

Automatización del procesamiento de encuestas de valoración de una empresa educativa

Kamal A. Romero S.
Memoria del Trabajo de Fin de Máster

Índice

1. Introducción	3
1.1. Cronograma del proyecto	4
1.2. Tecnologías empleadas	4
1.3. Estructura de los datos	4
1.4. Replicación del proyecto	5
1.5. Guía de uso del dashboard	6
1.6. Documentos adicionales	7
1.7. Resultados	7
2. Fases del Proyecto	8
2.1. Limpieza de datos	8
2.1.1. Implementación	9
2.2. Elaboración de reportes y el dashboard	10
2.3. Ejercicio empírico	11
2.3.1. Análisis factorial	11
2.3.2. Regresión logística	12

Índice de figuras

1. Aspecto del dashboard	6
2. Datos limpios	8
3. Datos sucios	8
4. Reporte de evaluación del profesorado	10
5. Cargas factoriales a partir de una rotación oblicua (promax)	13
6. Gráfico de Sedimentación (Scree Plot)	14

7.	Curva ROC	17
----	---------------------	----

Índice de cuadros

1.	Cargas factoriales a partir de una rotación oblicua (promax)	12
2.	Efectos marginales de la regresión logística	15
3.	Probabilidades relativas de obtener una valoración alta	16

1. Introducción

Las encuestas de valoración son una herramienta muy útil para cualquier empresa u organización, en la medida que permite un feedback entre la prestación del servicio y el usuario. En el caso de las empresas u organizaciones de educación superior es el modo más habitual que poseen para cuantificar la labor docente y/o el grado de satisfacción del alumnado.

No obstante, el procesamiento de las mismas no es trivial en el caso de organizaciones medianas y pequeñas, las cuales no suelen emplear los servicios de una empresa especializada. Dicho proceso involucra la elaboración, recolección y procesado de los resultados. En el presente proyecto nos concentramos en esta última.

En nuestro caso llamamos procesado al conjunto de acciones que van desde la limpieza del output de una lectora óptica hasta la presentación de dichos resultados, pasando por un análisis de los mismos. En el caso de pequeñas empresas, realizar dicha labor con el software propietario al cual se suele tener acceso habitualmente puede llegar a ser una labor casi artesanal, donde se procesan los datos de la lectora de forma manual, se consolidan en un software, se realiza el análisis estadístico en otro y la presentación en otro, siendo muy complicado y en ciertos casos imposibles la automatización.

En el presente proyecto se automatiza el procesado de un pequeño centro de educación superior, y posteriormente se realiza un análisis cuantitativo de los determinantes del grado de satisfacción global medido en las encuestas. Específicamente se automatiza el proceso de limpieza de los datos originales de la lectora, evitando cualquier manipulación manual. También se automatiza la realización y presentación de los informes de valoración individual de los docentes.

Mediante lo anterior se ha logrado reducir los tiempos de limpieza de datos y elaboración de reportes de 4 y 5 horas respectivamente a menos de un minuto (entre 13 segundos y 18 segundos dependiendo de la máquina) el primero y 4 minutos el segundo.

Empíricamente se han identificado dos factores latentes que resumen los items de valoración de la encuesta denominados *docencia* y *relación profesor-alumno*, como los principales determinantes de la probabilidad de obtener una valoración alta. Asimismo, se determina que ciertas características individuales de los profesores tales como el sexo, categoría, edad, etc. no influyen de manera determinante en el grado de satisfacción global del alumno.

1.1. Cronograma del proyecto

El proyecto se realiza en tres grandes fases que a su vez se subdividen en otras

1. Limpieza de datos

- a)* Bloque de valoraciones
- b)* Bloque de códigos administrativos

2. Elaboración de reportes

- a)* Generación automática de los reportes en pdf para todo el profesorado
- b)* Generación individual de reportes para una asignatura mediante un dashboard interactivo

3. Análisis empírico de los determinantes de las valoraciones

- a)* Análisis factorial exploratorio¹
- b)* Regresión logística

1.2. Tecnologías empleadas

Se emplea un código Python para la fase de limpieza de datos y un código R para la realización del análisis empírico y la elaboración de reportes para todo el profesorado con Rmarkdown. El dashboard para los reportes individuales se realiza con el paquete Shiny de R.

La memoria esta escrita en L^AT_EX.

1.3. Estructura de los datos

Los datos de entrada provienen de una lectora óptica, los campos están separados por punto y coma y se cargan en Python como un csv.

