Aplicación de Modelos de Machine Learning e Inteligencia Artificial a datos macroeconómicos de Latinoamérica

Miembros del equipo:

Este proyecto será desarrollado por estudiantes del curso Inteligencia Artificial y Machine Learning - septiembre 2023 de la Universidad de Chicago y se listan a continuación:

1. Camilo Andrés Cruz Nieves
2. Álvaro Ortuzar

Descripción del caso

El Fondo Latinoamericano de Reservas (FLAR), es el resultado de un acuerdo financiero regional entre Bolivia, Ecuador, Colombia, Perú y Venezuela que se constituyó con el propósito de que estos países tuvieran una institución financiera propia para facilitar el proceso integración regional y afrontar problemas derivados de desequilibrios externos a sus economías. Con los años y como resultado de su loable gestión su misión se ha extendido a otros países de Latinoamérica contando en este momento con nueve países miembros.

Como parte de los servicios financieros que el FLAR ofrece a sus países miembros corresponde a diferentes líneas de crédito de acuerdo a ciertas condiciones (por ejemplo, apoyo a la balanza de pagos), para poder estudiar y determinar "el comportamiento crediticio" de sus países miembros y de la región inicio la recolección, estandarización y publicación de los datos macroeconómicos mediante el Sistema de Información Económica SIE.

Es de gran interés tanto para la academia como para la operación del FLAR obtener insights basado en estos datos, sin embargo, a la fecha no se ha establecido un proyecto que implique el uso de machine learning o inteligencia artificial sobre estos datos, por lo cual es de interés para la organización poder experimentar y obtener conclusiones de los datos SIE.

En síntesis, contamos con alrededor de 90.000 valores observados de un indicador (dinero, porcentaje o índice), por país y por una fecha determinada (series de tiempo con datos periódicos según el indicador).

Proponemos realizar una exploración inicial de estos datos para determinar si es posible con esta información lograr alguno de estos dos posibles casos:

1. Agrupar los países de Latinoamérica, de acuerdo a su información macroeconómica y nos permita encontrar clúster de "países similares" que para el FLAR puedan tener prioridad o no en el proceso de membresía. (algoritmo de clasificación y clustering).
2. Predecir los próximos valores de cada indicador según los datos históricos que le permita al FLAR anticiparse a posibles situaciones de urgencia que puede caer un país y que requiera el acompañamiento del FLAR.
3. Algún otro resultado de impacto para el FLAR y que sea obtenido mediante machine learning e inteligencia artificial que encontremos en la exploración inicial de los datos.

Propuesta esquemática del trabajo

Para realizar el proyecto propuesto se han definido cinco fases que nos permitan seleccionar el caso, modelar, obtener resultados y documentar de la siguiente manera: Fase 1: Diseño del Proyecto, Fase 2: Machine Learning en Python, Fase 3: Documentación y Presentación, Fase 4: Evaluación y Revisión, Fase 5: Retroalimentación y Mejoras (Opcional).

Fase 1: Diseño del Proyecto

Recopilación de Datos y Exploración Inicial:

- Obtener datos del Sistema de Información Económica del FLAR (SIE).

El Fondo Latinoamericano de Reservas (FLAR), es el resultado de un acuerdo financiero regional entre Bolivia, Ecuador, Colombia, Perú y Venezuela que se constituyó con el propósito de que estos países tuvieran una institución financiera propia para facilitar el proceso integración regional y afrontar problemas derivados de desequilibrios externos a sus economías. Con los años y como resultado de su loable gestión su misión se ha extendido a otros países de Latinoamérica contando en este momento con nueve países miembros.

Así pues, tenemos dos grupos de paises divididos entre pertenecientes del Flar y no perteneciente. Describo a continuacion la division de los grupos:

- Países que pertenecen al FLAR (miembros y asociados):

['Bolivia', 'Colombia', 'Costa Rica', 'Chile', 'Ecuador', 'Paraguay', 'Peru', 'Uruguay', 'Venezuela']

- Países que no pertenecen al FLAR:

['Mexico', 'Argentina', 'Brazil', 'Honduras', 'Panama', 'El Salvador', 'Jamaica', 'Nicaragua', 'Guatemala', 'Republica dominicana']

En este apartado, se describe el proceso de obtención de datos del Sistema de Información Económica del Fondo Latinoamericano de Reservas (FLAR), conocido como SIE. Se destaca la importancia de recopilar, estandarizar y publicar datos macroeconómicos para comprender el comportamiento crediticio de los países miembros del FLAR. La obtención de estos datos es crucial para llevar a cabo análisis detallados y proporcionar información valiosa que pueda influir en la toma de decisiones y estrategias del FLAR.

- Realizar una exploración inicial para comprender la estructura y contenido de los datos.

En este apartado se realiza el analisis estadístico de datos para determinar las variables, llenado de vacíos y preprocesamiento de las dimensiones para el algoritmo de clasificación.

El propósito de nuestro trabajo es usar los datos del SIE y tulizar los indicadores (series de tiempo) de los 19 paises y clasificar los paises FLAR y no FLAR (es decir paises que son miembros o asociados y sus vecinos). Esto permitirá al Fondo aunar esfuerzo en mercadeo y oferta de valor a esos vecinos y priorizar. De este modo tenemos lo siguiente:

- Variables elementales:

Index (['serieID', 'refAreaID', 'refAreaName', 'indicatorID', 'indicatorNameEN',
'indicatorNameES', 'indicatorNameENShort', 'indicatorNameESShort',
'dataDomainID', 'categoryID', 'freqID', 'frequencyName', 'unitID',
'unitName', 'unitMultID', 'unitMultValue', 'timeFormatID', 'obsValue',
'timePeriod'].

	serieID	refAreaID	refAreaName	indicatorID	indicatorNameEN	indicatorNameES	indicatorNameENShort	indicatorNameESShort	dataDomainID	categoryID	...	unitName	unitMultID	unitMultValue	timeFormatID	MIEMBRO_FLAR	timeFormatName	timePeriod	Fecha_Estructurada	obsValue	obsValue_real
0	EC-NGDP_PA_XDC-Q	EC	Ecuador	NGDP_PA_XDC	National Accounts, Gross Domestic Product, Pro...	pib nominal, moneda nacional	Nominal Gross Domestic Product, National Currency	Producto interno bruto (PIB) nominal, moneda n...	NAG	1	...	US dollar	6	Millions	P3M	1	Quarterly	2000-Q1	2000-03-01	3819.126	3819126000
1	EC-NGDP_PA_XDC-Q	EC	Ecuador	NGDP_PA_XDC	National Accounts, Gross Domestic Product, Pro...	pib nominal, moneda nacional	Nominal Gross Domestic Product, National Currency	Producto interno bruto (PIB) nominal, moneda n...	NAG	1	...	US dollar	6	Millions	P3M	1	Quarterly	2000-Q2	2000-06-01	4402.479	4402479000
2	EC-NGDP_PA_XDC-Q	EC	Ecuador	NGDP_PA_XDC	National Accounts, Gross Domestic Product, Pro...	pib nominal, moneda nacional	Nominal Gross Domestic Product, National Currency	Producto interno bruto (PIB) nominal, moneda n...	NAG	1	...	US dollar	6	Millions	P3M	1	Quarterly	2000-Q3	2000-09-01	4906.653	4906653000
3	EC-NGDP_PA_XDC-Q	EC	Ecuador	NGDP_PA_XDC	National Accounts, Gross Domestic Product, Pro...	pib nominal, moneda nacional	Nominal Gross Domestic Product, National Currency	Producto interno bruto (PIB) nominal, moneda n...	NAG	1	...	US dollar	6	Millions	P3M	1	Quarterly	2000-Q4	2000-12-01	5190.343	5190343000
4	EC-NGDP_PA_XDC-Q	EC	Ecuador	NGDP_PA_XDC	National Accounts, Gross Domestic Product, Pro...	pib nominal, moneda nacional	Nominal Gross Domestic Product, National Currency	Producto interno bruto (PIB) nominal, moneda n...	NAG	1	...	US dollar	6	Millions	P3M	1	Quarterly	2001-Q1	2001-03-01	5904.082	5904082000

5 rows × 23 columns

Sin embargo, las columnas a escalar por medio de transformar la raw data en dimensiones para la clasificación, son lea siguientes:

- Transformación clasificada:

Index(['BCA_BP6_USD', 'BIP_BP6_USD', 'BIS_BP6_USD', 'BKF_BP6_USD',
'BMG_BP6_USD', 'BS_BP6_USD', 'BXG_BP6_USD', 'CG_DD_XDC', 'CG_DE_USD',
'CG_DE_XDC', 'CG_DT_USD', 'CG_DT_XDC', 'CG_GOB_XDC', 'CG_GPB_XDC',
'DS_USD', 'DS_XDC', 'D_USD', 'EA_IX', 'ENDA_XDC_USD_RATE',
'ENDE_XDC_USD_RATE', 'FASMB_XDC', 'FID_FX_USD', 'FID_FX_XDC',
'FID_NC_USD', 'FID_NC_XDC', 'FM2_XDC', 'FPOLM_PA', 'FSANL_PT',
'FSKERA_PT', 'FSNFC_USD', 'FSNFC_XDC', 'FSNNC_USD', 'FSNNC_XDC',
'LER_PT', 'LEUR_PT', 'LLF_PE_NUM', 'LUR_PT', 'NGDP_PA_R_XDC',
'NGDP_PA_USD', 'NGDP_PA_XDC', 'PCPI_IX', 'RAFA_USD', 'TMG_CIF_USD',
'TTT_BY_CP_A_IX', 'TXG_FOB_USD'].

Indicator-fecha	refAreaID	refAreaName	MIEMBRO_FLAR	BCA_BP6_USD 1980-03-01	BCA_BP6_USD 1980-06-01	BCA_BP6_USD 1980-09-01	BCA_BP6_USD 1980-12-01	BCA_BP6_USD 1981-03-01	BCA_BP6_USD 1981-06-01	BCA_BP6_USD 1981-09-01	...	TXG_FOB_USD 2023-01-01	TXG_FOB_USD 2023-02-01	TXG_FOB_USD 2023-03-01	TXG_FOB_USD 2023-04-01	TXG_FOB_USD 2023-05-01	TXG_FOB_USD 2023-06-01	TXG_FOB_USD 2023-07-01	TXG_FOB_USD 2023-08-01	TXG_FOB_USD 2023-09-01	TXG_FOB_USD 2023-10-01
0	AR	Argentina	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	4.925331e+09	5.239194e+09	5.734575e+09	5.898250e+09	6.261858e+09	5.414915e+09	6.060343e+09	5.910266e+09	5.751000e+09	NaN
1	BO	Bolivia	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	8.447497e+08	7.702248e+08	9.948223e+08	9.918004e+08	9.501458e+08	8.898648e+08	NaN	NaN	NaN	NaN
2	BR	Brazil	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	2.270570e+10	2.102480e+10	3.274670e+10	2.784390e+10	3.254200e+10	2.955240e+10	2.826600e+10	3.119320e+10	2.840630e+10	NaN
3	CL	Chile	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	8.979753e+09	8.314668e+09	9.813952e+09	7.609448e+09	7.836360e+09	7.939265e+09	7.393661e+09	8.063788e+09	7.341750e+09	7.725718e+09
4	CO	Colombia	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	3.696200e+09	4.202200e+09	4.431900e+09	3.739200e+09	4.497900e+09	4.021000e+09	4.133900e+09	3.946900e+09	NaN	NaN

5 rows × 16727 columns

Se tiene en cuenta que se debe eliminar previamente las columnas que tengan un porcentaje de valores nulo mayor a un umbral definido, personalizado además para tres espacios de estudio que vienen determinados por el numero enésimo de dimensiones que van desde 100, 50 y 0 respectivamente.

