UNIR

Impacto de la pandemia por COVID-19 a nivel mundial en el año 2020

AUTORES:

Maria Belén Benalcázar Tovar

Daniel Aguilar Noblecilla

Patricia Katherine Tigrero Quimi

Edith Pérez Tatamués

Cristian Jacinto Guanín Pilco

23 DE FEBRERO DE 2022

TABLA DE CONTENIDO

Impacto de la pandemia por COVID-19 a nivel mundial en el año 2020

1. Descripción del Problema

La enfermedad por coronavirus – COVID-19 – fue notificada por primera vez en Wuhan – China el 31 de diciembre de 2019 (OMS, 2022) y fue caracterizada como pandemia el 11 de marzo de 2020 debido a los alarmantes niveles de propagación y gravedad (OPS, 2020). A partir de esa fecha se ha continuado expandiendo por el mundo y ha causado gran impacto a nivel humano, social y económico.

La alta afectación en diferentes ámbitos, ha creado la necesidad de analizar esta pandemia en miras a identificar el comportamiento del virus y poder extraer conclusiones claras y objetivas que permitan conocer, evaluar y mitigar sus efectos. Hoy en día existen múltiples técnicas destinadas a estudiar la enfermedad del coronavirus, entre ellas, las que se enmarcan dentro del campo de la inteligencia artificial y por ende, dentro de Machine Learning han demostrado ser muy útiles para combatirla de manera eficaz según Pham et al. (2020).

El presente estudio pretende analizar el **Impacto de la pandemia por COVID-19 a nivel mundial en el año 2020**, mediante la identificación de similitud en la evolución del virus COVID-19 entre los países analizados con el fin de determinar las regiones y países mayormente golpeados durante el primer año de pandemia.

2. Metodología y Técnicas a utilizar

2.1. Descripción de los datos

El dataset base provisto para este estudio tomado de Data Europa, contiene datos del *número de nuevos casos (cases)* y *número de nuevas muertes (deaths)* por COVID-19 de 212 países en el período comprendido entre diciembre 2019 y diciembre 2020. Adicionalmente posee el valor de la población de cada país al 2019 *(popData2019)* y el dato *acumulado de casos para 14 días por cada 100.000 habitantes*.

Para complementar los datos provistos y determinar de mejor manera el impacto de la pandemia se usan datos de OurWorldInData (2022), cuyo dataset posee parámetros adicionales de los que se toma algunos factores que incluyen: *índice de restricciones, edad media, tasa de muerte por enfermedades cardíacas, prevalencia de la diabetes, expectativa de vida e índice de desarrollo humano*.

Se usan dos datasets suplementarios, uno para análisis de la tasa de incidencias y de muertes (CSSE, 2022) , pero sus valores no resultan ser representativos para el análisis y un segundo dataset geográfico para la representación gráfica (Kaggle, 2022).

Una vez consolidada la información precedente, el dataset final (Figura 1) posee 61.777 instancias con 16 atributos, siendo los dos últimos empleados en el análisis preliminar. En el caso del atributo cases_14d_100K se efectuará el proceso de limpieza de nulos únicamente cuando sea necesario, es decir, solo en los casos en donde se use dicho atributo, con el objetivo de no perder información para el entrenamiento.

Figura 1. Caracterización del dataset final

```
Int64Index: 61777 entries, 0 to 61776
Data columns (total 16 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   dateRep                 61777 non-null  datetime64[ns]
 1   cases                   61777 non-null  int64
 2   deaths                  61777 non-null  int64
 3   countryterritoryCode    61777 non-null  category
 4   popData2019             61777 non-null  float64
 5   continentExp            61777 non-null  category
 6   cases_14d_100K          59021 non-null  float64
 7   reproduction_rate       61777 non-null  float64
 8   stringency_index        61777 non-null  float64
 9   median_age              61777 non-null  float64
 10  cardiovasc_death_rate   61777 non-null  float64
 11  diabetes_prevalence     61777 non-null  float64
 12  life_expectancy         61777 non-null  float64
 13  human_development_index 61777 non-null  float64
 14  Incident_Rate           2160 non-null   float64
 15  Case_Fatality_Ratio     2160 non-null   float64
```

2.2. Limpieza y Análisis Exploratorio de Datos

En los campos *cases* y *deaths* se presentan números negativos, esto se debe a que en ocasiones los países reportan valores erróneos, cuya rectificación se realiza en días posteriores, aquello puede resultar en inconsistencias con respecto a los casos oficiales (ScienceDirect, 2021). Se corrige este inconveniente encerando el valor negativo y

restándolo sucesivamente de los días anteriores hasta obtener un valor mayor o igual a cero.

