

---

# 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). Adicionalmente, para la evaluación de performance se utilizará el área bajo la curva (AUC),  
11   en donde un valor de AUC de 0.5 indica que el modelo no tiene capacidad predictiva, mientras que  
12   un valor cercano a 1 indica que el modelo es capaz de predecir con alta precisión.

13   Con respecto al ajuste de hiperparámetros se utilizará la optimización bayesiana. La misma consistirá  
14   en 100 iteraciones en donde se buscará el valor óptimo de los siguientes hiperparámetros:

- 15     ■ Random Forest: profundidad máxima de los árboles (max\_depth) y la submuestra del ratio  
16        de columnas a considerar cuando se construye cada árbol (max\_features).
- 17     ■ XGBoost: la tasa de aprendizaje (learning\_rate) y el término de regularización L2 en los  
18        pesos (reg\_lambda).

19   Adicionalmente el parámetro que establece la cantidad de árboles creados (n\_estimators) quedará  
20   fijado en 1000.

21   Para evitar el data leakage, en cada iteración de la optimización bayesiana se utilizará la validación  
22   cruzada. Sin embargo, como se trabajará con una base de datos de panel, conviene utilizar una versión  
23   adaptada: el método *block-time-series cross-validation*, basado en Burman et al., 1994 y Racine,  
24   2000. El método aplicado en este caso consiste en generar 5 pares de entrenamiento y validación:  
25   1970 - 2009, 2010 - 2011; 1970 - 2011, 2012 - 2013; 1970 - 2013, 2014 - 2015; 1970 - 2015, 2016 -  
26   2017; 1970 - 2017, 2018 - 2019. Por lo tanto, cada set de entrenamiento consiste en observaciones  
27   desde 1970 hasta un año de corte (2009, 2011, 2013, 2015, 2017) y el set de validación contempla los  
28   dos años siguientes del mismo.

29   Una vez realizada la optimización bayesiana, se toman los valores de hiperparámetros que lograron  
30   maximizar el AUC y se entrena el modelo con el set de entrenamiento para intentar predecir los golpes  
31   de estado entre 2020 y 2022. Por último, para interpretar las variables más importantes en la predicción  
32   de golpes de estado, se utilizarán los valores Shapley (Strumbelj y Kononenko, 2010; Lundberg y  
33   Lee, 2017). Basado en la teoría de juegos, los valores Shapley buscan medir la contribución de cada  
34   predictor a la probabilidad de un golpe en relación con la probabilidad promedio de la muestra  
35   prevista de un golpe.

### 36 3. Análisis Exploratorio de Datos

37 Como primera aproximación a la base de datos de Varieties of Democracy o V-Dem (Coppedge  
38 et al., 2024), pasaremos a explicar la manera en que se construye la misma. Las variables centrales se  
39 obtienen a partir de encuestas suministradas a expertos sobre los distintos países. Inicialmente, se  
40 busca que cada país cuente con al menos cinco expertos. Actualmente, la institución cuenta con 22  
41 expertos promedio por país y 7,1 expertos por combinación de variable y país. Una vez obtenida las  
42 respuestas de los expertos, se pasa al proceso de agregación para así conformar una base de datos  
43 donde cada fila corresponde a un país en un año específico. De esta agregación obtienen diferentes  
44 versiones de la misma variable:

- 45     ■ Estimador del modelo (Variable sin sufijo): es la medida recomendada para su análisis.  
46     Corresponde a obtener la mediana del valor de la variable entre los expertos, reescalado a  
47     valores entre -5 a 5.
- 48     ■ Medidas de incertidumbre (\*\_codelow y \*\_codehigh): corresponden a un desvío estandar  
49     por encima y por debajo del estimador del modelo. Usadas conjuntamente, construyen un  
50     intervalo de confianza del 95 %.
- 51     ■ Escala original (\*\_osp): mediana de la variable, pero sin reescalar. Esta versión también  
52     cuenta con sus medidas de incertidumbre correspondientes.
- 53     ■ Media simple (\_mean): mediana de la variable, pero sin reescalar.
- 54     ■ Desvío estándar (\_sd): desvío estándar de la variable.
- 55     ■ Media simple (\_mean): media de la variable.
- 56     ■ Cantidad de expertos (\_nr): cantidad de expertos que respondieron por país, año y variable.

