Hola Equipo!!

Como os comenté ayer os describo lo que he estado haciendo, y que queda por hacer de mi lado, siempre con vuestro apoyo y ayuda que agradezco de veras.

Ya antes de las vacaciones me dediqué a crear un dataset por secciones de una elección en concreto, que es el de abril de 2019. El dataset lo tenemos en el drive, *gen_A19_unif_cols_prov_copia.txt*.

Además de los datos electorales, tiene columnas de población de cada sección, además de datos de paro, afiliación a la Seguridad Social y rentas por persona y hogar del municipio al que pertenece.

Estos dos días he estado trabajando en:

- **Hacer los datasets de las otras cuatro elecciones:** 11, 15, 16, y nov 19. Voy progresando. He comenzado por preparar los datos de paro y afiliación (los electorales ya están).

- **Equivalencia de las secciones en cada elección.** Es decir, cual es la sección de una elección que se correspondería con la de otra elección. Este es un tema delicado, ya que algunas aparecen, otras desaparecen, y desde luego cambian de definición geométrica. Lo que estoy haciendo es para cada sección en una elección seleccionar las más cercanas, dentro se su mismo municipio, en las otras cuatro elecciones. La distancia es la definida por los centroides respectivos, calculados con Geopandas. Es un proceso lento y complicado, pero voy avanzando aquí también.

- **Agregación de los datos por provincias.** Estoy en ello también; quizás el punto más complicado es el tema de las rentas, pues están puestas por persona, por lo que tendré que hacer una media ponderada para cada municipio en función de la población.

Dentro de la preparación de los datos brutos quedaría:

- **Crear los ficheros geojson para D3.js.** Aquí habría que añadir a los datasets una columna con la definición geométrica de las secciones. Geopandas, por fortuna, transforma los ficheros shapefile en geojson.

- **Quizás añadir columnas como superficie y/o códigos postales.** No creo que sean estos datos muy útiles, pero tampoco pienso que nos fuese a llevar mucho tiempo.

A partir de tener estos datos, ya podríamos empezar a hacer **modelos**. Lo curioso es que tenemos mucha flexibilidad, y podemos elucubrar muchas variables objetivo, tales como:

- Partido ganador (clasificación)
- Partido segundo (clasificación)
- Ganador entre PP y PSOE (clasificación)
- % de voto del PP, u otro partido (clasificación y regresión)

Combinando datasets de distintas elecciones podemos hacer modelos de los cambios de votos entre ellas, para eso necesitamos la equivalencia de las secciones en cada elección.

En cada modelo se pueden utilizar **algos** como LR, Random Forest, árboles de decisión, SVR, SVC, etc… además de redes neuronales, que por cierto lo piden que hagamos. También tendremos que aplicar algún algo de **reducción de dimensionalidad**, aquí incluyendo el encontrar qué secciones de cada provincia la representan mejor, como hemos ya comentado alguna vez. Se podrá hacer algún tipo de **clusterización** (que creo que también piden), tanto por secciones como por provincias.

Puede parecer mucho, pero pienso que una vez tengamos los datasets, deberíamos ir ya rápido. A lo que tengo algo más de respeto es a cómo trasladamos los cuadernos/código a GCP, y demás, en eso yo personalmente estoy bastante pez.