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

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

- Random Forest: profundidad máxima de los árboles (max_depth) y la submuestra del ratio de columnas a considerar cuando se construye cada árbol (max_features).
- XGBoost: la tasa de aprendizaje (learning_rate) y el término de regularización L2 en los pesos (reg_lambda).

Adicionalmente el parámetro que establece la cantidad de árboles creados (n_estimators) quedará fijado en 1000.

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

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

3. Análisis Exploratorio de Datos

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

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

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

- v2: variables de V-Dem.
- v3: variables pertenecientes a la base V-Dem histórica.
- v2x_: Índices principales e índices componentes.
- v2x[indicador de dos letras]: Índices específicos de ciertas áreas (ver más abajo).
- e_: variables no generadas por V-Dem y variables V-Dem en versión ordinal.

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

- ca: Espacio cívico y académico
- cl: Libertad civil
- dd: Democracia directa
- de: Demografía
- dl: Deliberación
- el: Elecciones
- ex: Ejecutivo
- exl: Legitimación
- ju: Poder judicial
- lg: Legislatura
- me: Medios de comunicación
- pe: Igualdad política
- ps: Partidos políticos
- sv/st: Soberanía/Estado

- x: Índice (calculado a partir de variables que también se incluyen en la base de datos)
- zz: Cuestionario posterior a la encuesta