El archivo viene con un numero indeterminado de columnas a priori, ya que depende de cuantas columnas desplace la lectora hacia la derecha en un error de lectura. El número correcto de columnas son 15, las cuales corresponden a la siguiente información de cada encuesta realizada:

- 1. **DIVISION**: la división académica correspondiente. En nuestro caso va del 1 al 4.
- 2. **CURSO**: el curso en el cual se imparte la asignatura evaluada, sus valores van del 1 al 5, donde 5 representa un máster

¹Agradezco al profesor Antonio Pita el haberme recomendado emplear esta técnica.

3. **GRUPO**: el grupo en el que se imparte la asignatura. Se encuentra identificado por una letra y su valor depende del grado o máster.
4. **CODIGO_ASIG**: el identificador numérico de la asignatura que se evalúa.
5. **COD_PROF**: el identificador numérico del profesor que imparte la asignatura.
6. **PREGUNTAS**: las evaluaciones correspondientes a los 10 ítems. Su valor va del 1 (insatisfecho) al 10 (muy satisfecho).

Las filas corresponden a las encuestas realizadas. El archivo original posee 3729 filas y 20 columnas.

Debido a que los impresos de las encuestas solo poseen opciones de grupos de la A a la D, hay dos grupos que al no poseer estas siglas han dejado esa casilla en blanco y no es posible determinar su pertenencia. Asimismo, hay un grupo que debido a un error al momento de pasar la encuesta no rellenaron su grupo correspondiente (D).

En el proceso manual esto se resolvía al momento de pasar las encuestas por la lectora, dado que la muestra que se posee no ha sido pasada en el mismo orden que en el procesado original, no es posible determinar el grupo correspondiente a estos registros. Se opta por eliminarlos de la muestra.

Debido a esto se pierden 187 registros (filas). Por lo que nuestro **archivo input definitivo tiene 3542 filas y 20 columnas**

1.4. Replicación del proyecto

La estructura de archivos del proyecto y el orden de ejecución es la siguiente:

Limpieza de datos Un código python en el archivo `limpieza.py`.

Inputs: un archivo de texto con la información original de las encuestas.

Outputs: un archivo con los datos de la encuesta *limpios*, `encuestas.csv`.

Asimismo se generan varios archivos .txt que indican las columnas que han sido modificadas en parte del proceso

Análisis empírico y elaboración de reportes Un código R en el archivo `encuestas.r` y un Rmarkdown en `Prueba_reporte.Rmd`

Para poder replicar el trabajo **se debe fijar el directorio de trabajo en el archivo**

Inputs: el archivo `encuestas.csv` generado en el paso anterior, un archivo en excel

`celitems.xlsx` que contiene las preguntas de la encuesta y otro archivo excel `datos_profesores_v2.xlsx` con los datos de las características de los profesores.

Outputs: 96 archivos .pdf que representan los reportes individuales de todos los profesores del centro.

Este proceso tarda unos 5-7 minutos aproximadamente.

Dashboard para la generación de reportes individuales Tres archivos `ui.r`, `server.r` y `global.r`.

En el `global.r` se cargan datos generados en el paso anterior. Se accede al dashboard en este link <https://kamecon.shinyapps.io/Reporte/>

Todos estos archivos se encuentran en la carpeta **Replicación** del Github.

1.5. Guía de uso del dashboard

En la figura 1 se observa el aspecto del dashboard.

Figura 1: Dashboard interactivo



En el lado izquierdo del panel se encuentra una pestaña desplegable con el nombre de las asignaturas, se despliega y escoge una. Una vez escogida la asignatura, en el lado derecho del panel aparecerá un gráfico con las distribuciones de frecuencias de cada uno de los items de valoración de la encuesta.

Lo verdaderamente interesante de este dashboard, es que permite descargarse los reportes en pdf correspondiente a la asignatura seleccionada, a través del botón *Download* situado debajo de la pestaña desplegable.

Una vez realizada la solicitud, se descarga un archivo .pdf al ordenador.

1.6. Documentos adicionales

Dado el limitado espacio de la memoria, se pueden ver con todo detalle y ampliamente documentados el proceso de limpieza de datos y el ejercicio empírico en el notebook `Tidy1.ipynb` y el html `analisis_empirico.html`² respectivamente.

1.7. Resultados

Se alcanza el objetivo de automatizar el procesado de las encuestas de valoración del profesorado, evitando cualquier manipulación manual de los mismos.

