Proyecto Data Science 2023

Entrega 24/01/2023 -Gaston-Roberto-Gimenez-ProyectoFinal-2023.ipynb

Profesor: Profesor: Octavio Lafourcade

Tutor: Néstor Jesús Ramírez Reyes

Alumno: Gastón Roberto Giménez

Índice:

1-Introducción:

El conjunto de datos para este análisis contiene datos de 9 países y una región administrativa especial (China, Francia, Alemania, Hong Kong, India, Japón, España, Reino Unido y Estados Unidos de América) desde 1980 hasta 2020. Incluye factores macroeconómicos importantes como inflación, desempleo, PIB, tipo de cambio (base USD) e ingreso per cápita. También contiene los precios de las acciones del principal índice bursátil del país respectivo que pueden ayudar a analizar el conjunto de datos para identificar el impacto de las principales variables macroeconómicas, el movimiento bursátil que modifique el precio del petróleo, también ofrece datos sobre como fueron los cambios durante el proceso de confinamiento debido al Covid-19, estos corresponden al primer dataset, mientras que la incorporación de un segundo dataset surge dado que la OPEP (La Organización de Países Exportadores de Petróleo que mediante la producción de este commodities incorpora cambios en el precio del petróleo y teníamos que tenerlo en cuenta para tener una mejor predicción de nuestra variable (Oil prices).

2- Versiones del Proyecto.

Versión	Fecha
0.1	**06/07/2022**
0.2	**18/07/2022**
0.3	**01/08/2022**
04	**17/08/2022**
0.5	**24/08/2022**
0.6	**31/08/2022**
0.7	**21/09/2022**
0.8	**28/09/2022**
0.9	**12/10/2022**
1.0	**24/10/2022**
1.1	**31/10/2022**
1.2	**14/11/2022**
1.3	**23/11/2022**
1.4	**30/11/2022**
1.5	**07/12/2022**
1.6	**24/01/2023**

3-Descripcion del problema

1. La aparición de nuevas fuentes de energía alternativas que pueden suponer una amenaza para la demanda del crudo, lo que pude poner en peligro las inversiones a largo plazo.

2. Los precios del petróleo pueden ser volátiles, tanto cuando son elevados como cuando son bajos. Los precios elevados pueden obstaculizar el crecimiento económico, ya que pueden crear una inflación y subidas de los tipos de interés superiores a los esperado. Los precios bajos, por otro lado, pueden hacer que los productos sean más accesibles.

3. Los precios varían continuamente debido a las fluctuaciones en la oferta y demanda a nivel mundial, por lo que hacer que sea difícil prever como se desarrollan los precios en el futuro. Por lo tanto, es importante tener en cuenta las

fluctuaciones del mercado a la hora de invertir en el petróleo mediante acciones bursátiles.

Teniendo en cuenta la volatilidad de los precios que corresponden al precio del petróleo

¿Cuáles son los factores que inciden en el precio del Petróleo?

¿El análisis macroeconómico puede llegar a obtener indicios de esos cambios?

¿Los Hechos históricos en transcurso del tiempo ocasionaron cambios que impliquen o determinen las fluctuaciones del precio del petróleo?

¿Podemos predecir precios a futuro con datos macroeconómicos?

4-Alcance

El Proyecto propuesto en este trabajo está destinado para ser usado principalmente por inversionistas, economistas, incluso puede ser destinado a uso académico. Por este motivo, se espera que el uso que se haga fundamentalmente operativo. Sin embargo, la investigación también proporciona información de índole estadística que puede servir para comunicar a los organismos de nivel superior en empresas del rubro del crudo, quienes pueden tomar decisiones más precisas a la hora de revisar futuras producciones sobre derivados este commodities.

5-Hipótesis

En el precio del petróleo no solo el índice de los movimientos bursátiles provoca cambios y otras variables, sino también el comportamiento, hechos geopolíticos globales como el grupo de países productores de este commodities (OPEP), que tiene como finalidad el volumen de producción que incide finalmente el precio y como consecuencia en la macroeconomía mundial teniendo en cuenta que es principal motor de la economía mundial.