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

3.1. Ánalisis de nulos

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

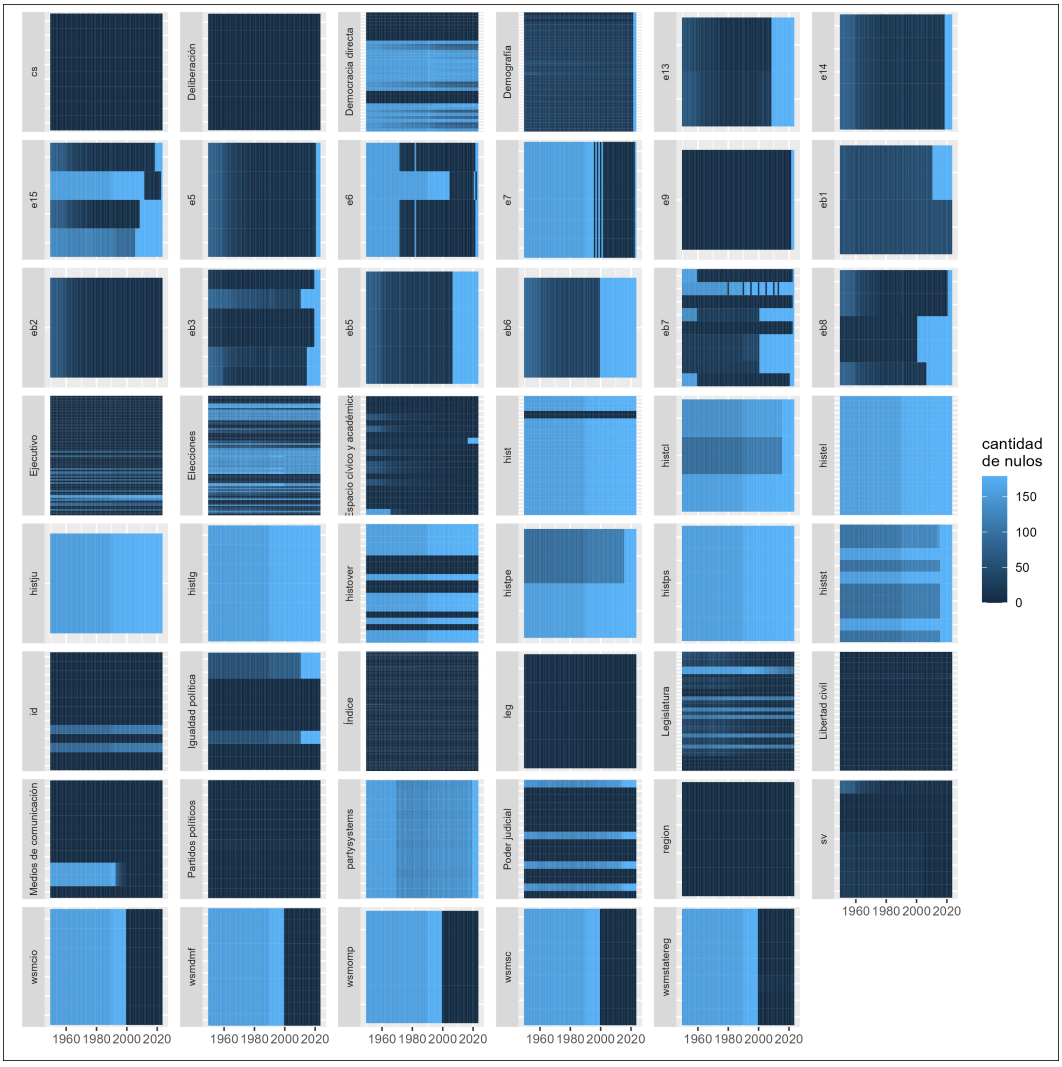

Figura 1: Conteo de nulos por año y agrupador de variables

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

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

Por último figuran los grupos de variables asociados a la base de datos histórica de v-dem (las que comienzan con hist), lo cual es lógico puesto que esta base busca tomar datos previos a 1900.

3.2. Análisis de variable objetivo

Es importante aclarar que en este trabajo no estamos contando la cantidad precisa de golpes de estado sucedidos en un período de tiempo, sino que simplemente relevamos si al menos un golpe de estado sucedió en un país y año determinado. Por lo tanto, si un país sufrió más de un golpe de estado en un año, el mismo será contabilizado una sola vez. Adicionalmente, en este trabajo también se consideran los golpes de estado que no fueron exitosos, es decir, que no lograron derrocar al gobierno en cuestión. De allí se desprende que países como Argentina, que en total ha tenido seis golpes de estado exitosos, figure con el doble de golpes en la figura 2.

Para realizar un paneo general de la variable objetivo, es decir, la presencia de estado a lo largo de los años, generamos un conteo y lo visualizamos en un planisferio. Destacamos que la mayor presencia de golpes se encuentra en el continente africano, en América del Sur y parte del Caribe, Medio Oriente y el Sudeste Asiático, con algunos casos de apenas un golpe en España, Rusia, Ucrania y Corea del Sur; así como dos y tres golpes en Grecia y Portugal, respectivamente.

Con mayor precisión, observamos que la región del Sahel se destaca con respecto a sus vecinos africanos. Los países en donde más golpes de estado se han producido son Bolivia (17), Sudán (14), Argentina (13), Ecuador (11), Iraq (11), Siria(11), Guatemala (10) y Tailandia (10).