El proceso de corrección manual de los errores de la lectora y humanos tarda aproximadamente 4 horas en la actualidad, el archivo `limpieza.py` logra **realizar el proceso en menos de 20 segundos**. El coste de esta mejora es que se pierden 10 registros debido a que no ha sido posible identificar el grupo o la asignatura con el algoritmo realizado.

El proceso de elaboración de los reportes tardaba aproximadamente una hora en hacer las tablas con un software propietario, y luego otras 3 aproximadamente en pasar dichas tablas a un editor de texto y convertirlo a formato pdf. Con nuestro código todo **este proceso se realiza en 4-5 minutos aproximadamente**.

Se aporta un dashboard preliminar que en principio permite acceder a los reportes individuales por asignatura ya en formato pdf con una breve descripción estadística. En la actualidad no existe ninguna herramienta que permita realizar algo similar.

Lo anterior representa una solución válida de negocio sin ninguna duda.

En el aspecto empírico se determina mediante el empleo de un análisis factorial exploratorio y una regresión logística, que dos variables latentes las cuales resumen los aspectos docentes y la relación entre el profesor y el alumno explican de manera adecuada la probabilidad de una valoración alta. Mientras que las características individuales de los docentes tales como el sexo, categoría y similares, no influyen de manera relevante la probabilidad de obtener valoraciones altas.

El modelo empleado presenta un R^2 ajustado de 0,63, una precisión del 91,5% y un AUC de 0,966, confirmando una buena capacidad para identificar las valoraciones altas de las bajas.

²Se puede descargar el html en esta dirección https://github.com/kamecon/TFM_Kschool/tree/master/Replicacion.

2. Fases del Proyecto

A continuación se describe lo realizado en cada fase del proyecto

2.1. Limpieza de datos

Esta ha sido fase más intensiva en términos de tiempo de implementación (mas de 2/3 del total) y de código (más de 200 líneas).

Esta fase ha sido mucho más compleja que la limpieza de datos estándar que consiste en imputar datos, eliminar y sustituir caracteres, y eliminar filas con datos incompletos. Para poder visualizar mejor el proceso, a continuación se presentan el formato *limpio* de los datos y se compara con una muestra de los datos brutos.

Idealmente para un proceso adecuado, los datos deberían tener el siguiente aspecto

Figura 2: Datos Limpios

DIVISION;	CURSO;	GRUPO;	CODIGO_ASIG;	COD_PROF;	ITEM1;	ITEM2;	ITEM3;	ITEM4;	ITEM5;	ITEM6;	ITEM7;	ITEM8;	ITEM9;	ITEM10
5;	3;	E;	003;	002;	10;	09;	09;	09;	08;	09;	10;	05;	10;	09
5;	3;	E;	003;	002;	06;	01;	05;	04;	04;	09;	06;	05;	01;	06
5;	3;	E;	003;	002;	07;	03;	06;	07;	08;	08;	10;	08;	10;	08
5;	3;	E;	003;	002;	06;	01;	05;	07;	04;	02;	07;	05;	02;	05
5;	3;	E;	003;	002;	10;	08;	07;	04;	10;	10;	10;	10;	10;	10
5;	3;	E;	003;	002;	09;	08;	08;	07;	07;	09;	10;	09;	08;	09
5;	3;	E;	003;	002;	07;	05;	08;	10;	04;	07;	09;	09;	03;	08
5;	3;	E;	003;	002;	08;	04;	03;	09;	06;	09;	10;	09;	03;	08
5;	3;	E;	003;	002;	07;	05;	09;	09;	08;	10;	10;	10;	07;	08
5;	3;	E;	003;	002;	07;	06;	08;	09;	05;	05;	10;	06;	03;	07
5;	3;	E;	003;	002;	07;	02;	05;	06;	06;	08;	09;	06;	06;	07

En este caso, se podrían cargar los datos en un data frame o incluso una hoja de cálculo sin problemas. No obstante los datos brutos suelen tener este aspecto