Se define el código correspondiente a tal efecto:

```
#funcion que  elimina las columnas que tengan un  porcentaje de valores nulo mayor a un umbral definido
def remove_columns_na_p(umbral,df):
    df=df.copy()
    df.drop('refAreaName', axis=1, inplace=True)
    porcentaje_nan = df.isnull().mean()
    # Selecciona las columnas que tienen un porcentaje de NaN menor o igual al umbral
    columnas_a_mantener = porcentaje_nan[porcentaje_nan <= umbral].index
    # Crea un nuevo DataFrame con las columnas seleccionadas
    df_filtrado = df[columnas_a_mantener]
    return df_filtrado
```

```
clafdata_50=remove_columns_na_p(0.5,clafdata)
clafdata_0=remove_columns_na_p(0,clafdata)
clafdata_100=remove_columns_na_p(1,clafdata)
```

```
clafdata_50.shape
```

```
(19, 3624)
```

```
clafdata_0.shape
```

```
(19, 121)
```

```
clafdata_100.shape
```

```
(19, 16726)
```

Posteriormente se sacan los indicadores de estudio, base sobre la cual se emparejan los paises objetivo según una serie de características compartidas, que determinaran las decisiones futuras al ejecutar los objetivos que marcan este trabajo y sobre los que se fundamentan las decisiones tomadas.
Estas son:

- Indicadores macroeconómicos: (series temporales)

    1.PIB (Producto interno bruto) ('NGDP_PA_USD', 'NGDP_PA_XDC')
    2. inflación ('PCPI_IX')
    3. déficit Cuentas corrientes (% del PIB) ('BCA_BP6_USD')
    4. déficit Fiscal (% del PIB) ('CG_GOB_XDC')
    5. Deuda (% del PIB) ('CG_DT_XDC', 'CG_DT_USD')
    6. Tasa Interes Politica Monetaria ('FPOLM_PA')
    7. Agregado monetarios M2 o M3 (%PIB). (FM2_XDC)

indicadores=['NGDP_PA_USD', 'NGDP_PA_XDC','PCPI_IX','BCA_BP6_USD','CG_GOB_XDC','CG_DT_XDC', 'CG_DT_USD','FPOLM_PA','FM2_XDC']

Aquí finaliza la primera parte de preparar los datos para una posible clasificación, pues cada fila corresponde a un país, y cada columna es la combinación de indicador y fecha (lo que nos da un numero o valor observado). Sin embargo, en el siguiente apartado vamos a detallar con más profundidad cada país (por indicador) y cada Indicador (por paises graficando las series de tiempo), dando paso a las series temporales con el potencial de poder realizar un nuevo apartado de interes, basado en el forecasting. Esta sección, aunque no presente en el trabajo por espacio, es fundamental para futuras presentaciones.

Definición de Objetivos:

- **Determinar claramente si el objetivo es clasificar o predecir datos.**

La fase de "Definición de Objetivos" es esencial para orientar el enfoque y los métodos utilizados en un proyecto de Machine Learning. Aquí, se busca establecer de manera clara y precisa lo que se espera lograr con el análisis de datos. Esta fase se compone de dos aspectos clave:

Clasificación: Si el objetivo es clasificar, significa que estamos tratando de asignar una etiqueta o categoría a los datos. Por ejemplo, identificar si un correo electrónico es spam o no.

Predicción: Si el objetivo es predecir, buscamos estimar o anticipar valores futuros basándonos en patrones históricos. Por ejemplo, predecir el precio de las acciones en el próximo mes.

De estas dos modalidades de estudio, lo que es de interes es la predicción. Se usarán variables en series temporales y técnicas de clusterización como técnicas matemáticas y de machine learning. Se probarán varios algoritmos de clasificación y clusterización; que me permita agrupar los paises en paises FLAR, NO FLAR y cuáles son los vecinos más cercanos a FLAR.

Básicamente se utilizarán técnicas de analisis de componentes principales PCA e incrustación de vecinos estocásticos distribuidos o T_SNE, para la visualización de conjuntos de datos de alta dimensionalidad. Se esperan observar distribuciones de nubes de puntos idénticas, fortaleciendo y extrapolando hacia nuevas aplicaciones de estudio.

- **Identificar qué se busca lograr con el análisis.**

Se trata de entender la finalidad o el propósito más amplio del análisis de datos.

Puede ser la toma de decisiones más informada, la identificación de patrones ocultos, la optimización de procesos, entre otros.

La identificación clara de este propósito guiará las etapas posteriores del proyecto, desde la selección de modelos hasta la interpretación de resultados.

Ejemplo Práctico:

Supongamos que estamos trabajando en un proyecto para el Fondo Latinoamericano de Reservas (FLAR) y tenemos datos macroeconómicos. Aquí, podríamos definir nuestros objetivos de la siguiente manera:

- Objetivo Principal: Predecir el comportamiento futuro de indicadores económicos clave (como PIB, inflación, déficit fiscal) para que el FLAR pueda anticiparse a situaciones de urgencia.
- Objetivo Secundario: Identificar patrones o similitudes entre países latinoamericanos basándonos en su información macroeconómica, lo que podría ser útil para priorizar acciones o membresías.

Estos objetivos claros proporcionan una guía para las siguientes fases del proyecto, desde la selección de modelos hasta la presentación de los resultados.

Preparación de Datos:

A tal efecto, se han seguido estos tres pasos hasta crear los documentos en formato tanto Excel con csv adecuados para poder cargarlos por medio de algoritmos de importación de datos. Tambien como dato se han filtrado las advertencias para ejecutar más cómodamente el código y en especial para ignorar varias advertencias específicas de Pandas.

- Limpieza y preprocesamiento de los datos, si es necesario.

Función que elimina las columnas que tengan un porcentaje de valores nulo o mayor a un umbral definido.

- Transformación de datos, como creación de características adicionales.

Función que transforma los datos originales en tipos de datos requeridos.
Función para transformar la raw data en dimensiones para la clasificación.
Creación de un nuevo Dataframe con las columnas seleccionadas.

- Manejo de series de tiempo y frecuencia de datos.

Graficado de series de tiempo para cada indicador y país.
Escalado de valores para solución de problemas de comportamiento relacionado de series temporales en variables.

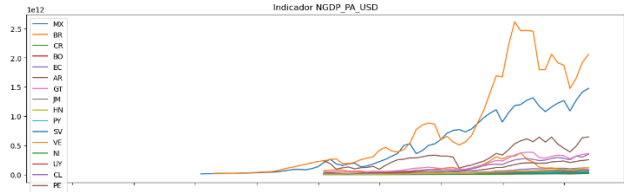
Fase 2: Machine Learning en Python

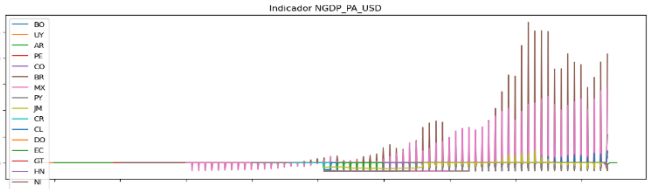
Selección de Modelo:

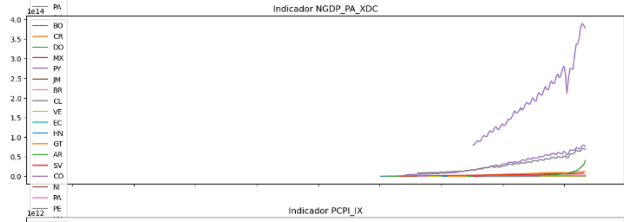
- Elegir algoritmos de machine learning adecuados en función de los objetivos y los tipos de datos.

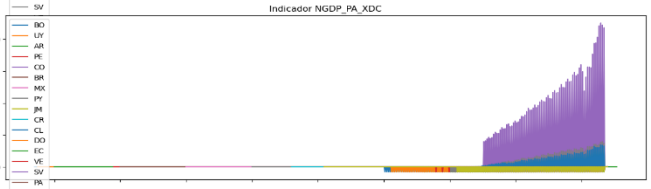
Siguiendo con la exploración inicial para comprender la estructura y contenido de los datos, tiene lugar el analisis de la estructura de las series temporales para los indicadores y a su vez las variables objeto de estudio principal, reunidas en cada país. De este modo paso a mostrar las series:

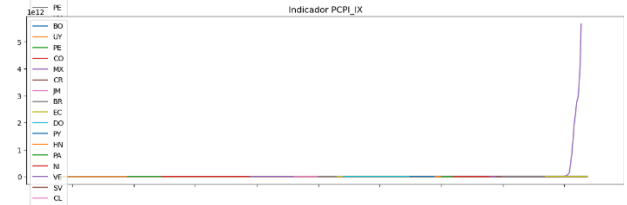
A todo esto, se sacan una serie de conclusiones generales donde algunos paises por su escala (por ejemplo, moneda nacional) tienen unas cifras bastante diferentes frente a otro paises, sin embargo, en general todas las series de tiempo tienen un comportamiento "relacionado" es por esto que proponemos realizar los mismos gráficos de series de tiempo escalando los valores. Izq.(serie), der(escalada)

