Minería de datos y modelización predictiva

Hugo Gómez Sabucedo
hugogomezsabucedo@gmail.com

Máster Big Data, Data Science & Inteligencia Artificial

Curso 2024-2025

Universidad Complutense de Madrid

Índice

1. Introducción

En esta práctica se nos pide, a partir de un archivo con datos sobre diferentes resultados electorales, seleccionar unas variables, con el objetivo de construir tanto un modelo de regresión lineal, a partir de una variable objetivo continua; como un modelo de regresión logística, a partir de una variable binaria. Para ello, se deberán asignar correctamente los tipos de los datos y realizar un análisis de los mismos, con el objetivo inicial de depurarlos, para asegurarnos que no tenemos variables incoherentes o que no se ajusten al modelo. A continuación, se corregirán los errores que se hayan detectado, así como los valores atípicos y perdidos. Una vez con los datos limpios, podremos analizar las relaciones entre las distintas variables input y las variables objetivo, para poder proceder con la creación de los dos modelos solicitados, para cada una de las variables.

Cada registro del archivo viene identificado por Name y CodigoProvincia (ya que, observando el conjunto de datos, se han encontrado municipios con el mismo nombre pero en provincias diferentes). Adicionalmente, contienen información sobre la CCAA a la que pertenecen. Por otra parte, tenemos las variables objetivo, que se dividen en las variables continuas (AbstentionPtge, el porcentaje de abstención; IdaPct, el porcentaje de votos a partidos de izquierdas; DchaPct, el porcentaje de votos a partidos de derechas; y OtrosPct, el porcentaje de votos a otros partidos); y las variables binarias o dicotómicas (AbstencionAlta, que vale 1 si el porcentaje de abstención es mayor al 30 % o 0 en otro caso; Izquierda, que toma valor 1 si la suma de votos a los partidos de izquierda es superior a la suma de votos a derchas y otros, y 0 en caso contrario; y Derecha, análoga a la anterior, pero para partidos de derechas). De estas, se escogerá **AbstentionPtge** para realizar la regresión lineal; e **Izquierda** para realizar la regresión logística.

Por otra parte, tenemos un conjunto de 29 variables explicativas, las cuales no entraremos a explicar en detalle, pero que se corresponden con aspectos demográficos o sociológicos de los distintos municipios, como puede ser el porcentaje de población por tramos de edad, el porcentaje de desempleo por edades o sectores, el número de empresas de los municipios por tipo de actividad y la actividad principal del mismo, el porcentaje de población respecto a su CCAA y provincia de nacimiento, o, evidentemente, el censo, población, superficie y densidad del municipio.

De esta forma, nuestro objetivo será construir un modelo de regresión lineal para la variable AbstentionPtge y un modelo de regresión logística para la variable Izquierda, que nos permitan en el primer caso predecir el porcentaje de abstención, y en el segundo caso, la probabilidad de que los partidos de izquierdas sean los más votados.

2. Importación de datos y análisis descriptivo

2.1. Importación

Para importar los datos, una vez establecido el directorio de trabajo a la carpeta correspondiente con `os.chdir`, se emplea la siguiente instrucción:

```
datos = pd.read_excel("DatosElecciones.xlsx", sheet_name
='DatosEleccionesEspaña')
```

Mediante `datos.head(5)` podemos ver las 5 primeras filas, y con `datos.dtypes` podemos ver los tipos de datos de las variables, lo que usaremos para comprobar que cada una de las variables tiene asignado el tipo que le corresponde (es decir, numérica o categórica). Creamos una lista con las variables, y las dividimos en categóricas y nnuméricas.

```
1 variables = list(datos.columns)
2 numericas = datos.select_dtypes(include=['int', 'int32',
  'int64','float', 'float32', 'float64']).columns
3 categoricas = [v for v in variables if v not in
  numericas]
```

3. Corrección de errores

4. Análisis de valores atípicos

5. Análisis de valores perdidos

6. Detección de relaciones entre variables

7. Regresión lineal

8. Regresión logística