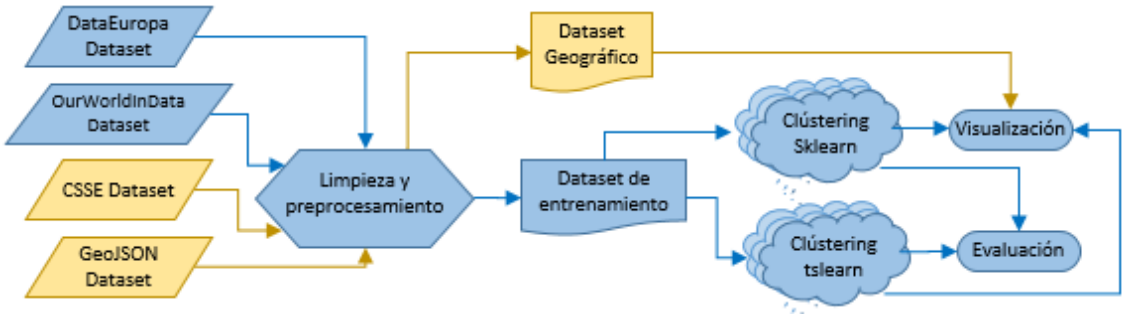
Para los parámetros que representan tasas globales, se efectúa el proceso de interpolación, para reducir los valores nulos y generar un análisis apropiado.

Al analizar la relación existente entre las variables, se determina una alta correlación entre las muertes (deaths) y el número de casos (cases), una conclusión evidente, y que no puede ser eliminada al ser las dos variables principales, que determinan la evolución de la pandemia, pero que podría ser de impacto al aplicar algunas técnicas de Machine Learning.

Se pretende aplicar diferentes técnicas de clustering para obtener una segmentación de los países y determinar la similitud existente entre ellos. La multicolinearidad no es un problema al modelar problemas de clustering (Revista de Ciencias Clínicas, 2015) como lo sería en técnicas de aprendizaje supervisado tales como clasificación, por lo que será la técnica principal usada para el presente estudio. No obstante, se utilizarán técnicas adicionales con el objetivo de validar la segmentación.

La Figura 2 muestra el esquema general del análisis. Se emplean dos grupos de técnicas de clustering. En primer lugar, usando la librería ***sklearn*** y medios tradicionales de clustering, y se contrasta con lo arrojado por los algoritmos de ***tslearn***, paquete especializado en series temporales.

Figura 2. Esquema general del análisis del estudio



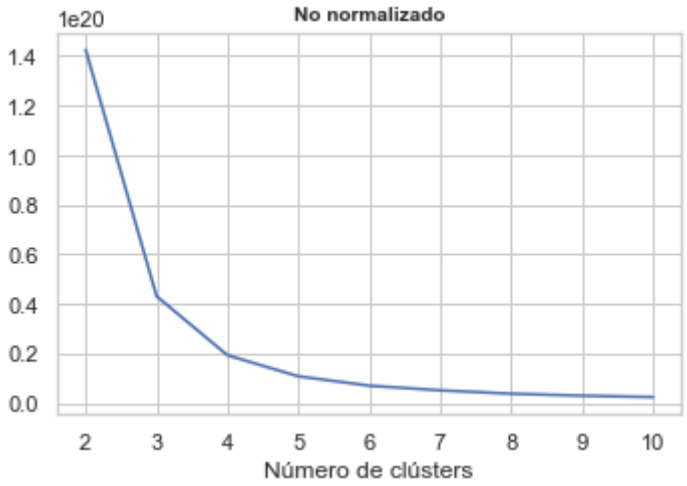
3. Detalle del Análisis

3.1. Análisis y segmentación a través de clustering

Determinación del número óptimo de clústeres

Una vez realizada la limpieza de los datos y empleando el método del codo (Sanz, 2018) se evalúa el número de clústeres óptimos para el presente estudio, obteniendo un total de cinco clústeres, valor obtenido usando sklearn. Se usa este número como constante para todos los algoritmos, con el objetivo de establecer un criterio comparativo entre ellos.

