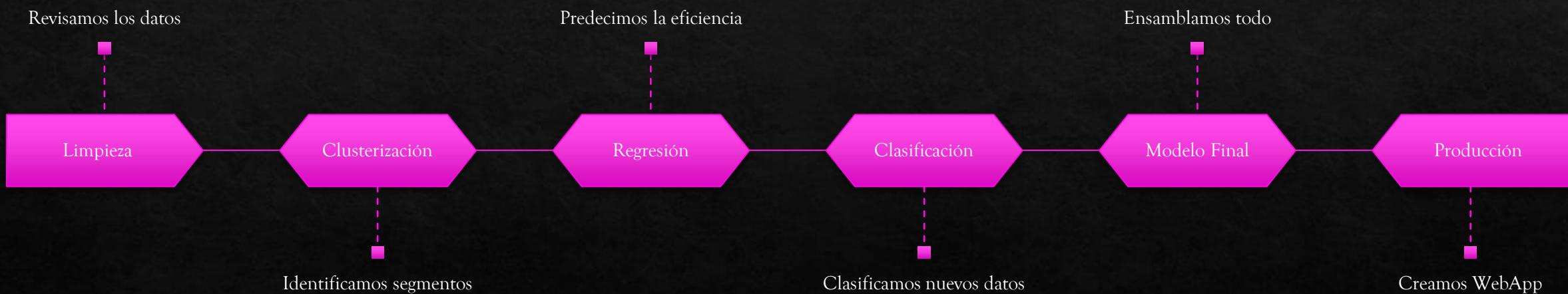


# CO<sub>2</sub> Emissions Machine Learning Project

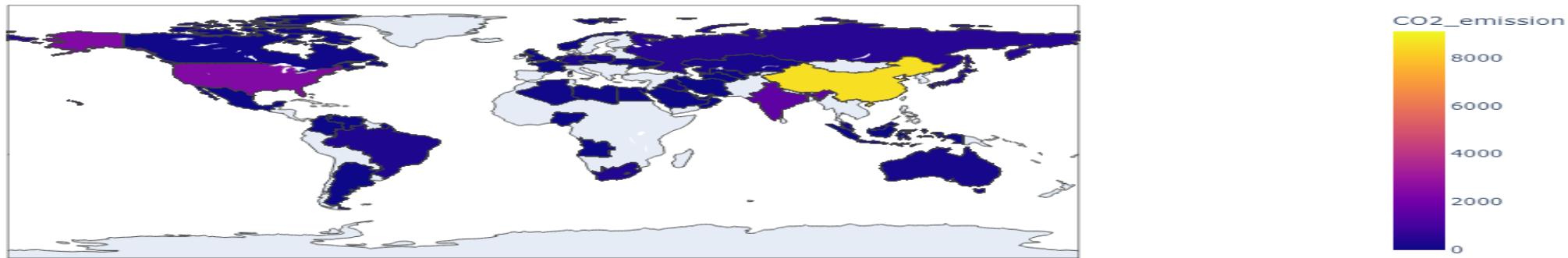
Fernando Sánchez Olmo

The Bridge Febrero 2022

# INDICE



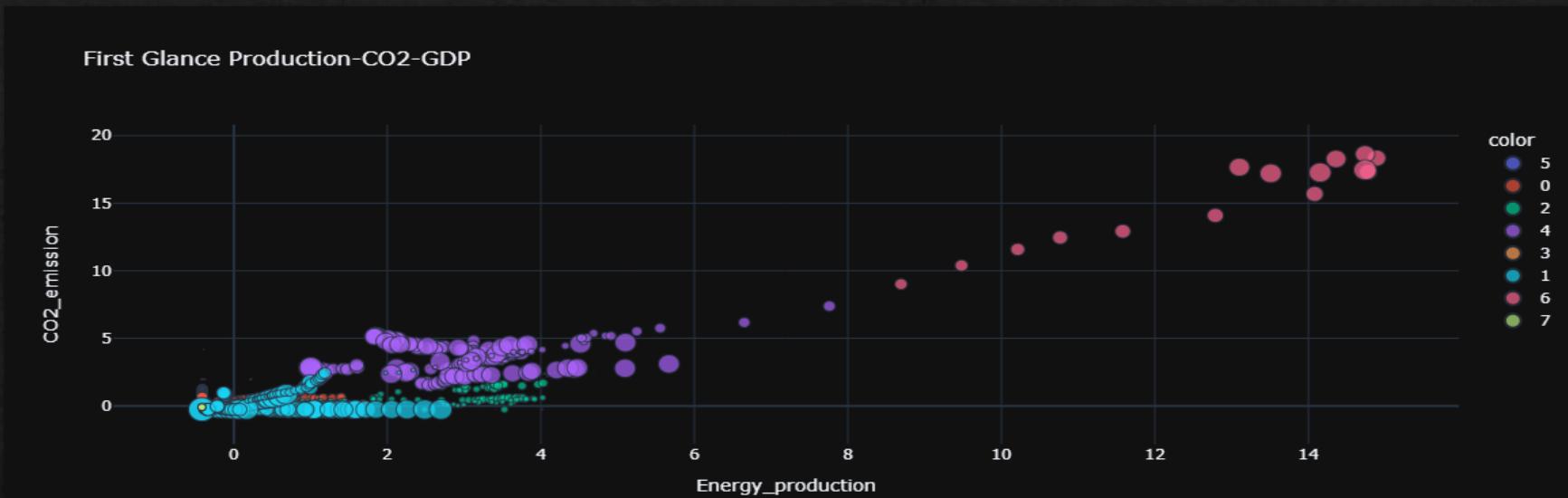
# LIMPIEZA



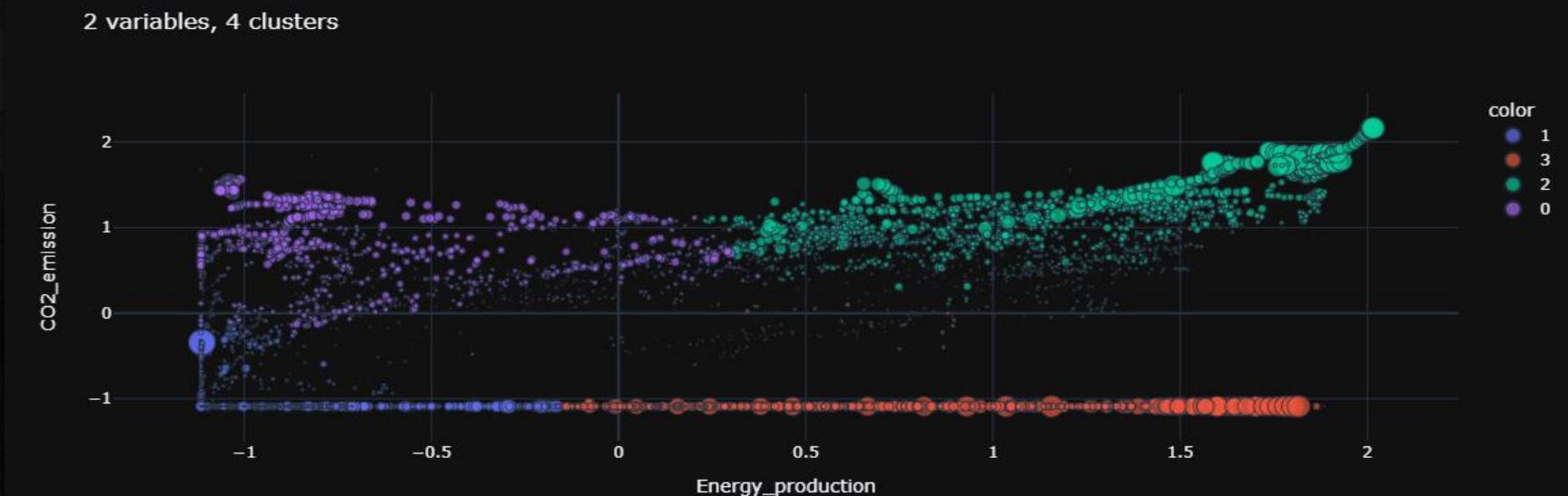
- ❖ En esta fase realizamos un exploración de los datos para ver si tenemos lo que necesitamos para desarrollar el proyecto de machine learning. Una vez revisado, realizamos los siguientes ajustes:
  - ❖ 1. Eliminamos la variable geometry, pues no nos va a ser necesaria.
  - ❖ 2. Encontramos que Yugoslavia es un país que se extinguió en el año 2013 y que más tarde se han ido creando diferentes países a lo largo del tiempo derivados de él. De esta manera, no nos es posible identificar que parte de los datos pertenece a cada uno de los actuales países en que se encuentra dividido por lo que se elimina del dataset.
  - ❖ 3. Checoslovaquia, fue extinta también pero se dividió solo en República Checa y Eslovaquia, por lo que realizamos la división de los datos en función del territorio de cada uno.
  - ❖ 4. Rellenamos los códigos internacionales de país faltantes

# Clusterización

❖ Situación de Partida



❖ Situación Final



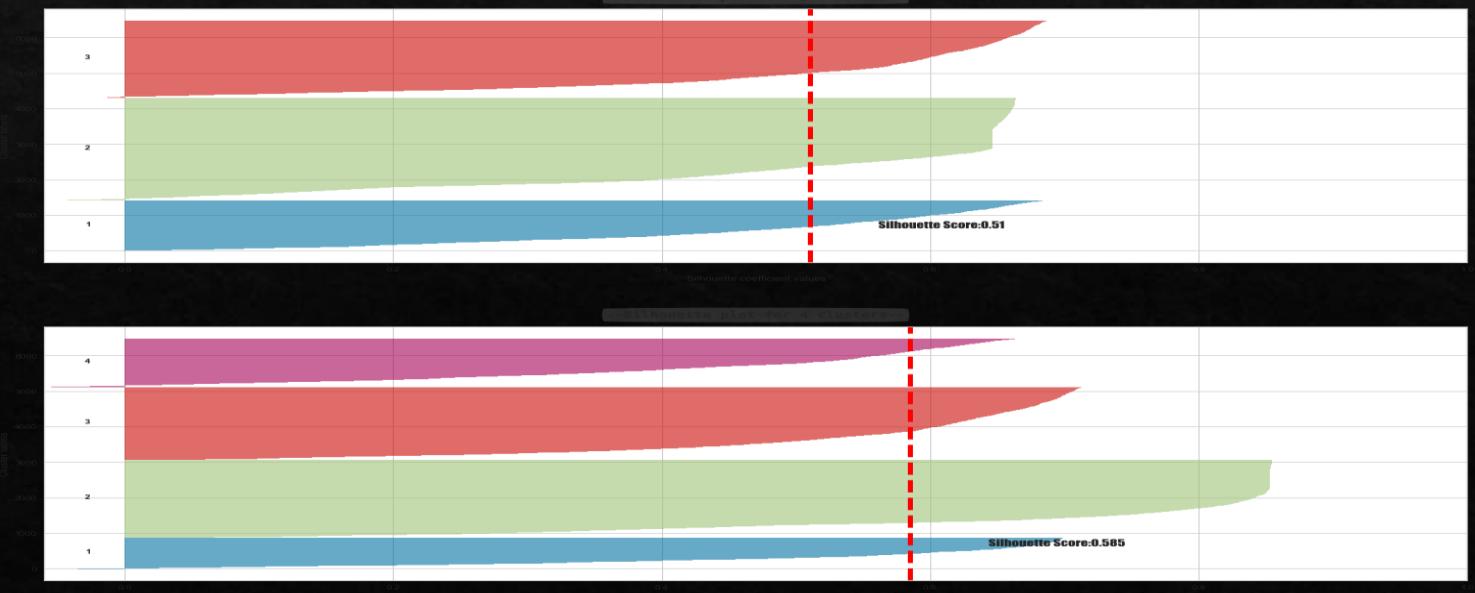
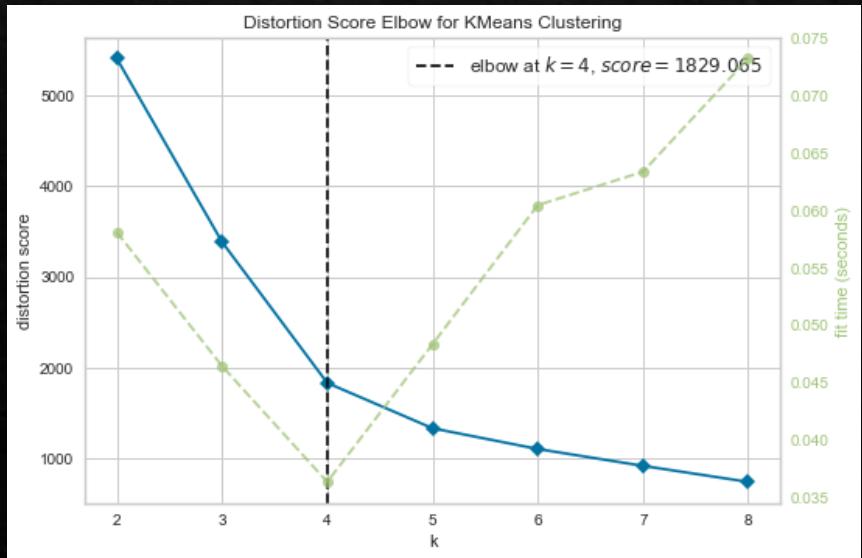
# Clusterización: ¿Como hemos llegado hasta ahí?