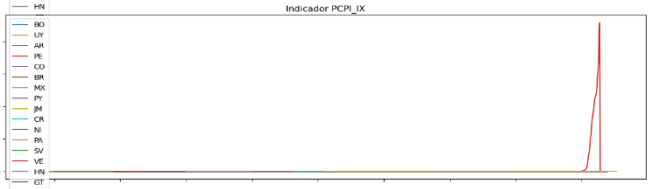

Indicador NGDP_PA_USD
1e12
MX
BR
CR
BO
EC
AR
GT
JM
HN
PY
SV
VE
NI
UY
CL
PE
CO
PA

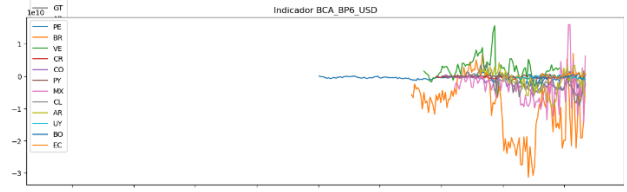

Indicador NGDP_PA_USD
BO
UY
AR
PE
CO
BR
MX
PY
JM
CR
CL
DO
EC
GT
HN
NI
PA
SV

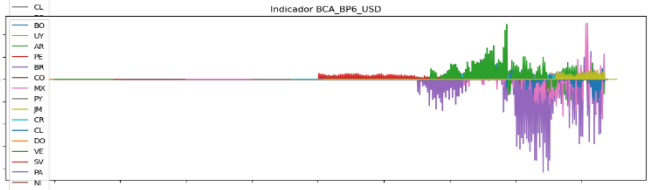

Indicador NGDP_PA_XDC
1e14
BO
CR
DO
MX
PY
JM
BR
CL
VE
EC
HN
GT
AR
SV
CO
NI
PA
PE

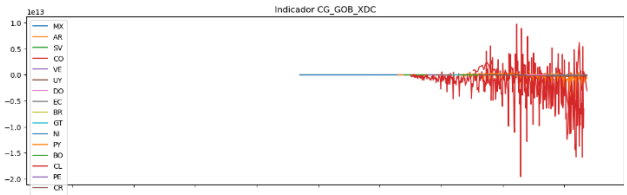

Indicador NGDP_PA_XDC
BO
UY
AR
PE
CO
BR
MX
PY
JM
CR
CL
DO
EC
VE
SV
PA
NI
HN

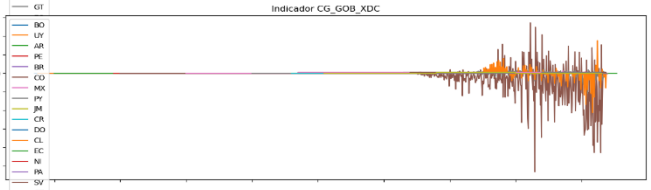

Indicador PCPI_IX
1e12
BO
UY
PE
CO
MX
CR
JM
BR
EC
DO
PY
HN
PA
NI
VE
SV
CL
GT

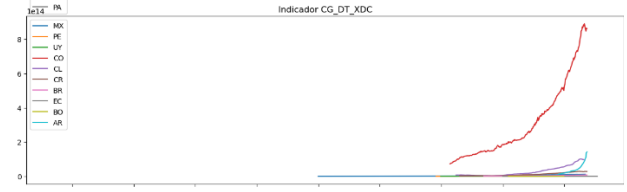

Indicador PCPI_IX
BO
UY
AR
PE
CO
BR
MX
PY
JM
CR
NI
PA
SV
VE
HN
GT
DO
CL

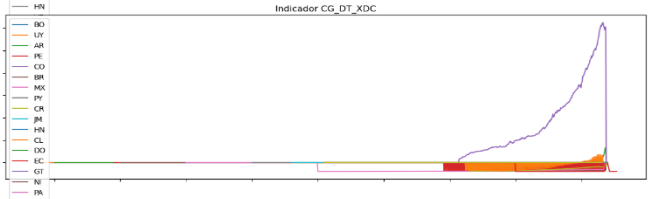

Indicador BCA_BP6_USD
1e10
PE
BR
VE
CR
CO
PY
MX
CL
AR
UY
BO
EC

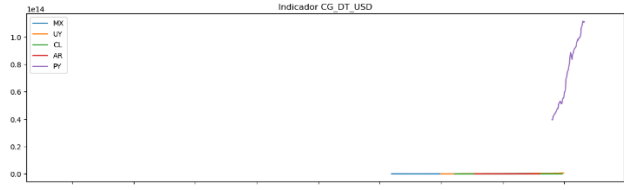

Indicador BCA_BP6_USD
BO
UY
AR
PE
BR
CO
MX
PY
JM
CR
CL
DO
VE
SV
PA
NI
HN
GT

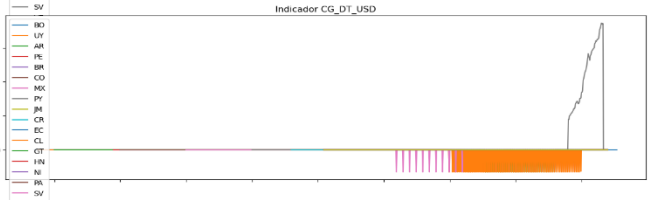

Indicador CG_GOB_XDC
1e13
MX
AR
SV
CO
VE
UY
DO
EC
BR
GT
NI
PY
BO
CL
PE
CR
HN
PA

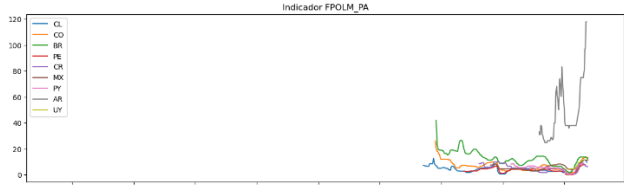

Indicador CG_GOB_XDC
BO
UY
AR
PE
BR
CO
MX
PY
JM
CR
DO
CL
EC
NI
PA
SV
GT
HN

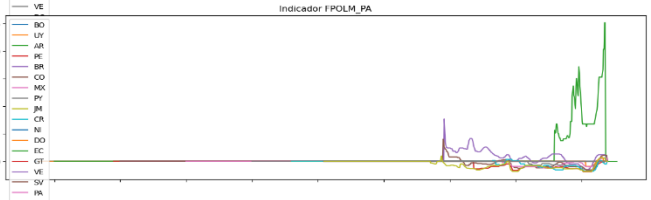

Indicador CG_DT_XDC
1e14
MX
PE
UY
CO
CL
CR
BR
EC
BO
AR

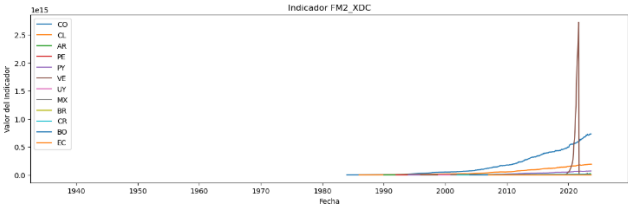

Indicador CG_DT_XDC
BO
UY
AR
PE
CO
BR
MX
PY
CR
JM
HN
CL
DO
EC
GT
NI
PA
SV

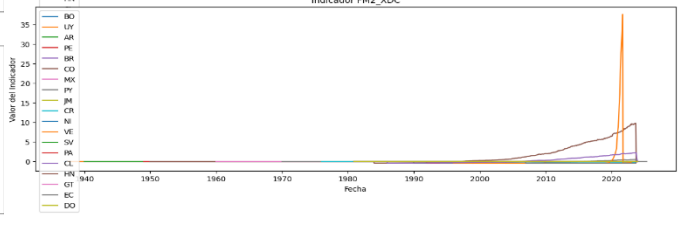

Indicador CG_DT_USD
1e14
MX
UY
CL
AR
PY


Indicador CG_DT_USD
BO
UY
AR
PE
BR
CO
MX
PY
JM
CR
EC
CL
GT
HN
NI
PA
SV
VE

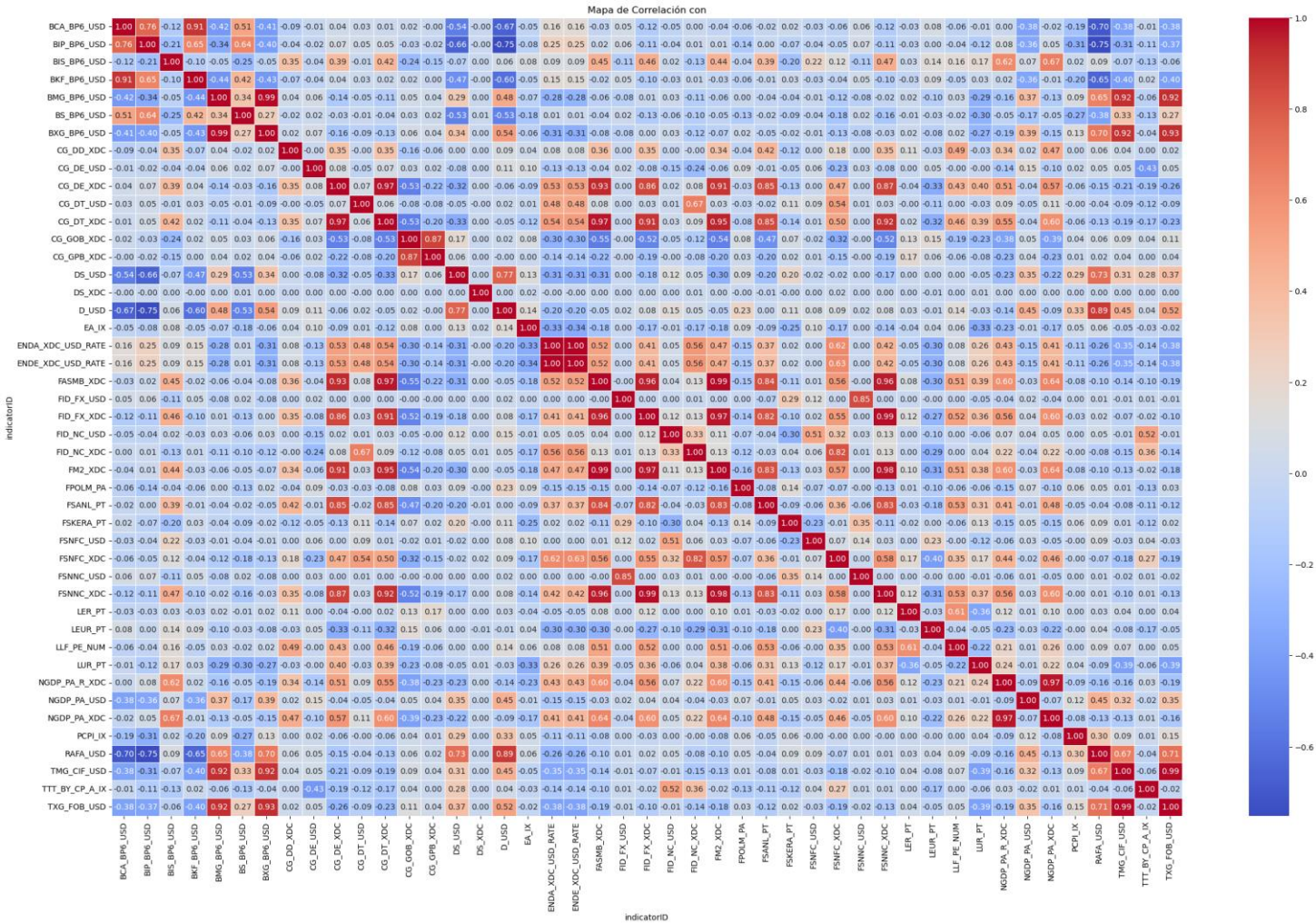

Indicador FPOLM_PA
CL
CO
BR
PE
CR
MX
PY
AR
UY


Indicador FPOLM_PA
BO
UY
AR
PE
BR
CO
MX
PY
JM
CR
NI
DO
EC
GT
VE
SV
PA
HN


Indicador FM2_XDC
1e15
CO
CL
AR
PE
PY
VE
UY
MX
BR
CR
BO
EC
Valor del Indicador
1940
1950
1960
1970
1980
1990
2000
2010
2020
Fecha


Indicador FM2_XDC
BO
UY
AR
PE
BR
CO
MX
PY
JM
CR
NI
VE
SV
PA
CL
HN
GT
EC
DO
Valor del Indicador
1940
1950
1960
1970
1980
1990
2000
2010
2020
Fecha

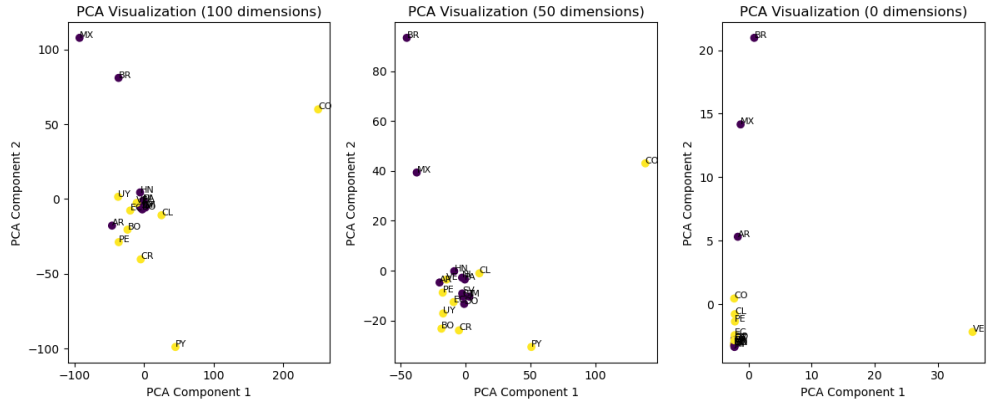
Tras la comprobación de que las series de tiempo tienen un comportamiento "relacionado" se propone realizar un mapa de correlaciones entre los indicadores, eso sí, después del escalamiento de datos.

