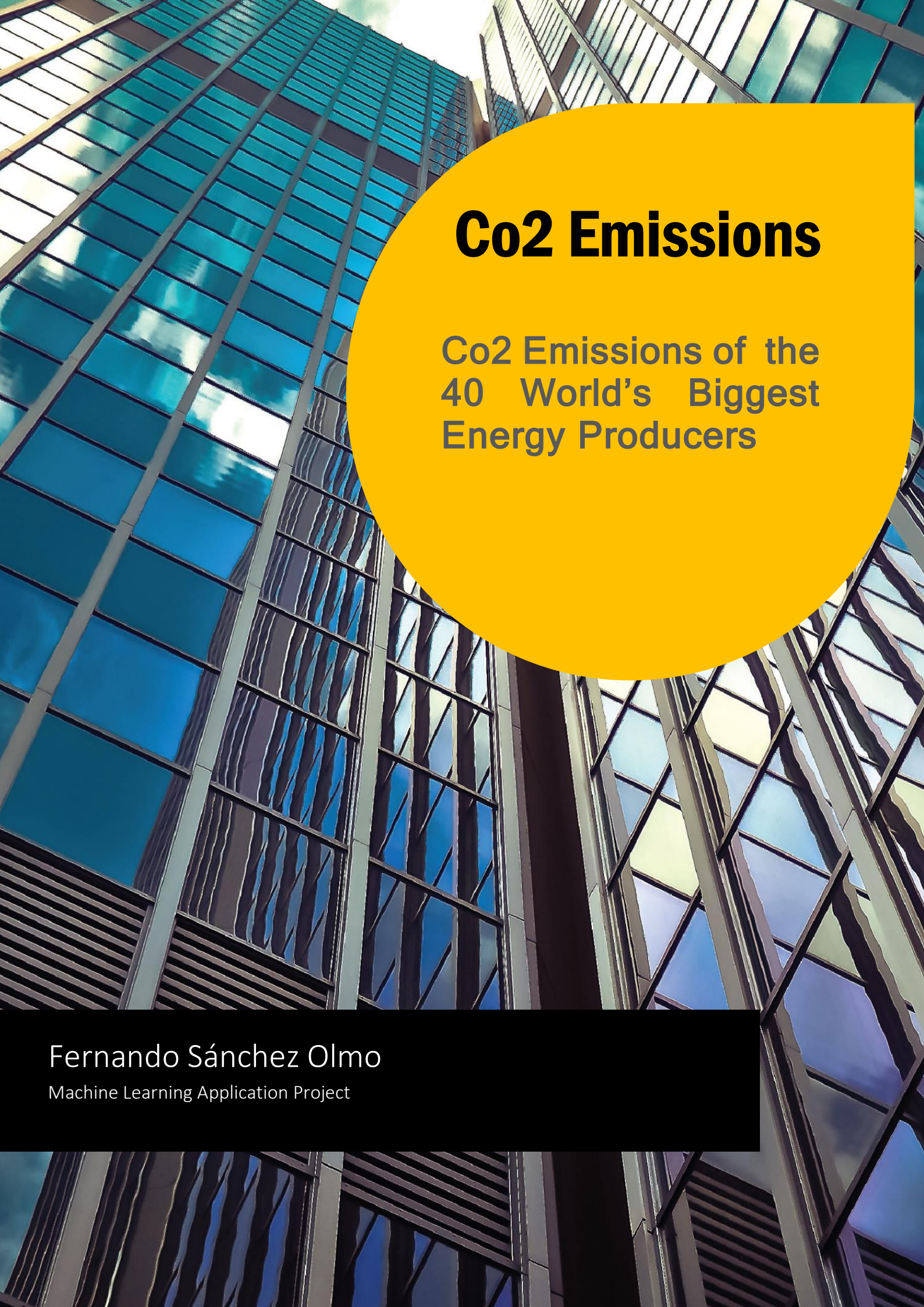

# Co2 Emissions

## Co2 Emissions of the 40 World's Biggest Energy Producers

Fernando Sánchez Olmo

Machine Learning Application Project

# INDICE

## Emisiones de CO2

**1. REVISIÓN DE DATOS**

Antes de empezar a preparar nuestros modelos de machine learning, adaptamos un poco el dataset a las necesidades que vamos a tener. Para ello eliminamos la columna geometry, que en esta ocasión no nos va a hacer falta. También rellenamos los códigos internacionales de algunos países que faltan para su representación en un gráfico y nos damos cuenta de que Yugoslavia es un país que se extinguió en el año 2013 en multitud de países. En este país, al no ser posible identificar que parte de los valores de cada variable corresponde con cada uno de los nuevos países ya creados, o creados y extinguidos en el proceso, debido a la constante aparición y desaparición de alguno de ellos, pues han tenido cambios de nombre etc., hemos decidido eliminarlo del dataset. Sin embargo, en el caso de Checoslovaquia, al ser sólo dos países y no mediar entre ellos otras formaciones, hemos podido diferenciar el resto de los datos mediante una regla proporcional en base a los metros cuadrados de territorio con el que a quedado cada uno con respecto a la antigua formación de países.

**2. CLUSTERIZACIÓN**

Para poder llevar a cabo el proyecto de machine learning lo primero que necesitábamos saber es si podíamos clasificar los países en segmentos diferentes. De esta manera, si podemos diferenciar a países por tener características diferentes, podríamos realizar regresiones para la eficiencia más fiables, y a posteriori, poder usar un modelo de clasificación para datos nuevos.