## KMEANS

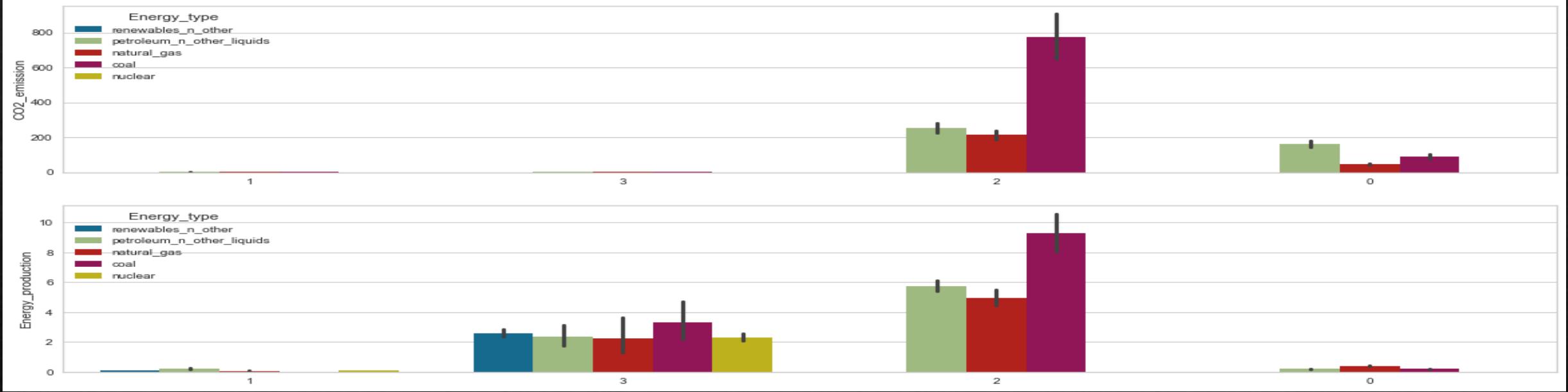
1. Baseline, Standard Scaler, Label Encoder, todas las variables: 7c(2>datos, 4 residuales), viable 4c, elbow 4c Ok(mal sil(0.3))
2. Standard Scaler, 2 variables (co2 y producción): var+repres, 7c inic, mejor reparto, elbow 3c Ok pero mayoría en clus 0.
3. Standard Scaler, 3 variables (eficiencia, producción y tipo de energía): Elbow Ok 4,buen sil (>0.7), no todos cl sil>si. Med.
4. Power Transformer, todas las variables, Label Encoder: distribución países coherente pero clusters mal definidos
5. Power Transformer, todas las variables, mapping: cluster mejor ajustados a realidad, sin reducción de complejidad solo Bl.
6. Power Transformer, 2 variables (co2 y producción), mapping, Label Encoder: GRÁFICOS DE ABAJO Y EL ANTERIOR

## DBSCAN

1. Baseline, Standard Scaler, todas las variables: 67 cl, sil -0.001, ruido 467 terminamos sil 0.6 y 4 cl pero usando todas vars
2. Standard Scaler, 2 variables (co2 y producción): sl 0.85 4cl pero casi todos en un solo cluster resto por debajo sl coef medio
3. Power Transformer, 2 variables (co2 y producción), mapping, Label Encoder: igual anterior pero datos poco inter. x dispers

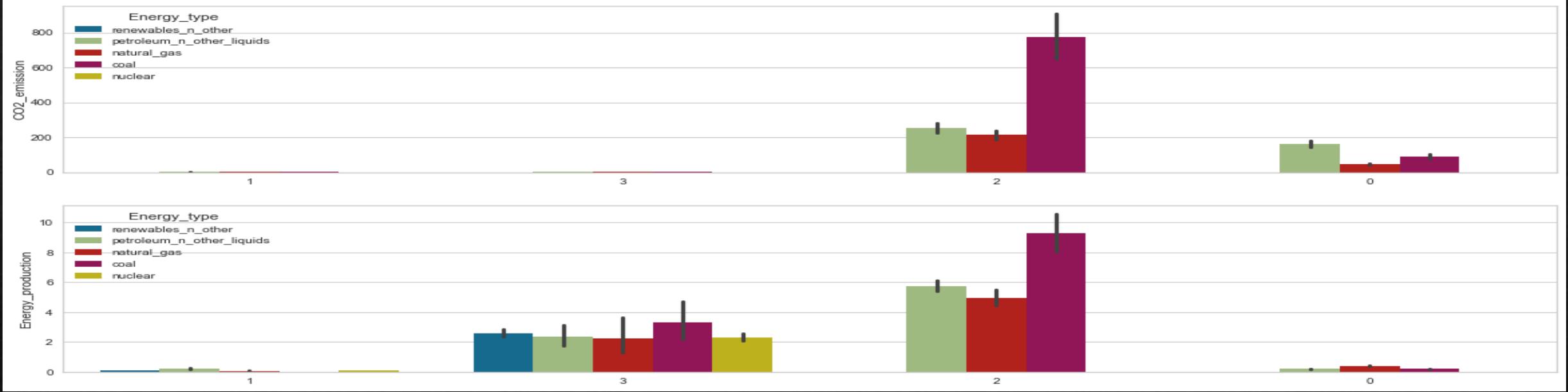


# Clusterización: Características de los clusters



1. El cluster 0: son países con una producción de energía pequeña, no siendo los que menos tienen, y basada en el petróleo, el gas natural y el carbón. Aunque son los terceros productores su contaminación es la segunda mayor. Por lo que podríamos resumir que son países con poca producción y bastantes emisiones de co2, nada eficientes.
2. El cluster 1 son países con poca producción y pocas emisiones de co2, son países que no resultan relevantes ni en producción ni en emisiones de co2, se basan en la nuclear, las renovables y el gas natural.

# Clusterización: Características de los clusters



3. El cluster 2 tiene a los países con mayor producción y mayor emisión de co2 del mundo. Su producción está basada en el carbón, el gas natural y el petróleo, siendo las tecnologías que más emiten co2 el carbón con diferencia, luego el petróleo y por último el gas natural.
4. El cluster 3 tiene a países con muy buenas producciones de energía mundiales, con un mix muy diversificado, pues usan todas las fuentes de energía en proporciones similares y sus emisiones de co2 son muy bajas, al nivel del cluster 1, para una producción mucho mayor. Diríamos que estos países son los más eficientes y serían el objetivo a seguir.

# Clusterización: Conclusión

Podemos establecer que DBSCAN trabaja mejor con StandarScaler y label encoder porque las datos están más agrupados, son más densos. Además, aunque consigue un mayor valor medio de silueta, es debido al valor tan alto que alcanza un solo 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. Por otra parte Kmeans con power transformer y mapping nos ofrece una buena relación entre interpretabilidad y calidad de segmentación, por lo que ha sido finalmente el elegido.

# Regresión: Creación de un contexto comparativo

- ❖ En la parte de Regresión una vez obtenidos los clusters hemos creado un contexto de comparación de tal manera que podamos tener una mejor noción de qué modelo es el mejor a la hora de predecir la eficiencia de cada uno de los países. En este contexto se establecen dos modelos extremos y uno central, como una aproximación media al objetivo.

## Modelo General

Un modelo para todos los países

## Modelo por Cluster

Un modelo para cada cluster

## Modelo individual

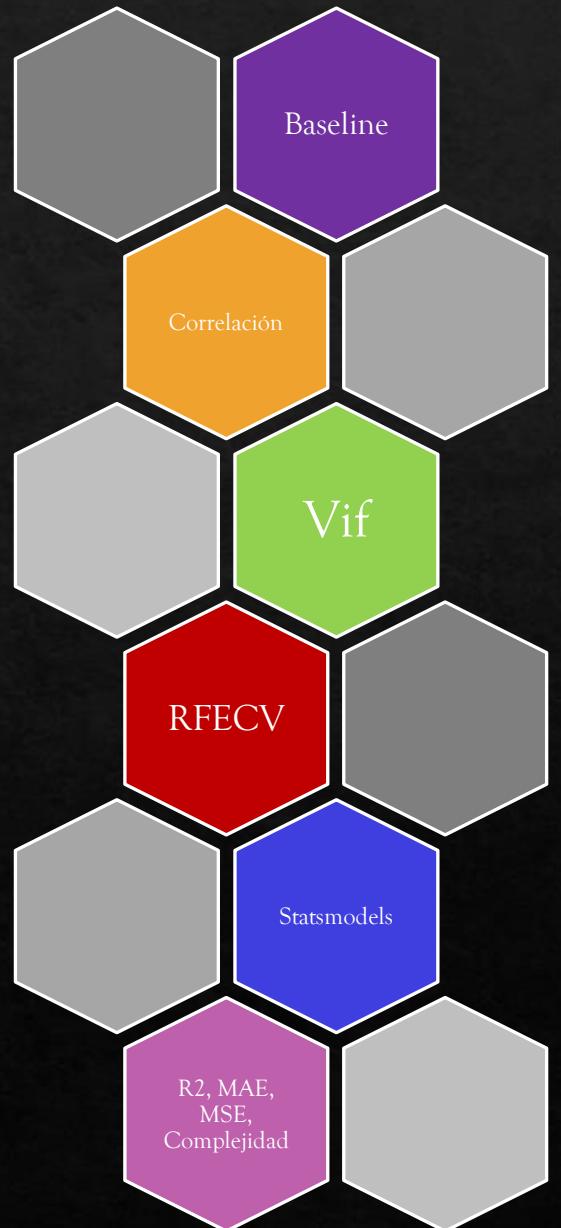
Cada país un modelo

# Esquema de Estudio

Durante esta primera aproximación establecemos la rutina a seguir en los tres modelos:

1. Creación de un Baseline
2. Selección de variables mediante correlación
3. Selección de variables mediante VIF (Variance Inflator Factor)
4. Selección de variables mediante RFECV (Recursive Feature Elimination Cross Validation)
5. Selección de variables mediante la librería Statsmodels
6. Comparación de R<sup>2</sup>, MAE ,MSE y complejidad de los modelos

Sin embargo, en el modelo general y en el individual no entramos demasiado en detalle puesto que sabemos que no van a ser opciones elegidas y simplemente nos sirven para establecer un contexto de referencia con el que poder comparar el modelo por clusters y con el que tener una noción de lo que pasaría en ese caso. Esto es así porque, el modelo general sería tratar a todos los países por igual, cosa que no tiene sentido al ser países con estructuras culturales, de producción y necesidades. Y por el contrario modelos individuales sería demasiado costoso y poco práctico llevarlo a cabo, y por eso se establecen como extremos a efectos de referencia.



# Modelo General: Statsmodels

Como vemos aunque el valor de R<sup>2</sup> no es que sea malo, aunque reducimos la complejidad del modelo, los valores de aic y bic no mejoran de una manera en que pueda ser relevante. Esto nos puede indicar que aunque el modelo es capaz de explicar la variabilidad de los datos concretos bien, es posible que no guarde esa misma capacidad a la hora de generalizar para otros datos.

## Todas variables

- Aic: 6307
- Bic: 6385
- R2: 0.802
- Resultado: eliminar energy intensity by gdp al ser estadísticamente igual a 0

## 11 variables

- Aic: 6305
- Bic: 6377
- R2: 0,803
- Resultado: eliminar use intesity per capita al ser estadísticamente igual a 0

## 10 variables

- Aic: 6303
- Bic: 6368
- R2: 0,803
- Resultado: Llegados a este punto hemos reducido algo la complejidad pero sin mejoras apreciables.