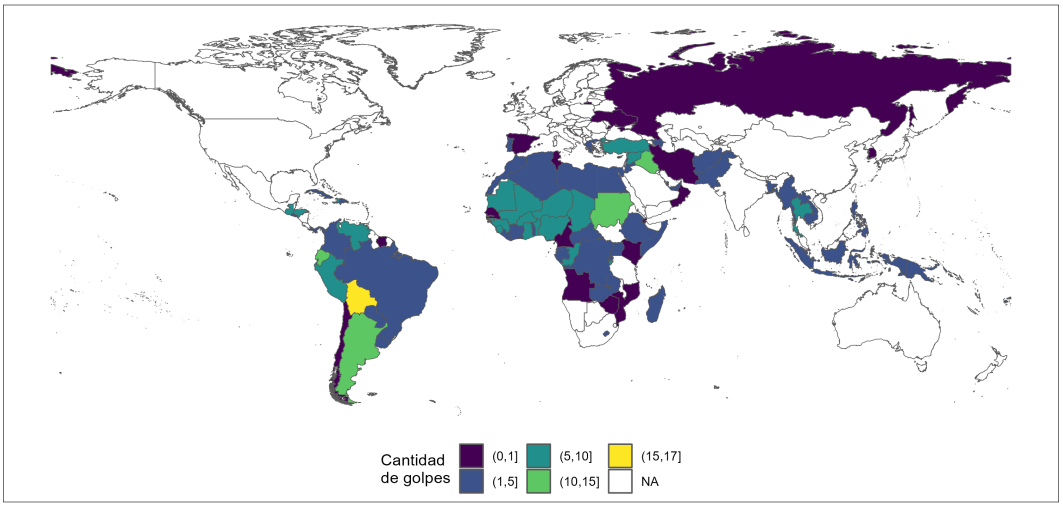

Figura 2: Conteo de golpes de estado en el mundo

Desagregando por década se observan algunos cambios, así como la persistencia en algunas regiones. La región del Sahel y varias naciones circundantes fueron persistentemente afectadas por golpes de estado desde los años 60. En América del Sur, en cambio, la presencia casi total de situaciones golpistas en la región se fue acotando a partir de los años 80 hasta finalmente desaparecer en el siglo xxi. Para observar con más detalle y discriminado por años y países se puede ver la figura 4.