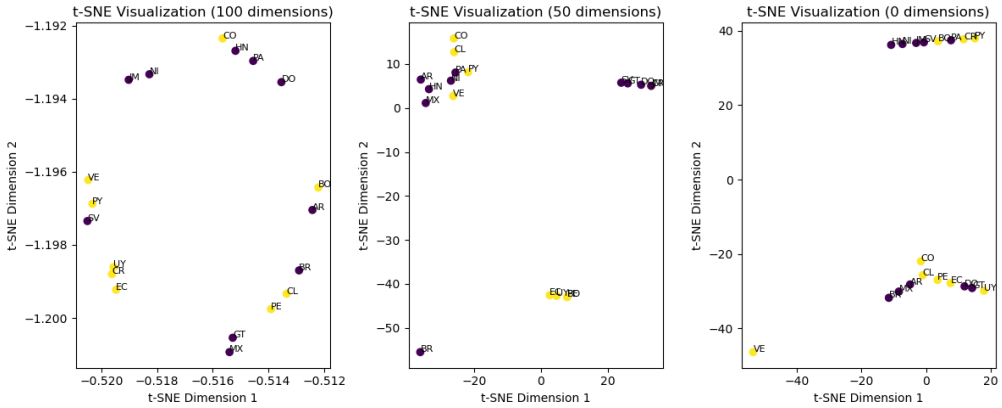


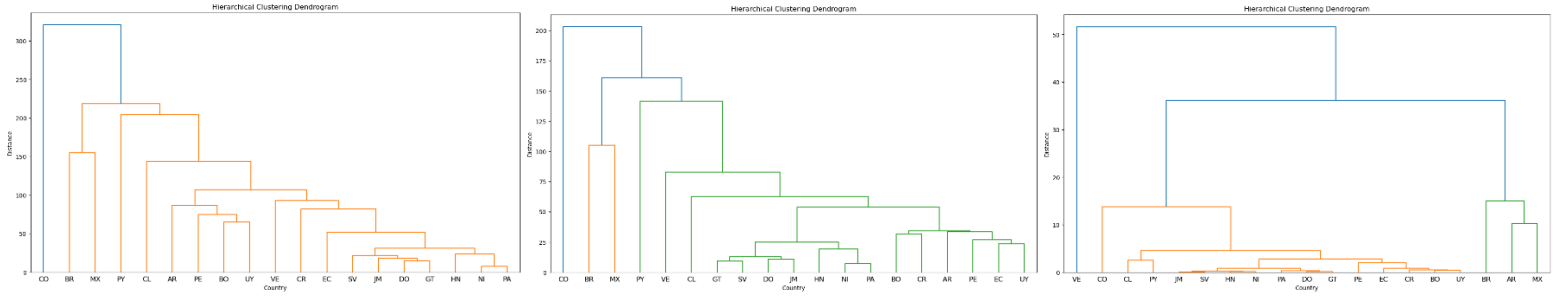
A continuacion y para finalizar el EDA podemos ejecutar técnicas de analisis de componentes principales PCA e incrustación de vecinos estocásticos distribuidos o T-SNE y verificar si los 19 paises son "clasificables" entre FLAR / NO FLAR. Todo ello para diferentes enésimos números de dimensiones que se centran en n=100, n=50, n=0.

Conclusiones locales.

Aplicar TSNE a los datos teniendo en cuenta 16.727,3624 y 122 dimensiones nos permite entender que hay valores en sus indicadores que permiten agrupar los paises FLAR en dos subgrupos (uno donde CO participa y otro donde BO participa) y alrededor de ellos hay unos paises que no son FLAR, pero comparten datos en sus indicadores, realizaremos PCA para encontrar el conjunto de datos







Conclusiones globales.

En términos de T-SNE:

Las coordenadas proporcionadas representan la ubicación de los países en un espacio bidimensional después de aplicar una reducción de dimensionalidad (PCA o t-SNE). Estos valores no tienen una interpretación directa en términos de latitud y longitud geográfica, pero ofrecen información sobre la proximidad relativa de los países en función de las características seleccionadas.

A continuación, algunas observaciones generales sobre la distribución de los países en cada escala.

Escala 100:

(Escala 100, reducción PCA) Brasil, Argentina y México, no participa ninguno.

Argentina (AR), Bolivia (BO), y Chile (CL) están relativamente cercanos entre sí. Brasil (BR) está bastante alejado de los otros países. Ecuador (EC) y Perú (PE) están cerca en el espacio.

Escala 50:

(Escala 50, reducción PCA) Brasil, Argentina y México, no participa ninguno.

Argentina (AR) y Brasil (BR) están más separados en comparación con la escala 100. Honduras (HN) y Nicaragua (NI) están más cercanos.

Escala 0:

(Escala 0, reducción PCA) Brasil, Argentina y México, no participa ninguno.

Los países muestran una distribución diferente en comparación con las escalas 50 y 100. Los países están dispersos en el espacio y no siguen un patrón claro. Estas observaciones sugieren que la distribución de los países en el espacio bidimensional varía según la escala utilizada. Es posible que las diferencias en las escalas reflejen cambios en la importancia relativa de las características utilizadas en el análisis.

En términos de PCA:

Aquí hay algunas observaciones sobre la distribución de los países en el espacio bidimensional después de aplicar PCA en diferentes escalas:

Escala 100:

(Escala 100, reducción PCA) Brasil, Argentina y México, no participa ninguno.

Argentina (AR), Bolivia (BO), y Chile (CL) están relativamente cercanos entre sí. Brasil (BR) está bastante alejado de los otros países. Ecuador (EC) y Perú (PE) están cercanos en el espacio.

Escala 50:

(Escala 50, reducción PCA) Brasil, Argentina y México, no participa ninguno.

Argentina (AR) y Brasil (BR) están más separados en comparación con la escala 100. Honduras (HN) y Nicaragua (NI) están más cercanos.

Escala 0:

Los países están distribuidos de manera más compacta y no siguen un patrón claro. Brasil (BR) tiene una posición diferente en esta escala.

(Escala 0, reducción PCA) Brasil, Argentina y México, no participa ninguno.

Estas observaciones indican que la distribución de los países en el espacio bidimensional cambia según la escala utilizada. Cada escala captura diferentes aspectos de la variabilidad en los datos. La interpretación de las coordenadas específicas puede depender del método exacto utilizado para el análisis de PCA y de las características seleccionadas. En general, las escalas más altas parecen capturar mejor las relaciones relativas entre los países en términos de las características consideradas.

Dendograma:

En un dendrograma, las ramas más cercanas indican una mayor similitud entre los elementos. Por lo tanto, si en la escala 0, México y Argentina están en ramas cercanas o se agrupan temprano en la formación del dendrograma, eso sugiere una mayor similitud entre ellos en términos de las características consideradas. No obstante, pertenecen a otra rama junto a Brasil y muy cercanos relativamente al resto en características objeto de estudio. Ninguno de estos paises pertenece al Flar y además no participan a cualquier escala tanto en PCA como en T-SNE.

A escala 100 como 50, existe relación muy cercana entre Brasil y México, estando Argentina apartada, de modo que estaría bien descubrir si existen más similitudes entre Brasil y México que entre Argentina y México.

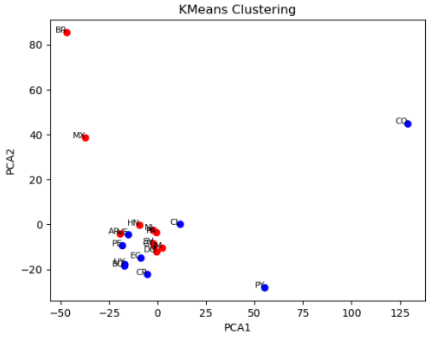
La relación México-Argentina demuestra este hecho, que ya de antemano todo apuntaba a una relación plausible o justificable en base a otros estudios.

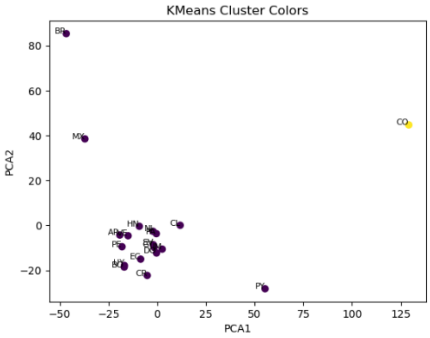
Entrenamiento del Modelo:

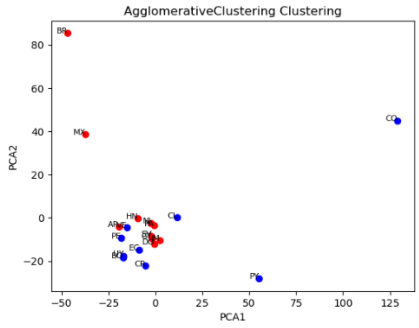
- Dividir los datos en conjuntos de entrenamiento y prueba.
- Entrenar el modelo utilizando los datos de entrenamiento.

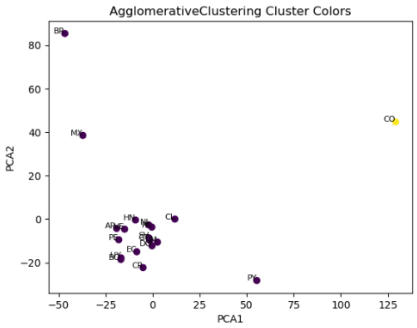
Para el entrenamiento y posterior evaluación del modelo entre diferentes modelos candidatos, se realizarán funciones para transformar la raw data en dimensiones para la clasificación, funciones que eliminan las columnas que tengan un porcentaje de valores nulo mayor a un umbral definido, escalado de los datos para normalizar las variables, visualización de los resultados en 2D ('PCA1', 'PCA2') utilizando PCA. Todo ello se representará mediante gráficos de dispersión de clusters basado en colores generados por algoritmo. Se trata se evaluación no supervisada de clustering de datos n dimensionales.

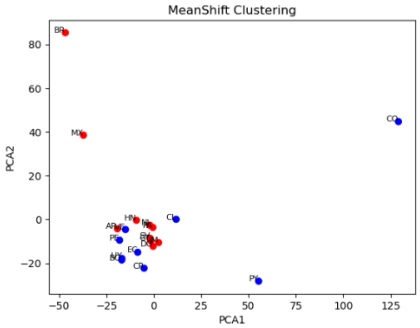
A modo de ejemplo muestro la distribucion con n=50 dimensiones.