# Modelo General: VIF-Statsmodels

10  
vars

- Aic: 6337
- Bic: 6403
- R2\_cv: 0,805

Statsmodels

Como el modelo ha empeorado usamos statsmodels para ver si podemos eliminar variables

8  
vars

- Aic: 6333
- Bic: 6386
- R2\_cv: 0,801
- No ha mejorado

# Modelo General: RFECV

**6 vars:**

R2: 0,76

Mae: 0,31

Mse: 0,24

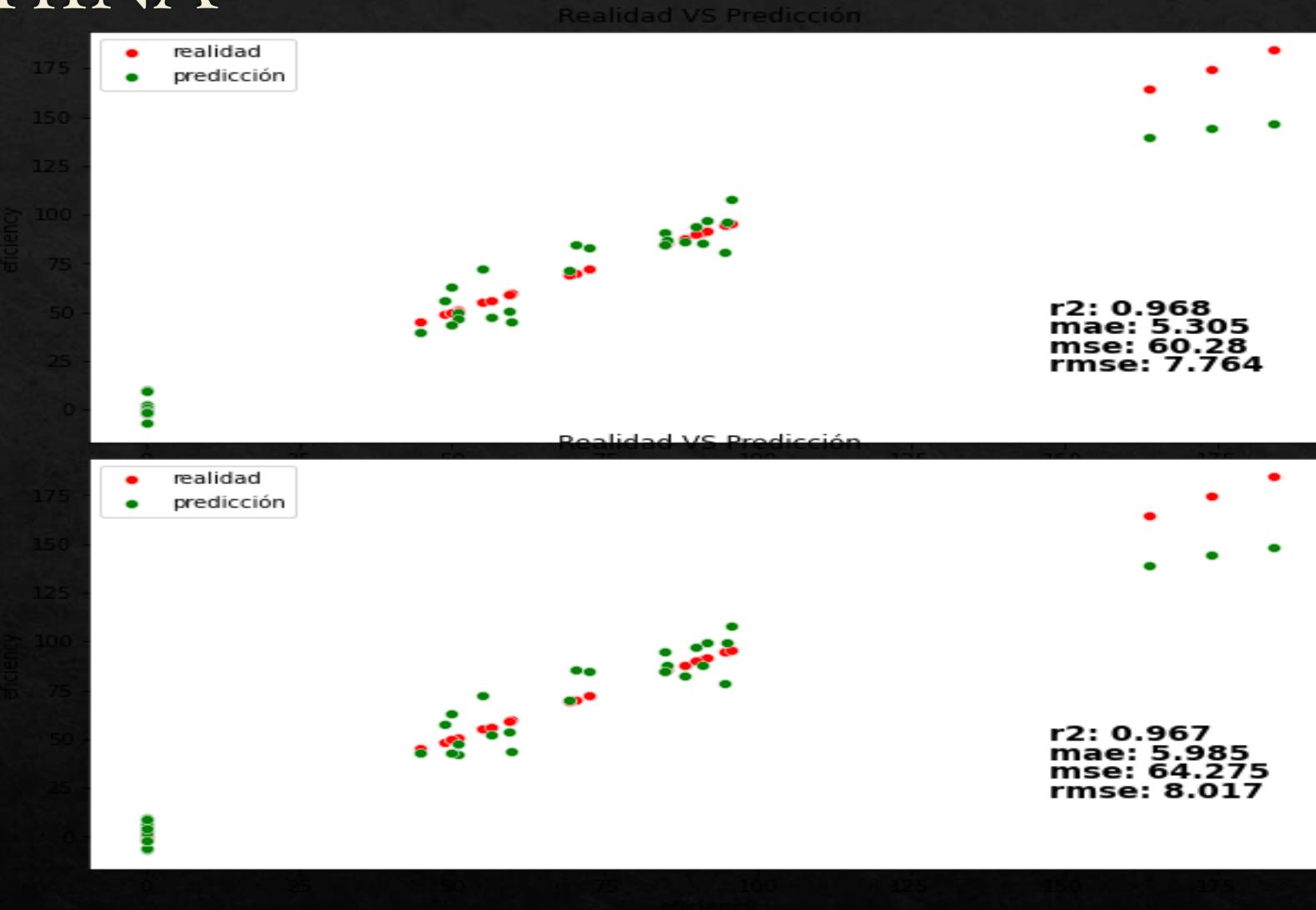
En este punto vemos que no vamos a mejorar mucho más y damos por establecido el primer extremo del contexto.

# Modelo Individual-CHINA

- USANDO TODAS LAS VARIABLES OBTENEMOS BUENOS DATOS PUES NO ESTAMOS MEZCLANDO LAS REALIDADES DE CADA PAÍS.

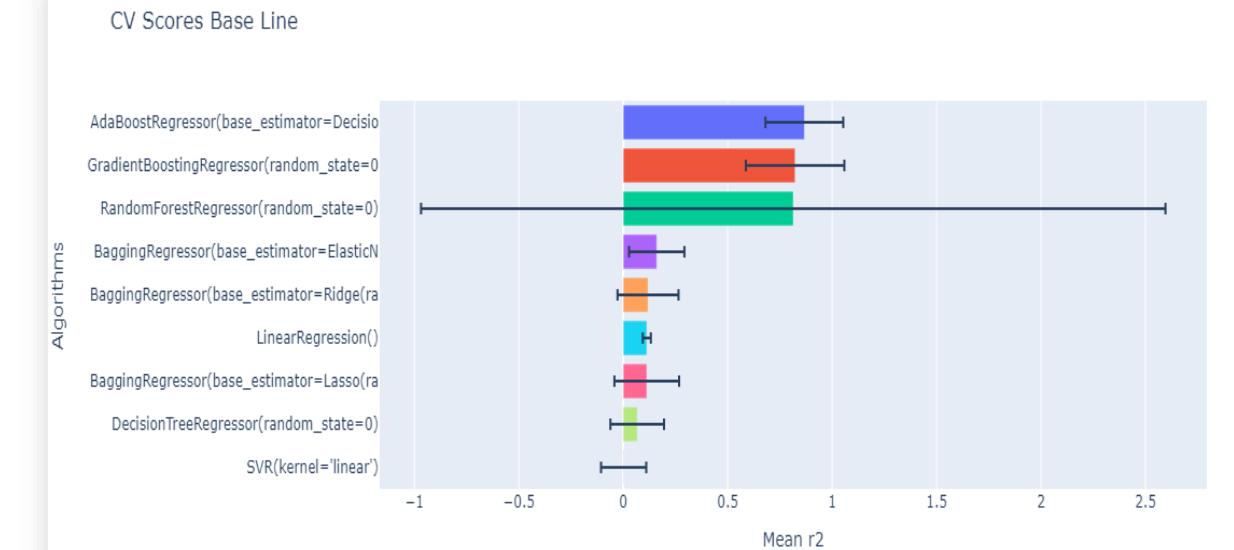
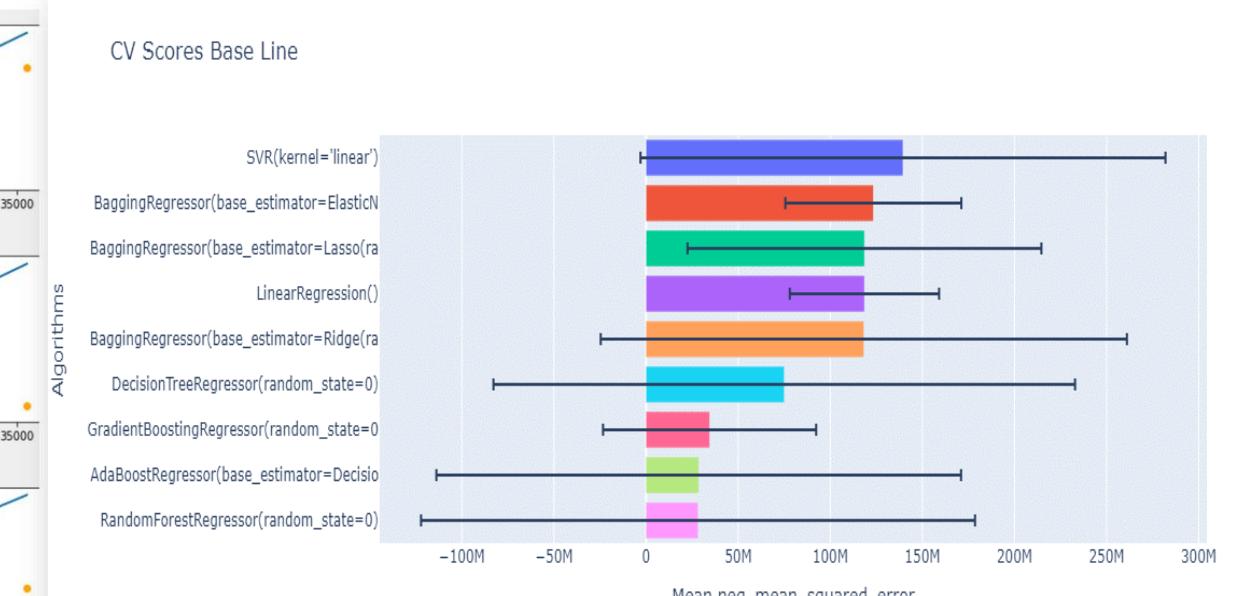
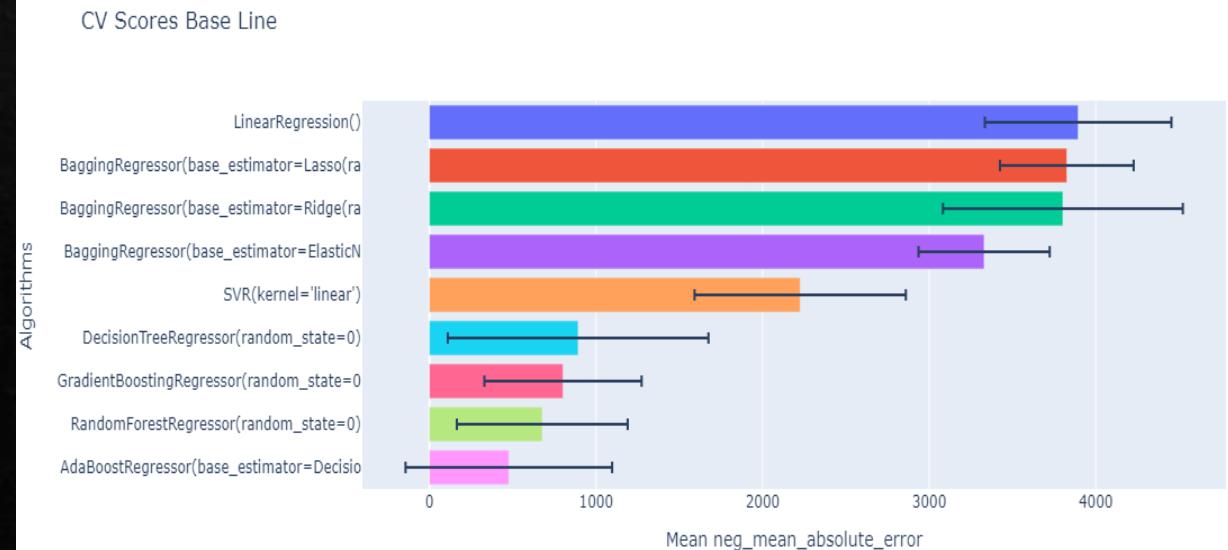
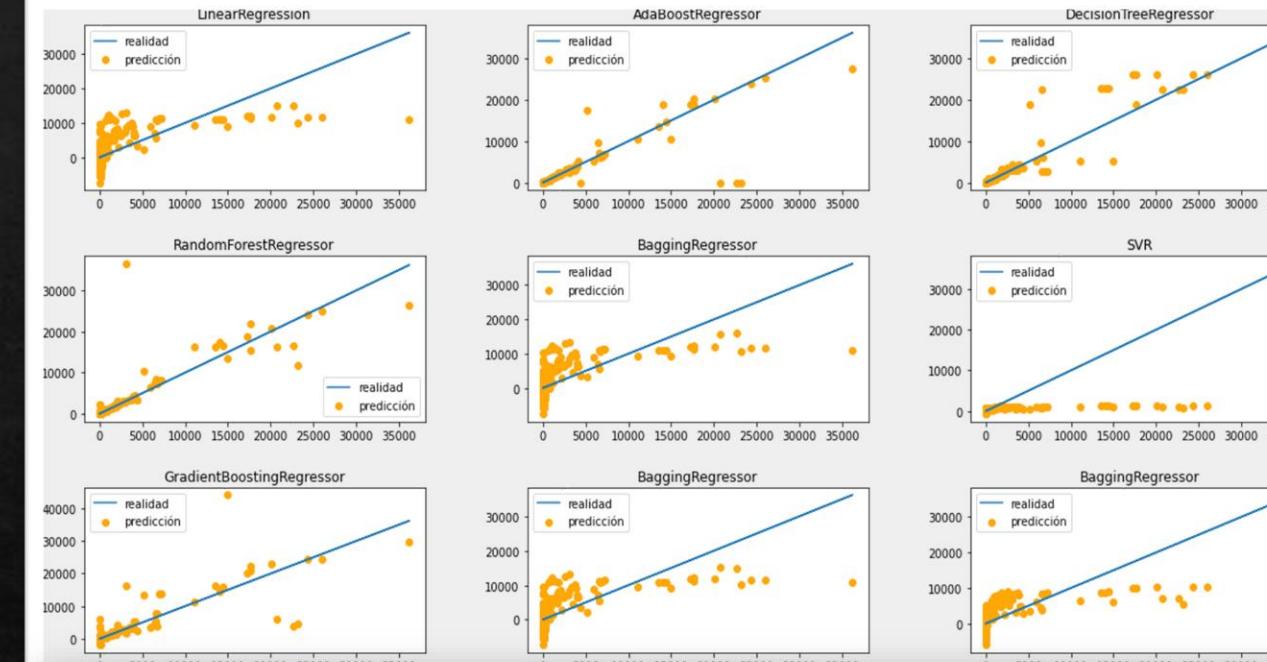
6 variables-correlación(mismas rfecv):