Para llevar a cabo esta fase hemos usado:

*KMEANS*

En primer lugar, empezamos realizando la clusterización con todas las variables que disponemos para generar un baseline con el que seguir avanzando y poder comparar ulteriores resultados. Dentro de los modelos vamos a usar varios tipos de escalados, y en concreto con la variable tipo de energía, vamos a probar con label encoder y con map de manera manual. Este punto está basado, en que dados los diferentes niveles de contaminación que cada una de las fuentes de energía producen, podemos establecer una relación de orden entre ellas, de manera que aquellas que producen más co2 se catalogan con un valor superior a aquellas que producen menos.

**PRIMERA CLUSTERIZACIÓN:**

*standard scaler, label encoder, todas las variables*:

Encontramos que tenemos una división en 7 clusters, de los cuales 2 contienen la mayoría de los datos y 4 podrían considerarse residuales, los otros 2 restantes contienen datos en similar

magnitud y se podrían considerar como clusters pequeños. De esta manera, consideramos que es viable el poder obtener una clusterización de 4 países de manera aceptable y que encaja con la idea general del inicio del proyecto, que, aunque se basaba en 3 cluster, es posible que se nos esté diciendo que existe un segmento de países adicional que no estábamos considerando.

Para poder corroborar la idea anterior llevamos a cabo la métrica denominada elbow, la cual en base a la inercia de los puntos respecto a los centroides, nos indicará cual puede ser el óptimo en nuestro dataset. Vemos que la variación de la inercia de un punto a otro comienza a descender de manera menos pronunciada a partir del cluster 4, esto nos indica que el óptimo se encuentra en 4 cluster y por tanto podemos afirmar la consideración inicial como acertada, dado que la suma de los cuadrados de las diferencias entre los puntos de cada cluster con su centroide (inercia) empieza a disminuir en menor proporción a partir del cluster 4. Lo que nos quiere decir esta menor disminución, es que si lo vemos desde el punto de vista de una tasa porcentual, es que la disminución porcentual que se produce a partir del cuarto cluster empieza a ser menor con cada cluster que se añade, indicándonos que las diferencias que separan un cluster de otro empieza a difuminarse, y por tanto a perder sentido la clusterización. Para poner en contexto esta idea, podemos llevarlo al extremo de que siguiéramos haciendo clusteres hasta que la inercia fuera 0. En este punto, lo que tendríamos sería un cluster por cada una de las entradas de nuestro dataset y, por tanto, cada dato de manera individual sería un cluster. En este punto el objetivo de la clusterización se ha difuminado por completo y no tiene sentido, pues no hemos agrupado ningún dato en función de características similares, aunque nuestra inercia sea 0. Hemos perdido la información agregada de agrupar nuestros datos en segmentos diferenciados por características comunes. Además, en una segunda prueba elbow, en la que añadimos el tiempo de ejecución, vemos que con 4 clusters obtenemos el segundo mejor tiempo en coste computacional, lo que es un valor añadido a tener en cuenta, pues ahorra dinero.

Esta prueba inicial nos lleva a replantearnos la idea inicial que teníamos de división de los países en poco contaminantes, contaminantes y muy contaminantes a:

* poco contaminantes
* contaminación moderada
* contaminantes
* muy contaminantes

Realizamos la clusterización comentada y vemos que de nuevo obtenemos 2 clusters con la mayoría de los datos, uno pequeño y uno residual. Este residual puede ser debido a que, dado el conocimiento que tenemos de los datos por la realización del EDA, nuestro dataset contiene 3 países en concreto que tiene datos muy alejados del resto, como son China, Estados Unidos y Rusia, pues son los mayores productores mundiales de energía. Ellos solos proveen en conjunto la mayoría de la energía en el mundo. Por lo que seguiremos realizando pruebas a ver si este punto se confirma.

**SEGUNDA CLUSTERIZACIÓN:**

*standard scaler, 2 variables: emisiones de co2 y producción de energía:*

Empezamos realizando la clusterización sin indicar número de clusters y con estas dos variables nos vuelve a realizar 7 cluster por defecto. Estas dos variables han sido elegidas en base a que son las más representativas del objeto de este proyecto y, por tanto, aquello que queremos representar con nuestros clusters. Sin embargo, ahora los datos están más repartidos entre los cluster, quedando como residuales solo dos que, aun siendo residuales, tienen más datos que los anteriores. Como hicimos en la anterior prueba, realizamos nuestra métrica del codo para revisar si estos clusters son óptimos y vemos que ahora el óptimo se sitúa en 3 clusters. Los realizamos y nos queda una división en la que prácticamente todos los datos se sitúan en el cluster 0, por lo que consideramos que no es una solución válida para nuestro modelo. Lo que ha pasado es que el cluster 0 y el 1, obtenido anteriormente con 4 clusters se han fusionado y, por tanto, como estos clusters correspondían a los países poco contaminantes y con contaminación moderada, ahora han quedado todos como poco contaminantes.

**TERCERA CLUSTERIZACIÓN**

*Power Transformer, 3 variables: eficiencia, producción de energía y tipo de energía:*

Realizamos un nuevo modelo con 3 variables en las cuales incluimos la eficiencia de cada uno de los países, por incluir información sobre la emisión de co2, la producción de energía y el tipo de energía con el que se ha generado la misma. Elegimos estas variables, dado que son representativas de lo que se quiere representar con los clusters. En esta prueba, vemos que el óptimo mediante la métrica elbow vuelve a ser 4 clusters con una separación entre ellos parecida a la inicial con todas las variables, siendo la diferencia principal que con tres variables los cambios en la inercia son más acusados y por tanto es más fácil distinguir el óptimo. Esto nos da a entender que sí hay una diferencia importante entre los cluster y por tanto que el óptimo se pueda establecer en 4.