Figura 3: Datos Sucios

DIVISION;	CURSO;	GRUPO;	CODIGO_ASIG;	COD_PROF;	PREGUNTAS;
2;	1;	A;	00	;006;07;07;06;06;06;08;07;	;08;08;
2;	1;	A;	002;	006;10;10;10;10;10;10;10;10;10;	
2;	1;	A;	002;	;08;07;07;06;07;08;07;	;07;08;
2;			005;004;	10;07;08;09;09;09;10;08;09;10;	
2;	1;	A;	005;	073;06;02;07;05;07;07;08;05;03;05;	
2;	1;	A;	005;	004;07;08;07;07;06;08;07;06;06;07;	
2;			005;004;	08;09;07;09;08;07;09;07;09;08;	
		A;		;08;05;08;04;09;10;10;08;08;09;	
2;			005;004;	06;01;05;05;03;??;	;08;09;04;
2;	1;	A;	004;	003;10;08;06;07;09;05;10;04;	;07;07;
2;	1;	C;	003;	002;08;04;06;07;07;06;05;04;06;05;04;04;	
2;	1;	C;	003;	002;07;	;06;05;04;
1;	1;	A;	204;	204;08;08;08;08;07;07;08;08;07;09;	
			4	;09;	;09;
4;	4;		419;	446;08;09;08;	
4;	4;		419;	419;10;	
			250;	273;02;??;	01;04;

La anterior es una muestra sesgada de los datos para precisamente mostrar los elementos a limpiar, pero no es una muestra exhaustiva de los errores que nos podemos encontrar. Dichos errores son una combinación de errores humanos y de lectura de la lectora óptica. A continuación se enumeran algunos:

1. Los usuarios dejan casillas en blanco
2. Colocan mal alguno de los códigos
3. No rellenan bien la casilla
4. La lectora desplaza algunos campos
5. La lectora interpreta mal algún código y lanza un símbolo en lugar del número
6. etc

Dado que parte de estos errores se corregían de forma manual, afortunadamente tenemos una amplia comprensión de su origen y solución.

El proceso de limpieza de datos se divide en dos fases:

- La primera, en la cual se limpian los datos correspondientes a los ítems de evaluación. Primero se corrige un desplazamiento de columnas realizado por la lectora óptica, posteriormente se eliminan caracteres generados por la lectora y finalmente se sustituyen las valoraciones NAN por la mediana del resto.
- El segundo paso es el más complejo, y requiere la limpieza de un bloque de códigos administrativos (división, curso, asignatura y profesor). El reto de esta sección es que no se pueden emplear tácticas comunes de eliminación y sustitución por imputación de valores.

Dado lo extenso y en ocasiones complejo del proceso, **se refiere al usuario revisar el notebook Tidy1.ipynb para los detalles de la estrategia de limpieza**

2.1.1. Implementación

La implementación de esta fase del proyecto se encuentra en el archivo `limpieza.py`, el cual recibe como input un archivo de texto con la información de las encuestas y como output un archivo `.xls` y un archivo `.csv` con los datos *limpios*.

Como se ha mencionado arriba, el proceso de limpieza no es trivial, el archivo `limpieza.py` posee más de 200 líneas de código y 8 funciones personalizadas. El código se encuentra documentado, pero **para entender el proceso de limpieza con detalle es necesario revisar el notebook Tidy1.ipynb**.

2.2. Elaboración de reportes y el dashboard

La elaboración de reportes se hace mediante el empleo de Rmarkdown, se puede observar el procedimiento en la última sección del archivo `encuestas.r` y `Prueba_reporte.Rmd`. Este último recibe dos inputs generados en el primero, un data frame `cabecera` y una lista `cuadro`.

El proceso de elaboración de reportes se puede dividir en dos partes.


La primera es la elaboración de un cuadro en el cual aparezcan las frecuencias de valoraciones por ítem obtenidas por cada profesor. Esto se realiza mediante un bucle que recorre todas las encuestas, las agrupa por asignatura primero y posteriormente por profesor, y construye una tabla de frecuencias y calcula estadísticos básicos.

La segunda, genera un reporte en formato pdf por cada profesor, mediante el empleo de un bucle que recorre todas las asignaturas y los profesores que la imparten. El bucle accede a las tablas generadas en el paso anterior, y construye un data frame con el nombre del profesor y la asignatura para que sirva de cabecera del informe.

Para guardar la confidencialidad del profesorado se generan nombres de profesores y asignaturas ficticios con el empleo de la librería `charlatan`.