- datos ligeramente superiores al anterior pero reducción de complejidad significativa
- aic y bic sean de -98,24 y -80,18 comparado con -150 y -117. conservamos buena capacidad explicativa general y representa bien los datos concretos.
- MAE,MSE Y RMSE parecidos al modelo de todas variables y mejor que el vif-5vars



VEMOS QUE ENTRE LOS DOS EXTREMOS ES VIABLE CONSEGUIR UNA REDUCCIÓN DE LA COMPLEJIDAD DEL MODELO IMPORTANTE MANTENIENDO UNA BUENA CAPACIDAD EXPLICATIVA DE LA TARGET Y QUE PUEDA GENERALIZAR BIEN PARA NUEVOS DATOS

# Modelo por Cluster - Baseline Cluster 0



# Modelo Individual-Cluster 0-Selección modelo y variables

Hemos decidido que vamos a trabajar con el AdaBoostRegressor por las razones expuestas anteriormente y por tanto, vamos ahora a decidir las variables que se usarán. Para ello probamos con las obtenidas mediante RFECV, Correlación, Statsmodels y VIF y los mejores resultados los obtenemos con:

ADABOOSTREGRESSOR

BASE-ESTIMATOR: DECISIONTREEREGRESSOR

VARIABLES:

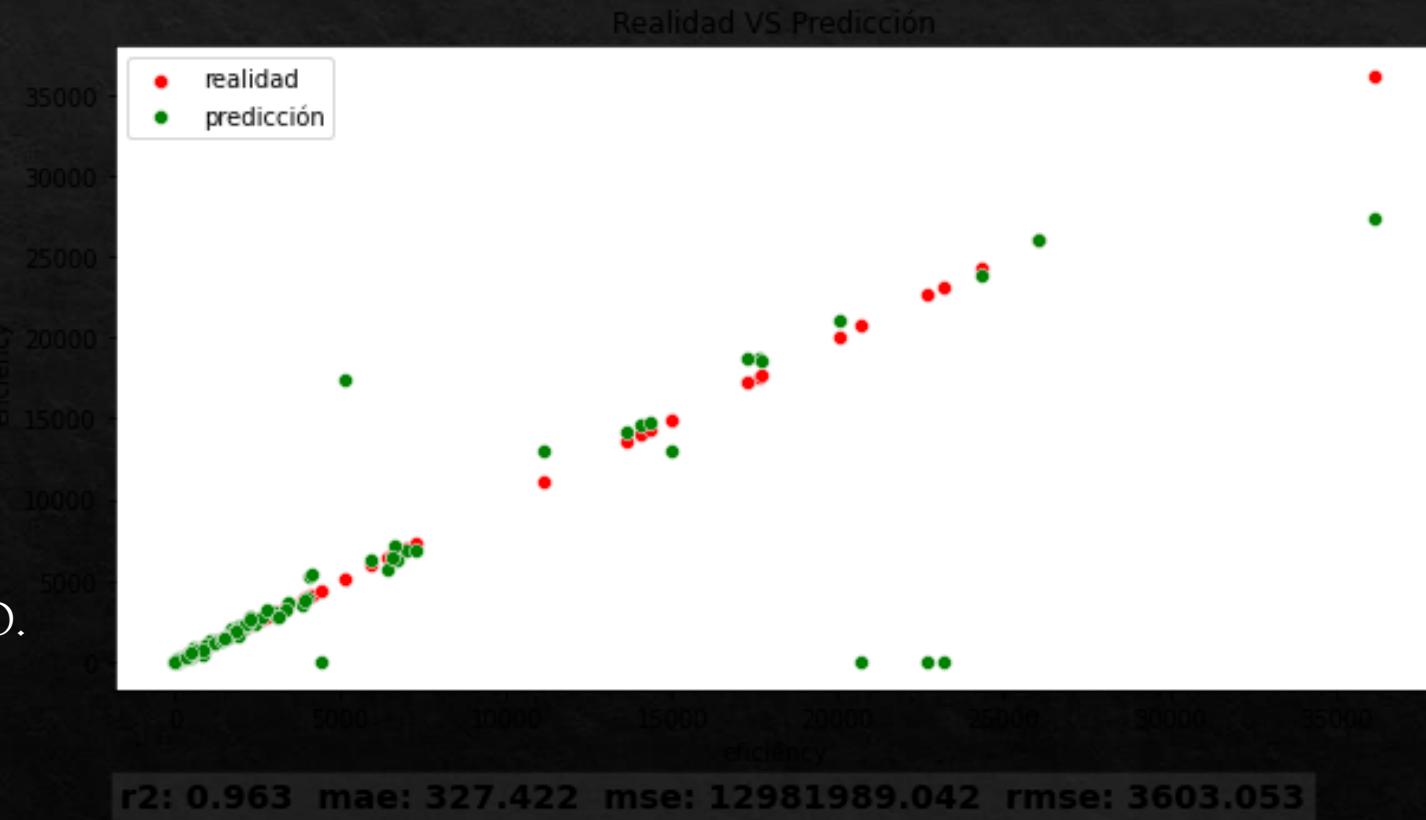
ENERGY PRODUCTION

ENERGY CONSUMPTION

CO2 EMISSION

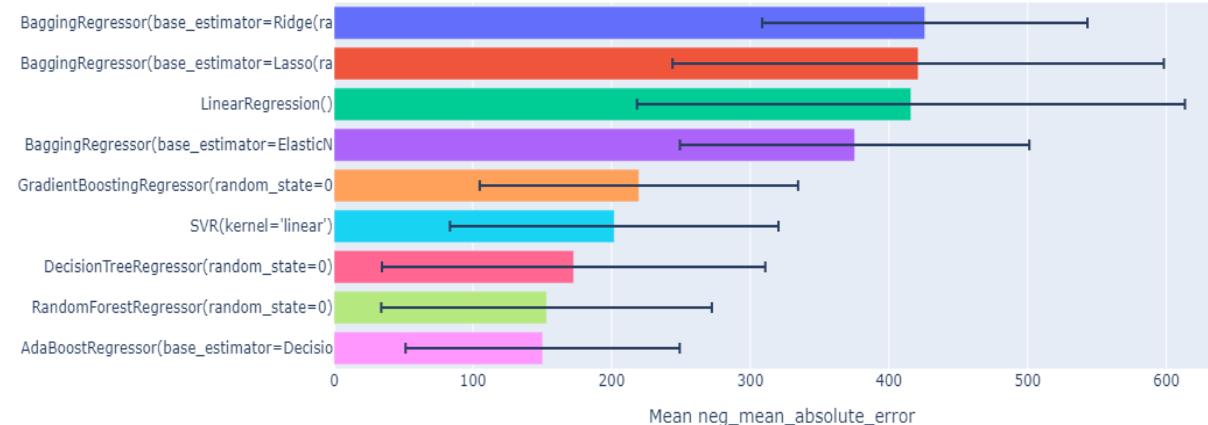
BALANCE

VEMOS QUE CON 4 VARIABLES HEMOS  
CONSEGUIDO UN BUEN AJUSTE DEL MODELO.

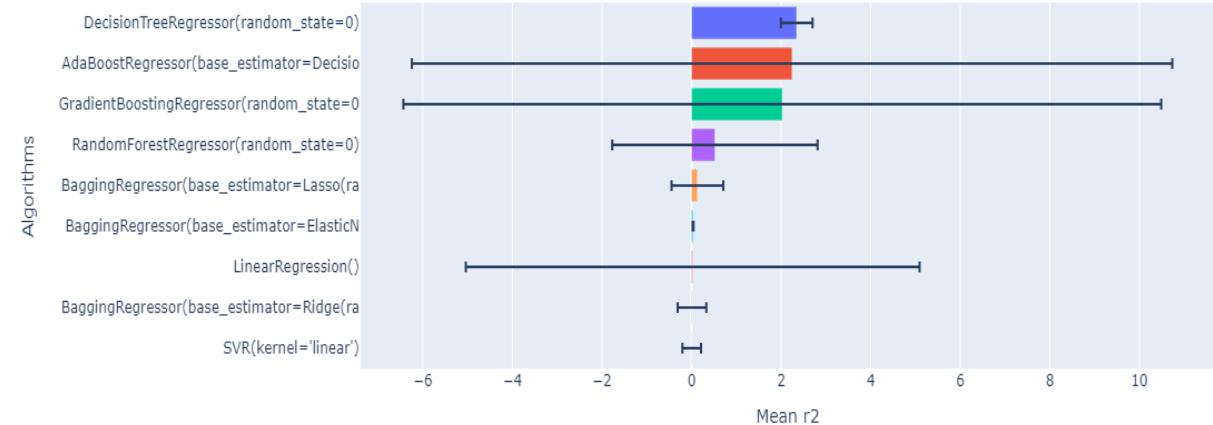


# Modelo por Cluster - Baseline Cluster 1

CV Scores Base Line



CV Scores Base Line

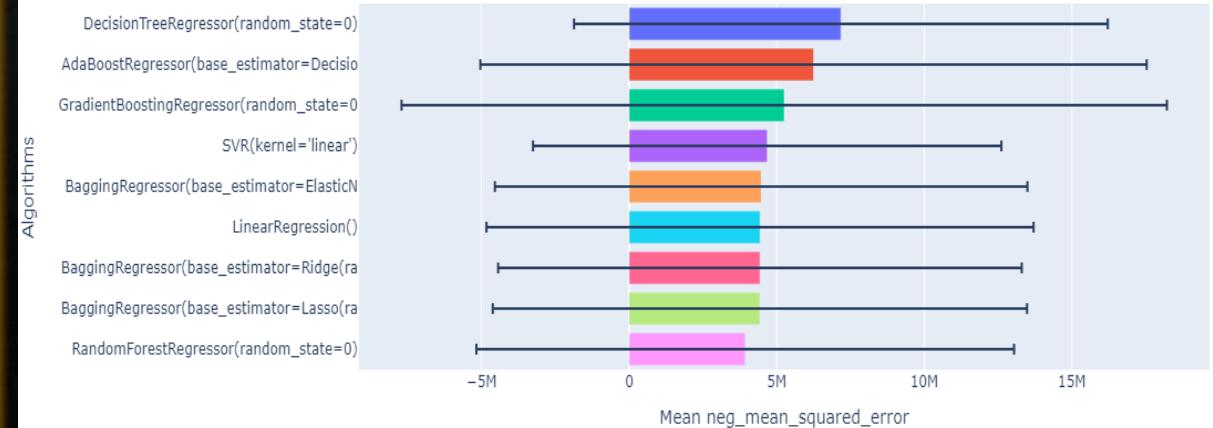


En este cluster los ajustes de todos los modelos eran muy malos con Power Transformer por lo que probamos todos los modelos con diferentes scalers. El que mejores datos nos dio fue Robust Scaler sin eliminar los valores extremos y en un reducido grupo de estimadores, que fueron:

- \* AdaBoost-DecisionTree
- \* DecisionTree
- \* RandomForest
- \* GradientBoostingRegressor

Por tanto la decisión de estimador, quedará entre ellos.