Figura 3. Resultado del método del codo para determinación del número de clústeres



Segmentación sklearn con parámetros principales

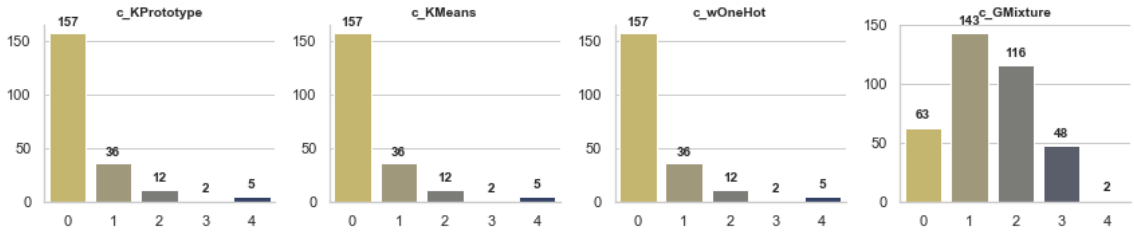
Dentro de los algoritmos basados en distancias, se entrena ***KPrototypes***, que usa un esquema similar al algoritmo *KMeans* de *sklearn*, pero permite el uso de datos combinados: numéricos y categóricos (PyPi, 2022). Para este algoritmo se selecciona el método de inicialización ***Cao*** *que «selecciona prototipos para cada objeto de datos en función de la densidad del punto de datos y el valor de disimilitud»* (Zazueta, 2020). Este algoritmo presenta resultados robustos – a nivel de cohesión – pero tiene un alto costo computacional.

Se prueba el algoritmo ***KMeans*** de ***sklearn*** con la ***distancia euclidiana***, usando diferentes modelos del dataset: valores absolutos, valores acumulados, valores normalizados. Se obtiene la misma distribución para valores absolutos normalizados y

sin normalizar. Para la evaluación, se usa la métrica ***silhouette,*** que permiten verificar la cohesión de los grupos, en todos los casos tienen valores aceptables superiores a 0.5 en cada clúster. Cuando se usan tasas de incidencia en lugar de valores absolutos, la cohesión de los segmentos es menor.

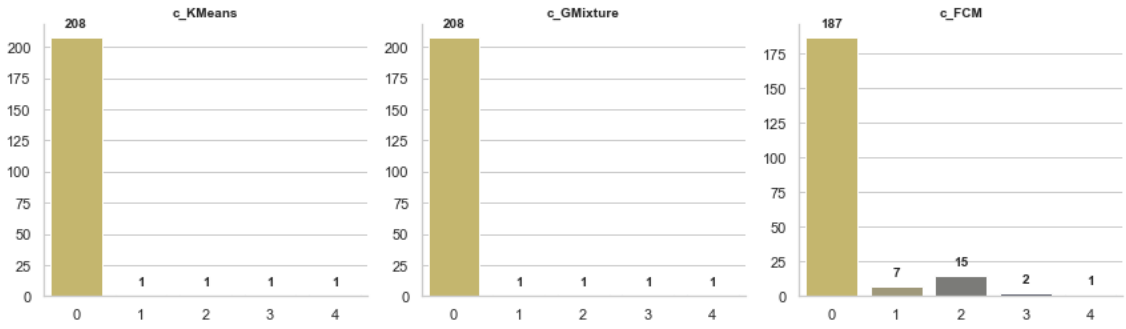
Se realizaron pruebas también con el algoritmo **GMixture,** el cual permite obtener clústeres de forma variada, pues toma en cuenta la varianza, sin embargo, no se obtuvieron resultados favorecedores. Usando el tipo de covarianza esférico (considera que cada componente tiene su varianza única) se obtiene una distribución distinta a los algoritmos basados en distancia, pero hay que tener en cuenta que el nivel de cohesión es más bajo.

Figura 4. Distribución de los clústeres usando dataset con valores absolutos



Considerando que la temporalidad implícita en el dataset puede incidir en los resultados, se realizan pruebas con las medias de los países, generando un dataset agrupado. La mejor segmentación en este caso se logra con el algoritmo **Fuzzy C Means**, que utiliza un método de agrupamiento suave, es decir, en lugar de asignar una pertenencia rígida a un grupo, le asigna una probabilidad (Towards Datascience, 2021). Sin embargo, aunque los valores de cohesión son altos, los resultados muestran que al realizar este cambio se omite gran cantidad de información y causa que la mayor parte de los países se asignen a un solo grupo.