El aspecto del reporte se muestra en la figura

Figura 4: Reporte de Evaluación

ENCUESTA DE EVALUACIÓN DEL PROFESORADO													
CURSO 2016-2017													
													
Asignatura							Profesor						
Hydrogeologist							Boyle						
	Media	Desviación Típica	1	2	3	4	5	6	7	8	9	10	
El profesor/a explica con orden y claridad.	7.18	2.15	2	0	0	2	3	3	10	8	7	4	
El profesor/a logra mantener mi atención en clase.	6.74	2.30	2	2	0	0	6	4	8	8	7	2	
La forma de dar clase y la metodología empleada por el profesor/a facilitan la comprensión de los contenidos de la asignatura.	7.00	2.26	2	1	0	3	1	4	8	11	6	3	
El profesor/a fomenta la participación en clase y anima a los alumnos/as a plantear preguntas y dudas en clase.	7.26	2.35	2	0	2	0	2	8	1	13	4	7	
Los criterios de evaluación establecidos permiten que el profesor/a se forme una visión realista y detallada del nivel de aprendizaje alcanzado por cada alumno/a.	6.87	2.20	2	0	1	4	0	6	8	9	7	2	
El profesor/a acude a clase puntualmente.	7.64	2.23	2	0	1	0	0	7	5	7	10	7	
El profesor/a atiende adecuadamente las dudas y/o preguntas que los alumnos le planteamos en clase.	8.08	2.02	1	0	1	0	3	0	6	7	12	9	
Cuando yo o alguno de mis compañeros/as, envía un correo electrónico al profesor/a, o se ha dirigido al profesor/a a través del campus virtual, el profesor/a ha respondido de forma rápida y eficaz.	8.21	2.13	2	0	0	0	1	1	7	5	12	11	
Resultado de interés asistir a sus clases para preparar adecuadamente la asignatura.	7.69	2.33	2	1	0	1	0	2	9	8	7	9	
Evalúa del 1 al 10 tu grado de satisfacción global con respecto al profesor/a.	7.85	2.10	2	0	1	0	0	2	6	9	15	4	

El dashboard se elabora en `shiny`. Se emplean los mismos procedimientos descritos arriba para la elaboración del reporte. Asimismo se emplean el mismo procedimiento que la sección

de análisis descriptivo del archivo `encuesta.r` para realizar el gráfico.

2.3. Ejercicio empírico

Posterior a la limpieza de datos se realiza un ejercicio empírico empleando los datos de valoraciones de las encuestas, el cual consiste en analizar el efecto de ciertas variables sobre la valoración global. Para ello procedemos en dos pasos: primero realizamos un análisis factorial para reducir la dimensión de los items de valoración (9)³ y condensarlos en dos factores que interpretamos como *docencia* y *relación alumno-profesor*.

Posteriormente empleamos estos factores junto a información del profesorado (sexo, categoría, etc.) y de las asignaturas (cuantitativa vs no-cuantitativa) como variables explicativas en una regresión logística cuya variable dependiente (target) es la valoración global del profesor.

Lo anterior nos permitirá comprender como afectan las variables postuladas y los factores a los que hemos reducido los items de valoración, a la valoración del profesorado

Todo el ejercicio se encuentra descrito con detalle en el archivo `analisis_empirico.html`. Se anima al lector a revisar el mismo si desea analizar en profundidad el procedimiento, a continuación solo se describen los resultados obtenidos.

2.3.1. Análisis factorial

El análisis factorial tiene sentido en la medida que las variables en estudio se encuentren altamente correladas, existen tests previos que permiten contrastar la hipótesis de alta o baja correlación entre las variables, en nuestro caso los items de la encuesta.

El test de Barlett rechaza la hipótesis nula de ausencia de correlación entre los items con un nivel de significatividad del 1 %, y el valor del test KMO es de 0.9337983⁴, por lo que nuestra muestra puede ser calificada de adecuada para un análisis factorial.

La hipótesis de partida, basada en la estructura de preguntas de la encuesta, es la existencia de dos factores que representan la labor docente y relación entre el profesor y el alumno. Procedemos a estimar los factores con el uso de la función `factanal` la cual emplea el método de máxima verosimilitud.

Estimamos el modelo sin rotaciones, con una rotación ortogonal (varimax) y una rotación oblicua (promax). Determinamos la existencia de dos factores en el modelo con rotación oblicua, las cargas se representan en la tabla 1 y el gráfico 5.

El primer factor tiene una alta carga en los items 1,2,3 y 9, mientras que el factor 2 posee

³En la sección anterior se ha mencionado que tenemos 10 items, el item 10 corresponde a la valoración global de la asignatura, que en nuestro ejercicio es la variable dependiente o target. Por eso aplicamos el análisis factorial a los 9 factores restantes.

⁴Mientras más cercano a uno, existe grado de asociación de las variables debido a factores comunes.