6-Objetivos

- Comprender la macroeconomía mediante el análisis de datos.
- Incorporar el uso de Machine Learnig a los movimientos bursátiles y hechos geopolíticos que inciden en la macroeconomía.
- Aportar información consistente a los clientes para la toma de mejores decisiones.

7-Diccionario de variables

STOCK INDEX	Índice bursátil: Un **índice bursátil** corresponde a un registro estadístico compuesto usualmente de un número, que trata de reflejar las variaciones de valor o rentabilidades promedio de las acciones que lo componen. Generalmente, las acciones que componen el índice tienen características comunes tales como: pertenecer a una misma bolsa de valores, tener una capitalización bursátil similar o pertenecer a una misma industria. Estas son usualmente usadas como punto de referencia para distintas carteras, tales como los fondos mutuos.
COUNTRY	Países que integran el estudio de variables económicas de los últimos 40 años. El conjunto de datos contiene datos de 9 países y una región administrativa especial (China, Francia, Alemania, Hong Kong, India, Japón, España, Reino Unido y Estados Unidos de América) desde 1980 hasta 2020. Incluye factores macroeconómicos importantes como inflación, desempleo, PIB, tipo de cambio (base USD) e ingreso per cápita, precios de las acciones del principal índice bursátil del país respectivo que pueden ayudar a analizar el conjunto de datos para identificar el impacto de las principales variables macroeconómicas en el movimiento de los precios del índice bursátil y el índice de la variable target que es el precio del petróleo.
INDEX PRICE	Un **índice de precios** es un número índice calculado a partir de los precios y cantidades de un período. El más utilizado es el Índice de precios al consumo, que mide cómo evoluciona el gasto de una familia media.
LOG_INDEX PRICE	Es un registro, son datos logarítmicos, logaritmo, logaritmo natural de los precios para tener una mejor visualización de los datos y ver con detalles los periodos ej: log10 100= 2, ln 100= 4.56 => e**4.56=100
INFLATION RATIO	La tasa de inflación es el **coeficiente que muestra la variación porcentual de los precios de un determinado territorio, durante un periodo determinado**. La tasa de inflación recoge las variaciones que experimentan los precios en un periodo determinado, en un lugar determinado.
OIL PRICES	Precio del petróleo en el transcurso de los últimos 40 años el **West Texas Intermediate** (WTI) es una corriente de crudo producido en Texas y el sur de Oklahoma y es utilizado como punto de referencia en la fijación de precios del petróleo.
	El concepto de razón de cambio se refiere a la **medida en**

