Entrega 2: Metodología y EDA

José Saint Germain
joesg998@gmail.com

1. Introducción

El objetivo de esta entrega es realizar una breve descripción de las metodologías que se utilizarán durante el trabajo final de especialización, así como realizar un análisis exploratorio de los datos (EDA), para comprender mejor la estructura de los datos que se trabajarán.

2. Metodología

Como lo que buscamos realizar es experimentar con diferentes datos el mismo trabajo realizado por el FMIm (Cebotari et al., 2024), vamos a replicar las mismas técnicas de optimización de hiperparámetros, así como los mismos algoritmos de entrenamiento y de intepretación de resultados.

Los algoritmos que se utilizarán serán Random Forest (Breiman, 2001) y XGBoost (Chen y Guestrin, 2016). Para ajustar los hiperparámetros se utilizará la optimización bayesiana junto al método de block-time-series cross-validation. Por último, la métrica a optimizar y que se utilizará para comparar predicciones será el área bajo la curva (AUC). Adicionalmente, se realizará un análisis exploratorio de datos de manera introductoria al trabajo y se buscará utilizar valores Shapley para analizar los resultados de cada algoritmo.

3. Análisis Exploratorio de Datos

Como descripción general de la base de datos de VDEM, podemos mencionar que cuenta con 27734 filas y 4607 columnas. Como es una base de datos de panel, se tiene información de 202 países durante 235 años. Para comprender la estructura de la información, es importante destacar que la base original cuenta con información brindada por distintos expertos para cada país en cada año. Para poder procesar y obtener la base final, se agrega la información de diferentes maneras. Es por este motivo que, además de la información identificactoria de cada país (la cual se repite en cada año), la mayoría de las variables sustantivas cuentan con diferentes versiones, por cada tipo de variable de agregación generada. Por ejemplo, una variable puede contar con su versión princial, la cual es un promedio reescalado del 1 al 5, sumado a una versión con la media simple (con sufijo $_{m}ean$); $unaversinconelvalormximoymnimoexpresadoporunexperto(_{c}odehighycodelow, respectivamente); yunavers$