57 Podemos mencionar que la base cuenta con 27734 filas y 4607 columnas. Como es una base de datos  
58 de panel, se tiene información de 202 países durante 235 años. Las variables cuentan con un tipo de  
59 codificación particular que permite identificar el origen de la variable. En primer lugar, el primer  
60 prefijo es indicativo de si fue producido por V-Dem o no:

- 61     ■ v2: variables de V-Dem.
- 62     ■ v3: variables pertenecientes a la base V-Dem histórica.
- 63     ■ v2x\_: Índices principales e índices componentes.
- 64     ■ v2x[indicador de dos letras]: Índices específicos de ciertas áreas (ver más abajo).
- 65     ■ e\_: variables no generadas por V-Dem y variables V-Dem en versión ordinal.

66 El nombre de la variable también permite identificar la área temática a la que pertenece:

- 67     ■ ca: Espacio cívico y académico
- 68     ■ cl: Libertad civil
- 69     ■ cs: Sociedad civil
- 70     ■ dd: Democracia directa
- 71     ■ de: Demografía
- 72     ■ dl: Deliberación
- 73     ■ el: Elecciones
- 74     ■ ex: Ejecutivo
- 75     ■ exl: Legitimación
- 76     ■ ju: Poder judicial
- 77     ■ leg: Legislatura
- 78     ■ lg: Legislatura
- 79     ■ me: Medios de comunicación
- 80     ■ pe: Igualdad política

- 81     ■ ps: Partidos políticos  
 82     ■ sv: Soberanía  
 83     ■ st: Estado  
 84     ■ x: Índice (calculado a partir de variables que también se incluyen en la base de datos)  
 85     ■ zz: Cuestionario posterior a la encuesta  
 86     ■ ws: Encuesta de sociedad digital

87 A la base original obtenida desde la librería de V-Dem, se le realizaron los siguientes filtros: en primer  
 88 lugar, se removieron todas las variables que no sean las principales, es decir, que no cuenten con  
 89 sufijo. De esa manera, se busca reducir el tamaño de la base y así poder agregar nuevas columnas  
 90 mediante ingeniería de atributos. En segundo lugar, se filtraron los años superiores a 1950, para  
 91 adecuarnos al periodo utilizado en el artículo del FMI. De esa manera, la base filtrada cuenta con  
 92 12208 filas y 1460 columnas.

### 93 3.1. Análisis de nulos

94 Debido a la alta cantidad de variables, no es posible realizar un análisis pormenorizado de la presencia  
 95 de nulos en cada una. Por ese motivo, se decidió visualizar la misma mediante los agrupadores de  
 96 variables con las que cuenta el codebook de V-Dem. El mismo, discrimina las variables a partir de  
 97 sus temas en común. De esa manera, en la figura 1 la cantidad de nulos por categoría de variable  
 98 expresado en un mapa de calor, donde cada fila es una variable individual y las columnas los diferentes  
 99 años del panel.