Por otra parte, representando los valores en un gráfico encontramos que la distribución de estos no tiene una forma circular. Por ello, como standard scaler nos hace un escalado en el que se centran los datos entorno a una media 0 con desviación típica unitaria, pero le afectan los datos extremos, los datos escalados conservan la distribución original, que no es parecida a la normal. Así, lo que vamos a hacer es usar powertransformer con el transformador yeo-johnson, pues funciona con datos positivos y negativos, que nos escala los datos con media cero y desviación típica unitaria, y los transforma haciendo que la distribución de estos se asemeje a una distribución Gaussiana, reduciendo la asimetría, que en nuestros datos es muy pronunciada al parecerse sus distribuciones, en su mayoría, a una distribución gamma. La consecuencia de todo esto es que la varianza de nuestras variables es muy elevada y powertransformer nos ayudará a estabilizarla.

Antes de realizar estos pasos para nuestro Kmeans, echamos un vistazo a la métrica silueta que es una métrica usada para medir la bondad del ajuste en clusterizaciones. Su valor oscila entre -1 y 1, siendo una métrica aceptable a partir de un score de 0.5. Obtener un 1 significaría que los clusters se diferencian bien unos de otros, 0 que la distancia entre los clusters no es significativa y por tanto es indiferente separarlos y -1 es que los clusters han sido asignados de manera incorrecta.

Sabiendo esto, hemos comprobado que el número óptimo de clusters está en 4 porque obtiene un score medio de silueta de 0.7, siendo de 0.3 en las demás clusterizaciones. Además, para 3 clusters no todos ellos están por encima de este score medio, mientras que en 4 sí lo están. Esto reafirma la elección de 4 clusters, pero, aunque 4 clusters con standard scaler obtiene el mejor compromiso entre score general e individual de los clusters, queremos seguir probando con powertransformer y mapping para el tipo de energía.

**CUARTA CLUSTERIZACIÓN**

*Power Transformer, todas las variables, label encoder:*

Obtenemos 7 clusters como al inicio del estudio. Además, mirando cual ha sido el cluster medio que se ha otorgado a cada país a lo largo de todos los años del dataset vemos que nos concuerda con las conclusiones extraídas del EDA previo. Así, mirando una muestra de estos países podemos ver lo siguiente:

Russia, Estados Unidos y China:

Han sido clasificados en media con un cuatro y se encuentran en la punta derecha de la gráfica del estudio donde están los países con mayores producciones y emisiones de co2, aunque no los que más.

Angola, África del sur:

son países con poca producción y pocas emisiones de co2, se han representado con un 1, y tiene sentido en relación con los datos del EDA.

Alemania, Reino Unido, Brasil:

Calificados con un 2 o 3 son países con producción media y emisiones de co2 media, tiene sentido.

Sin embargo, la representación en el gráfico no es homogénea y es difícil diferenciar los clusters. Esto nos está indicando que la definición de estos no es buena y por tanto resultará en un coeficiente de silueta pobre.

**QUINTA CLUSTERIZACIÓN:**

*Power Transformer,mapping,todas las variables*

Usando igualmente todas las variables y asignando los pesos del encoding en la variable tipo de energía en función de la emisión de co2 media que tiene cada una de las fuentes de obtención de la energía vemos que:

China, Rusia y Estados Unidos:

se clasifica a China con casi un 4 de media que corresponde al carbón y es muy contaminante, a Rusia con casi un 2.78, que está a medio camino entre el gas y el petróleo, pero no llega al 3 porque produce más gas, y a Estados Unidos con un 3.6 que está entre el petróleo y el carbón. Esta clasificación se ajusta más a la realidad de los datos, si vemos que además sus medias no están muy alejadas las unas de las otras, como ocurre en su producción y emisión de co2 en los datos que tenemos, situándolos como países con altas producciones y emisiones de co2, pero sin llegar a ser los más contaminantes en términos de eficiencia.

Nigeria, Colombia, Argentina:

se sitúan en torno al cluster 6 y 7 y en el anterior gráfico estaban en el cluster 2. Los 6 y 7 tienen producciones de energía en su mayoría pequeñas, aunque se encuentran a lo largo de todo el gráfico, y emisiones moderadas. En el anterior gráfico el 2 representaba producciones y emisiones altas por lo que en este sentido el mapping está haciendo mejor trabajo en la definición de los países de cada cluster, aunque 7 clusters es demasiado elevado para los objetivos buscados con la clusterización.

Mirando los clusters óptimos mediante la métrica elbow vemos que sigue siendo 4 aunque el coste computacional ahora es el más elevado de los clusters.

**SEXTA CLUSTERIZACIÓN**

*2 variables: emisión de co2 y producción de energía,power transformer,mapping y label econder*

Tanto con mapping como con label encoder la reducción de la complejidad del modelo a solo dos variables delimita claramente los clusters y obtenemos un gráfico muy interpretable. Sin embargo, con map es más clara que con label encoder pues podemos diferenciar mejor los clusters en base a las dos características elegidas y se ajusta mejor a la realidad obtenida en el EDA, donde hay más concentración de países en los clusters de emisiones bajas y menos concentración en los clusters de emisiones co2 y producción altas.

En conclusión, dada la mayor interpretabilidad del modelo con 2 variables, la delimitación clara que se produce con kmeans y el mapping, para la variable tipo de energía y el buen coeficiente de silueta obtenido en este modelo resulta ser el elegido.