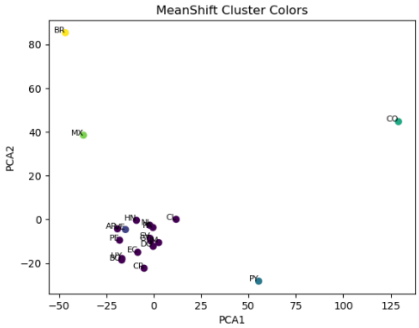

KMeans Clustering
PCA1
PCA2

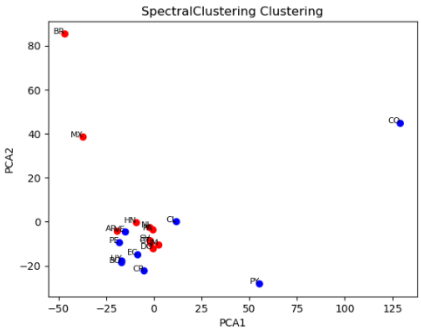

KMeans Cluster Colors
PCA1
PCA2

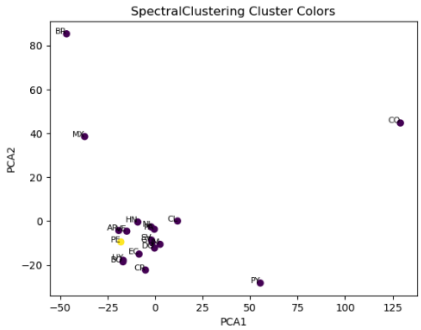

AgglomerativeClustering Clustering
PCA1
PCA2

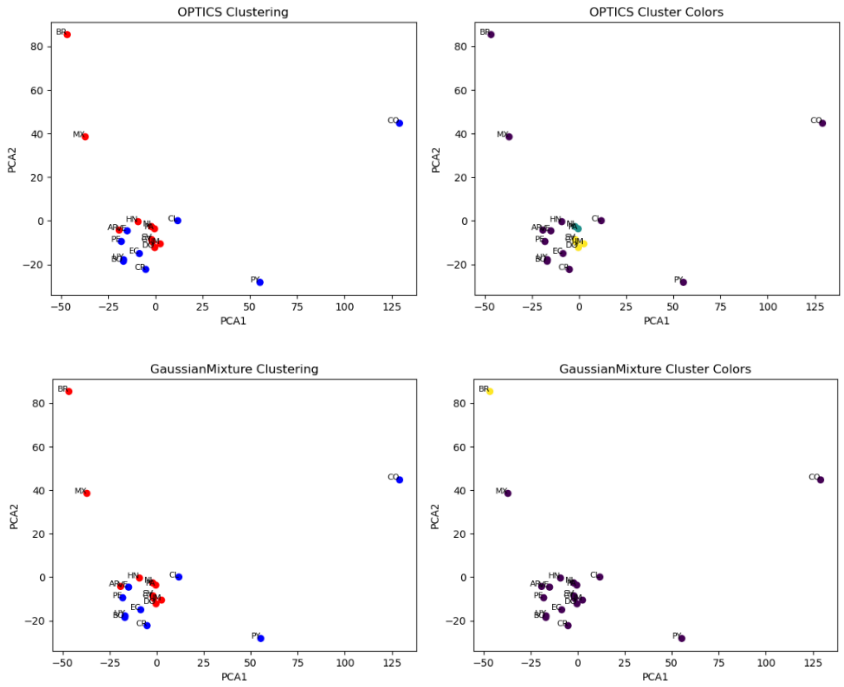

AgglomerativeClustering Cluster Colors
PCA1
PCA2


MeanShift Clustering
PCA1
PCA2


MeanShift Cluster Colors
PCA1
PCA2


SpectralClustering Clustering
PCA1
PCA2


SpectralClustering Cluster Colors
PCA1
PCA2



Estos resultados representan métricas de evaluación para diferentes algoritmos de clustering aplicados a los datos. Aquí hay una breve interpretación:

KMeans y AgglomerativeClustering

Silhouette Score: 0.6059

Davies-Bouldin Index: 0.2451

Interpretación: Ambos algoritmos muestran una buena cohesión interna y separación entre los clústeres, indicado por el Silhouette Score alto y el bajo Davies-Bouldin Index. Esto sugiere una estructura clara en los datos.

MeanShift

Silhouette Score: 0.4027

Davies-Bouldin Index: 0.2315

Interpretación: Aunque el Silhouette Score es menor que en KMeans y AgglomerativeClustering, aún indica una separación decente entre los clústeres. El Davies-Bouldin Index también es bajo, lo que es favorable.

SpectralClustering

Silhouette Score: -0.2642

Davies-Bouldin Index: 1.3198

Interpretación: El Silhouette Score negativo sugiere que hay solapamiento entre clústeres o que algunos puntos podrían estar mal asignados. El Davies-Bouldin Index más alto indica que los clústeres no están tan bien definidos.

OPTICS

Silhouette Score: -0.1762

Davies-Bouldin Index: 4.4500

Interpretación: Ambas métricas indican un rendimiento inferior en comparación con otros algoritmos. El Silhouette Score negativo y el alto Davies-Bouldin Index sugieren que OPTICS podría no ser la mejor opción para estos datos.

GaussianMixture

Silhouette Score: 0.4905

Davies-Bouldin Index: 0.3424

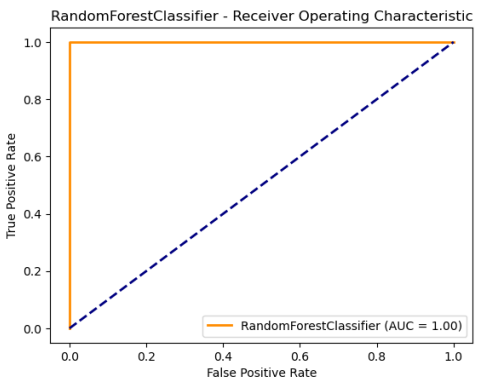
Interpretación: Muestra un rendimiento sólido con un Silhouette Score decente y un Davies-Bouldin Index bajo, indicando clústeres bien definidos y separados.

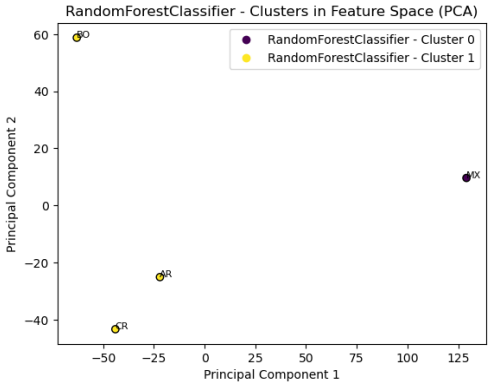
En resumen, KMeans, AgglomerativeClustering y Gaussian Mixture parecen ser opciones prometedoras para la tarea de clustering en los datos, mientras que otros algoritmos podrían no ser tan adecuados.

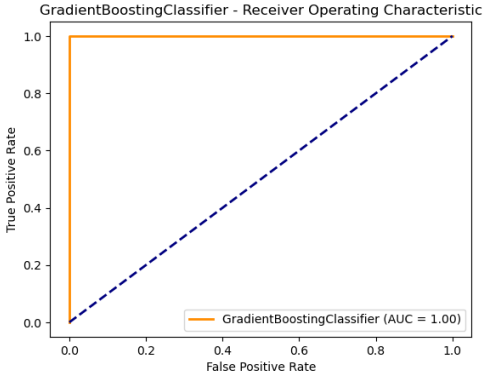
Sin embargo, queda determinar cuál es el modelo optimo a utilizar bajo una serie de criterios de optimalidad aplicables al caso de estudio. Se han determinado al menos dos criterios, precisión que se usaría para dimensiones concretas, por ejemplo, si nos ceñimos a n=0 dimensiones o n=50 dimensiones y por otro lado el criterio general, que seria valido independientemente de la n dimensión utilizada, pudiendo ser valido para estudios de analisis cruzado que incluyan diferentes panoramas de estudio. No obstante, no hay resultado mejor ni peor si elegimos entre n dimensión o n dimensiones, solo diferentes sistemas de cálculo. Sin embargo, dentro de una determinada n dimensión, ahí sí que existiría mejor resultado en términos de modelo a utilizar.

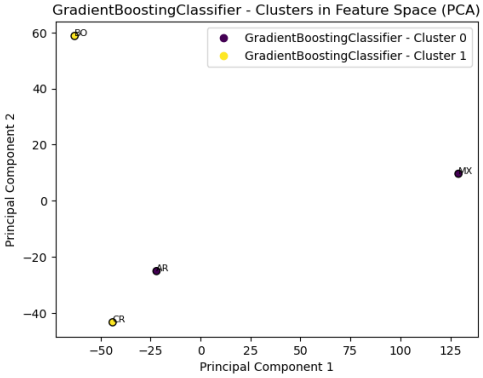
Evaluación del Modelo:

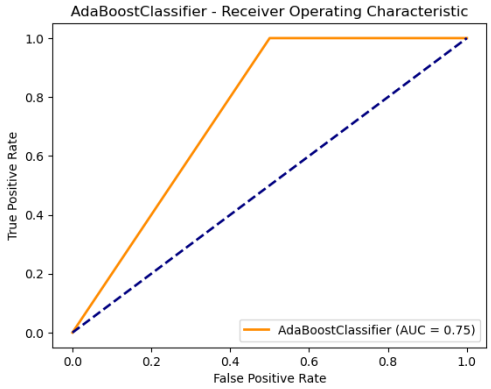
- Evaluar el rendimiento del modelo utilizando métricas apropiadas (Precisión, F1-score, Recall, Accuracy).
- Ajustar el modelo si es necesario.

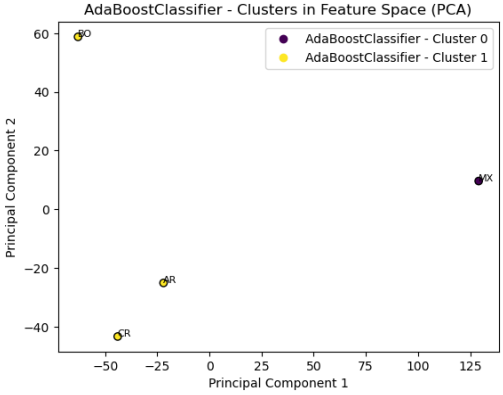

RandomForestClassifier - Receiver Operating Characteristic
True Positive Rate
False Positive Rate
RandomForestClassifier (AUC = 1.00)

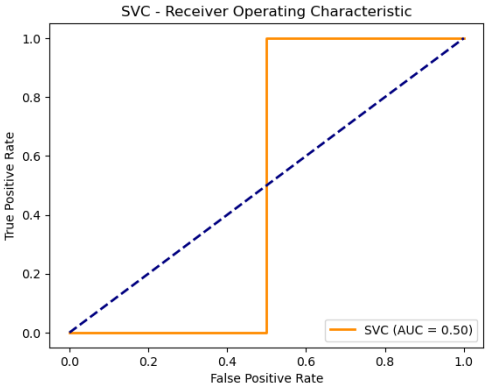

RandomForestClassifier - Clusters in Feature Space (PCA)
RandomForestClassifier - Cluster 0
RandomForestClassifier - Cluster 1
Principal Component 2
Principal Component 1
BO
MX
AR
CR

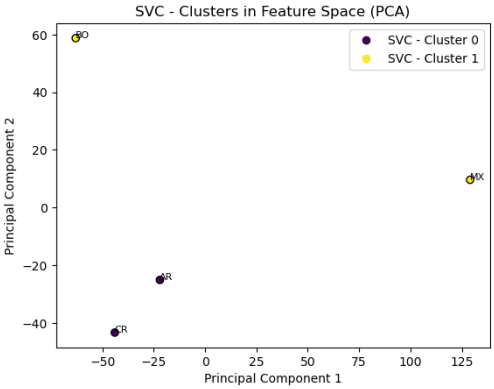

GradientBoostingClassifier - Receiver Operating Characteristic
True Positive Rate
False Positive Rate
GradientBoostingClassifier (AUC = 1.00)