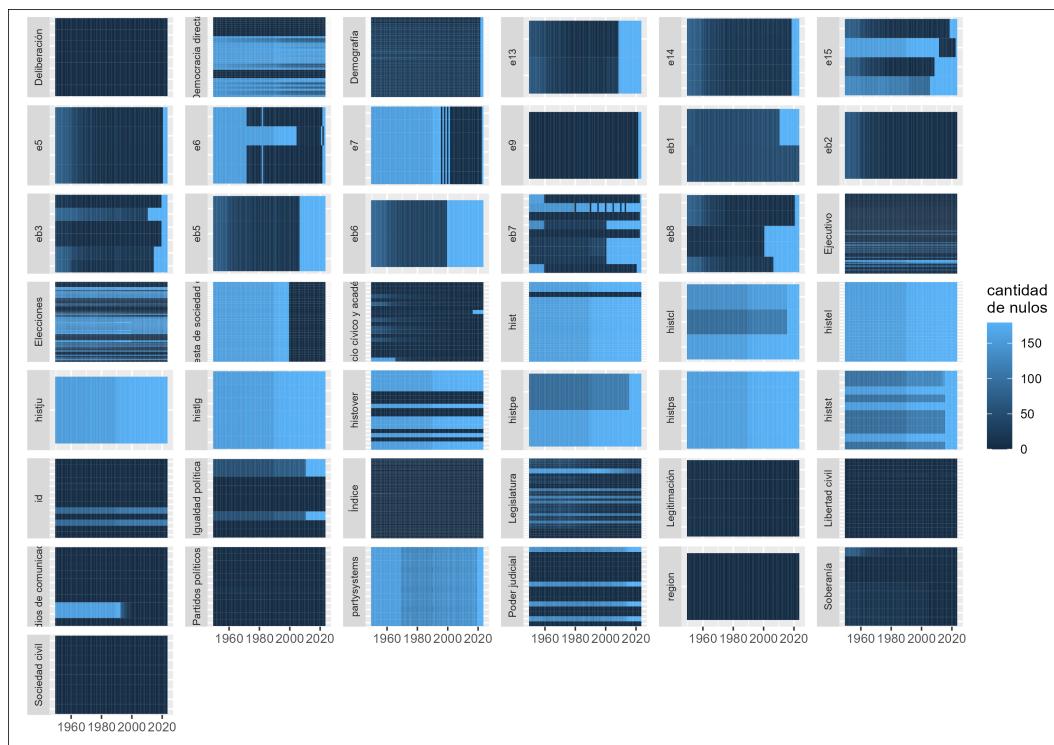


Figura 1: Conteo de nulos por año y agrupador de variables

100 De este gráfico podemos aprehender ciertos patrones sobre la presencia de nulos en algunos grupos  
 101 de variables: En primer lugar, observamos variables que, anteriormente a un año puntual, no cuentan  
 102 con información. En este ejemplo caen las variables sobre gobernanza otorgadas por el banco mundial  
 103 (e7), las preguntas pertenecientes a la encuesta de sociedad digital (wsmcio), variables referentes a la

104 libertad en medios digitales (wsmdmf), las referentes a la polarización en medios online (wsmomp) y  
105 las referentes a clivajes sociales (wsmsc).  
106 En segundo lugar, figuran casos contrarios, en donde a partir de determinado año la cantidad de datos  
107 faltantes salta a la totalidad de los casos. En este grupo figuran las variables asociadas a instituciones  
108 y eventos políticos (e13), cuya fuente es un artículo de Przeworski de 2013; las variables cuya  
109 fuente es la base de datos polity V (e14); las variables sobre educación (aumentan los nulos en  
110 algunas variables) (eb1); las variables sobre recursos naturales (eb5), cuya fuente tiene datos hasta  
111 2006; las variables sobre infraestructura (eb6); y las relacionadas a conflictos (eb8). En general, esta  
112 discontinuidad sucede debido a que la información de estas variables provienen de fuentes externas  
113 no gestionadas por V-Dem, las cuales finalizaron su serie en un año puntual.  
114 Por último figuran los grupos de variables asociados a la base de datos histórica de v-dem (las que  
115 comienzan con hist), lo cual es lógico puesto que esta base busca tomar datos previos a 1900.