Hemos podido corroborar que Kmeans trabaja bien con valores más dispersos, que es lo que realiza powertransformer cuando los asemejamos a la distribución normal y que aunque obtiene un score medio de silueta algo inferior a DBSCAN sus conclusiones son más robustas dado que todos los valores de silueta de sus cluster están por encima del score medio y por

encima de 0.5 lo que nos indica que los segmentos de la población se diferencian suficientemente bien y son coherentes con las conclusiones sacadas del EDA. Además, la silueta o cluster más importante tiene un score cercano a 0.9 como pasa en DBSCAN reflejando el hecho de que existe un cluster mayoritario en el dataset.

***DBSCAN***

En las representaciones de los datos usando las emisiones de co2 y la producción de energía, hemos podido apreciar que existe una gran densidad de datos, si usamos Standard Scaler o no los escalamos, y por tanto creemos que DBSCAN como algoritmo de clusterización puede hacer un buen trabajo dado que se basa en esta densidad de datos para separar los clusters. Se encarga de separarlos identificando regiones menos densas estableciendo un radio alrededor de cada cluster y un mínimo de puntos que debe contener estas regiones densas. Además, es un algoritmo que trabaja muy bien con valores extremos.

*standard scaler, todas las variables:*

Establecemos el baseline con todas las variables y usando standard scaler. Con ello obtenemos 67 clusters, 473 puntos de ruido y un coeficiente de silueta de -0.001 lo que nos indica que esta clasificación no es válida. Para mejorar esto, lo que vamos a hacer es intentar buscar un valor adecuado para el parámetro eps, que es el radio establecido en el cual se van a buscar las regiones densas de puntos. Es decir, los datos deben estar dentro de ese radio para ser considerados un cluster. Este dato es el más importante en DBSCAN. Mediante ello hemos encontrado que el óptimo lo encontraríamos en 4 clusters con un eps de 3.18 aproximadamente y un ruido de 47 datos. Para llegar a este punto, hemos buscado además cuál sería el valor óptimo de puntos mínimos a incluir en cada cluster mediante iteración y hemos obtenido un valor de 9 lo que nos ha dado como resultado un valor de silueta de 0.6.

*usando dos variables: emisiones de co2 y producción de energía, standard scaler:*

Si usamos solamente las dos variables más representativas del proyecto, llegamos a conseguir un coeficiente de silueta medio de 0.85, por lo que nuestros clusters quedan muy bien definidos, usando 4 clusters. Sin embargo, el cluster 0 aglutina a prácticamente todos los datos del dataset y dado su elevado coeficiente, en la media este pesa mucho y consigue ese dato tan bueno. Lo que está pasando es que uno de los clusters tiene muchos datos, pero el resto muy pocos y ese cluster tiene tanto peso que nos da una imagen general que parece muy buena, pero no nos convence, pues el resto tiene un score muy malo y no es representativo de nuestro dataset esta división de grupos.

*2 variables, power transformer mapping y label encoder*

aunque obtenemos mejor coeficiente de silueta medio que con Kmeans para cuatro clusters, la dispersión que existe entre ellos es muy elevada, puesto que está apoyada en el gran peso que tiene uno solo de los cluster, el cual queda por encima del score medio de silueta, pero el resto de ellos está muy por debajo.

En conclusión, podemos establecer que DBSCAN trabaja mejor con StandarScaler y label encoder porque los datos están más agrupados y por tanto tiene una distribución más densa, y es justamente donde este algoritmo lo hace mejor. Pero los resultados que obtenemos no se corresponden enteramente con las conclusiones extraídas del EDA. Además, aunque consigue un mayor valor medio de silueta, es debido al valor tan alto que alcanza uno solo del cluster dejando el resto por debajo del valor medio y por tanto nos dice que identifica muy bien un solo cluster, pero el resto los diferencia muy mal.

**3. REGRESIÓN**

A. *Modelo General*

En principio eliminamos algunas de las variables que no vamos a necesitar para nuestro modelo de regresión y codificamos la variable energy type con map, de manera que nos permita asignarle los números correspondientes a cada uno de los labels de la variable, y por tanto los pesos que tendrán cada uno de ellos, en base al co2 emitido.

Revisamos la distribución de cada una de las variables y vemos que la gran mayoría tienen poca similitud con una distribución normal. En general son distribuciones asimétricas, con cola a la derecha, dado que tenemos países con valores muy grandes en relación con los demás, que por tanto tiran de la media hacia valores altos, dejando a esta a la derecha de la mediana, excepto en la variable balance. En esta variable podemos ver que es más simétrica que las otras, pero tiene una curtosis bastante acentuada, siendo por tanto leptocúrtica. De forma general, lo que se aprecia es que:

1. existen variables con escalas muy diferentes entre ellas, así podemos ver que GDP tiene valores en cientos de miles, per cápita production en cientos y en miles y co2_pc en unidades

2. que los datos tienen bastantes valores extremos en general. Por esta razón, hemos decidido que vamos a escalarlos para que los algoritmos puedan trabajar mejor con los datos. Dadas las características de nuestros datos establecemos que vamos a usar uno de estos dos:

- PowerTransformer: se usará yeo-jhonson debido en que algunas variables tenemos valores negativos y box-cox no los soporta, y nos gusta porque es adecuado cuando la varianza de los datos no es constante o muy grande, ayudando a estabilizarla resultando en una distribución parecida a la Gaussiana, cuando existe una gran asimetría y es buena cuando existen valores extremos.