EXCHANGE RATIO	**la cual una variable se modifica con relación a otra**. Se trata de la magnitud que compara dos variables a partir de sus unidades de cambio. En caso de que las variables no estén relacionadas, tendrán una razón de cambio igual a cero. El cociente de las diferencias **∆y ∆x = f(x2) − f(x1) x2 − x1** se llama razón de cambio promedio de y con respecto a x en el intervalo [x1,x2] y se puede interpretar como la pendiente de la línea secante PQ. con respecto al tiempo).
GDP PERCENT	El PIB (**P**roducto **I**nterno **B**ruto) es un indicador económico que muestra el valor (en dinero) de todos los bienes y servicios en una economía. El **PIB** mide la riqueza económica de un país. Cuanto más crece este indicador, mayor es la capacidad de esa economía para generar empleo e inversión. PBI=C+G+I+(X-IM)
PERCAPITA INCOME	La **renta per cápita**, **PIB/PBI per cápita** o **ingreso per cápita** es un indicador macroeconómico de productividad y desarrollo económico, usado para entregar una visión respecto al rendimiento de las condiciones económicas y sociales de un país, esto en consideración del crecimiento real y la fuerza laboral. Generalmente también se utiliza como indicador de bienestar social. Es la relación que hay entre el PIB y la cantidad de habitantes de un país. Para obtenerlo, hay que dividir el PIB de un país entre su población.
UNEMPLOYMENT RATIO	La tasa de **Desempleo se** calcula como la el cociente entre la población desocupada (conjunto de personas que, sin tener ningún trabajo, buscaron uno en forma activa en la semana de referencia) y la población económicamente activa (conjunto de personas que tienen una ocupación o que sin tenerla la busca activamente).
MANOFACTURING OUT_PUT	La manufactura es la **actividad del sector secundario de la economía, también denominado sector industrial, sector fabril, o simplemente fabricación o industria**. El sector manufacturero está estrechamente relacionado con la ingeniería y el diseño industrial.
USTREASURY	El **Departamento del Tesoro de los Estados Unidos** (en inglés ,*United States Department of the Treasury*) es un departamento del Gobierno Federal de los Estados Unidos encargado de administrar el tesoro público de Estados Unidos. Creado mediante el acta de fundación por el Congreso en 1789 para recaudar apoyos económicos al gobierno inicial de Estados Unidos.
Hecho histórico	Corresponde a los hechos históricos mas importantes o mas relevantes por año desde el periodo de 1980-2020.
Close	Es el precio de cierre que informa la (API)yahoo financial.
Volume	Es el volumen de producción que informa la (API)yahoo financial.

8-EDA (análisis exploratorio de datos)

Recorremos el dataset que incorporamos de página “Kaggle” para hacer una visualización preliminar de los datos macroeconómicos, llamamos al dataframe “datos_económicos.csv” y vemos las características del mismo mediante un ‘head’, donde nos devuelve datos de los 9 países con variables macroeconómicas con una línea de tiempo comprendida en los últimos 40 años, para su posterior análisis que nos guie a encontrar patrones que nos ayuden con nuestro propósito que es la predicción del precio

del petróleo dirigido a accionistas para su posterior asesoramiento, evaluación en toma de decisiones de la compra de este activo financiero que representan una parte alícuota del capital social de diferentes compañías.

9-Data-Wrangling

#Se cambia la columna year de float a (int)
#Ajustamos a un decimal los valores
#Se agrega una 'variable sintética'que identifica el hecho mas relevante en ese año, que p ueda darnos indicios de un cambio.
#Se duplica la columna 'year' para luego reemplazarla por otros valores.
#Convierte la columna'Hecho_historico' y lo devuelve en un array aplanado para ser reem plazado numericamente.
#El presedente histórico se reemplaza con valores del 1-
9, como hecho mas relevante en el año en este diccionario.
#14 Variables o dimensión de nuestro análisis.
#tipos de datos mediante un (info).
#Datos object,float,int son los que comprende este dataset mediante un (dtypes).
#Tamaño del Dataset mediante un (shape).
#Vemos el producto de las filas x las columnas mediante un (size).
Se corrige y se remplaza los datos NaN de la columna 'manufacturingoutput' por (0).
#Limpieza del dataset para su posterior trabajo.
#Datos NULL Y NaN mediante un (isna),(isnull)para su posterior corrección.
#Se corrigen los datos Null y NaN de las siguientes variables y se reemplazan por (0).
#52 NULL Y NaN 'index price'
#Se corrigen los datos Null y NaN de las siguientes variables y se reemplazan por (0).
#2 NULL Y NaN 'exchange_rate'.
#Se corrigen los datos Null y NaN de las siguientes variables y se reemplazan por (0).
#43 NULL Y NaN 'inflationrate'.
#Se corrigen los datos Null y NaN de las siguientes variables y se reemplazan por (0).
#19 NULL Y NaN 'gdppercent'.
#Se corrigen los datos Null y NaN de las siguientes variables y se reemplazan por (0).
#21 NUUL Y NaN 'unemploymentrate'

#Se corrigen los datos Null y NaN de las siguientes variables y se reemplazan por (0).
#4 NUUL Y NaN 'tradebalance'.
#1 NUUL Y NaN 'manufacturingoutput'.