CV Scores Base Line



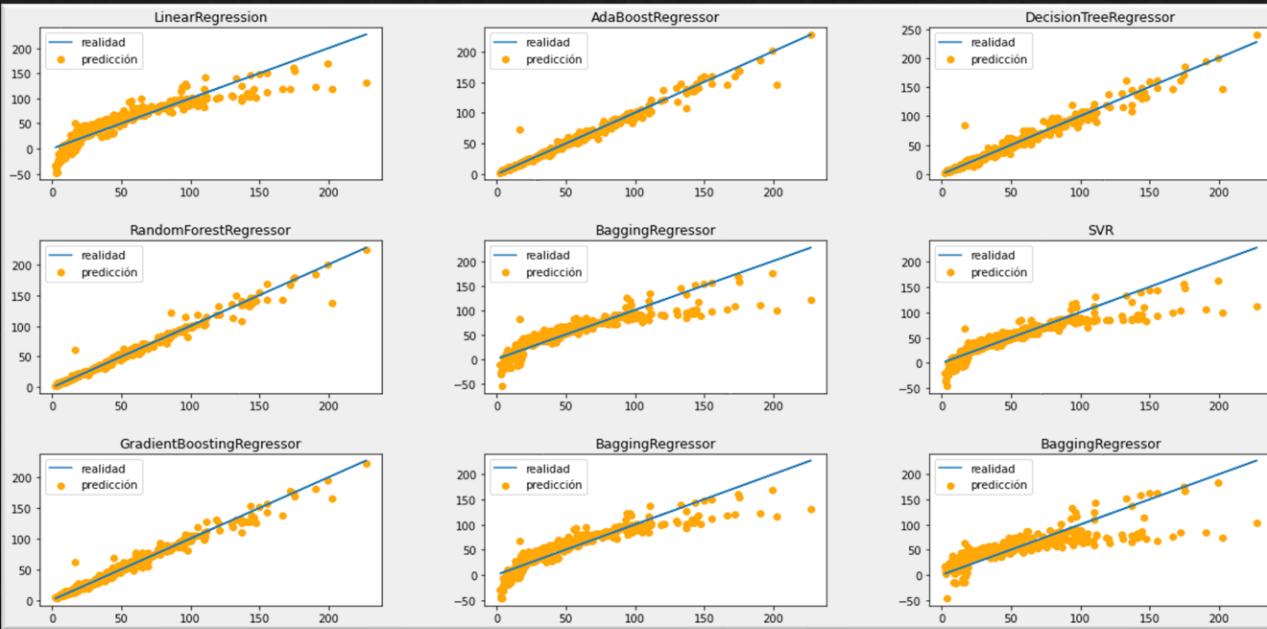
# Modelo Individual-Cluster 1-Selección modelo y variables

estimador	R2	Mae	Rmse	n_variables
DecisionTree-vif	0.917	46.46	1126.91	11
ADA-Ols	0.929	44.01	1100.19	11
DecisionTree-rfecv	0.925	44.73	1110.34	4

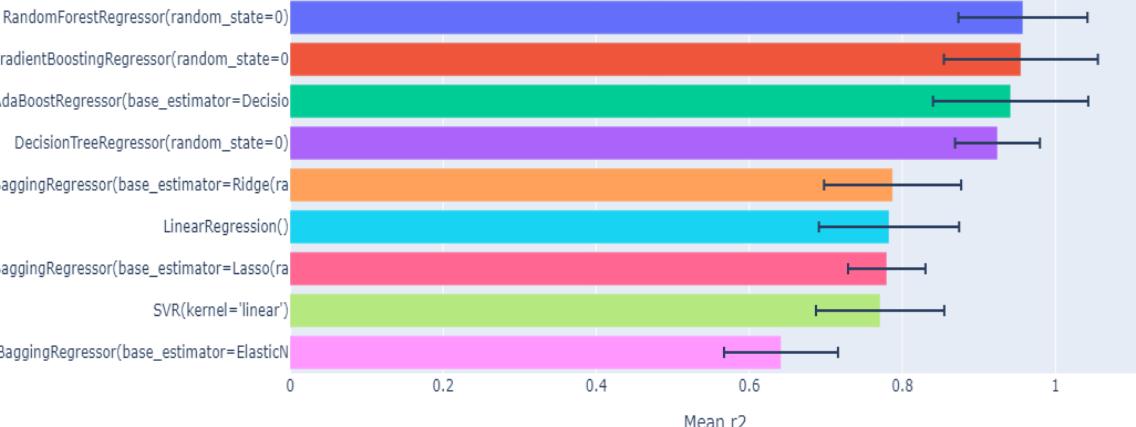
```
['CO2_emission', 'co2_pc', 'Energy_consumption', 'Energy_intensity_by_GDP']
```

Al final se elige DecisionTree con la selección mediante Recursive Feature Elimination Cross Validation que ofrece un punto intermedio entre todas las métricas con una simplificación del modelo muy elevada. OFRECE LA MEJOR RELACIÓN SIMPLICIDAD-MÉTRICAS

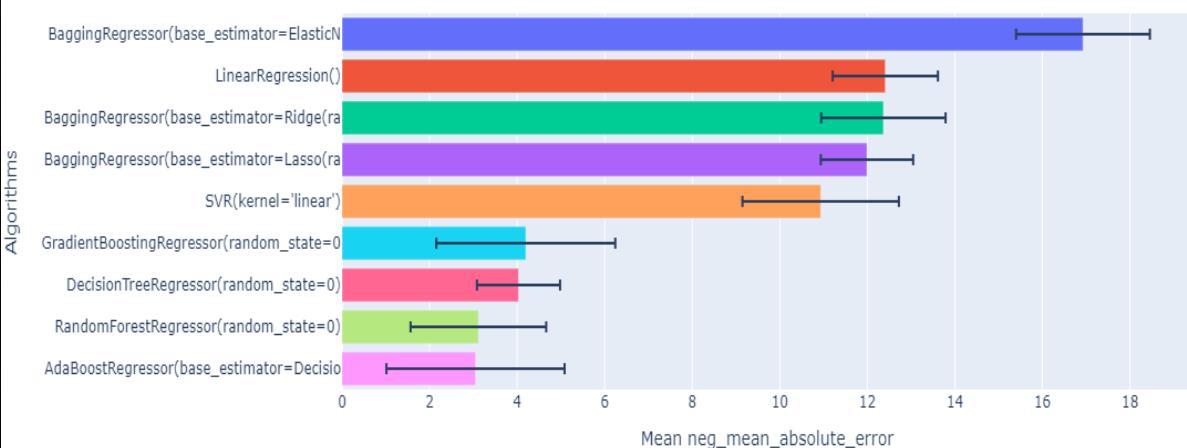
# Modelo por Cluster - Baseline Cluster 2



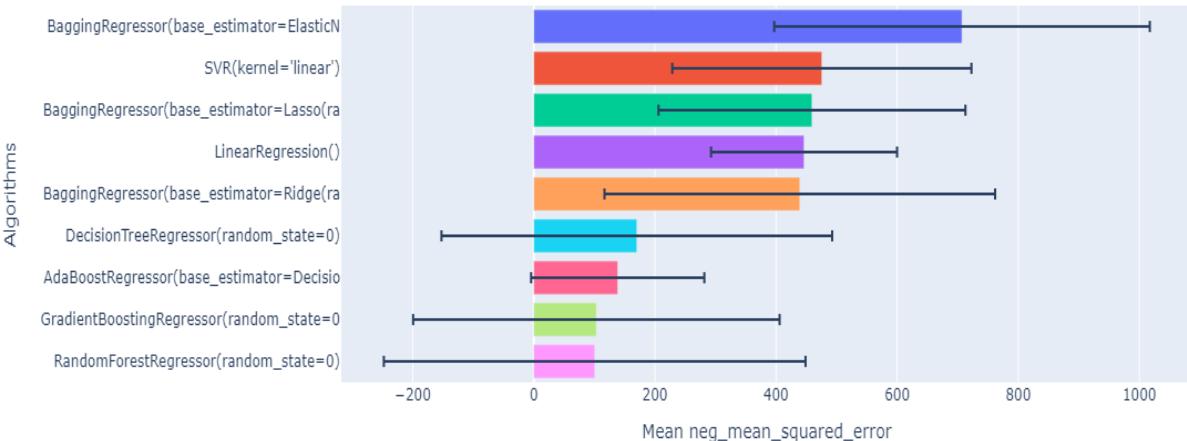
CV Scores Base Line



CV Scores Base Line

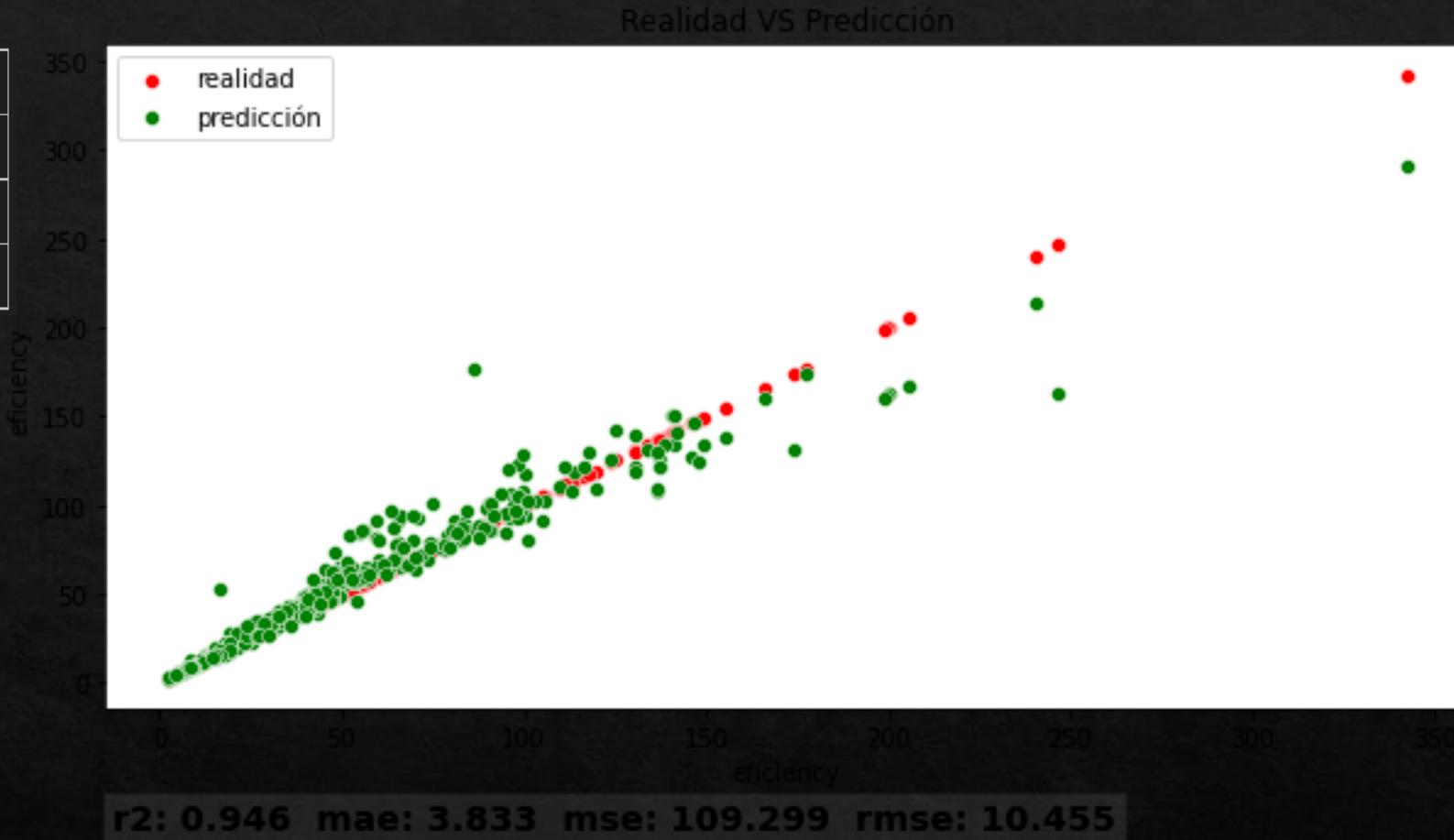


CV Scores Base Line



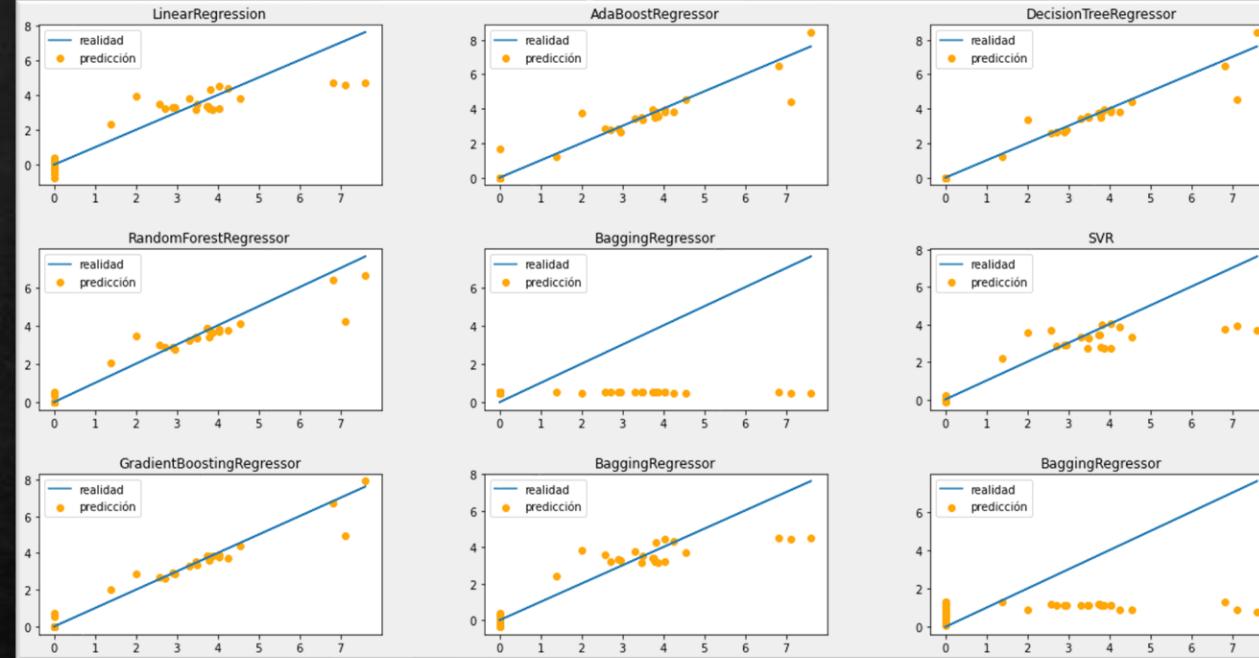
# Modelo Individual-Cluster 2-Selección modelo y variables

