
Entrega 2: Metodología y EDA

José Saint Germain
joesg998@gmail.com

1 **1. Introducción**

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

5 **2. Metodología**

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

9 Los algoritmos que se utilizarán serán Random Forest (Breiman, 2001) y XGBoost (Chen y Guestrin,
10 2016). Para ajustar los hiperparámetros se utilizará la optimización bayesiana junto al método de
11 block-time-series cross-validation. Por último, la métrica a optimizar y que se utilizará para comparar
12 predicciones será el área bajo la curva (AUC). Adicionalmente, se realizará un análisis exploratorio
13 de datos de manera introductoria al trabajo y se buscará utilizar valores Shapley para analizar los
14 resultados de cada algoritmo.

15 **3. Análisis Exploratorio de Datos**

16 Como descripción general de la base de datos de VDEM (Coppedge et al., 2024), podemos mencionar
17 que cuenta con 27734 filas y 4607 columnas. Como es una base de datos de panel, se tiene información
18 de 202 países durante 235 años. Para comprender la estructura de la información, es importante
19 destacar que la base original cuenta con información brindada por distintos expertos para cada país en
20 cada año. Para poder procesar y obtener la base final, se agrega la información de diferentes maneras.
21 Es por este motivo que, además de la información identificatoria de cada país (la cual se repite en
22 cada año), la mayoría de las variables sustantivas cuentan con diferentes versiones, por cada tipo de
23 variable de agregación generada. Por ejemplo, una variable puede contar con su versión principal, la
24 cual es un promedio reescalado del 1 al 5, sumado a una versión con la media simple (con sufijo
25 _mean); una versión con el valor máximo y mínimo expresado por un experto (_codehigh y _codelow,
26 respectivamente); y una versión con el desvío estándar (_sd), en caso de buscar conocer el grado de
27 'acuerdo' entre los expertos respecto a la situación del país.

28 A la base original obtenida desde la librería de VDEM, se le realizaron los siguientes filtros: en
29 primer lugar, se removieron todas las variables que no sean las principales, es decir, que no cuenten
30 con sufijo. De esa manera, se busca reducir el tamaño de la base y así poder agregar nuevas columnas
31 mediante ingeniería de atributos. En segundo lugar, se filtraron los años superiores a 1950, para
32 adecuarnos al periodo utilizado en el artículo del FMI. De esa manera, la base filtrada cuenta con
33 12208 filas y 1460 columnas.

34 **3.1. Análisis de nulos**

35 Debido a la alta cantidad de variables, no es posible realizar un análisis pormenorizado de la presencia
36 de nulos en cada una. Por ese motivo, se decidió visualizar la misma mediante los agrupadores de

41 De este gráfico podemos aprehender ciertos patrones sobre la presencia de nulos en algunos grupos
42 de variables: En primer lugar, observamos variables que, anteriormente a un año puntual, no cuentan
43 con información. En este ejemplo caen las variables sobre gobernanza otorgadas por el banco mundial
44 (e7), las preguntas pertenecientes a la encuesta de sociedad digital (wsmcio), variables referentes a la
45 libertad en medios digitales (wsmdmf), las referentes a la polarización en medios online (wsmomp) y
46 las referentes a clivajes sociales (wsmsc).

47 En segundo lugar, figuran casos contrarios, en donde a partir de determinado año la cantidad de datos
48 faltantes salta a la totalidad de los casos. En este grupo figuran las variables asociadas a instituciones
49 y eventos políticos (e13), cuya fuente es un artículo de Przeworski de 2013; las variables cuya
50 fuente es la base de datos polity V (e14); las variables sobre educación (aumentan los nulos en
51 algunas variables) (eb1); las variables sobre recursos naturales (eb5), cuya fuente tiene datos hasta
52 2006; las variables sobre infraestructura (eb6); y las relacionadas a conflictos (eb8). En general, esta
53 discontinuidad sucede debido a que la información de estas variables provienen de fuentes externas
54 no gestionadas por VDEM, las cuales finalizaron su serie en un año puntual.

55 **3.2. Análisis de variable objetivo**

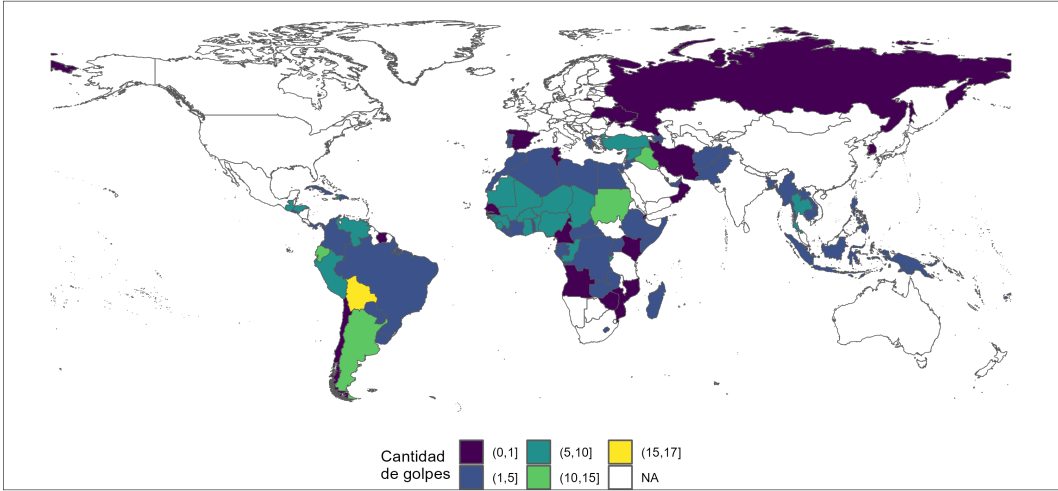


Figura 2: Conteo de golpes de estado en el mundo

56 **4. Preprocesamiento de los datos**

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

60 **Referencias**

61 Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32. <https://doi.org/http://doi.org/10.1023/A:1010933404324>
62
63 Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the*
64 *22nd international conference on knowledge discovery and data mining*, 785-794. <https://doi.org/https://doi.org/10.48550/arXiv.1603.02754>
65
66 Cebotari, A., Chueca-Montuenga, E., Diallo, Y., Ma, Y., Turk, R., Xin, W., & Zavarce, H. (2024).
67 *Political Fragility: Coups d'État and Their Drivers*. IMF Working Paper 24/34. <https://doi.org/https://doi.org/10.23696/mcwt-fr58>
68

69 Coppedge, M., Gerring, J., Knutsen, C. H., Lindlberg, S. I., Teorell, J., Altman, D., Angiolillo, F.,
70 Bernhard, M., Borella, C., Cornell, A., Fish, S. M., Fox, L., Gastaldi, L., Gjerløw, H.,
71 Glynn, A., God, A. G., Grahn, S., Hicken, A., Kinzelbach, K., . . . Ziblatt, D. (2024). *V-Dem*
72 *Dataset v14* Varieties of Democracy (V-Dem) Project* (Report). [https://doi.org/https:](https://doi.org/https://doi.org/10.23696/mcwt-fr58)
73 [//doi.org/10.23696/mcwt-fr58](https://doi.org/10.23696/mcwt-fr58)