### 116 **3.2. Análisis de variable objetivo**

117 Es importante aclarar que en este trabajo no estamos contando la cantidad precisa de golpes de estado  
118 sucedidos en un período de tiempo, sino que simplemente relevamos si al menos un golpe de estado  
119 sucedió en un país y año determinado. Por lo tanto, si un país sufrió más de un golpe de estado en un  
120 año, el mismo será contabilizado una sola vez. Adicionalmente, en este trabajo también se consideran  
121 los golpes de estado que no fueron exitosos, es decir, que no lograron derrocar al gobierno en cuestión.  
122 De allí se desprende que países como Argentina, que en total ha tenido seis golpes de estado exitosos,  
123 figure con el doble de golpes en la figura 2.  
124 Para realizar un paneo general de la variable objetivo, es decir, la presencia de estado a lo largo de los  
125 años, generamos un conteo y lo visualizamos en un planisferio. Destacamos que la mayor presencia  
126 de golpes se encuentra en el continente africano, en América del Sur y parte del Caribe, Medio  
127 Oriente y el Sudeste Asiático, con algunos casos de apenas un golpe en España, Rusia, Ucrania y  
128 Corea del Sur; así como dos y tres golpes en Grecia y Portugal, respectivamente.  
129 Con mayor precisión, observamos que la región del Sahel se destaca con respecto a sus vecinos  
130 africanos. Los países en donde más golpes de estado se han producido son Bolivia (17), Sudán (14),  
131 Argentina (13), Ecuador (11), Iraq (11), Siria(11), Guatemala (10) y Tailandia (10).

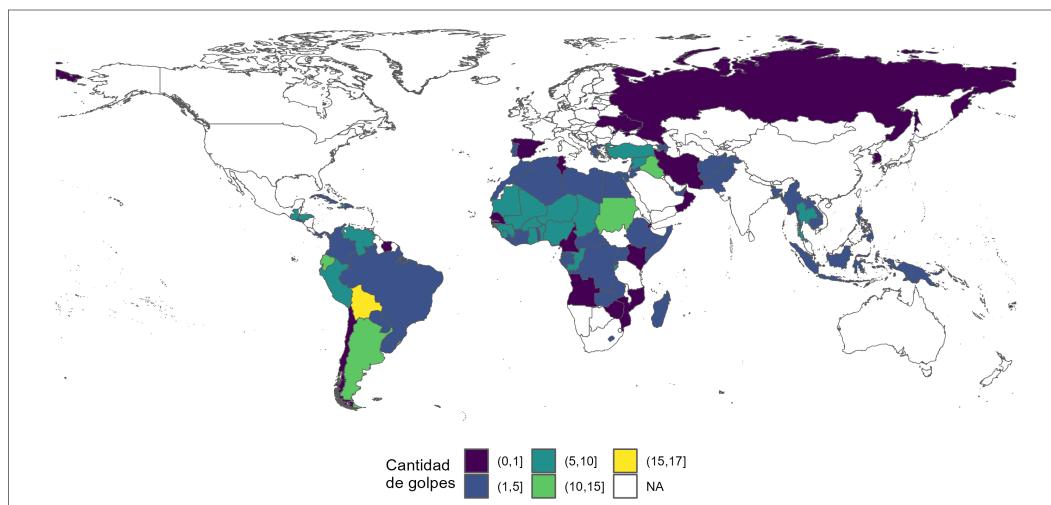


Figura 2: Conteo de golpes de estado en el mundo

132 Desagregando por década se observan algunos cambios, así como la persistencia en algunas regiones.  
133 La región del Sahel y varias naciones circundantes fueron persistentemente afectadas por golpes  
134 de estado desde los años 60. En América del Sur, en cambio, la presencia casi total de situaciones

135 golpistas en la región se fue acotando a partir de los años 80 hasta finalmente desaparecer en el siglo  
136 xxi. Para observar con más detalle y discriminado por años y países se puede ver la figura 5.