estimador	R2	Mae	Rmse	n_variables
RandomForest-Correlación	0.946	3.833	10.455	5
RandomForest-Vif	0.999	0.171	1.697	8
RandomForest-Rfecv	0.966	2.817	8.361	6

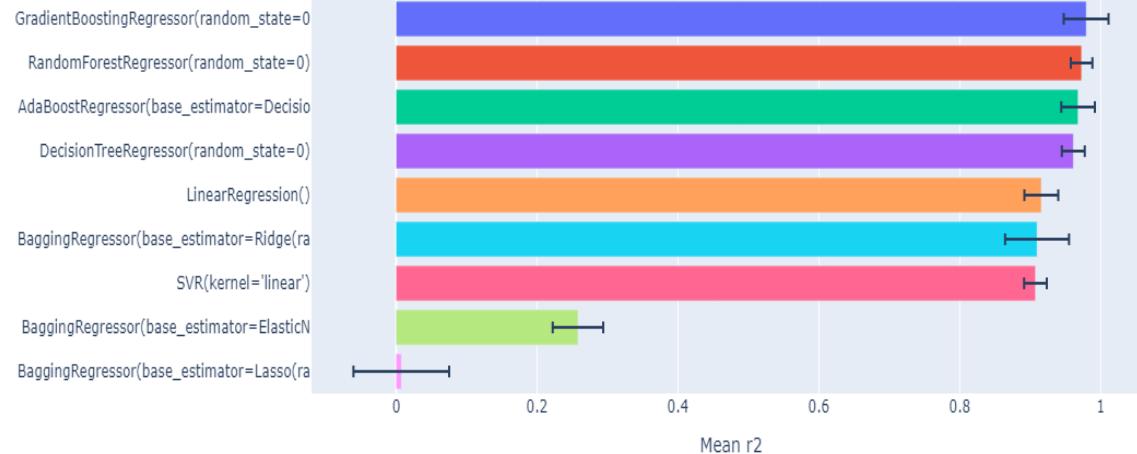


Vemos que el modelo se ajusta bastante bien a la realidad y que aunque no tiene los mejores valores en las métricas, los consideramos suficientes dada la mayor simplicidad que ofrece.

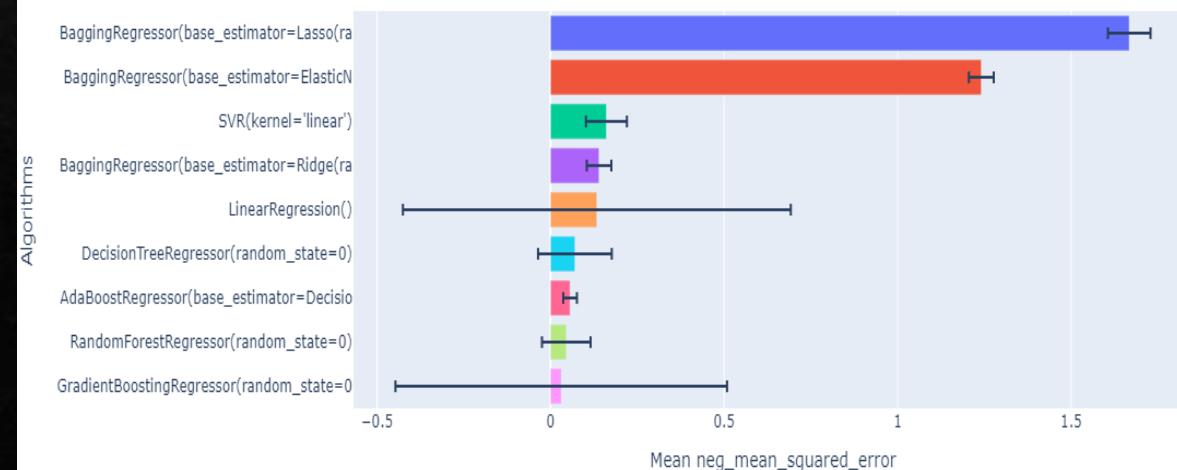
# Modelo por Cluster - Baseline Cluster 3



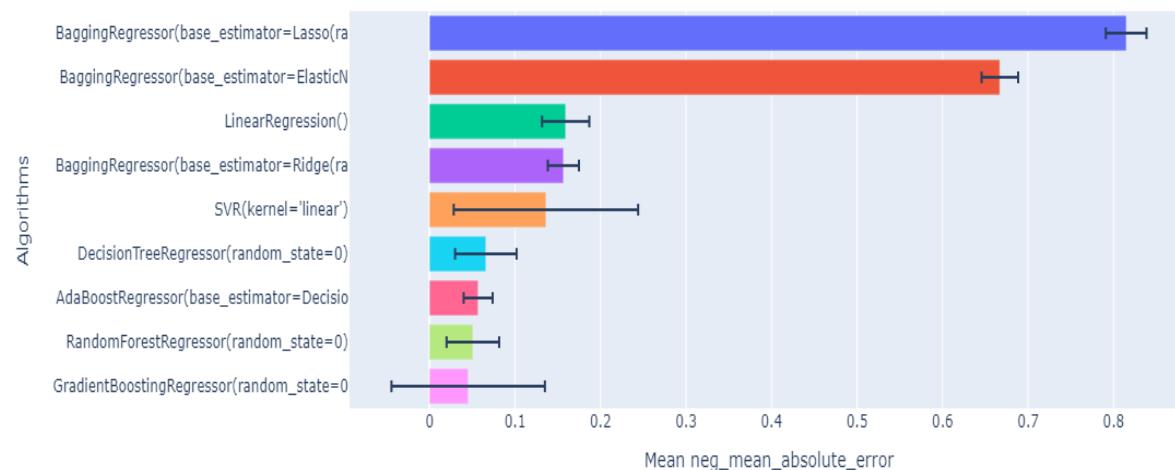
CV Scores Base Line



CV Scores Base Line

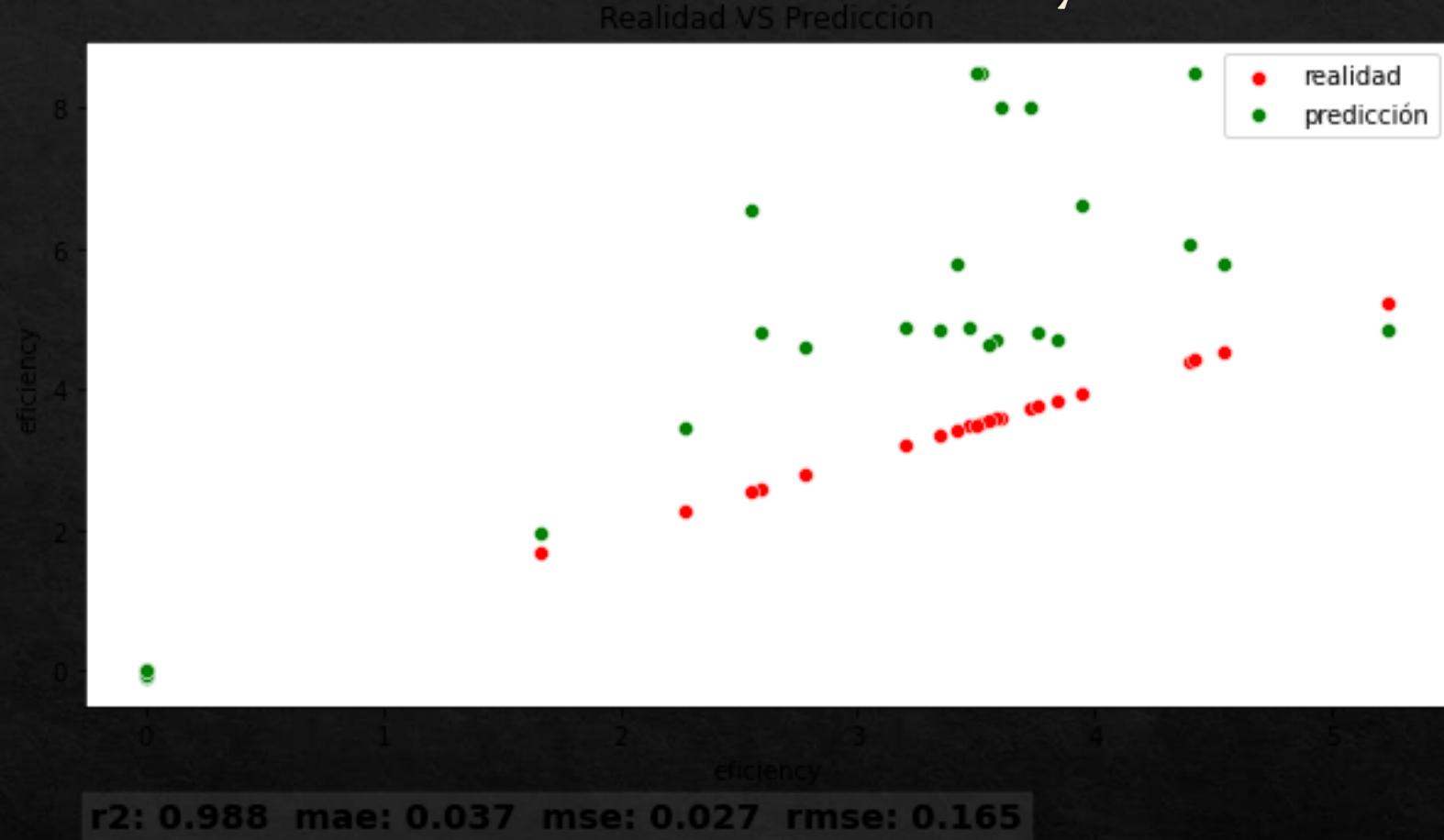


CV Scores Base Line



# Modelo Individual-Cluster 3-Selección modelo y variables

Las diferencias en métricas son mínimas entre ellos por lo que hemos optado por el que ofrece menor complejidad y por tanto por GradientBoostingRegressor con la selección de variables mediante RFECV

