Proceso de ETL de la Práctica Final – Grupo Electomedia

En este documento describimos muy brevemente el proceso de ETL aplicado a los datos base de la Práctica Final del grupo Electomedia. Este proceso de ha llevado a cabo utilizando 11 cuadernos (notebooks) que tratan y finalmente engloban bases de datos procedentes de distintas fuentes que finalmente se engloban para formar los datasets que se utilizan para su análisis y creación de modelos. Todos los cuadernos contienen comentarios que describen el proceso de tratamiento de los datos, por lo que no nos extenderemos en demasía en este fichero.

Podemos dividir el proceso de ETL en estos pasos, que expondremos en mayor detalle a continuación:

- **Datos electorales**
- **Datos sociodemográficos**
- **Unificación de datos electorales y sociodemográficos**
- **Similitud de las secciones electorales**

Datos electorales

Los datos electorales son la base de la Práctica, y constan de los resultados electorales detallados de las secciones censales en los comicios generales de 2011, 2015, 2016, y los dos celebrados en 2019. Los datos se han extraído de la página web del Ministerio del Interior. Se trata de, para cada elección, de un conjunto de ficheros .DAT. Los dos fundamentales son los que definen las distintas mesas electorales, y el resultado de cada partido en ellas.

En el cuaderno ***Transformacion_inicial_elecciones*** cargamos el fichero .DAT, y colocamos los votos recibidos por cada partido en columnas. Los códigos y nombres de cada partido los hemos definido nosotros para simplificar y poder compararlos con los de otras elecciones. Posteriormente, añadimos datos comunes a las mesas electorales, tales como censo o número de votantes, que proceden de otro fichero .DAT. Finalmente, agrupamos los datos por secciones censales, que contienen una o varias mesas electorales. Se crea un dataset para cada elección, cinco en total, cuyas filas son los resultados en cada sección, unas 36 mil, aunque el número no es fijo.

El siguiente paso es el uniformizar las columnas (votos a cada partido) para todas las elecciones. Ello lo ejecutamos mediante el cuaderno ***Reforma_Resultado_secciones***, que básicamente crea las columnas para cada partido que no se ha presentado en cada elección. Posteriormente

ordenamos las columnas de forma común, con lo que conseguimos una estructura igual para los cinco datasets, uno por cada elección.

Para mejorar la interpretabilidad de los datos, introducimos los nombres de las CCAA, provincias y municipios donde se sitúa cada sección. Tomamos los datos de los ficheros de cada elección del Ministerio del Interior, y del INE. El hecho que ambos tengan distintos códigos para las comunidades autónomas ha sido una fuente de problemas. La inserción de estos nombres se lleva a cabo en el cuaderno ***Identificacion_Mun_Prov***. A través de los tres cuadernos, los datos electorales ya están listos para unificar con los datos sociodemográficos.

Datos sociodemográficos

Hemos complementado los datos electorales de las secciones con datos sociodemográficos, también a nivel de sección censal cuando fue posible, o a nivel de municipio en su defecto. El origen de los datos ha sido el INE, el Ministerio de Inclusión Social y Migraciones, y el Ministerio de Trabajo y Economía Social.

El primer conjunto de datos es la distribución de edades de la población de cada sección censal, que se pueden descargar aquí de la web del INE. Los ficheros son de formato .PX, por lo que hubo que transformarlo en dataframes mediante un sencillo cuaderno en R. El dataframe resultante lo tratamos en el cuaderno ***Secciones_edades***, que básicamente lleva a cabo un método pivot() y pone en columnas los tramos de edad y la distribución por sexos. Los datos del INE tienen periodicidad anual, por lo que utilizamos los de cada año para cada elección.

El INE está produciendo una serie de estadísticas experimentales por secciones censales de la renta per cápita, por hogar, y su distribución según su origen, que hemos descargado de este link. Su formato es complicado, aparte de contener bastantes datos en blanco. A ser ficheros únicos que contienen datos a partir de 2015, hemos utilizado sus mismos datos para todas las elecciones. Hemos dividido su tratamiento en dos cuadernos: ***Rentas_INE*** y ***Distribucion_rentas_INE***. A los ficheros Excel que se han descargado les hemos sometido a un ligero pre-pretratamiento, básicamente eliminado las filas iniciales.

Los datos de carácter más socio-económico son los de afiliación a la Seguridad Social y los de los demandantes del paro, que se pueden descargar de aquí y aquí, respectivamente. Al ser datos mensuales, éstos se ajustan perfectamente en el tiempo a las elecciones, pero, sin embargo, tienen solo información a nivel de municipio. Hemos tratado estos datos en un solo cuaderno, ***Afiliacion_SS_Paro***.

Unificación de datos electorales y sociodemográficos

Una vez que tenemos preprocesados los datos de distinto tipo, debemos unificarlos en una sola base de datos, ya que todos ellos están referenciados en la práctica a las secciones censales. Ello lo llevamos a cabo en el cuaderno ***Unificacion***. El proceso es relativamente simple: partiendo de los datos electorales vamos aplicado el método .merge() de pandas, tomado como columna de unión la del código de sección o del municipio.

Dentro de este cuaderno creamos nuevas columnas que pensamos que nos serán útiles de cara a correlacionarlas con los datos electorales de cada sección. En su mayor parte se trata de ratios, no de datos absolutos, tales como porcentaje de censados por encima de los 64 años de edad. De esa forma obtenemos nuevos datos que son independientes del tamaño de cada sección. Obtenemos finalmente los datasets con los que podremos modelizar las secciones electorales, compuestos por cerca de 100 columnas.

Hemos llevado a cabo el mismo proceso conjunto de tratamiento de datos con las provincias, aunque no hemos incluido los cuadernos al no tener gran interés, ya que finalmente no hemos utilizado los datasets obtenidos.

Similitud de las secciones electorales

A priori se podría pensar que las secciones censales son iguales y fijas para todas las elecciones, pero no es así. Al contrario, muchas varían de definición geográfica, al menos una vez al año. El INE informa de ello anualmente, y distribuye una definición geométrica de todas ellas en su página web, estando en formato SHAPE. Estas secciones se pueden mostrar en un mapa mediante la librería Geopandas, y ello es lo que hacemos en el cuaderno del mismo nombre, ***Geopandas_Ejemplo***. En este cuaderno hacemos un merge con los datasets correspondientes a cada elección que obtenemos del cuaderno Unificación, con lo que conseguimos una visualización por secciones, no solo de datos electorales, sino también sociodemográficos.

Una utilidad de Geopandas es que proporciona métodos para hallar el centroide de cada sección. Con ese dato podemos determinar la sección equivalente de elección en elección, es decir, con qué sección en la elección, pongamos, de 2015, deberíamos comparar en la elección de 2016. El criterio es la mayor proximidad geográfica, dentro del mismo municipio o provincia. El dataframe de las equivalencias de cada sección a lo largo de las 5 elecciones lo hemos calculado en el cuaderno ***Evolucion_Secciones***. Como se ha mencionado anteriormente, la confusión entre los códigos de las CCAA nos hizo variar el dataframe en el cuaderno ***Evolucion_Secciones_reforma***, que no es en la práctica más que la aplicación de un mapeo.

En resumen, tras poner en marcha estos cuadernos estamos listos para determinar los modelos a un sinfín de variables objetivo que podemos definir.