Figura 5. Distribución de clústeres usando dataset de promedios por país.



Segmentación sklearn usando parámetros adicionales

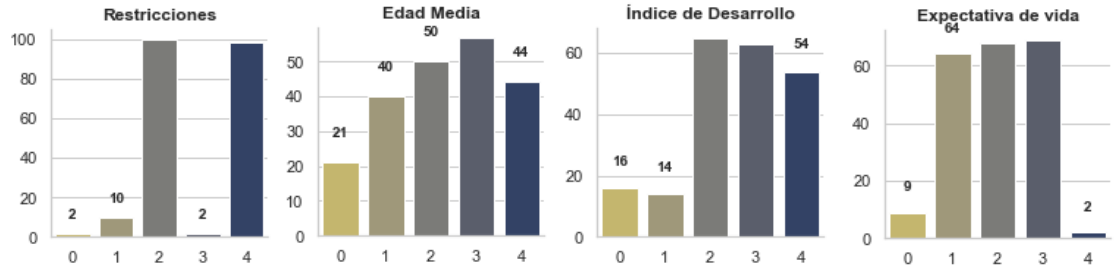
De acuerdo al coeficiente de variación (Tabla 1) no se considerarían representativos los atributos *stringency_index*, *median_age*, *human_development_index* y *life_expectancy*, por sus valores bajos, pues estas tasas presentan valores constantes para la mayoría de países. Sin embargo, se realiza el entrenamiento de los algoritmos de clustering en miras a conocer la influencia en estas tasas.

La distribución en los clústeres es diferente (Figura 6) a la obtenida en el apartado anterior, pero los coeficientes de cohesión son negativos. Por lo que, para los siguientes análisis, solamente se toman en cuenta los atributos más representativos.

Tabla 1. Coeficiente de variación de los atributos de entrada

Coeficiente	%
stringency_index	*0,59%*
median_age	*0,49%*
human_development_index	*0,45%*
life_expectancy	*0,23%*

Figura 6. Resultados obtenidos entrenados con diferentes parámetros



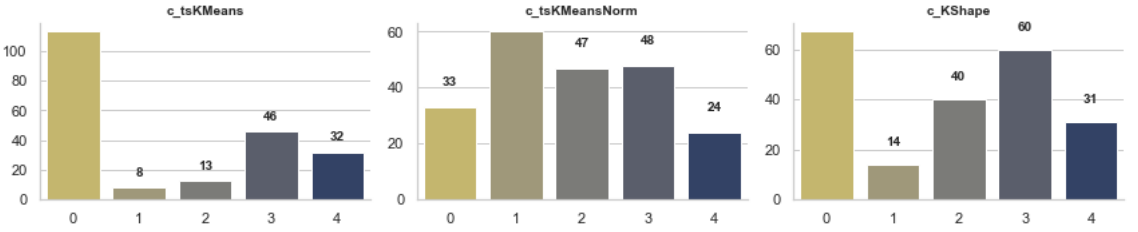
Segmentación usando tslearn

Teniendo en cuenta que en el análisis con sklearn, se ha descartado la temporalidad, un factor muy importante en la pandemia, se usan algoritmos de clustering del paquete **tslearn** – especializado en series temporales – para evaluar el contraste en la segmentación. Se prueban tres algoritmos: **TimeSeriesKMeans**, **KernelKMeans** (ambos basados en distancias) y **KShape** que se basa en la correlación cruzada de las series.

Para los algoritmos basados en distancia, se selecciona como medida de distancia ***DTW (Dynamic Time Warping Distance),*** la cual se ajusta de mejor manera a las series temporales. Con estos algoritmos se obtiene una distribución más balanceada (Figura

7), aunque la cohesión de los grupos es más baja que la obtenida con los algoritmos con sklearn. Los mejores resultados de ***silhouette*** se obtienen con el algoritmo **KShape**.