- RobustScaler: está basado en percentiles y por tanto no está influenciado por outliers. Esto resulta en que el rango resultante de valores transformados sea mayor que otros transformadores, pero similar al de la variable, pues mantiene la distribución. Es bueno cuando no se quieren eliminar los valores extremos.

A continuación, miramos el Variance Inflator Factor (VIF) que nos sirve para detectar la multicolinealidad entre las variables. Esta se produce cuando dos variables están muy correlacionadas y por tanto contienen información similar entre ellas. El vif pone en relación la varianza del modelo entero con la varianza de ese mismo modelo, pero solo usando la variable a estudiar, y nos indica cuánto contribuye a la varianza total del modelo en su conjunto. La multicolinealidad puede no interferir en el rendimiento del modelo, pero afecta negativamente a la interpretación de los predictores utilizados para construir el modelo. De esta manera, aquellas variables que tengan un vif superior a 5 serán eliminadas una a una, y en cada eliminación se reevaluará el modelo, de esta manera nos aseguramos de perder la mínima información posible. Con la selección de estas variables realizaremos un modelo.

Por otra parte, usamos también la regresión lineal de statsmodels que nos provee pruebas paramétricas de significancia estadística para crear otro modelo y decidir que variables lo van a componer, con ello podremos comparar los modelos creados y seleccionar las variables adecuadas.

Otra línea de selección de variables que vamos a seguir durante la fase de regresión será usando la herramienta que nos provee sklearn denominada RFECV (Recursive Feature Ellimination Cross Validation). Esta herramienta lo que hace es realizar un ranking de las variables con mejor desempeño y con peor. De esta manera, va evaluando el modelo y eliminando las peores hasta quedar con el modelo que mejor métrica arroja. Mientras realiza esta selección recursiva de variables lleva a cabo un cross validation, de tal manera que la métrica sea lo más consistente posible y no fruto de una prueba al azar.

En esta primera fase, hemos expuesto el método que vamos a seguir en los próximos notebooks de regresión y que conforman el baseline de estudio.

La conclusión a la que hemos llegado en esta primera parte es que vamos a pasar a estudiar a cada uno de los países por separado ya que predecir en base a todos los países en conjunto no tiene mucho sentido y de ahí que al realizar cada una de las pruebas mencionadas anteriormente no se obtengan mejoras significativas en cada una de las fases de estudio, ni eliminando variables.

B. *Modelo Individual*

En esta fase, pasamos a estudiar el extremo opuesto al anterior, de manera individualizada en vez de general, para ver qué datos obtenemos. Para el estudio hemos elegido a China, por ser uno de los países más importantes del dataset y que fue objeto de estudio durante la fase de EDA. Para ello, lo hemos escalado con powertransformer por las razones expuestas en la fase general y hemos establecido como índice la fecha de cada una de las entradas.

Al realizar un primer modelo con LinearRegression mejoramos los resultados del modelo general rápido, pasando de obtener valores de r^2 de 0.968 mientras en el general rondaba el 0.8 con cualquier método o selección de variables.

Siguiendo las fases anteriores vamos a seleccionar las variables mediate statsmodels, RFECV, VIF, mejor correlación con el target, eficiency. De estas pruebas hemos podido corroborar que, realizar un estudio individualizado por países mejora bastante el modelo predictivo ya que estamos teniendo en cuenta la realidad de ese país, no mezclándola con la de otros países. De ahí que hemos sido capaces de simplificar el modelo hasta 6 variables y obtener un r2 de 0.967 con las variables mejor correlacionadas con el target. Mirando el AIC y el BIC de este modelo llegamos a la conclusión de que hemos reducido la complejidad del modelo bastante y hemos conservado una muy buena capacidad explicativa general del target, por lo que el modelo representa bien los datos concretos del estudio y además es capaz de generalizar para otros datos.

Las métricas de mae, mse y rmse son parecidas al modelo con todas las variables y mejor que un modelo con 5 seleccionadas mediante vif. Por lo que nos indica que es un punto óptimo y pensamos que marca un punto de referencia para la siguiente fase debido a que por ahora hemos estudiado solamente dos extremos del mismo estudio a efectos comparativos. Esto es así porque, igual que realizar un modelo general donde se mezclan realidades de países completamente diferentes no tiene sentido a efectos de predicción. Un modelo individualizado para cada uno de los 40 países que tenemos no es práctico y es posible obtener buenos resultados si unificamos estas características individuales mediante la detección de algún patrón que los haga similares y entonces realizar un modelo para cada segmento de país.

Por otra parte, en la fase final del estudio de regresión vamos a tratar de seleccionar el mejor estimador para cada uno de los clusters obtenidos.

C. *Modelo por Clusters*

En este punto estudiamos el punto medio entre los dos extremos anteriores, la generación de un modelo de regresión por cluster. Este punto está basado en que es posible obtener un modelo de predicción bueno en base a la segmentación de los países por características similares entre ellos y que sea un proceso práctico. Para llevarlo a cabo vamos a utilizar las mismas técnicas de selección que en el modelo de regresión individual, pero además vamos a comparar diferentes métodos de regresión.

El primer método que llevamos a cabo es la selección de variables mediante correlación, estableciendo que serán elegibles todas aquellas que tengan un valor absoluto superior a 0.49 (dato sacado del hecho de que los valores de correlación entre las variables objeto de estudio no son muy elevados en general), y podemos comprobar que prácticamente se repiten las mismas variables que se venían usando en el modelo individual.