Cuadro 1: Cargas Factoriales

	Factor1	Factor2
Item_1	0.79	0.11
Item_2	0.89	0.02
Item_3	0.86	0.06
Item_4	0.39	0.47
Item_5	0.40	0.35
Item_6	-0.07	0.67
Item_7	-0.01	0.88
Item_8	-0.01	0.73
Item_9	0.69	0.20

cargas relativamente mayores en los items 6, 7 y 8. En el gráfico se observa una separación más clara entre los items que corresponden a cada factor, los correspondientes a la labor docente y a la relación profesor alumno.

Estos resultados confirman nuestra hipótesis de partida.

Finalmente en la figura 6 analizamos el gráfico de sedimentación para verificar el número de factores a tomar.

Observando la tasa a la cual deja de caer la varianza explicada, pareciera que dos factores es una buena alternativa, a pesar que el segundo valor propio se encuentra debajo de uno.

Decidimos mantener dos factores debido a su facilidad de interpretación y asociación con los factores no observables que se han postulado al inicio del análisis.

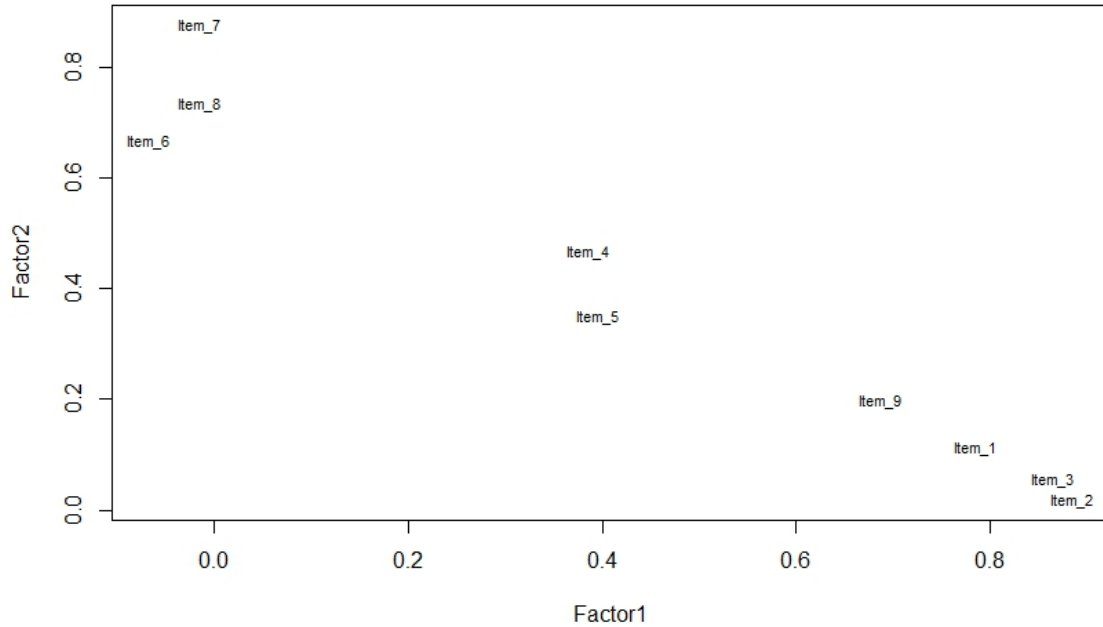
2.3.2. Regresión logística

En esta sección se construye una regresión logística con el objeto de analizar el efecto de una serie de variables sobre la valoración de una asignatura.

La variable dependiente (respuesta, target) es la valoración de una asignatura. Las variables independientes o regresores son:

- Los factores obtenidos en el apartado anterior, y los cuales resumen las características latentes evaluadas en la encuesta, *docencia* y *relación alumno-profesor*
- Características individuales de los profesores:
 - Edad
 - Sexo
 - Tipo de asignatura (cuantitativa o no)
 - Categoría (licenciado, doctor, titular o catedrático)