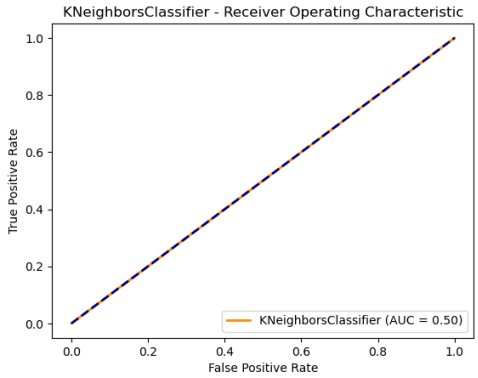

GradientBoostingClassifier - Clusters in Feature Space (PCA)
GradientBoostingClassifier - Cluster 0
GradientBoostingClassifier - Cluster 1
Principal Component 2
Principal Component 1
BO
MX
AR
CR

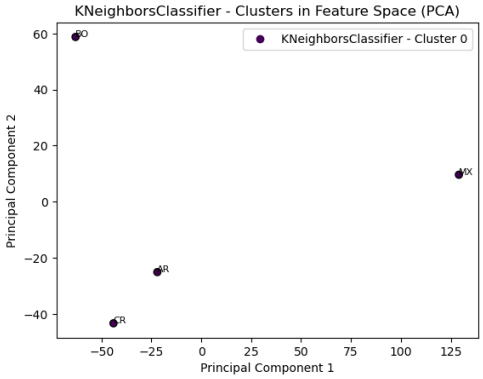

AdaBoostClassifier - Receiver Operating Characteristic
True Positive Rate
False Positive Rate
AdaBoostClassifier (AUC = 0.75)

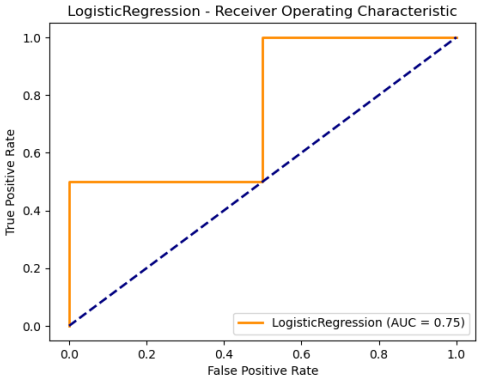

AdaBoostClassifier - Clusters in Feature Space (PCA)
AdaBoostClassifier - Cluster 0
AdaBoostClassifier - Cluster 1
Principal Component 2
Principal Component 1
BO
MX
AR
CR

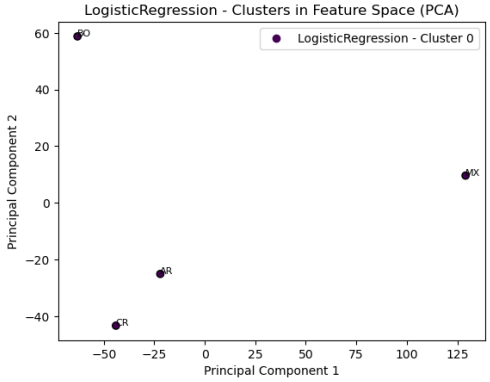

SVC - Receiver Operating Characteristic
True Positive Rate
False Positive Rate
SVC (AUC = 0.50)

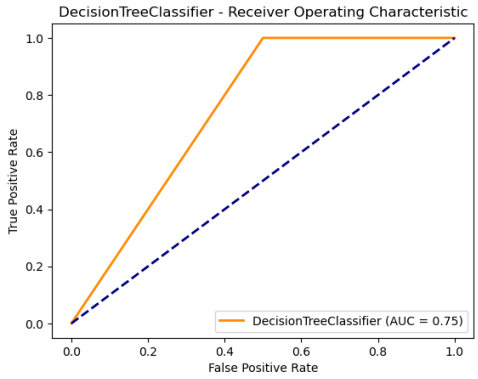

SVC - Clusters in Feature Space (PCA)
SVC - Cluster 0
SVC - Cluster 1
Principal Component 2
Principal Component 1
BO
MX
AR
CR

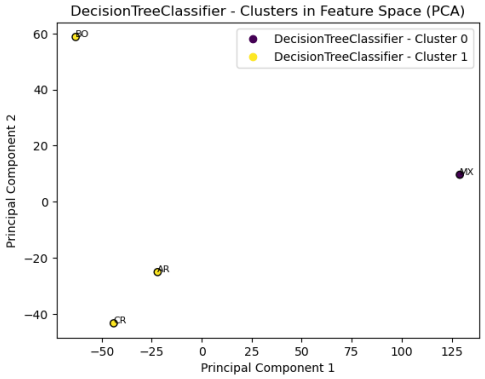

KNeighborsClassifier - Receiver Operating Characteristic
True Positive Rate
False Positive Rate
KNeighborsClassifier (AUC = 0.50)

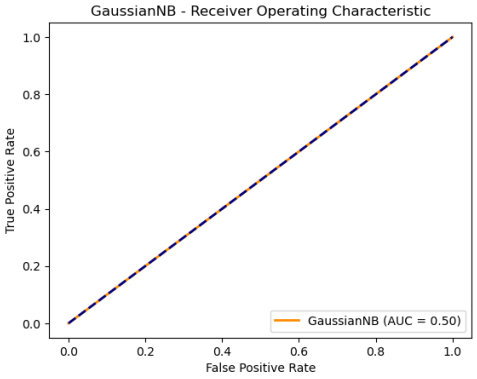

KNeighborsClassifier - Clusters in Feature Space (PCA)
KNeighborsClassifier - Cluster 0
Principal Component 2
Principal Component 1
BO
MX
MR
CR

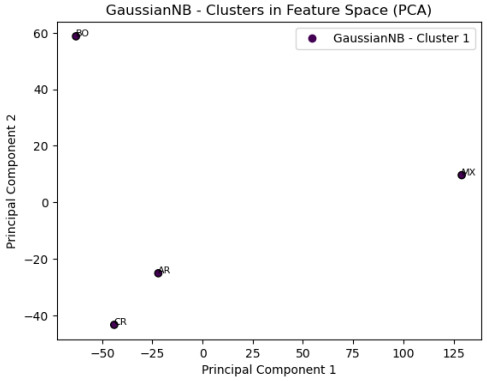

LogisticRegression - Receiver Operating Characteristic
True Positive Rate
False Positive Rate
LogisticRegression (AUC = 0.75)


LogisticRegression - Clusters in Feature Space (PCA)
LogisticRegression - Cluster 0
Principal Component 2
Principal Component 1
BO
MX
MR
CR


DecisionTreeClassifier - Receiver Operating Characteristic
True Positive Rate
False Positive Rate
DecisionTreeClassifier (AUC = 0.75)


DecisionTreeClassifier - Clusters in Feature Space (PCA)
DecisionTreeClassifier - Cluster 0
DecisionTreeClassifier - Cluster 1
Principal Component 2
Principal Component 1
BO
MX
MR
CR


GaussianNB - Receiver Operating Characteristic
True Positive Rate
False Positive Rate
GaussianNB (AUC = 0.50)


GaussianNB - Clusters in Feature Space (PCA)
GaussianNB - Cluster 1
Principal Component 2
Principal Component 1
BO
MX
MR
CR

	Classifier	Accuracy	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1-Score (Class 0)	F1-Score (Class 1)	AUC
0	RandomForestClassifier	0.75	1.0	0.666667	0.5	1.0	0.666667	0.800000	1.00
1	GradientBoostingClassifier	1.00	1.0	1.000000	1.0	1.0	1.000000	1.000000	1.00
2	AdaBoostClassifier	0.75	1.0	0.666667	0.5	1.0	0.666667	0.800000	0.75
3	SVC	0.50	0.5	0.500000	0.5	0.5	0.500000	0.500000	0.50
4	KNeighborsClassifier	0.50	0.5	0.000000	1.0	0.0	0.666667	0.000000	0.50
5	LogisticRegression	0.50	0.5	0.000000	1.0	0.0	0.666667	0.000000	0.75
6	DecisionTreeClassifier	0.75	1.0	0.666667	0.5	1.0	0.666667	0.800000	0.75
7	GaussianNB	0.50	0.0	0.500000	0.0	1.0	0.000000	0.666667	0.50

Aquí hay algunas conclusiones basadas en los resultados de la evaluación anterior de los modelos candidatos.

Mejores Modelos:

GradientBoostingClassifier muestra un rendimiento excepcional en todas las métricas, con una precisión del 100% y AUC de 1.0. Parece ser el modelo más robusto en este conjunto de datos.

Rendimiento Intermedio:

RandomForestClassifier y AdaBoostClassifier también tienen un rendimiento decente, con una precisión del 100% para la clase 0 y un rendimiento sólido para la clase 1.

Modelos con Rendimiento Bajo:

Los modelos SVC, KNeighborsClassifier, LogisticRegression y GaussianNB tienen un rendimiento limitado en términos de precisión y AUC. Estos modelos podrían necesitar ajustes o considerarse menos adecuados para este conjunto de datos.

Consideración Especial para KNeighborsClassifier y LogisticRegression:

Estos modelos muestran una precisión del 50%, lo que podría indicar que están tomando decisiones casi aleatorias. Pueden requerir una revisión más profunda o ajustes en los hiperparámetros.

Ajuste del Modelo:

Considera ajustar los hiperparámetros de los modelos para mejorar aún más el rendimiento. El ajuste podría realizarse mediante validación cruzada u otras técnicas de búsqueda de hiperparámetros.

Importancia de la Clase 0:

La precisión para la clase 0 es del 100% en varios modelos, lo que podría deberse a un desequilibrio en las clases. Asegúrate de que el modelo no esté simplemente prediciendo la clase mayoritaria.

Observaciones sobre GaussianNB:

GaussianNB tiene un rendimiento mixto. La precisión para la clase 1 es del 50%, pero el F1-score es más alto. Esto podría indicar un problema con el umbral de decisión.

Generalización del Modelo:

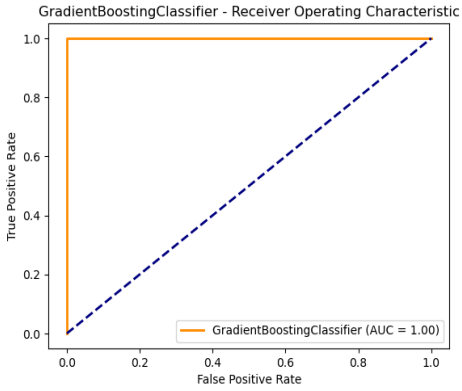
Asegúrate de evaluar estos modelos en un conjunto de datos de prueba separado para garantizar la generalización del rendimiento.

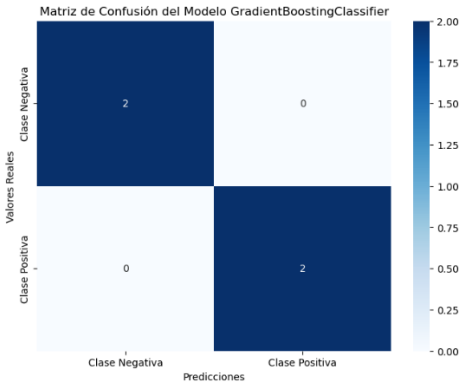
En resumen, el GradientBoostingClassifier destaca como el modelo más prometedor en este conjunto de datos, lo cual podría utilizarse para todos los n dimensiones propuestos anteriormente con requisitos de optimalidad en general, mientras que otros modelos podrían beneficiarse de ajustes adicionales, este es el caso para requisitos locales de optimalidad, dado el caso, el modelo gaussiano para n 0 dimensiones.