Figura 7. Distribución de clústeres con los algoritmos de series temporales

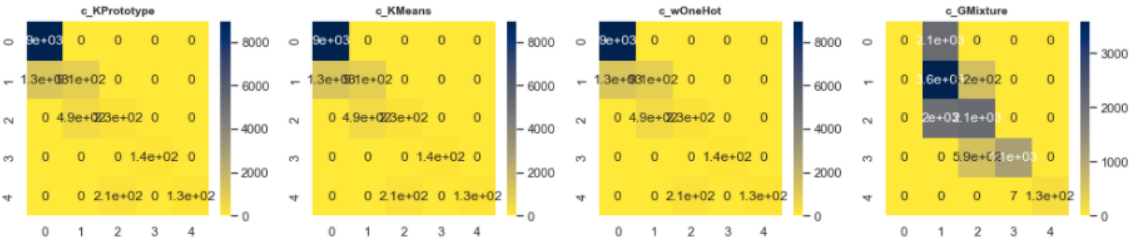


3.2. Comprobación mediante clasificación

Se usan algoritmos de clasificación para comprobar la distribución obtenida en el apartado anterior y poder seleccionar uno de los algoritmos para análisis.

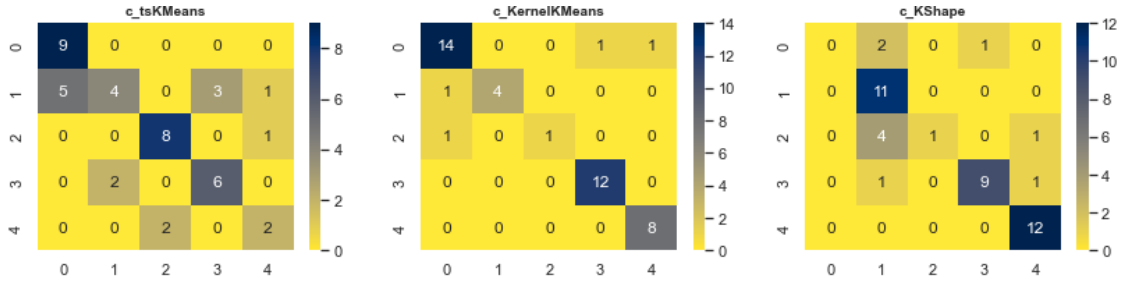
En el caso de los algoritmos sklearn se usa **LogisticRegression** para la validación. La exactitud global es aceptable, pero al no existir balanceo entre las clases, existen grupos específicos cuyas métricas son bajas, por lo que no se podría considerar una clasificación confiable.

Figura 8. Matrices de confusión obtenidas por los algoritmos sklearn



En el caso de los algoritmos tslearn, se utiliza el algoritmo **Early Classification.** Las métricas obtenidas son superiores a 0.67. El algoritmo más robusto a nivel de clasificación es **Kernel KMeans** con una exactitud de 90%, seguido del algoritmo **KShape** con un valor de 77%.

Figura 9. Matrices de confusión obtenidas por el algoritmo tslearn



4. Resultados y conclusiones

Este estudio buscaba examinar el impacto de la pandemia a nivel mundial en el año 2020 y conocer la similitud entre la evolución en los diferentes países. Una vez analizado el número de clústeres óptimos, se entrenaron diferentes modelos y se evaluaron en base al parámetro de cohesión ***silhouette***. El análisis con los atributos más significativos para la evolución de la pandemia como lo son: casos, muertes y la población de cada país; es más óptimo y permite una mejor segmentación que usando parámetros adicionales.

Aunque el experimento ha permitido determinar que no existe un comportamiento uniforme en la evolución de la pandemia, ha sido posible segmentar por países en cinco grupos, clasificando su impacto en grados del 1 al 5. Tras el análisis de métricas y la evaluación de clasificación se acepta como resultado más óptimo, el obtenido por el algoritmo KShape de tslearn, con un valor de silhouette de 0.79 y una exactitud en la clasificación de 0.77.

En la Figura 10 se muestra la distribución de cada uno de los clústeres.