3.1. Ánalisis de nulos

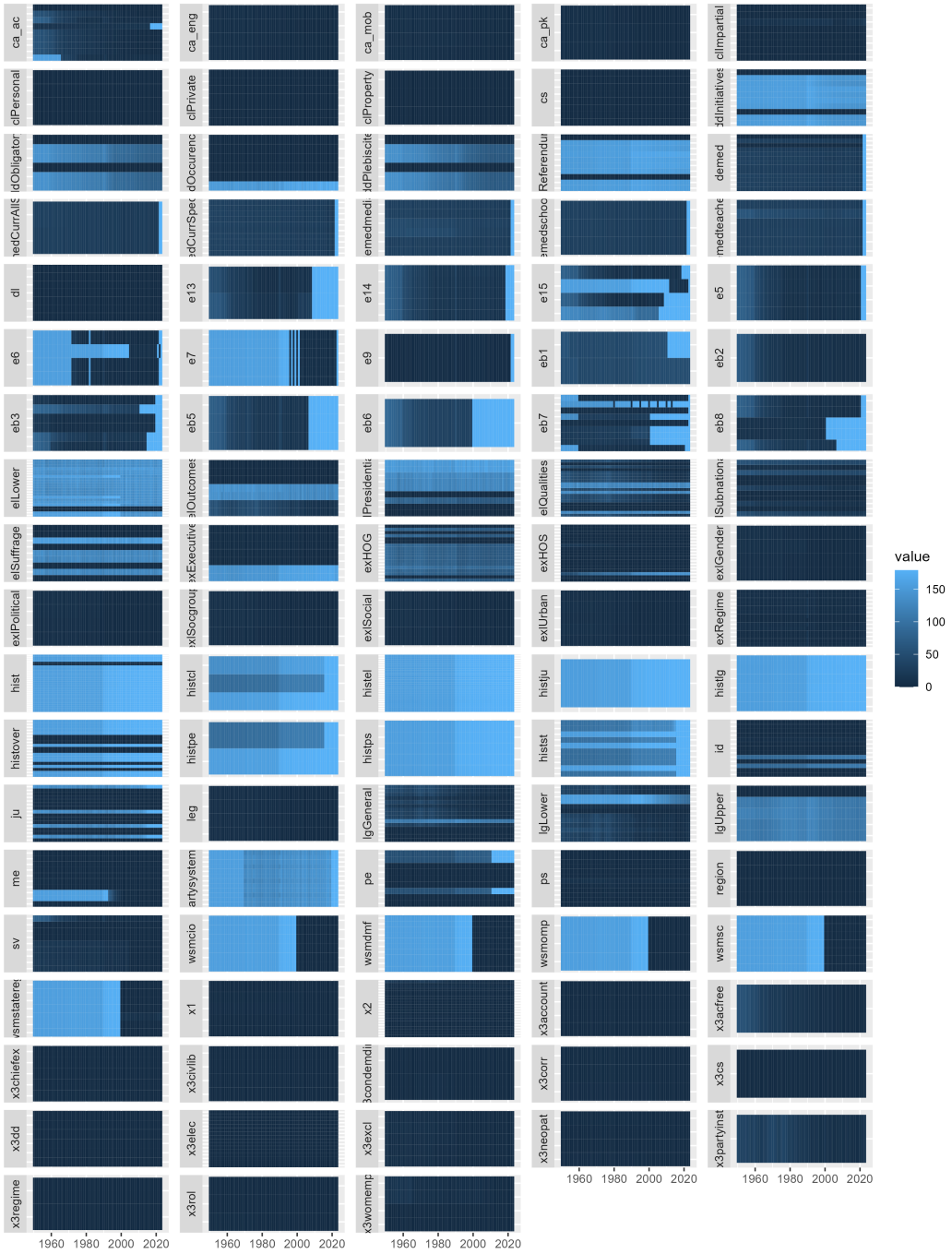

Figura 1: Conteo de nulos por año y agrupador de variables

3.2. Análisis de variable objetivo

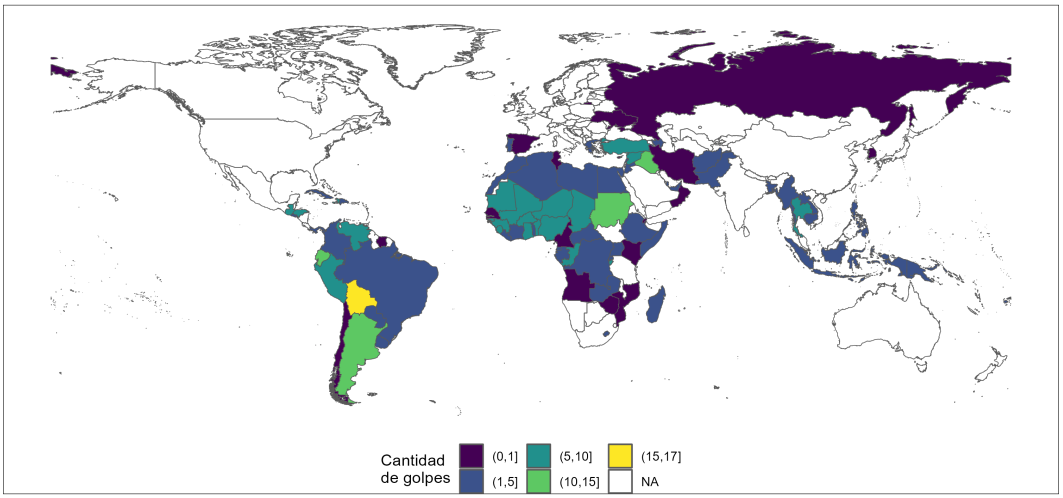

Figura 2: Conteo de golpes de estado en el mundo

4. Preprocesamiento de los datos

Cargar el dataset con los datos para cada sujeto y los nombres y coordenadas de las regiones cerebrales a las que se les registró la actividad. Reportar cuántos sujetos y cuántos estados de sueño se observan en el conjunto de datos.

Referencias

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd interntional conference on knowledge discovery and data mining*, 785-794.

Cebotari, A., Chueca-Montuenga, E., Diallo, Y., Ma, Y., Turk, R., Xin, W., & Zavarce, H. (2024). *Political Fragility: Coups d'État and Their Drivers*. IMF Working Paper 24/34. https://doi.org/https://doi.org/10.23696/mcwt-fr58