#Verificamos datos faltantes de nuestras varibles mediante un ´count´ en caso de borrar o agregar faltantes mediante un (count) con dato faltante.
#Verificamos datos duplicados.
#Datos enriquecidos.

10-Datos enriquecidos por medio de una (API) pública.

#Crude Oil
#Fuente
#https://finance.yahoo.com/quote/CL%3DF?p=CL%3DF
#Asignamos un nuevo dataframe
#Hacemos un discribe de esta nueva base de datos.
#Graficamos en histogramas todas las variables numericas del nuevo dataframe.
#Creamos un nuevo dataframe llamado (df_nuevo).
#Probamos la función merge para concatenar los dos dataframe he incorporar datos nuev os hasta la actualidad (2022).
#Creamos un nuevo dataframe ya concatenado llamado (result_total).
#Tamaño del Dataset concatenado.
#Hacemos un shape para saber el tamaño del dataset
#Vemos el producto de las filas x las columnas.
#Incorporamos, rellenamos datos faltantes a los años (year).
#Verificación la transformación del dataframe ´result_total´.
#tipos de datos mediante (info).
#Verificamos el tipo de dato mediante (dtypes).
#Verificamos datos NaN mediante (isna).
#Verificamos datos Null mediante (isnull).
#Eliminamos los campos que no necesitamos.
result_Total= result_Total.drop(columns= ['stock index'])

```
result_Total= result_Total.drop(columns=['country'])
result_Total= result_Total.drop(columns=['log_indexprice'])
result_Total= result_Total.drop(columns=['USTreasury'])
result_Total= result_Total.drop(columns=['Open'])
result_Total= result_Total.drop(columns=['High'])
result_Total= result_Total.drop(columns=['Low'])
result_Total= result_Total.drop(columns=['Dividends'])
result_Total= result_Total.drop(columns= ['Stock Splits'])
#Agregamos como indice los años 'year'.
#Ajustamos a un decimal los valores
#Completamos datos faltantes he incorporamos la std, median y (0)
según el caso, debido a los outliers.
#Se cambia la columna percapitaincome de float a (int).
#Se cambia la columna Volume de float a (int).
```

11-Análisis macroeconómico de los 9 países.

```
#Primer análisis macroeconómico de los 9 países.
#Análisis Uni variado
#Gráfico de líneas que
nos indican los movimientos en el indice al consumidor de los ultimos 40 años (1980-
2020).
#Gráfico nos indican los movimientos en el indice al consumidor de los ultimos 40 años
(1980-2020), transformados a logaritmos para ver movimientos precisos.
#Hacemos un describe para saber los datos estadísticos.
#Creación de histogramas para saber si exíste una distribución normal.
#Agrupamos las variables a utilizar
('year','country','percapitaincome','oil prices','inflationrate','index price','tradebalance','manu
facturingoutput','Hecho_historico').

#Graficamos en histogramas todas las variables numericas del nuevo dataframe.
#Hecho histórico variable sintética.
#Corresponde a los Hechos Historicos que tuvieron lugar en los ultimos 40 años 1) GUER
RA. 2) CAIDA DE MERCADOS. 3) CAIDA DE REGIMEN. 4) CAMBIO POLITICO Y MONE
```