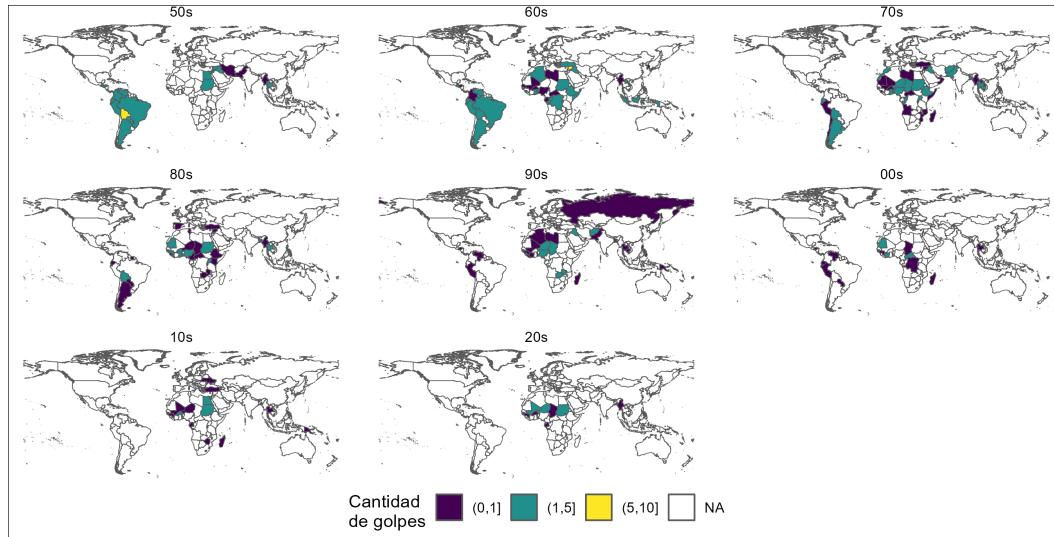


Figura 3: Conteo de golpes por década

#### 137 4. Correlación entre variables

138 Puesto que el dataset (previo filtrado de variables que no se utilizarán para el entrenamiento) cuenta  
139 con alrededor de 1000 columnas, se vuelve imposible realizar un mapa de calor que cruce todas las  
140 variables entre ellas (en la figura 6 dentro del anexo se muestra una tabla de correlación de las variables  
141 individuales, agrupadas por el grupo de variable al que pertenecen). Por lo tanto, recurriremos al  
142 agrupamiento de variables provisto por el codebook para reziliar un gráfico de correlación entre los  
143 promedios de grupos de variables, como se puede apreciar en la figura 4.