**Cluster 0**

En primer lugar, realizamos un base line usando todas las variables para ver que estimador lo hace mejor a priori. En este cluster, el que mejores datos ha sacado ha sido en r^2 AdaBoostRegressor con el valor más alto y la menor dispersión entre cada métrica, obtenida

del cross validation. Mirando el error absoluto medio, también ha sido el mejor, con el menor valor. En rmse, el mejor ha sido esta vez Random Forest, pero AdaBoost ha sido el segundo mejor, con una dispersión similar.

Pasando a mirar cómo se ajustan las predicciones a la realidad, podemos apreciar que los mejor ajustados son Random Forest, AdaBoost y Gradient Boosting que son los que obtienen mejores métricas de error absoluto medio, error medio cuadrático y r^2. De esta manera, podemos intuir que este será el estimador a elegir. Vemos que al final el estimador a elegir para este cluster es el AdaBoostRegressor con base en DecisionTreeRegressor y usando las variables proporcionadas por la mejor correlación y RFECV, pues ambos han llegado a la misma selección. Estas han sido la producción de energía, el consumo de energía, las emisiones de co2 y el balance. Por tanto, estas son las características que mejor definen al cluster 0.

**Cluster 1**

El cluster 1 es el que más quebraderos de cabeza nos ha dado, pues para él hemos tenido que seleccionar un escalado diferente, que nos daba mejores resultados en las métricas, este ha sido RobustScaler. Pensamos, que viene determinado porque en este cluster existen valores extremos, por tanto, hemos probado a quitarlos, ya que afectan a las medias y la varianza, pero no mejoraban las métricas, obteniendo los mejores resultados usando solo el escalado. Una cosa que caracteriza a este cluster también es que sus variables son las que menos correlación tienen con la eficiencia de todos los clusters, y pensamos que debe ser un punto importante que afecta a los modelos de este cluster en general. Así las cosas, hemos desechado todas aquellas combinaciones de estimador-variables que nos daban valores de r2 iguales a 1 y mae-rmse cercanos a 0 debido a que son modelos que tienen mucha varianza y poco bias, lo que los hace modelos que captan muy bien los datos del modelo en concreto y de forma específica pero que no sirven y lo hacen muy mal a la hora de predecir y generalizar para datos nuevos, overfitting. Esto lo podemos ver en que en las gráficas la realidad sigue un patrón y las predicciones otro completamente diferente. Al igual que hemos descartado estos modelos, también hemos descartado los modelos con valores de r^2 negativos o cercanos a 0 y valores de mae-rmse muy altos debido a que son modelos los cuales sus variables independientes no son capaces de explicar la variabilidad de la variable dependiente y cometen en consecuencia errores enormes de predicción, no siendo mejores que una media, y haciendo que su uso sea del todo inadecuado. De esta manera, la constante positiva ha sido que los modelos AdaBoost, DecisionTree, RandomForest y GradientBoostingRegressor han sido los estimadores que de manera consistente han obtenido buenos resultados en r^2 y con valores de mae y rmse dentro de la normalidad y que es posible considerarlos normales. De ellos, el mejor ha sido el DecisionTree con la selección de variables mediante RFECV, que ha conseguido unos valores de R2 de0.925, mae 44.73 y rmse 1110.34. Las variables que han caracterizado a este grupo han sido las emisiones de co2, las emisiones de co2 per cápita, el consumo de energía, y la intensidad de energía por euro de pib.

**Cluster 2**

Para este cluster, a priori el estimador que mejor lo hace es AdaBoostRegressor con base en DecisionTree ya tiene el mejor dato de mse y rmse y tiene su r2 prácticamente al nivel de GradientBoostingRegressor que es el mejor en ese aspecto. Igualmente, sus dispersiones en las

métricas indican que es un estimador robusto y confiable. Sin embargo, analizado cada par estimador-variables al final se selecciona RandomForestRegressor con las variables obtenidas mediante correlación ya que obtiene valores similares a los demás, pero con una complejidad de modelo inferior que los demás. Así frente a las 8 variables que usa adaboost y 6 de RandoForest con selección mediante RFECV, este modelo obtiene resultados similares con 5 variables.

Aun así, la estimación del modelo mediante el base line en el cual nos indicaba a AdaBoost no iba desencaminada, puesto que ADA ha sido uno de los mejores y estaba basado en DecisionTrees que al final es el algoritmo que hemos elegido, pero mediante Bagging en vez de Boosting.

Las variables que caracterizan a este modelo son el GDP, la población, el consumo de energía, la emisión de co2 y el balance.

**Cluster 3**

En el cluster 3 el baseline indica que GradientBoostingRegressor es el mejor en todas las métricas, si a valor medio nos referimos, aunque no en dispersión o consistencia de resultados. Sin embargo, en este último tampoco es el peor, pues tiene valores aceptables y pensamos que es posible que sea el que se obtenga definitivamente. Vista esta primera impresión, al pasar a estudiar los pares estimador-selección de variables obtenemos que este modelo usará GradientBoostingRegressor con las variables seleccionadas mediante RFECV pues con solo 4 variables consigue datos iguales o mejores que el resto de estimadores con más variables.

Las variables que caracterizan a este cluster son la emisión de co2, la producción de energía, el balance y el consumo de energía.

Para finalizar este apartado pondremos en relación las conclusiones obtenidas en la parte clusterización con la de regresión. De esta manera, el cluster 0 se caracterizaba por contener países que tienen poca producción, pero emiten bastantes emisiones de co2, resultando poco eficientes y los menos eficientes de todos. Si relacionamos esta descripción con las variables que caracterizan su modelo podemos ver que tenemos:

+ balance

+ energy_consumption

+ energy_production

+ co2_emission

es decir, las variables que más caracterizan a esta descripción. Al ser países con poca producción y mucha emisión de co2, sabiendo la producción y el co2 podemos saber si cumplen esta cualidad. Además, al ser países que producen poco, son países con un consumo de energía superior a la producción y por tanto con un balance negativo.