Figura 5: Cargas Factoriales



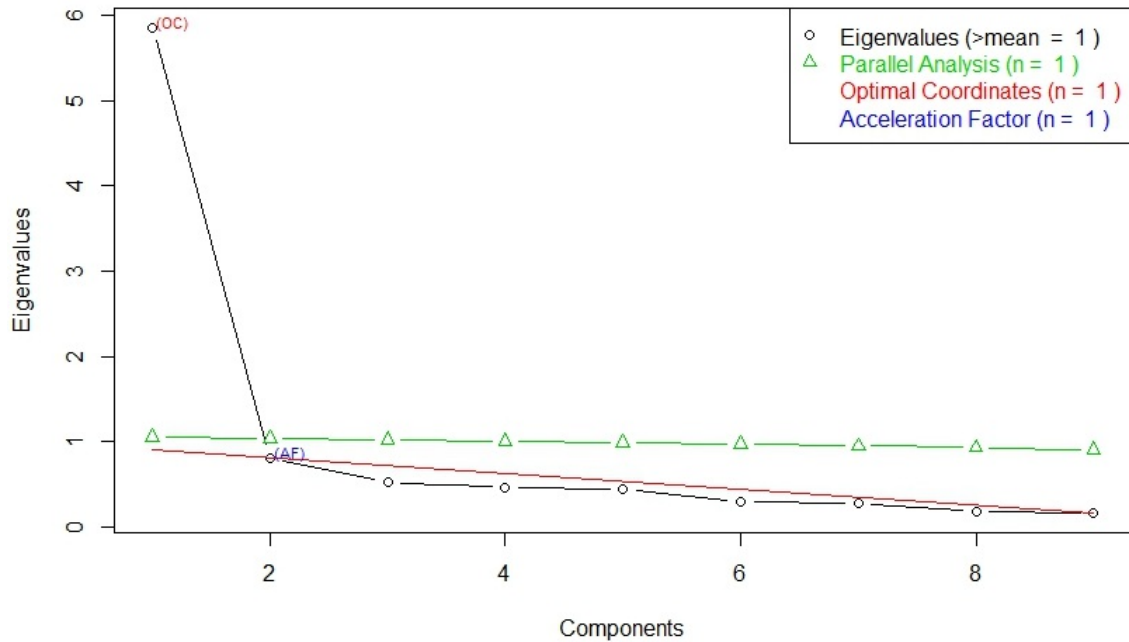
- Tener o no un cargo administrativo dentro de la universidad
- Estar en posesión o no de una acreditación de la ANECA

Los factores y la edad son variables continuas, el tipo de asignatura, cargo administrativo y la acreditación son binarias 0-1 y la categoría es una variable categórica con 4 niveles del 1 al 4.

Las variables se encuentran codificadas del siguiente modo:

- Sexo: 1 si es hombre y 0 si es mujer
- Cuantitativa: 1 si el profesor imparte alguna asignatura de tipo cuantitativo y 0 en caso contrario
- Categoría: 1 licenciado, 2 doctor, 3 titular y 4 catedrático
- Tareas de gestión: 1 si se dedica a ello y 0 si no lo hace
- Acreditación: toma el valor 1 si el individuo tiene algún tipo de acreditación y 0 en caso contrario (los funcionarios están incluidos como acreditados)

Figura 6: Gráfico de Sedimentación (Scree Plot)



En el cuadro 2 se muestran los resultados de la regresión logística expresados como los efectos marginales⁵.

Solo los dos factores y la edad son estadísticamente significativos. La mayoría de las características individuales de los profesores no parecieran ser relevantes a la hora de que el alumno evalúe la asignatura, mientras que los factores que engloban el desempeño de la actividad docente, la relación profesor-alumno y la edad del instructor si lo son.

Según estos resultados, un incremento de una unidad del factor 1 (docencia) incrementa la probabilidad de obtener una valoración alta en un 24%, mientras que un alza unitaria del factor 2 (relación profesor-alumno) aumenta la probabilidad de una valoración alta en un 22%. Mientras que una variación de la edad no parece tener ningún efecto sobre dicha probabilidad.

Otra manera de interpretar los coeficientes es calculando la exponencial de los mismos. El exponencial de los coeficientes nos indican que tan más probable es en términos relativos obtener una valoración alta cuando varía x_i en una unidad manteniendo el resto de la x 's constante.

En el cuadro 3 se muestran los coeficientes calculados de esta manera. Manteniendo el resto de variables constante, ante un incremento unitario del factor 1 (docencia), es 50,14

⁵La versión original de la salida de la regresión logística se puede ver en el archivo [analisis_empirico.html](#)

Cuadro 2: Efectos Marginales	
	Modelo
Factor1	0,24*** (0,02)
Factor2	0,23*** (0,02)
edad	0,00*** (0,00)
sexo1	0,00 (0,01)
cuantitativa1	0,00 (0,01)
categoría2	-0,00 (0,01)
categoría3	-0,04 (0,07)
categoría4	-0,03 (0,04)
Gestion1	0,00 (0,01)
Acreditacion1	0,01 (0,01)
Num. obs.	3292
Log Verosimilitud	-678.20
Deviance	1356.41
AIC	1378.41
BIC	1445.50

*** $p < 0,001$, ** $p < 0,01$, * $p < 0,05$

veces más probable tener una valoración alta. Una variación del factor 2 (relación profesor-alumno) en una unidad hace 38 veces más probable obtener una valoración alta. Mientras que la variable sexo no parece tener efecto alguno debido a que su valor es uno.

