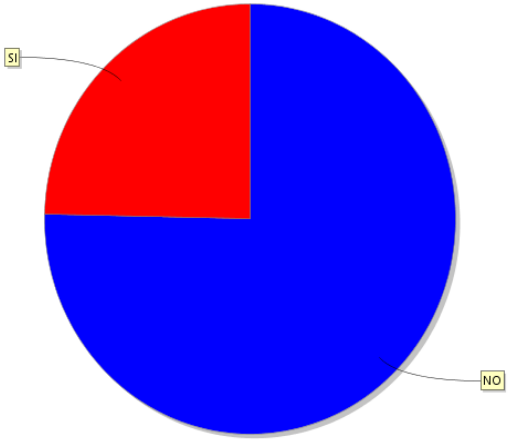
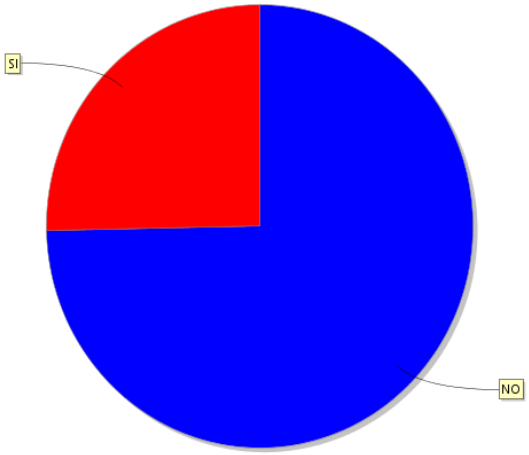
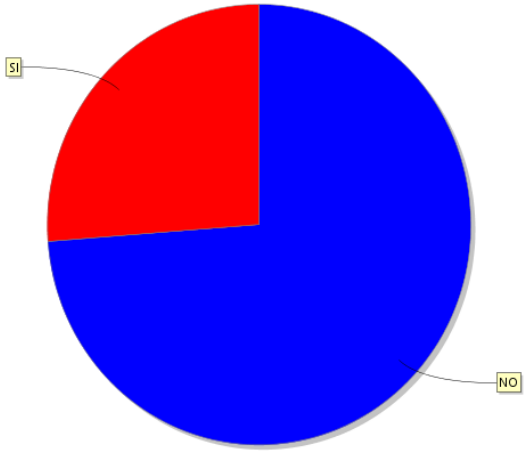
- Se quiere estudiar, si las similitudes sobre las matrículas de las asignaturas que agrupa a los individuos en cada grupo tienen relación con el abandono de la carrera

- Resultado:
  - cluster_0: 3522 estuds.
  - cluster_1: 13705 estuds.
  - cluster_2: 1325 estuds.



Clusters

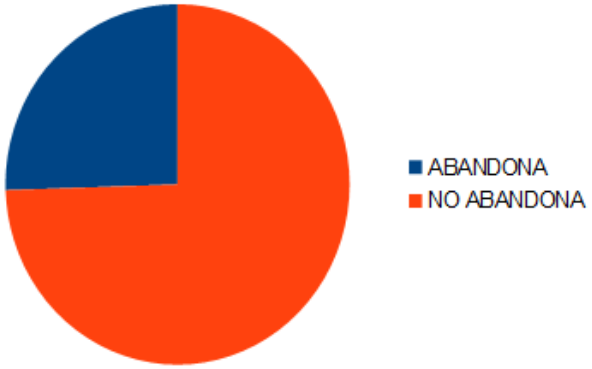
- Distribución del abandono en cada cluster:

Cluster_0	Cluster_1	Cluster_2
		

Evaluación del modelo 1

- Se aprecia en los gráficos que el abandono en los tres clusters se distribuye de una forma similar.

- En los datos de partida de este estudio se observa que los alumnos que abandonan son 4.687 y los que no abandonan son 13.685.



- Por tanto la distribución obtenida en cada uno de los clusters es prácticamente la misma que había en los datos originales, con lo que no se descubre una propensión hacia el abandono o no abandono en ninguno de los clusters mayor que la que existe entre el conjunto total de estudiantes.

Modelo 2: Árboles a partir de los clusters del modelo 1

- Se construyen árboles de decisión a partir de los datos de los clusters del modelo 1.
- Se quiere ver si existe algún patrón de comportamiento en alguno de los clusters que pueda predecir el abandono

Evaluación del modelo 2

- Se encuentran algunos casos en los que el error cometido es bajo
- Pero en la mayoría de los casos el error cometido es de alrededor del 25%.
- La distribución de los datos originales revela que el abandono de los estudiantes es de alrededor del 25%, por tanto los árboles obtenidos no tienen una calidad suficiente

Modelo 3: Árbol a partir de los datos del respositorio original

- Aprovechando la información obtenida en la realización de los clusters del modelo 1 se pretende seleccionar los campos mas significativos a la hora de que un estudiante abandone.
- Observando la distribución de los valores en los cluster y sus centroides se seleccionan 6 asignaturas.
- Con las campos seleccionados se realiza un árbol de decisión a partir de los datos originales.

Evaluación del modelo 3

- La mayoría de los caminos seguidos para recorrer el árbol acaban en NO ABANDONO
- Para menos del 0,1% del total de alumnos el modelo predice que abandonan
- En los casos que se predice el abandono el error cometido es de alrededor del 40%

Conclusiones

- La decisión de abandonar o no los estudios no parece que tenga relación con las asignaturas de las que se matriculan los estudiantes.

- Quedan abiertas otra vías de estudio en las que se prueben otros modelos con otros datos en busca de unos resultados mejores.

Posibles mejoras

- Se podría ampliar la información que hay en la base de datos sobre los alumnos.

- Es posible que por medio de encuestas pudieran obtenerse datos que puedan ser de mayor utilidad como:

  - Si el estudiante trabaja y tiene que compaginar el trabajo y los estudios
  - El tiempo libre del que dispone
  - Las cargas familiares
  - Los ingresos que tiene
  - etc.

Gracias

Víctor Aguilera Arranz