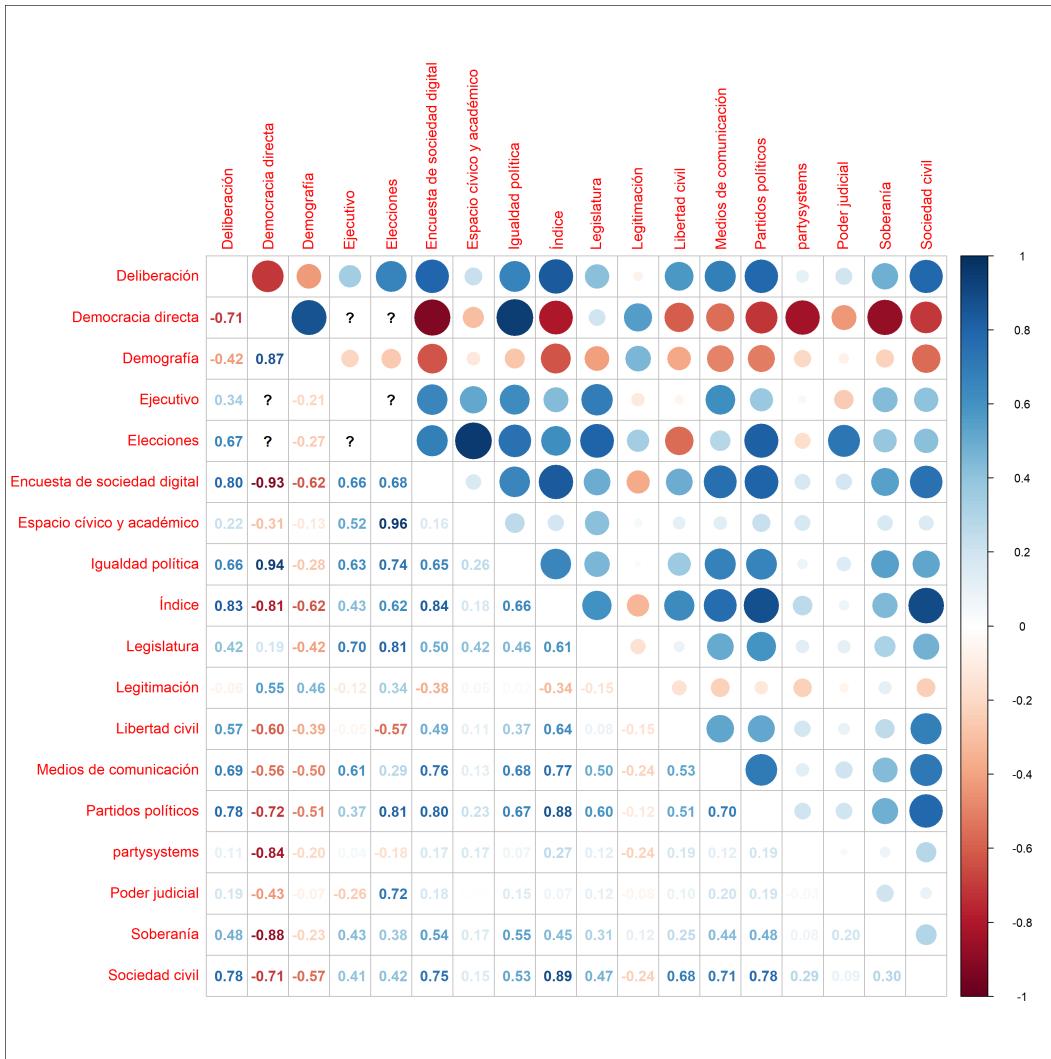


Figura 4: Correlación entre promedios de grupos de variables

144 El primer comentario a la figura surge de la fuerte correlación negativa que el grupo 'democracia  
145 directa' tiene con varios grupos: 'encuesta de sociedad digital', con el grupo de los 'índices', el  
146 de 'partidos políticos', el de 'partysystems', el de 'soberanía' y con el de 'sociedad civil'. Por otro  
147 lado, correlaciona fuertemente pero de manera positiva con los grupos de 'demografia' e 'igualdad  
148 política'. Otras fuertes correlaciones a mencionar son entre el grupo 'elecciones' y 'espacio cívico y  
149 académico', entre el de 'sociedad civil' y el de 'índice', entre 'partidos políticos' e 'índice'.