Por otra parte, el cluster 1 está representado por países con poca producción y pocas emisiones de co2 basados en el petróleo, las renovables y el gas natural, siendo los segundos menos eficientes de todos. Las variables que más los caracterizan para predecir su eficiencia han sido:

+ co2_emission

+ co2_pc

+ per_capita_production

+ Energy_consumption

En estos países, pensamos que el modelo ha preferido las variables per cápita de las emisiones de co2 y de la producción porque contiene valores más pequeños, dados los valores también pequeños de las emisiones de co2 y la producción en valores absolutos. Aun así, como son países que los caracterizan mucho las pocas emisiones de co2 y el poco consumo de energía, en valores absolutos, el modelo los ha seleccionado como predictores de la eficiencia.

El cluster 2 tiene a los países con mayor producción y emisión de co2 del mundo, muy por encima de los otros cluster, pero siendo los segundos más eficientes de los clusters. Su producción está basada en el carbón, el gas natural y el petróleo. Dada esta descripción, las variables que han caracterizado el modelo de regresión de la eficiencia han sido:

+ GDP

+ Population

+ Energy_consumption

+ CO2_emission

+ balance

Al ser países tan grandes en media y con tanta producción, son países con grandes productos interiores brutos y población. Además, son países que tienen un consumo elevado de energía debido a sus altos niveles de PIB y las emisiones de co2 son las más elevadas, llevando a que tengan balances de energía positivos.

En el cluster 3 nos encontramos a países con las segundas mayores producciones de energía mundiales, que usan todas las fuentes de energía en proporciones similares en media, y sus emisiones de co2 son muy bajas, cercanas al cluster 1 para una producción mucho mayor.

Estos países son los más eficientes con diferencia y serían la referencia que seguir. Los países que más se encuentran a lo largo de los años del estudio en este cluster han sido:

1. Rusia
2. Estados Unidos
3. Canadá
4. Francia
5. Japón

Francia y Japón son dos países altamente enfocados en la energía nuclear(75,5% y 54,1% de su producción) y Estados Unidos la usa también, aunque en menor medida que estos dos países. Por otra parte, Rusia junto con Canadá usan como fuentes prioritarias de producción el gas natural y el petróleo y en menor medida el carbón. Este medio de generación es el tercero para Rusia y el cuarto para Canadá que prioriza las renovables con un 21% de su producción total siendo la nuclear residual para ellos.

Como variables para predecir la eficiencia de este cluster han sido seleccionadas:

+ co2_emission
+ energy_production
+ balance
+ energy_consumption

que son variables muy parecidas a la de los clusters anteriores e iguales a las del cluster 0.

Para concluir podemos decir que la eficiencia es posible predecirla en todos los clusters de manera general con una estructura de variables estable, y por tanto estas variables son las que más definen a la eficiencia. En este caso serían, la emisión de co2, la producción de energía, el consumo de energía y el balance entre la producción y el consumo de esta.

**4. CLASIFICACIÓN**

Como hemos ido siguiendo en las anteriores etapas del proyecto, primeramente, hemos creado un baseline usando todas las variables del dataframe a efectos comparativos y para tener una primera sensación de qué modelo va a tener mayores posibilidades de salir elegido. En este caso, hemos comprobado que de manera reiterada en todas las métricas los mejor posicionados han sido RandomForest, GradientBoosting y DecisionTree, además en el orden especificado en todas. Por tanto, hemos decidido que serán los tres que vamos a tratar con más detalle, seleccionando para cada uno de ellos las variables a usar. A priori, el que más nos gusta es DecisionTreeClassifier ya que, aunque es el tercero de todos, su dispersión entre las

métricas de cada uno de los cross validation efectuados es la menor de todas, lo que nos indica que es un estimador robusto y fiable.

Por otra parte, dada la experiencia en la selección de variables de RFECV, que la mayoría de las veces ha sido el que mejores métricas ha obtenido, en esta ocasión solo vamos a usar esta opción para la selección de variables.

Además, hemos revisado el balance de clases que existe en el dataset y hemos visto que no existe un gran desbalance entre ellas, aunque sí que hay diferencias, pero que no hemos considerado que fuera necesario realizar un resampling.

Vemos que a la hora de clasificar a nuestros países en un cluster han sido elegidas las variables más importantes la emisión de co2 y la energía de producción. Estos resultados son coherentes con todo el estudio que se viene realizando, ya que son las variables más relacionadas con el objetivo que se persigue, que es en función de la producción considerar si los países emiten mucho o poco co2 y por tanto cuanto son de contaminantes. Además, las dos que más les importan son el consumo de energía, que tiene una relación directa con la producción de energía y por tanto con el nivel de emisiones de co2 y la eficiencia que es resultado de todas las anteriores. Estas variables son además la base de la caracterización de los clusters de nuestro estudio.

En general en nuestros modelos no necesitaríamos hacer un hyperparameter tuning, a excepción del realizado manualmente en la fase de clusterización con DBSCAN, dado que los resultados son muy buenos.