TARIO. 5) CATASTROFE. 6) CAIDA DEL PRECIO DEL PETROLEO. 7) TRATADOS Y A CUERDOS ECONOMICOS. 8) TERRORISMO. 9) PANDDEMIA GLOBAL.
#Inflación acumulada de los últimos 40 años.
#Promedio de Salarios Por País.
#Agrupamos en un nuevo DataFrame df_2 y analizamos los hechos historicos reemplazan dolos por strings para tener unas etiquetas más visibles en nuestro gráfico.
#Agrupamos las variables a utilizar.
#Reemplazamos mediante un diccionario.
#Maximo de precios según el Hecho Histórico y cuanto influye en el precio del petróleo.
#Cantidad de eventos por Hecho_historico.
#Cantidad de veces que la inflación estuvo en un determinado valor durante los últimos 40 años.
#Evolución del índice Per cápita he índice de Precios.
#Inflación.
#Índice Per Cápita.
#Evolución del Precio del Petróleo.
#Precios del consumo (Familias).
#Balanza Comercial-Relación entre las Importaciones y las exportaciones.
#Producción Manufactura/mes
#Análisis Multivariado del Petroéo, Precio al consumidor y Hecho Historico.
#Vemos que los outliers por alto consumo y precio del petróleo promedio, corresponde a u n hecho bélico y otro a un proceso político monetario, mientras que el mayor precio del pe tróleo se da en hechos bélicos-
político y monetario en el transcurso de los últimos 40 años.
#FacetGrid para analizar el comportamiento.
#Análisis Multivariado Precio del Petróleo, Balanza comercial y los Hechos Históricos.
#Respecto a la balanza comercial se nota que la mayor cantidad de eventos vs mayor pre cio del petróleo corresponde a caidas de mercados, seguido de eventos bélicos.
#FacetGrid para analizar el comportamiento.
#Análisis Multivariado Precio del Petróleo, inflación y los Hechos Históricos.
#Vemos que
uno de los patrones mas altos que corresponden a muy alta inflación se da con un precio muy bajo del petróleo y los valores mas altos del petróleo corresponden catástrofes.
#FacetGrid para analizar el comportamiento.

#Análisis Multivariado Precio del Petróleo, Per cápita y los Hechos Históricos.
#Aqui vemos que uno de los valores mas altos per capita y precio promedio del petróleo c
orresponde a hechos bélicos y cambios políticos y monetarios, como así también
mayor precio del petróleo.
#FacetGrid para analizar el comportamiento.
#Análisis Multivariado Precio del Petróleo, manufactura y los Hechos Históricos.
#En el caso de la manofactura también se ve que los altos niveles de producción correspo
nden a hechos bélicos, cambios políticos y monetarios y a mayor precio del petróleo pred
omina los hechos bélicos y y las catástrofes.
#FacetGrid para analizar el comportamiento.
#Creamos un Pairplot del dataset.
#Pairplot según los hechos históricos.
#Correlación de Variables Económicas del primer dataset (datos_economicos).
#Vemos cuales de las variables se correlacionan mas que otras para la posterior eliminaci
ón.
#Correlación del segunto DataFrame concatenado (result_total).
Al explorar la heatmap del dataset concatenado vemos que las correlaciones por encima del 40% corresponden al precio del del petróleo con el índice per cápita y el hecho_histórico y por encima del 50% la manufactura con la balanza comercial, también notamos que las variables incorporadas 'Close' y 'Volume' tienen correlación negativa con respecto a las variables del primer dataset.
#Hacemos un JoinPlot de todas las variables del dataframe con respecto al precio del petr
óleo.
#JoinPlot para analizar el comportamiento.
#Vemos la relacion con las variables independiente,
calificaciones de lista, de los nuevos datos incorporados.
#para poder ver todos los datos menos outliers. modificamos un cierre de precio mayor a 30 USD.
#Vemos la relacion con las variables independiente, y notamos que a menor volumen de p
etroleo mayor es precio.
#De este análisis surgen dos variables “Close” y “Volume” para incorporar a nuestro mode
lo, para sumar a un futuro análisis más completo.

12-Estandarizacion de los datos.