Figura 10. Clústeres creados por el algoritmo KMeans con valores absolutos de casos y muertes

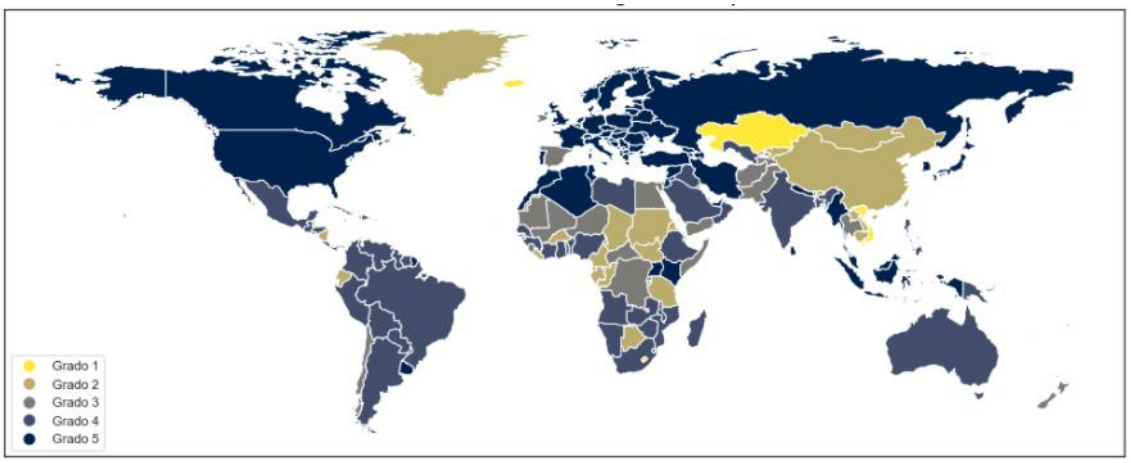


Figura 11. Resultados de la segmentación con el algoritmo KShape de tslearn

c_KShape	cases_mean	deaths_mean	count
4	6.914794e+07	1.796378e+06	67
3	6.320386e+07	1.609302e+06	55
2	1.555991e+07	4.909972e+05	29
1	1.461242e+06	7.521742e+04	53
0	3.681225e+06	4.418562e+04	8

El esquema de la Figura 11 muestra los resultados de la segmentación de este algoritmo, obteniendo cinco grados nombrados desde el más leve al más grave. Los clústeres 3 y 4, correspondientes a los grados 4 y 5 respectivamente, registran medias similares superiores al 1.5 millones de muertes y 60 millones de casos. En la distribución, estos suman 122 países correspondientes al 57.5 %, que serían los países de más *alto impacto* de la pandemia. El nivel *intermedio* está formado por 29 países en los cuales la media de casos supera los 15 millones y las muertes tienen un promedio menor a 500K. En los dos grupos restantes, que los denominaremos de *impacto moderado*, se asignan 61 países, con una media de muertes superior a 44K y una media de casos entre 1.4 y 3.7 millones.

5. Bibliografía

- CSSE. (2022). *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins*. Obtenido de https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports
- Kaggle. (2022). *Human Development Reports*. Obtenido de https://www.kaggle.com/sudhirnl7/human-development-index-hdi
- OMS. (2022). Brote de enfermedad por coronavirus (COVID-19). págs. https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019. Obtenido de https://www.who.int/es/news/item/28-11-2021-update-on-omicron
- OPS. (13 de Marzo de 2020). La OMS caracteriza a COVID-19 como una pandemia. Obtenido de https://www.paho.org/es/noticias/11-3-2020-oms-caracteriza-covid-19-como-pandemia
- OurWorldInData. (2022). Obtenido de https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv
- Revista de Ciencias Clínicas. (2015). Conglomerados como solución alternativa al problema de la multicolinealidad en modelos lineales. *ELSEVIER*. Obtenido de

https://www.elsevier.es/es-revista-ciencias-clinicas-399-articulo-conglomerados-como-solucion-alternativa-al-S1665138315000166

- ScienceDirect. (2021). Dataset of COVID-19 outbreak and potential predictive features in the USA. Obtenido de https://www.sciencedirect.com/science/article/pii/S2352340921006429
- Q. Pham, D. C. Nguyen, T. Huynh-The, W. Hwang and P. N. Pathirana, "Artificial Intelligence (AI) and Big Data for Coronavirus (COVID-19) Pandemic: A Survey on the State-of-the-Arts," in IEEE Access, vol. 8, pp. 130820-130839, 2020, doi: 10.1109/ACCESS.2020.3009328
- Sanz, A. (2018).Evaluación de una nueva metodología para la estimación de transvase de votos entre elecciones.