En esta primera parte de clasificación existe prácticamente una relación matemática entre las variables y el output. Esto es debido a que los clusters están hechos en base a las emisiones de co2 y la producción de energía, y la eficiencia es la división entre las emisiones de co2 y producción de energía. Por lo tanto, para llevar a cabo la clasificación, mediante estas dos variables, los estimadores puede identificar correctamente a que cluster pertenecen. Debido a esto, lo que vamos a hacer es que vamos a quitar las variables co2_emission,eficiency y Energy_production y vamos a ver como lo hace sin ellas. Además de esto, vamos a incorporar como variables onehotencoded el continente al que pertenece cada país, para ver si nos aporta alguna información relevante al modelo.

Generado el baseline con las nuevas características descritas anteriormente podemos ver que, los continentes no aportan ninguna información relevante y por tanto son prescindibles. Esto lo podemos ver en que, aunque las métricas seleccionadas de evaluación, como son el accuracy, el recall, precission y f1 score han bajado algo, cosa que era esperable, los datos siguen siendo muy buenos y RFECV no ha seleccionada a estas variables en ningún momento como importantes. Además, podemos apreciar que los clasificadores hacen un trabajo muy consistente, pues en el top 3 se mantienen los mismos que en el notebook anterior, teniendo DecisionTree la menor dispersión de todos y siendo el tercero de ellos, pero con la diferencia que ahora se han intercambiado las posiciones, siendo ahora RandomForesst el mejor y GradientBoosting el segundo.

En esta ocasión la estructura del modelo ha cambiado y necesita más variables para realizar un buen trabajo y se da cuenta que lo importante para la predicción es la emisión de co2 y la producción de energía, porque todas las variables que selecciona la mayoría de las veces tienen relación con ellas. Así las cosas, la estructura básica de los tres es:

+ Energy_consumption

+ Population

+ balance

+ co2_pc

+ energy_type

+ per_capita_production

Siendo el elemento diferencial el uso de las variables que tienen algo que ver con PIB del país.

Mediante el cross validation medimos el accuracy, la precisión, el recall y el f1 score de todos ellos y apreciamos que las diferencias son pequeñas, pues todos hacen muy buen trabajo. Sin embargo, el que mejor lo hace es RandomForest que consigue un f1 score de 0.978 con una dispersión de 0.003. La dispersión para esta métrica es en todos igual pero su valor es ligeramente superior al resto. Esto nos indica que en general tiene un equilibrio entre precisión y recall superior al resto de estimadores.

En las métricas de precisión, recall y accuracy obtiene el mismo valor, 0.978, superior al de los otros dos que tienen, 0.975 GradientBoosting y 0.969 Decision Tree. La dispersión también es menor entre sus métricas en Random Forest, siendo especialmente buena en accuracy al obtener el doble menos de dispersión, con un valor de 0.002 frente a 0.004 para los otros dos.

Si miramos las mismas métricas por clase, es decir, los valores que obtiene cada uno de los estimadores prediciendo la clase 0, l 1, la 2 y la 3 podemos apreciar que:

* Precisión: de los casos positivos que tenemos, el que más acierta en el cluster 0 es Random Forest, en el cluster 1 es decision tree, en el 2 es Random Forest y en el 3 es Gradient Boosting. Todos los valores están cercanos, pero Random Forest al acertar en más clusters hace un trabajo mejor general en términos de precisión.

* Recall: del total de casos positivos reales el que más positivos detecta, y por tanto mayor sensibilidad tiene, en los cluster 0,1,2 es Random Forest con un 0.985. Sin embargo, para el cluster 3 todos los estimadores lo hacen igual con 0.964. Por tanto, de nuevo en recall Random Forest lo vuelve a hacer mejor.

* f1 score: Como el valor de f1 es una medida combinada de la precisión y el recall era de esperar que Random Forest fuera el que mejor valor tuviera en todos los cluster, siendo el cluster 2 el que mejor se le da clasificar con un valor de 0.99 y el cluster 0 el que peor con un valor de 0.969.

En conclusión, hemos podido comprobar que en términos generales el mejor clasificador es Random Forest, tanto en términos de accuracy, precisión, recall y f1. Además, de manera individualizada por clusters, vemos que Random Forest suele tener los mejores datos, aunque en algunas ocasiones otros estimadores lo hagan mejor que él en clusters determinados. Es así

con el cluster 1 y 3 donde Decision Tree y Gradient Boosting lo hacen mejor que Random Forest en términos de precisión. Lo que realmente se le da bien a Random Forest es detectar una gran cantidad de positivos(clusters que se corresponden con la realidad dada), es decir tiene una gran exhaustividad o recall, y por eso es el mejor en todos los clusters, y de ahí que tenga el mejor f1 score en todos los clusters también.

## 5. MODELO FINAL

El modelo final se ha planteado como una clase en la cual se asignan sus atributos como las variables que van a entrar en el modelo, así como el resto de las variables, que se establecen como atributos pero que son calculadas a partir de las introducidas por el usuario. Una vez que la clase ha sido instanciada con sus correspondientes valores se crean las funciones que compondrán la base del modelo final. De esta manera, se divide el proceso de predicción en fases y se ensamblan en una única función que ejecuta el modelo final completo.

### PCA

Hay que indicar que, dados los resultados, nuestros modelos no necesitan muchas variables para hacer un buen trabajo y, por tanto, dada la dificultad añadida que un principal component analysis conlleva a la hora de explicar los modelos, aunque lo hemos realizado, no hemos visto necesario utilizarlo. Mediante este principal component analysis llegamos a la conclusión de que con ocho componentes lográbamos captar el 90.07% de la variabilidad del modelo, lo que no supone una mejora en la reducción de la dimensionalidad de los modelos, dado que con menos variables hemos sido capaces de obtener buenos resultados.