#Escalamiento, Normalización y Estandarización para no utilizar los outliers.
#Agrupamos las variables a utilizar.
#Graficamos los datos en escala original.
#Como se distribuyen los datos.
#Agrupamos variables para un nuevo dataframe llamado (df_agrup2).
#Renombramos la columna “oil Price# por “Oil” que vamos a utilizar.
#Se cambia la columna close de float a (int).
#Se cambia la columna Volume de float a (int)
#Decidimos sacar Index price de nuestro dataset para un mejor análisis y entrenamiento.
#Homogeneización de los datos, dataframe resul_total.
#Escalar en función de la Normalización del mínimo y máximo.
#Normalización en función de la norma del vector.
#Estandarización (desv_std=1, media=0)
#StandardScaler utiliza la media de centralidad, Consiste en cambiar la distribución de los datos para que tengan una media= 0 y un std= 1, para aplicar mejor el ML por que trabaja bajo el supuesto de que los datos se distribuyen de forma normal.
Por lo tanto, nuestros datos están más distribuidos.
#Análisis Multivariado del Petroéo, Precio al consumidor y Hecho Historico.
#Vemos que los outliers por alto consumo y precio del petróleo promedio, corresponde a u n hecho bélico y otro a un proceso político monetario, mientras que el mayor precio del pe tróleo se da en hechos bélicos-
político y monetario en el transcurso de los últimos 40 años.
#FacetGrid para analizar el comportamiento.
#RobustScaler utiliza el rango intercuartílico por lo tanto nuestros datos van a estar menos ditribuidos y de esa manera podemos sacar datos atípicos.
Vectores con Escalamiento estándar y el escalamiento Robusto.

13-Cross-validation (o Validación Cruzada)

#Fase 1
#Random Forest (Bosque Aleatorio) para Clasificación con Scikit-Learn y Python.
#Fase 1
#Aleatoriedad, selección de características para divisiones.
#Creación de bosque aleatoreo, modelo de clasificación supervisado.

14-Modelo de Regresión lineal.

#Fase 2

#Entrenamiento de los datos.

#Se utiliza el 80% de los datos para entrenamiento y 20% de los datos para la validación de los datos.

#Se entrena sin columna de la variable a predecir oil prices.

#Modelo de regresión.

#Vemos que se ha entrenado este modelo con los datos de test.

#**Predijo [52.50829063].**

#Mean Absolute Error: 1.887670409916462

#Mean Squared Error: 39.25363343356687

#**Root Mean Squared Error: 6.26**

#**mape: 5.618432366017228**

#r2_score: 0.31951174179891795

#Generamos un nuevo evento del precio del pretróleo para corroborar la capacidad del m odelo.

#**Predijo [52.50829063].**

#Preparamos datos

#Empezamos a crear nuestro modelo

#Entrenando el modelo

from sklearn.linear_model import LinearRegression

#Recuperamos la intersección

26.504356309520258

#La pendiente

[-0.00365698]

#Hacemos nuestras predicciones

[26.00682439, 26.21593041, 26.2907522, ..., 26.37691062, 26.28219487, 26.19749924])

#Convertimos en df la salida

#Evaluación del modelo:

#mape: 6.059737223519276

#Evaluamos el rendimiento del algoritmo. Este paso es particularmente importante para comparar qué tan bien funcionan los diferentes algoritmos en un conjunto de datos en particular.

#El error absoluto medio (MAE) El error cuadrático medio (MSE) Root Mean Squared Error (RMSE)

#Mean Absolute Error: 1.9383842411457666
#Mean Squared Error: 41.88946621767839

#Root Mean Squared Error: 6.472207213746976

REGRESIÓN LINEAL MÚLTIPLE:

#En la regresión lineal multivariable, el modelo de regresión tiene que encontrar los coeficientes más óptimos para todos los atributos.

#Mean Absolute Error: 2.0987024944291077
#Mean Squared Error: 56.346090533781165

#Root Mean Squared Error: 7.506403302100225

#r^2 : -0.00046557665470614573
-0.00046557665470614573

#RMS Debe ser menor para indicar si funciona el modelo.
#PCA dataframe
#Verificamos el coef.

interseccion (b) 5.506706202140776e-13
Pediente(m) [1.]
Diagrama de dispersion.
#datos=pd. DataFrame(datos)
#Vemos que tiende a aumentar la energía a mayor producción.
#from sklearn.preprocessing import StandardScaler

Agrupamos y creamos un nuevo dataframe (df_agrupar)

#Matriz de covarianza, cálculo de covarianza y visualización.