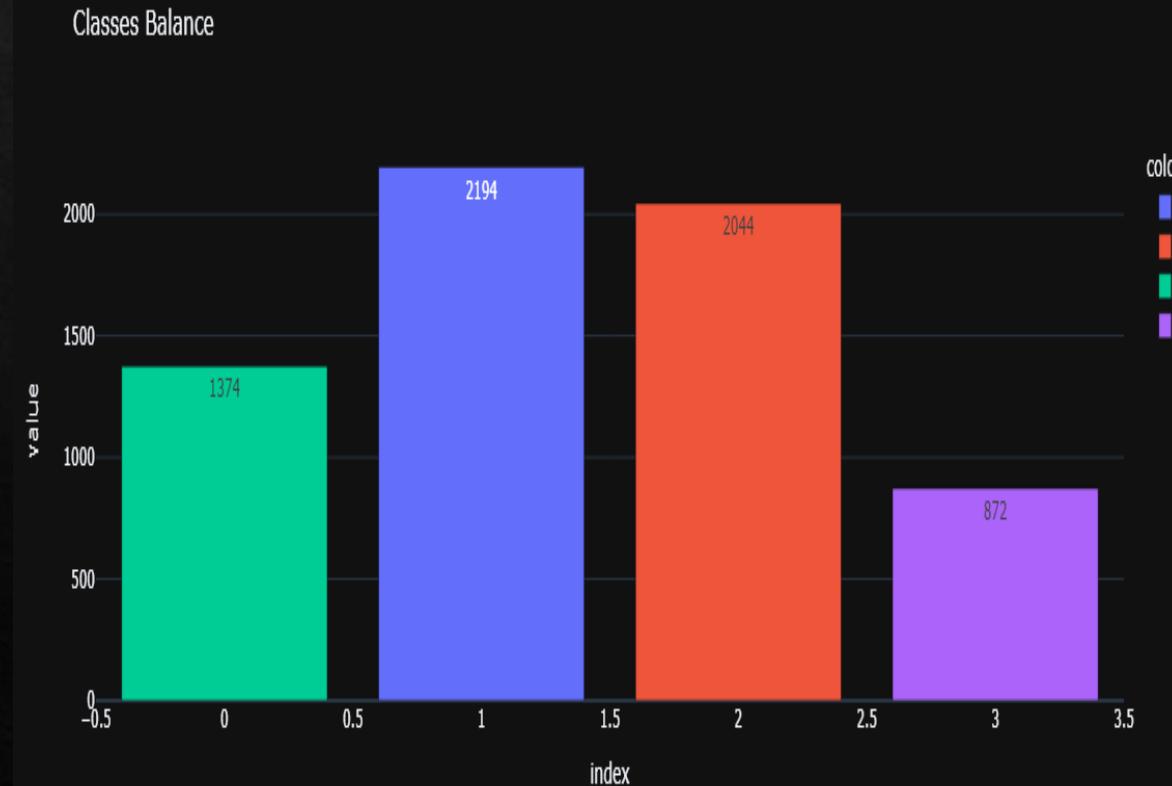
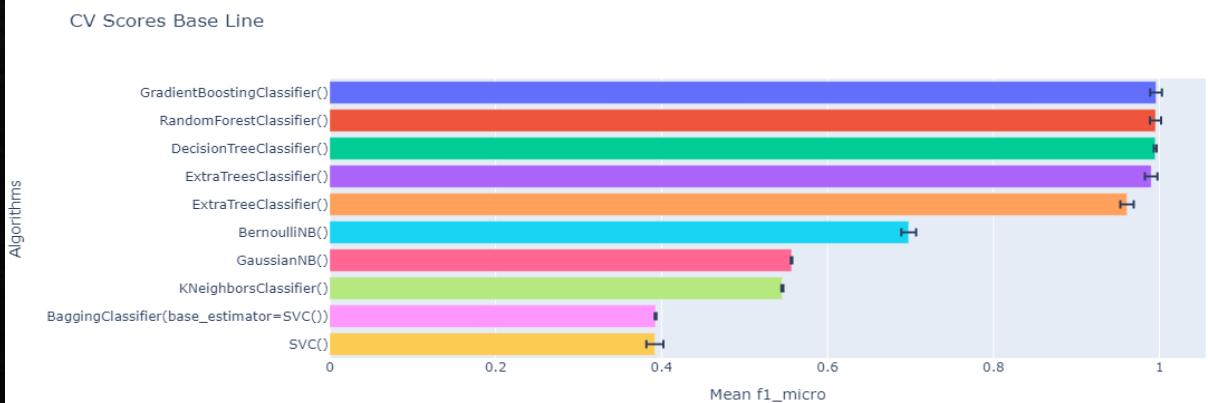
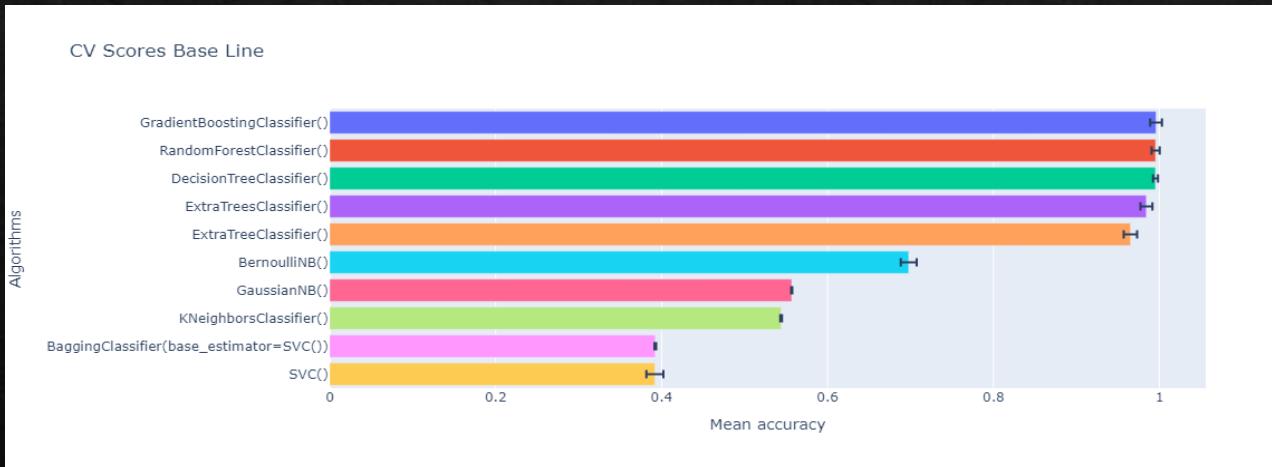


estimador	R2	Mae	Rmse	n_variables
GradientBoostingRegressor-Correlación	0.986	0.041	0.173	5
RandomForest-Vif	0.997	0.011	0.081	9
GradientBoostingRegressor-Rfecv	0.988	0.037	0.165	4
GradientBoostingRegressor-OLS	0.998	0.009	0.06	12

# Clasificación - Parte 1

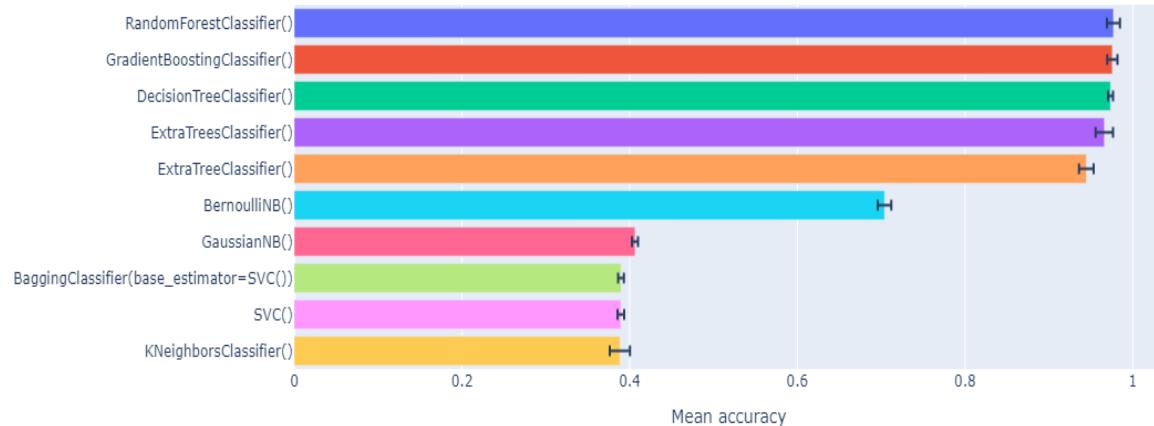
Consideraciones iniciales:

1. Dados los buenos resultados obtenidos por RFECV en los modelos anteriores, solo se usará este método de selección.
2. No se usarán las variables C02\_emission ni Energy\_production ya que los clusters están realizados en base a ellos y los modelos las usan para obtener valores de predicción casi perfectos.
3. La variable eficiency tampoco será usada al ser la división entre las emisiones de co2 y la producción de energía.