150 5. Anexo

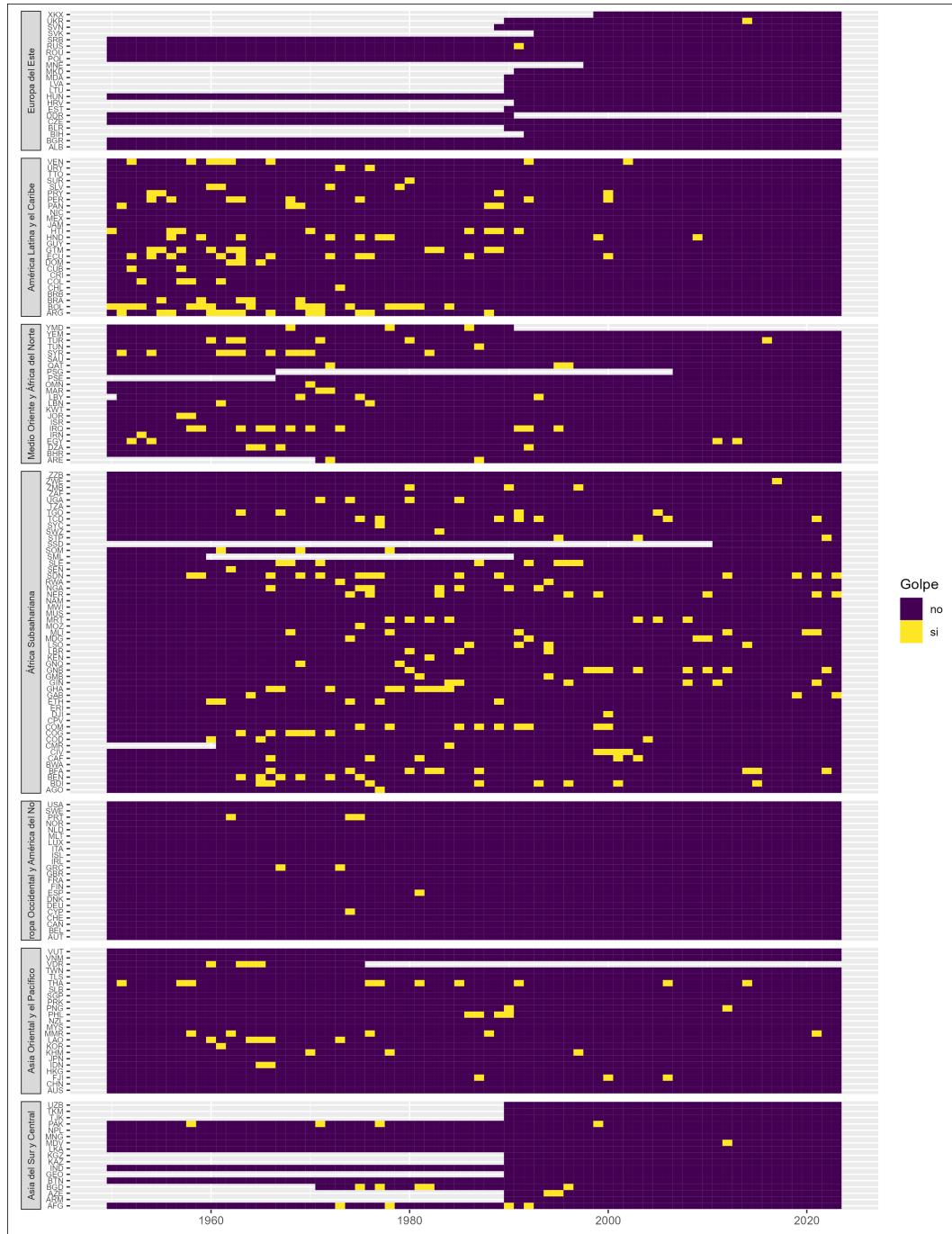


Figura 5: Conteo de golpes por año y región

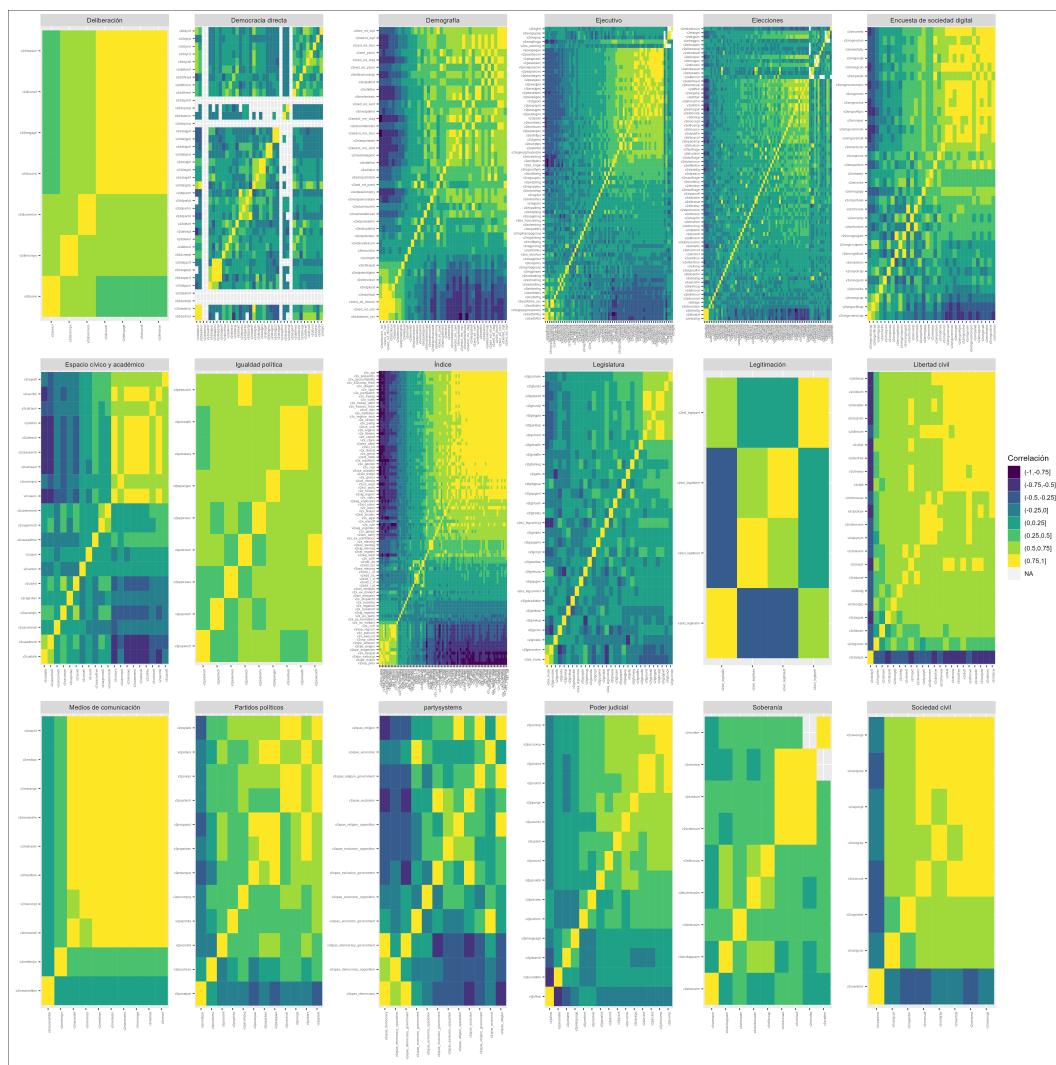


Figura 6: Correlación entre variables por grupo

## 151 Referencias

- 152 Burman, P., Chow, E., & Nolan, D. (1994). A Cross-Validatory Method for Dependent Data. *Biometrika*, 81(2), 351-358. Consultado el 1 de mayo de 2024, desde <http://www.jstor.org/stable/2336965>
- 153 Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, 99(1), 39-61. [https://doi.org/https://doi.org/10.1016/S0304-4076\(00\)00030-0](https://doi.org/https://doi.org/10.1016/S0304-4076(00)00030-0)
- 154 Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32. <https://doi.org/http://doi.org/10.1023/A:1010933404324>
- 155 Strumbelj, E., & Kononenko, I. (2010). An Efficient Explanation of Individual Classifications using Game Theory. *The Journal of Machine Learning Research*, 11, 1-18.
- 156 Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd international conference on knowledge discovery and data mining*, 785-794. <https://doi.org/https://doi.org/10.48550/arXiv.1603.02754>
- 157 Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett, Eds.). 30. [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf)

- 169 Cebotari, A., Chueca-Montuenga, E., Diallo, Y., Ma, Y., Turk, R., Xin, W., & Zavarce, H. (2024).  
170      *Political Fragility: Coups d'État and Their Drivers*. IMF Working Paper 24/34. <https://doi.org/https://doi.org/10.23696/mcwt-fr58>  
171  
172 Coppedge, M., Gerring, J., Knutsen, C. H., Lindberg, S. I., Teorell, J., Marquardt, K. L., Medzhorsky,  
173      J., Pemstein, D., Fox, L., Gastaldi, L., Pernes, J., Rydén, O., von Römer, J., Tzelgov, E.,  
174      Wang, Y.-t., & Wilson, S. (2024). "V-Dem Methodology v14" Varieties of Democracy (V-Dem)  
175      Project (Report). <https://v-dem.net/data/reference-documents/>