Metodo cov
#Eigenvalores & Eigenvectores
#Calcular eigenvalores y eigenvectores

#Visualizacion
#Grafica de los datos
#Graficando los Eigenvectores

15-Matriz de Confusión:

Métricas sobre las etiquetas de datos reales vs Datos predichos.

#confusion_matrix

Esta función toma dos conjuntos de etiquetas y devuelve una matriz de confusión, que es una tabla que muestra cuántas veces se predijeron correctamente o incorrectamente cada clase

#from sklearn.metrics import accuracy_score#Esta función toma dos conjuntos de etiquetas y devuelve la precisión, que es la proporción de predicciones correctas sobre el total de predicciones
#from sklearn.metrics import precision_score#Esta función también toma dos conjuntos de etiquetas y calcula la precisión de clase, que es la proporción de predicciones correctas de una clase sobre el total de predicciones de esa clase.
#from sklearn.metrics import recall_score# Es una métrica que se puede usar para evaluar el rendimiento de un modelo de clasificación. Se define como el número de predicciones verdaderas positivas realizadas por el modelo dividido por el número total de etiquetas positivas en los datos.
#from sklearn.metrics import f1_score.

16-Futuras agregaciones al proyecto.

Para ampliar el proyecto a futuro, para la ayuda de nuestros clientes se pensó en la realización de una (App) que este haciendo un relevamiento en tiempo real sobre este commodities en los principales sitios web financieros, incorporando hechos históricos de relevancia, como noticias de los mercados y cambios geopolíticos en los principales medios de comunicación para tener información precisa y amplia, con chat de ayuda para nuestros clientes sobre herramientas analíticas y pedidos de informes, como así también tablas y archivos csv, etc.

16-Conclusión:

Al incorporar nuevas variables que pertenecen a un (API) de Yahoo Financial, una variable sintética 'Hecho_histórico', sirvió para predecir mucho mejor a nuestra variable target (oil prices), el trabajo consistió una vez hecho el EDA y Data Wrangling, sumar informacion robusta, hicimos limpieza, eliminación, transformaciones y rellanamos datos faltantes, el enriquecimiento de los datos permitió sacar mejores conclusiones para nuestro modelo, utilizamos Regresión lineal para luego verificar el Coef de varianza, Matriz de covarianza, cálculo de covarianza y visualización, #Eigenvalores & Eigenvectores, estandarizar, graficar y aportar con el PCA el % de varianza que explica cada elemento. Primeramente analizamos el primer dataset y notamos que mediante los outliers y los movimientos del petróleo, sobre todo las variables macroeconómicas, la incorporación de la variable sintética "hechos_históricos"(Hechos Bélicos, cambios monetarios, tratados internacionales, pandemia, catástrofes) nos muestra que la manufactura(producción),la inflación el índice per cápita como conjunto de variables influyen en el precio del petróleo sumado al segundo análisis con la incorporación del cierre de precio (Close) y el volumen (Volume) que son factores determinantes en la suba y baja de este commodities. Luego de corroborar cuantas variables íbamos a utilizar decidimos eliminar y quedarnos con una agrupación de variables, estas mismas fueron utilizadas para el modelo. Al incorporar nuevos datos al modelo se logró una predicción bastante cercana a la realizad que ante nuevos eventos el modelo puede predecir muy bien para futuros reportes a nuestros clientes, con un seguimiento mediante una (app) que este actualizada con los mercados y los hechos geopolíticos en tiempo real, más aún tener certeza a la hora de invertir en este commodities tan particular.

Fuentes:

Kagge: 9 Countries (1980-2020).csv

API Pública: https://finance.yahoo.com/quote/CL%3DF?p=CL%3DF

Herramientas usadas:

Word, PDF, Excel, Power Point,

Entorno de uso: Google Colab.