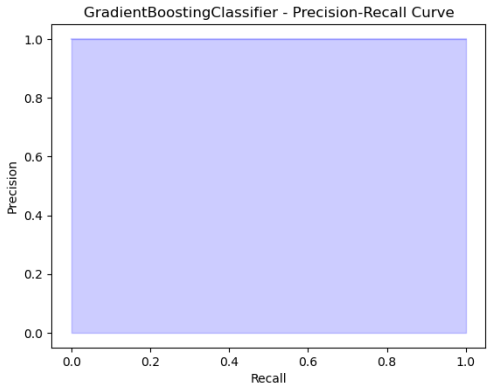
La elección final del modelo será subjetiva y dependerá de los objetivos específicos y del rendimiento deseado en diferentes métricas.

Si nos centramos para el caso querer centrarnos en el criterio específico, es decir, elección criterio de modelo con mayor precisión, candidato a elegir, tendríamos lo siguiente.

Si se valora más la precisión y se considera que es la métrica más crítica al problema, entonces se debe mirar la precisión en la clase que nos interesa (posiblemente la clase 1, dependiendo de tu problema específico). En este caso, el modelo GradientBoostingClassifier tiene una precisión del 100% para la clase 1, rendimiento optimo, mientras que RandomForestClassifier tiene una precisión del 66.67% para la misma clase, rendimiento intermedio. Elegiré el primero.

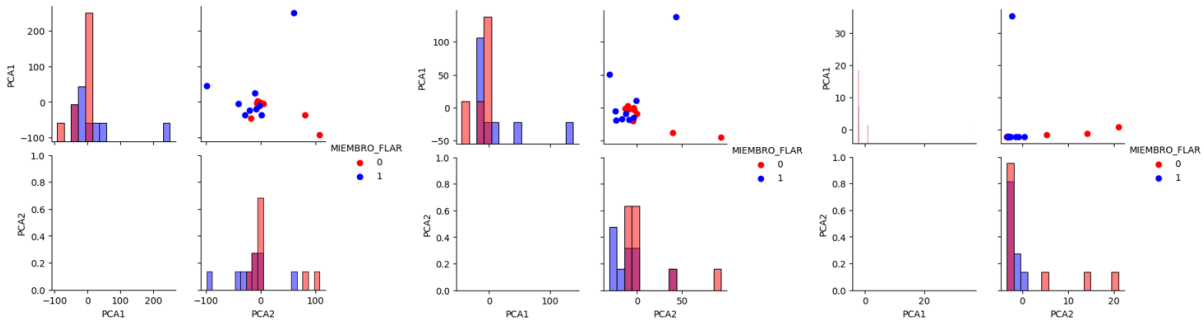


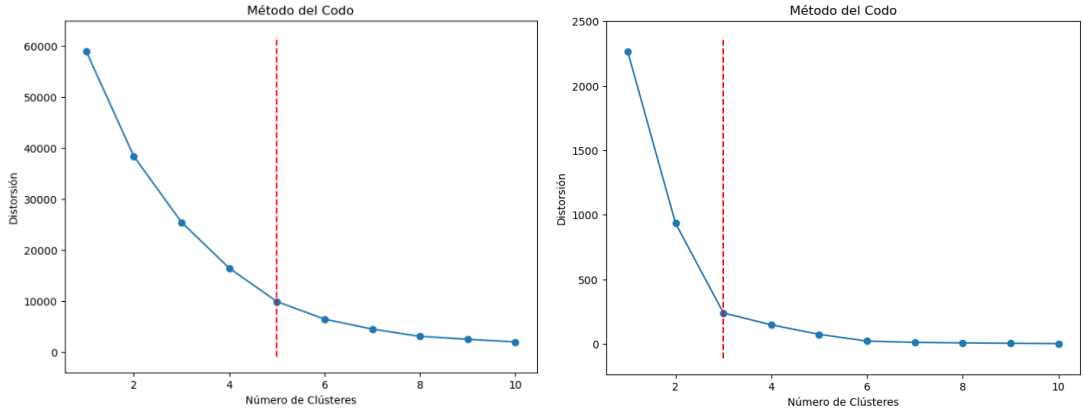




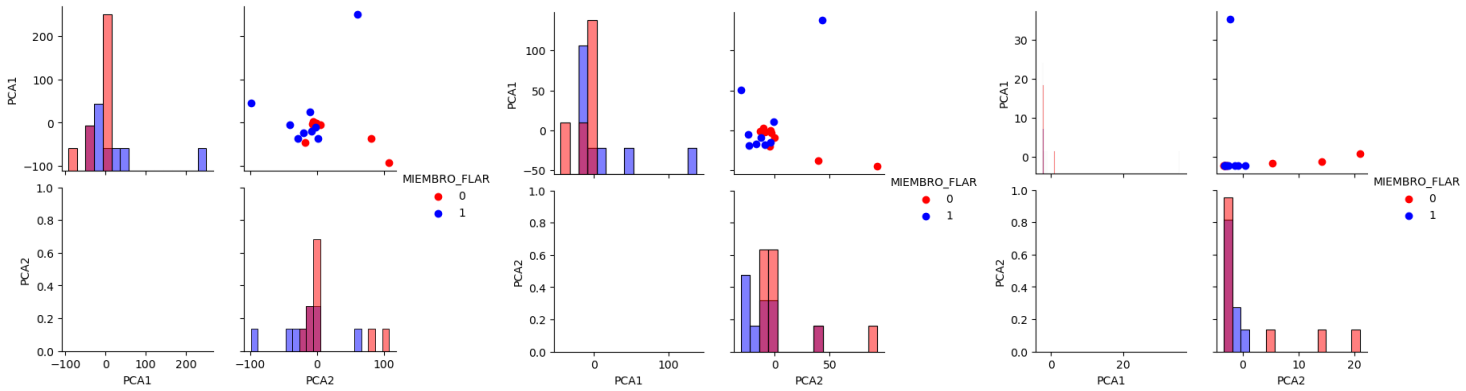
Visualización de Resultados:

- Generar visualizaciones que ayuden a interpretar los resultados.
- Utilizar gráficos de series temporales si corresponde.

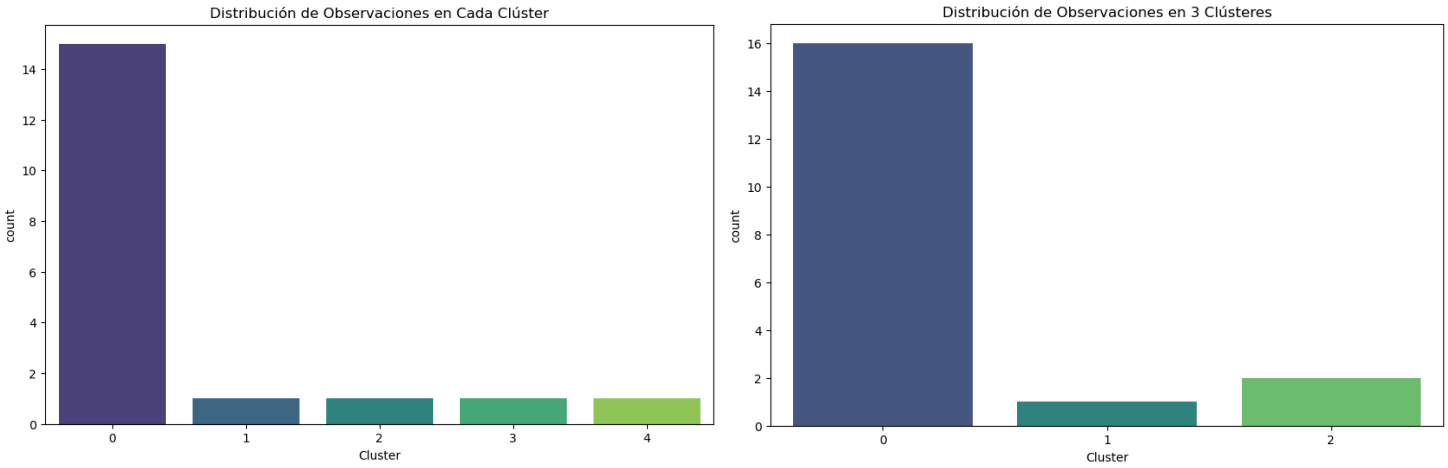




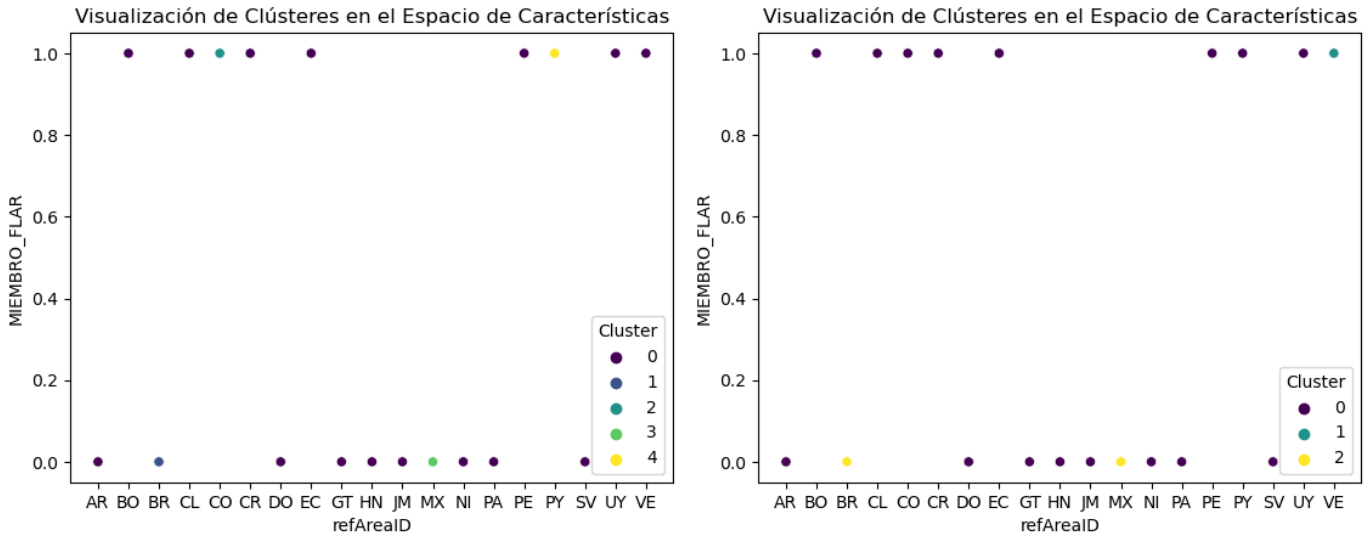
Ajuste (fitting) del modelo KMeans a los datos y entrenamiento del Modelo KMeans. (N=100 izquierda, n=50 centro, n=0 derecha). Como prueba de que los resultaddos varian en function de la dimensionalidad.