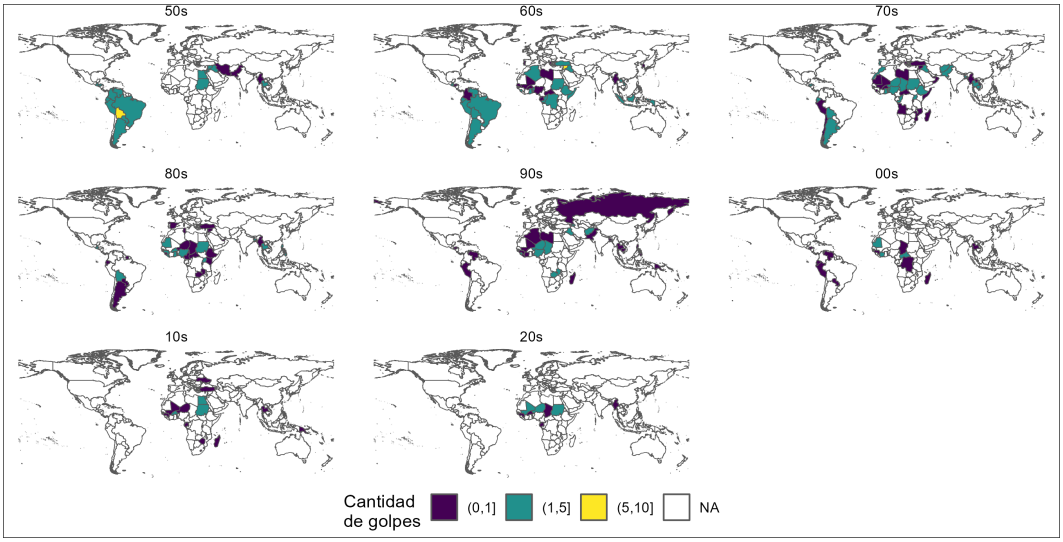

Figura 3: Conteo de golpes por década

4. Correlación entre variables

5. Anexo

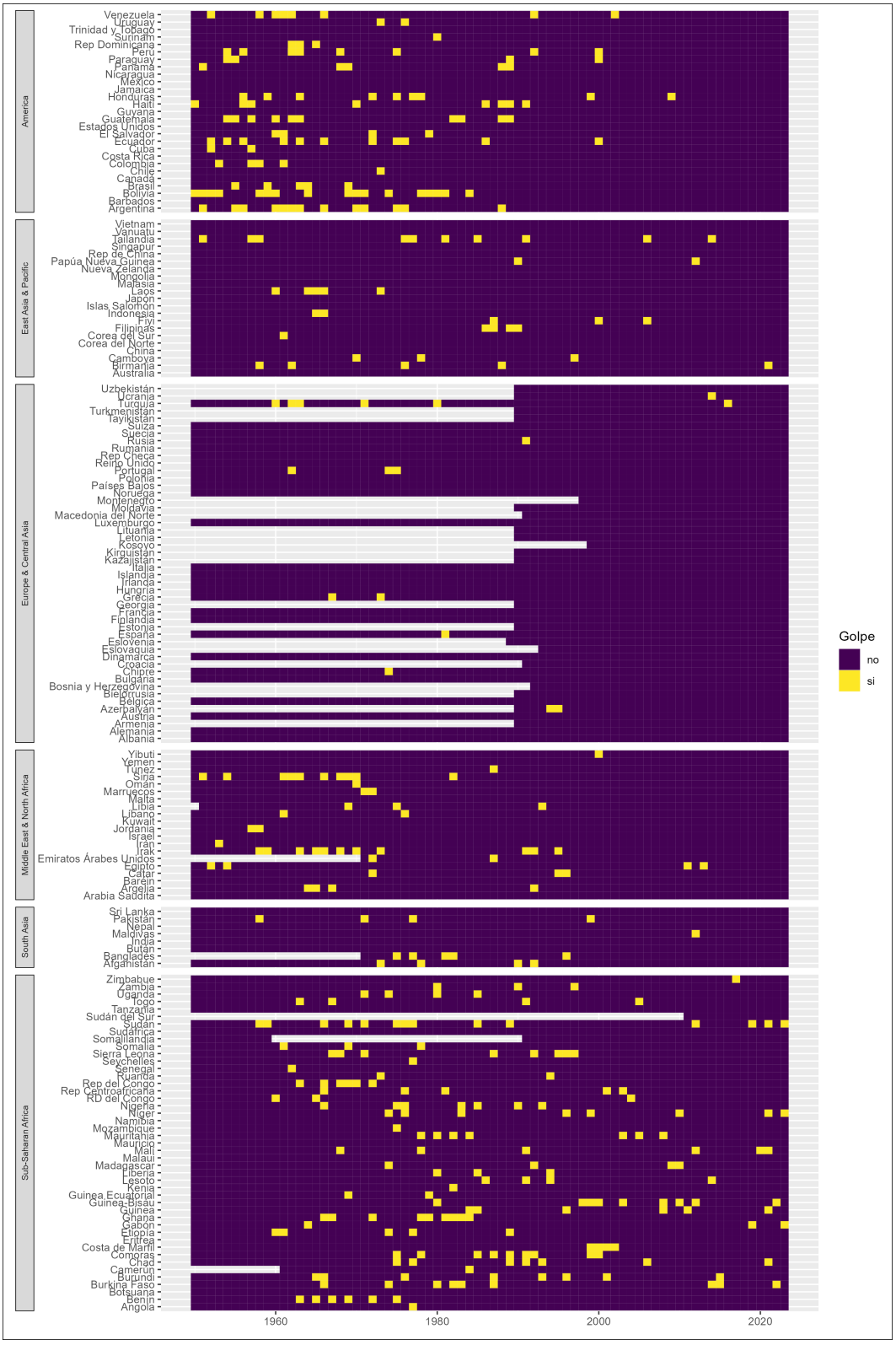

Figura 4: Conteo de golpes por año y región

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

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

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

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32. https://doi.org/http://doi.org/10.1023/A:1010933404324

Strumbelj, E., & Kononenko, I. (2010). An Efficient Explanation of Individual Classifications using Game Theory. *The Journal of Machine Learning Research*, *11*, 1-18.

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. https://doi.org/https://doi.org/10.48550/arXiv.1603.02754

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

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

Coppedge, M., Gerring, J., Knutsen, C. H., Lindberg, S. I., Teorell, J., Marquardt, K. L., Medzihorsky, J., Pemstein, D., Fox, L., Gastaldi, L., Pernes, J., Rydén, O., von Römer, J., Tzelgov, E., Wang, Y.-t., & Wilson, S. (2024). *"V-Dem Methodology v14"Varieties of Democracy (V-Dem) Project* (Report). https://v-dem.net/data/reference-documents/