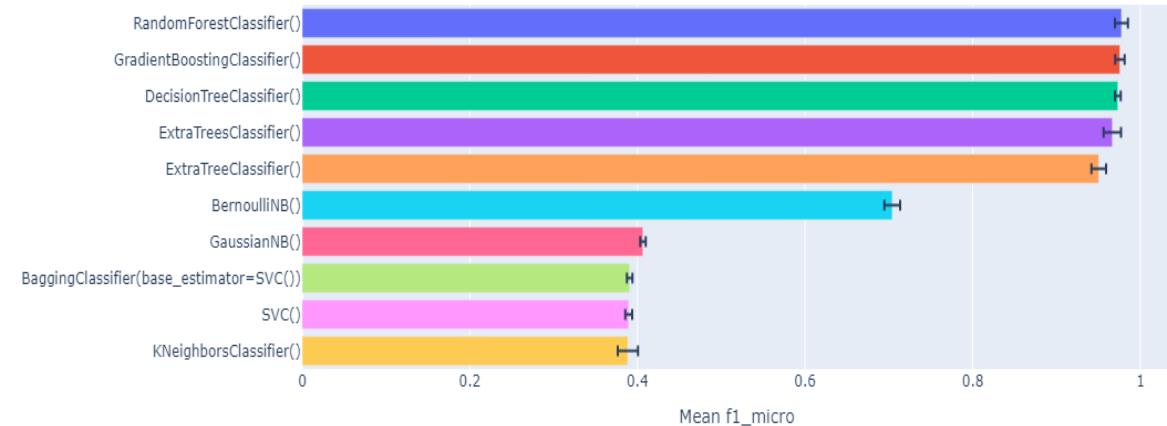


# Clasificación - Parte 2 - Baseline

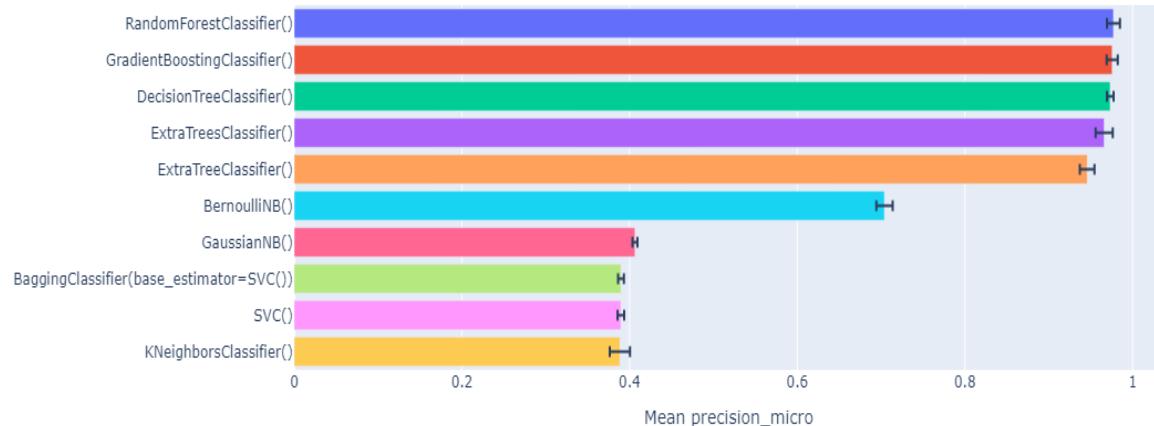
CV Scores Base Line



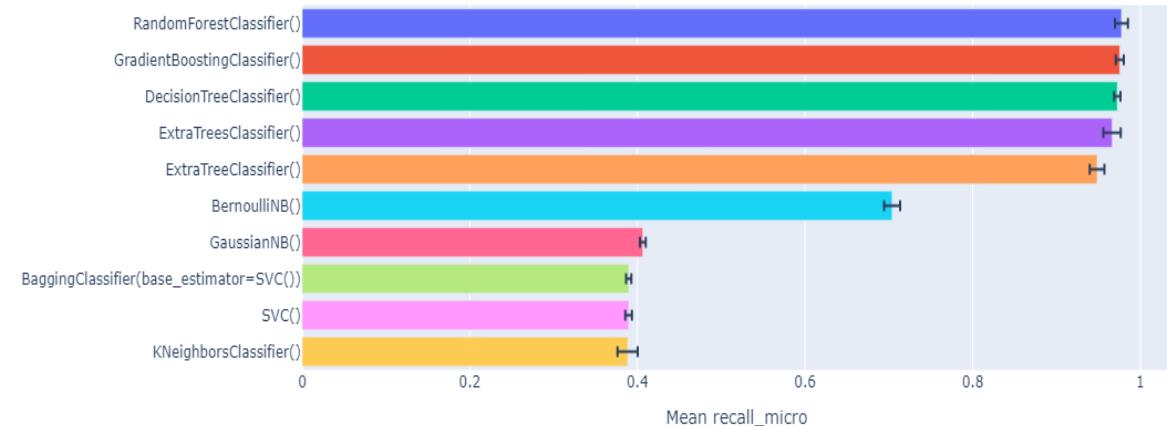
CV Scores Base Line



CV Scores Base Line

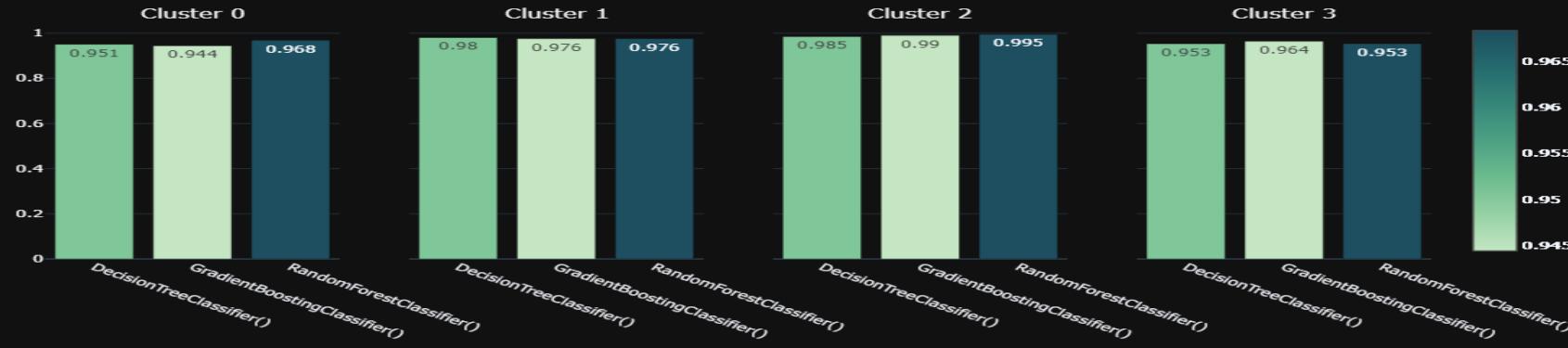


CV Scores Base Line



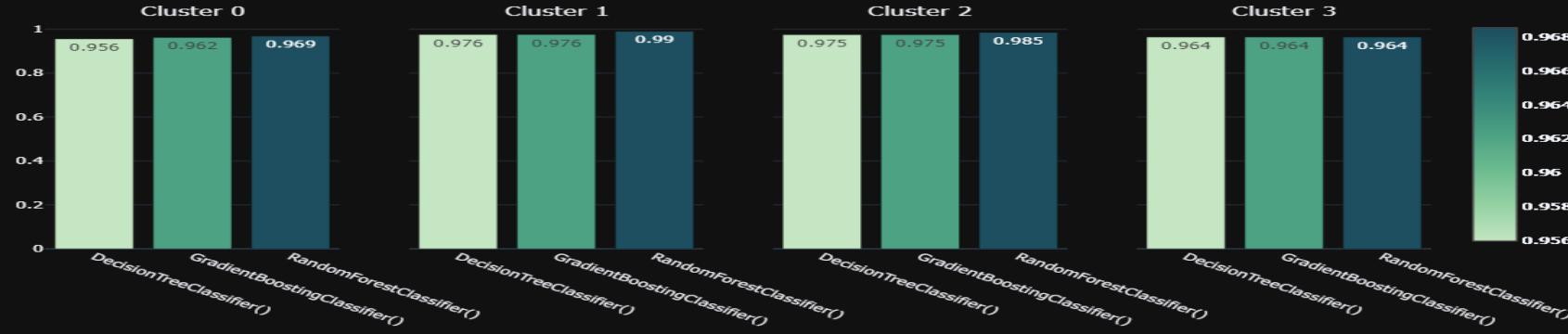
# Clasificación - Métricas Individualizadas

Mean Precision Score for 10 folds



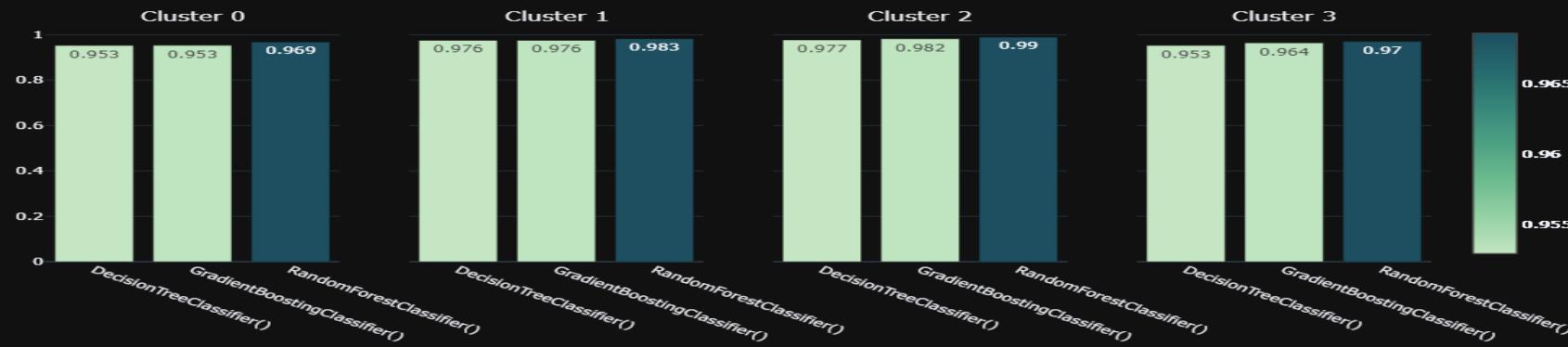
De la predicción del cluster 0, el modelo que más acierta es RF. Para el 1 DT, para el 2 RF y para el 3 GB.

Mean Recall Score for 10 folds



De los casos que son cluster 0, el modelo que mejor los identifica es RF. Para el 1 RF, para el 2 RF y para el 3 todos por igual.

Mean F1 Score for 10 folds



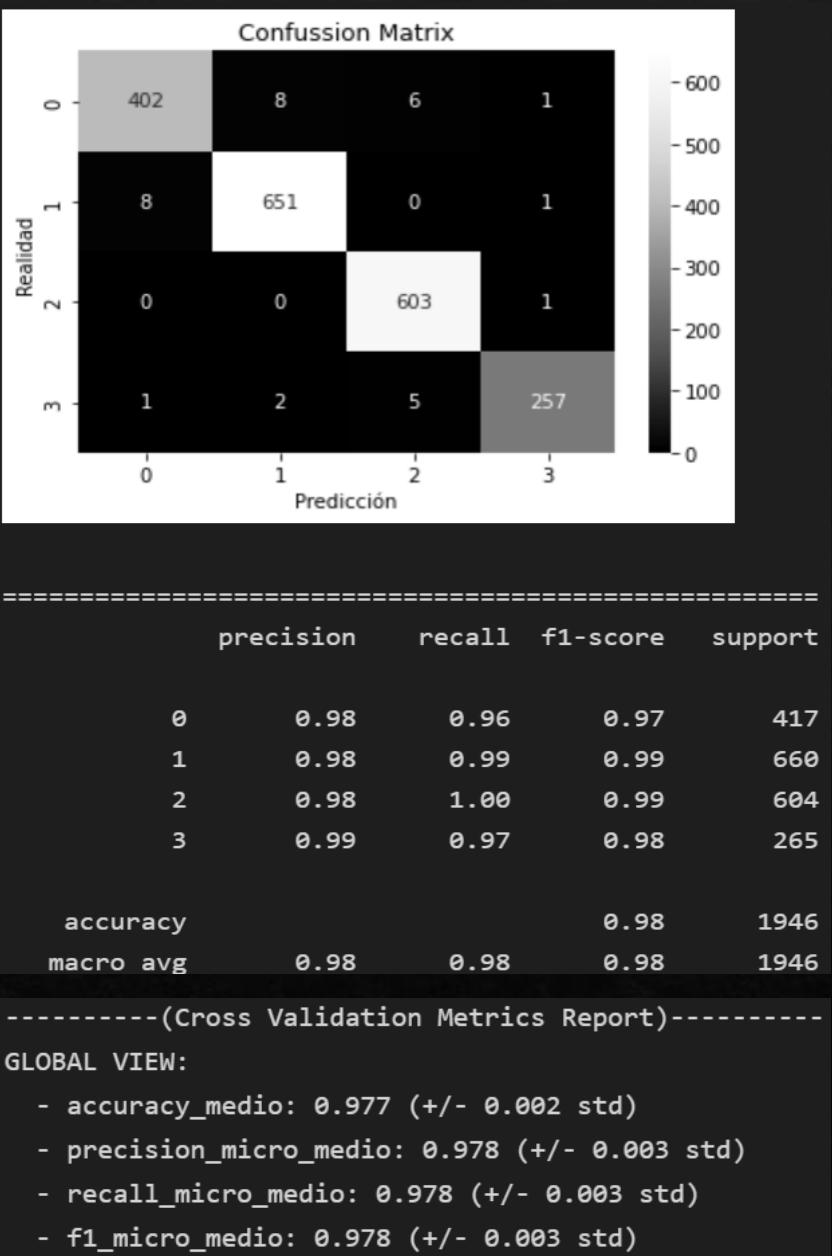
A efectos comparativos por cluster de las dos métricas en conjunto, Rf tiene el mejor score en todos ellos

# Clasificación-Selección de modelo y variables

Ahora ha cambiado la combinación de variables usadas. Lo que podemos ver es que en general necesitan más variables que antes y que han intentado elegir aquellas que tienen algo que ver con la emisión de co2 o con la producción de energía. 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.



ELECCIÓN :  
RANDOM FOREST

La calidad del modelo es muy buena puesto que del total de predicciones positivas en cada clase acertamos el 97,8% de media.

De la cantidad de positivos reales somos capaces de identificar también el 97,8% de media.

La clase que más confunde es el 0 con el 1 y el 2 y el 3 con el 2

# Producción - WebApp

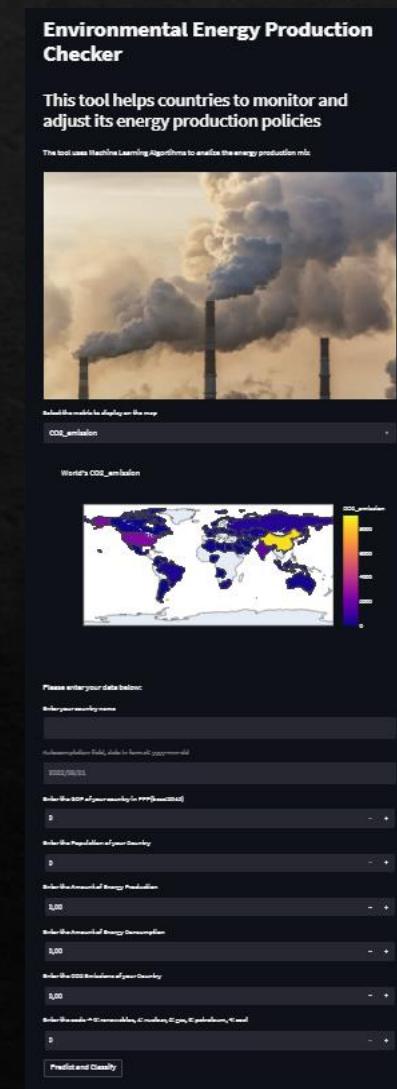
HEROKU - WEBAPP :

The screenshot shows a web browser window with the title "Environmental Country Checker". The URL is <https://co2-ml-app.herokuapp.com/>. The page has a dark blue header with the title "ENERGY PRODUCTION ANALYZER FOR REDUCING CO2 EMISSIONS". Below the header, there is a section titled "Please enter your data below:" with a table-like structure. The columns are "Attribute" and "Value". The attributes listed are: Country Name, yyyy-mm-dd, GDP, Population, Energy Production, Energy Consumption, Co2 Emitted, and a dropdown menu showing options: 0:renewab.,1:nuclear,2:gas,3:petrol,4:coal. A pink "Predict" button is at the bottom. The background of the page is white.

URL: <https://co2-ml-app.herokuapp.com/>

REpositorio GitHub:  
<https://github.com/fersaol/heroku-flask-webapp2>

STREAMLIT - WEBAPP :



URL: <https://fersaol-co2-streamlit-webapp-app-uh1z3w.streamlitapp.com/>

REPOSITORIO GITHUB:  
<https://github.com/fersaol/co2-streamlit-webapp>

GRACIAS POR LA ATENCIÓN