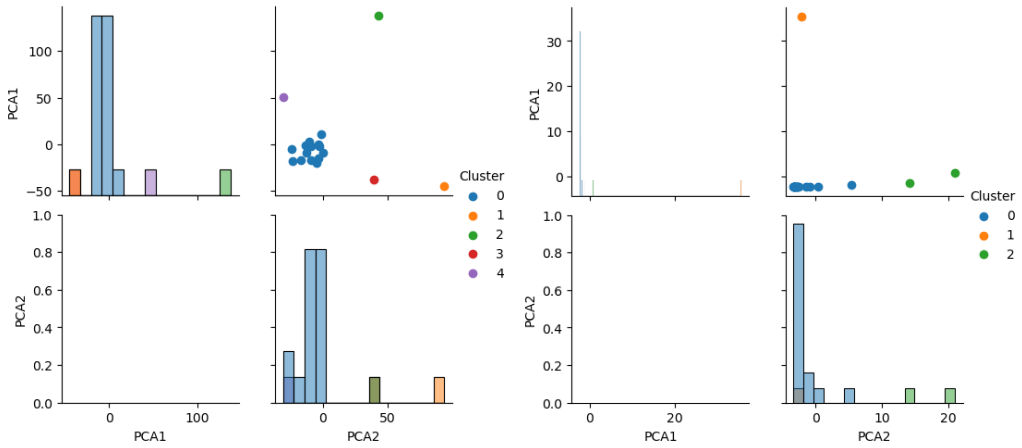
Visualización de la Distribución de Clústeres y Análisis Estadístico por Clúster. N=50 derecha y n=0 izquierda.



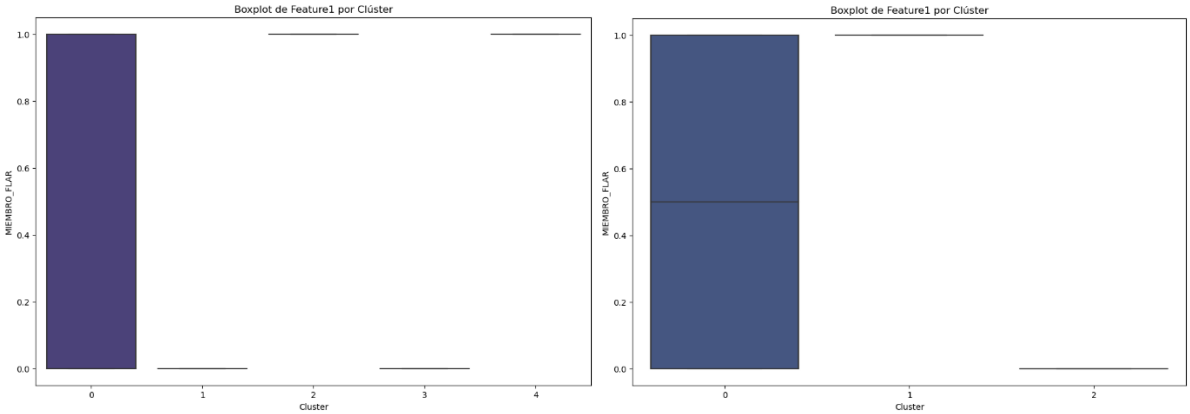
Visualización de Características Importantes por Clúster (por ejemplo, para pares de características específicas). N=50 derecha y n=0 izquierda.



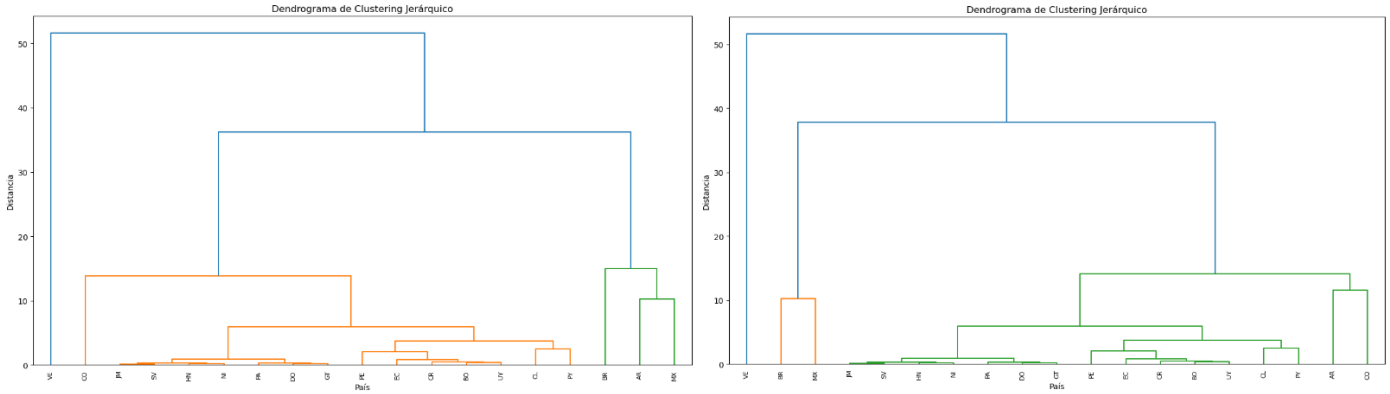
Visualización de PCA con Colores por Clúster. N=50 derecha y n=0 izquierda.



Visualización de Boxplots por Clúster (para una característica general y específica).



Visualizaciones Adicionales por dendograma.



Observaciones:

Se ha trabajado con datos para el "clafdata_50_scaled", al no estar por defecto. Parece que la columna 'Cluster' no se ha agregado al Dataframe clafdata_100_scaled, ni clafdata_0_scaled después de realizar el clustering, quiere decir que, para estudiar cada caso por dimensionalidad, hay que plantear dentro del mismo esquema de estudio, otra serie de algoritmos compatibles a la nueva realidad. En este caso, nos llevaría al Dendograma de en medio, el de 50 dimensiones. Sin embargo, he escogido el de debajo de 0 dimensiones para establecer una comparativa.

Durante el proceso de modelado, se estaban probándose varios modelos primero los supervisados y luego los no supervisados y los resultados no daban diferente que el TSNE (Sobre todo MX)", pudiendo ser porque el resultado comparativo del T-SNE en n dimensiones no fuese el mismo al de Dendograma para mismas dimensiones (por ejemplo, en 0 dimensiones donde argentina está cerca de México), ejemplo T-SNE clafdata_0_scaled vs Dendograma clafdata_0_scaled, 50 vs 50, 100 vs 100.

Conclusiones finales:

Conclusiones finales del PCA y TSNE en base al archivo data base. Tras una visualización tanto de PCA como de T-SNE y Dendograma para cada escala o número de dimensiones (sobre las que ya operaban).

En términos de PCA, Destacar que Argentina-México-Brasil están muy parejos en características comunes y sin pertenecer ninguno al Flar, ninguno participa a ninguna escala(color rojo, valor 0). Sin embargo, a medida que el numero de dimensiones disminuye, la relación de argentina con el resto se distancia teniendo menos características en común, aumentando otra vez cuando disminuye de las 50 dimensiones.

En términos de T-SNE la relación entre Argentina y México aumenta a medida que disminuye el numero de dimensiones, siendo Brasil un satélite que aumenta para luego disminuir.

En el modelo optimo bajo condiciones generales de optimalidad, México y Argentina comparten Cluster y diferenciándose del otro formado por Bolivia, Costa Rica, sin embargo, bajo condiciones de precisión, comparten Cluster todos. Es decir, cuando somos mas restrictivos a la hora de tener criterios de clasificación de paises, mas de estos comparten características, y por lo tanto mas atencion a la hora de tomar decisiones habría que tener. Tal vez un modelo mas ajustado nos traiga mayor incertidumbre.

En cuestión de la tipología del Cluster, decir que independientemente de la dimensionalidad, existe un gran conjunto de paises muy homogéneo en características intrínsecas a las variables macroeconómicas estudiadas de entres14 a 16 paises de 19), sin embargo, los que faltarían poseen características propias muy notorias, que se homogeneizan en numero a medida que el número de clusters aumenta.

Es decir, que, si quieres seleccionar algún país objetivo de un montón, es mejor trabajar en bajas dimensiones o reducción de clústeres. Colombia, México, Paraguay, Venezuela, Brasil serian objetivos a observar. Dando especial importancia a Argentina, México, Brasil, Colombia.

Consideraciones:

A la pregunta que se planteó de "Si la conclusión sigue siendo MX y AR" y puedo identificar cuales indicadores ahora preprocesados y escalados, tenemos la evidencia que argumenta el proyecto de clasificación (segunda parte de la documentación y sustentación del proyecto) siendo una conclusión posible y con enfoque argumentado.

A la pregunta que se planteó de en el t-SNE Brasil está alejado del grupo, y en los modelos de clustering aparecen otros países que pueden tener prioridad (HN), decir que en términos de Dendograma, esta bastante emparejado a México y esa separación solo se da en n=50 dimensiones, siendo mas referente en n=0 dimensiones. Es cierto que en los modelos previa optimización esta muy separado del resto y que posteriormente no figura en una clusterización optimizada bajo cualquier criterio. Tal vez puedan tener prioridad Costa Rica y Bolivia.

Entonces para la conclusión final Argentina y México y en su caso Brasil, ¿serian objeto de atencion para el Flar?

Si, en especial los dos primeros, siendo Brasil un candidato muy presente pero que no cumple condiciones para ser observado con plena garantía de éxito. Lo contrario sucedería con Costa Rica y Bolivia, que, sin estar muy presentes, bajo condiciones de optimalidad aparecen en clusters muy diferenciados del resto de paises.

Fase 3: Documentación y Presentación

Creación del Informe Final:

- Documentar todo el proceso, desde la recopilación de datos hasta la evaluación del modelo.
- Incluir hallazgos clave y decisiones tomadas.

Entrega del Cuaderno de Jupyter:

- Preparar un cuaderno Jupyter que muestre el código y los pasos seguidos.
- Incluir un archivo "predicciones.csv" si corresponde. Sobre todo, a partir de las series temporales. Seria una segunda parte del trabajo.

Presentación de PowerPoint:

- Preparar una presentación de PowerPoint concisa con hasta 10 diapositivas.
- Destacar los aspectos más importantes del proyecto.

Fase 4: Evaluación y Revisión

Revisión del Proyecto:

- Evaluar el proyecto en términos de claridad, originalidad, contenido y corrección.
- Asegurarse de que todas las partes del proyecto estén bien documentadas.

Presentación del Proyecto:

- Preparar para la presentación del proyecto, vestirse de manera profesional y estar preparado para responder preguntas.

Entrega Final:

- Enviar todas las entregas en el plazo establecido.

Fase 5: Retroalimentación y Mejoras (Opcional)

Análisis de Retroalimentación:

- Considerar cualquier retroalimentación recibida y aprender de la experiencia.

Identificar Mejoras:

- Identificar áreas de mejora en el proyecto y en las habilidades de machine learning.

Roles de cada miembro del equipo

Camilo Cruz (Acceso a datos y experto en la información): Responsable de conseguir, procesar, explorar y otras actividades iniciales de los datos del SIE, así como establecer las reuniones o aclaraciones con los economistas expertos de la dirección de estudios económicos del FLAR, esto implica acompañamiento a la selección del caso y análisis de los resultados del modelo.

Álvaro Ortuzar (Científico de datos): Responsable de la prueba y selección de los modelos, evaluación y muestra de resultados de los modelos.

Camilo Cruz y Álvaro Ortuzar (Documentadores): responsable de la creación de la documentación definida en la fase 4

Esta definición de roles ayuda a determinar la responsabilidad de cada miembro del equipo, pero también nos ilustra que hay algunas actividades transversales donde todos seremos participes por ejemplo las fases de Evaluación y retroalimentación y mejoras deberíamos intervenir todos.