En cuanto a la bondad de ajuste, el modelo estimado presenta un pseudo R^2 de 0.63, el cual es un valor relativamente alto tomando en cuenta el problema planteado. Realizando un contraste de significatividad global, similar al contraste F empleado en los modelos lineales, se rechaza la hipótesis nula de que el modelo estimado y un modelo con solo el intercepto sean iguales, por lo que el test de significatividad conjunta del modelo apunta a que el mismo es significativo.

Para evitar la estimación de un modelo sobreajustado, repetimos el ejercicio anterior con una muestra de entrenamiento (65 %) y otra de prueba (35 %). Observando la significatividad individual y conjunta, los efectos marginales y el pseudo R^2 , el modelo estimado (entrenado)

Cuadro 3: Probabilidades relativas

	<i>Variable dependiente:</i>
	Valoración global
Factor1	50.141*** (0.161)
Factor2	37.947*** (0.156)
edad	1.038*** (0.010)
sexo1	1.041 (0.154)
cuantitativa1	1.020 (0.205)
categoría2	0.954 (0.194)
categoría3	0.518 (1.027)
categoría4	0.655 (0.598)
Gestion1	1.002 (0.186)
Acreditacion1	1.180 (0.189)
Constant	2.102 (0.507)
Observaciones	3,292
Log Verosimilitud	−678.204
Akaike Inf. Crit.	1,378.407
<i>Nota:</i>	*p<0.1; **p<0.05; ***p<0.01

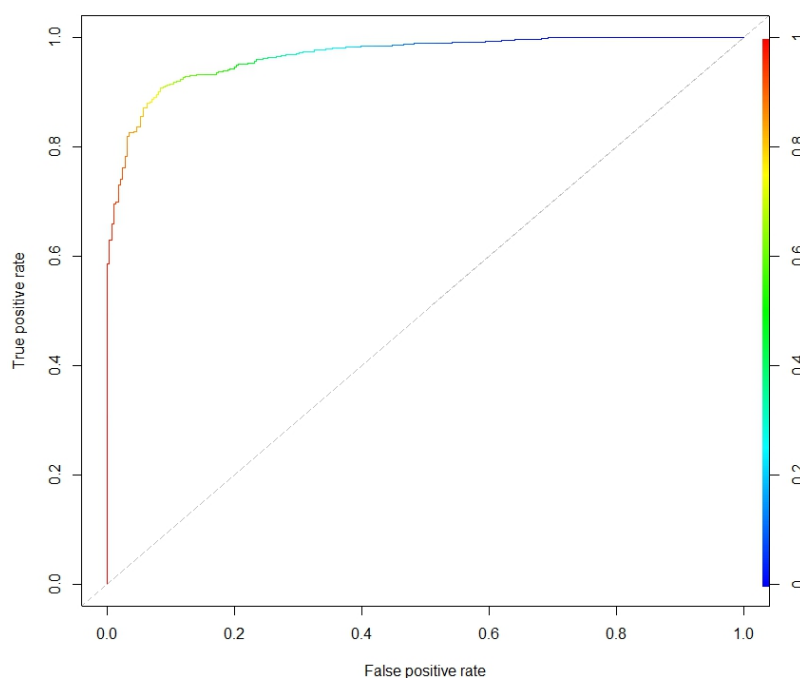
en la muestra de entrenamiento no presenta ningún cambio a resaltar, en efecto son bastantes similares.

Se calcula la matriz de confusión empleando como límite para determinar el *éxito* (obtener una valoración alta) no el 0,5 que trae por defecto R, sino el que maximiza la precisión del modelo.

En este caso tenemos 3 valores que maximizan la precisión: 0.6074202, 0.6025771 y 0.5942560. Empleamos el mayor de los valores.

La precisión es del 91,5%, la tasa de aciertos de las valoraciones bajas 88,11 % y la de valoraciones altas del 92,6 %. En el gráfico 7 se puede ver la curva ROC del modelo

Figura 7: Curva ROC



Se obtiene una curva ROC que muestra una alta capacidad de clasificación del modelo. Calculamos el área bajo la curva ROC (**AUC**) y el valor es cercano a uno (**0,966**), confirmando la buena capacidad del modelo para discriminar las valoraciones altas de las bajas.