**## :minidisc: Exploración de datos :minidisc:**

**#### :scroll:Resampling y lectural del icv**

Inicialmente observamos que la serie de tiempo del índice de cartera vencida (icv) no está en la misma frecuencia temporal que las variables suministradas, puesto que el icv está mensual y las variables están trimestrales.

:exclamation: Por lo anterior, nos vemos obligados a tomar la primer decisión importante en este proyecto: **\*\*¿cual de las dos bases de datos modificar?\*\*** ya que lo más conveniente es que ambas series de tiempo tengan la misma frecuencia temporal:hourglass:. Para el caso de las variables, al ser datos macroeconómicos trimestrales, es más dificil convertirlas a mensuales, por lo tanto se decide realizar un **\*\*resampling\*\*** al icv, para convertirlo de mensual a trimestral.

![](figuras\_y\_tablas/figura\_1.PNG)

**\*\*Figura 1\*\*** se carga la información sobre el icv en **\*\*indice\_df\*\*** y luego se realiza el resampling o cambio de frecuencia temporal a trimestral, y se guarda en el dataframe **\*\*quarterly\_resampled\_indice\_df\*\***

**\*\*NOTA:exclamation::\*\*** inicialmente, se entrenarán los modelos con datos trimestrales, ya que así se tendrán más datos y esto mejora la robustez del modelo; luego, más adelante en este documento, se mostrará cómo *\*se calcula el valor proyectado del icv del sistema financiero a un año\**

**#### :book: Lectura y ajuste de variables**

:green\_book: A continuación se carga la base de datos de variables y se eliminan las 3 primeras filas y la última, esto con el fin de que puedan ser agrupadas correctamente con los datos del icv y formar un sólo dataframe y así explorar más fácil los datos.

**\*\*El objetivo es predecir cual será el icv promedio de los próximos 3 meses usando predictores o variables de la fecha actual.\*\***

![](figuras\_y\_tablas/Figura\_2.PNG)

**\*\*Figura 2\*\***

:green\_book: Luego, se deben unir los datos del ICV con las variables en un solo dataframe, además de agregar una columna numérica incremental para posteriores análisis, también se usa el comando **\*\*info()\*\*** para conocer los tipos de datos de las columnas y presencia de valores nulos

![](figuras\_y\_tablas/Figura\_3.PNG)

**\*\*Figura 3\*\***

:green\_book: De la figura 3 se puede inferir que hay 4 columnas con valores nulos que deben ser modificados posteriormente mediante imputación.

**#### :bar\_chart:Gráfico de correlaciones entre variables:chart\_with\_upwards\_trend:**

:green\_book: Usando el comando **\*\*sns.pairplot()\*\*** se genera el siguiente gráfico, el cual muestra en gráficos de disepersión, la correlación existente entre todas las variables, inclusive el icv.

![](figuras\_y\_tablas/Figura\_4.png)

**\*\*Figura 4\*\***

:green\_book: De la figura 4 se observa que la variable **\*\*icv\_cartera\_total\*\*** tiene una correlación positiva con la variable **\*\*Desempleo\*\***, :bangbang:**\*\*pero su correlación con las demás variables no es fuerte\*\***:bangbang:, hecho que se evidencia más adelante en este proyecto cuando se analicen las **\*\*relaciones y coeficientes del modelo de regresión lineal\*\***, el cual es uno de los varios modelos posteriormente entrenados.

**## :memo:Imputación usando Regresión lineal:fountain\_pen:**

:green\_book: En la Figura 4 también se puede observar que las variables **\*\*PIB\*\*** e **\*\*IPC\*\*** tienen una alta correlación positiva con el tiempo (a medida que pasan los años, el valor del PIB e IPC aumenta); por esta fuerte correlación, se ha tomado la decisión de usar un modelo de Regresión Lineal para reemplazar los datos faltantes en estas dos columnas; para esto, se desarrolló la función **\*\*regression\_imputer()\*\*** la cual hace el proceso de imputación automático.

:green\_book: El código de la función se muestra a continuación, en dónda básicamente se entrena el modelo con los valores de la columna que no tengan valores nulos y luego de entrenado el modelo, éste se usa para predecir y sustituir los valores nulos por las predicciones realizadas por el modelo.

![](figuras\_y\_tablas/Figura\_5.png)

**\*\*Figura 5\*\***

:green\_book: Y a continuación se enseña cómo, usando la función **\*\*regression\_imputer()\*\*** y un ciclo for, se reemplazan los valores nulos de las columnas **\*\*IPC\*\*** y **\*\*PIB\*\***.

![](figuras\_y\_tablas/Figura\_6.png)

**\*\*Figura 6\*\***

**## :pencil2:Dividir datos entre 'entrenamiento' y 'testeo':scroll:**

:arrow\_right:Esta etapa es crucial en el desarrollo de cualquier modelo predictivo de Machine Learning, ya que **\*\*NUNCA\*\*** debe ser usado el 100% de los datos disponibles, sino realizar una partición y utilizar una parte en entrenar el modelo y la otra en testearlo para evaluar éste cómo se comporta haciendo predicciones con datos desconocidos.

**\*\*NOTA:exclamation::\*\*** **\*\*esta etapa es fundamental para garantizar la robustez estadística de cualquier modelo\*\***

**\*\*NOTA 2:exclamation::\*\*** **\*\*Existen métodos más sofisticados de entrenamiento de modelos (CROSS VALIDATION) y optimización de hiperparametros:exclamation: (GRIDSEARCHCV), pero debido a la poca cantidad de datos disponibles para entrenar los modelos en este proyecto, se ha optado por una única partición de datos 'entrenamiento' - 'testeo'\*\***

:arrow\_right:A continuación, se procede a guardar en **\*\*x\*\*** la matriz de predictores de nuestros modelos y en **\*\*y\*\*** el vector de etiquetas o **\*\*'labels'\*\***, en este caso el icv, y a partir de estas dos bases de datos, creamos las particiones de **\*\*'entrenamiento'\*\*** y **\*\*'testeo'\*\*** usando la función de sklearn **\*\*train\_test\_split()\*\***.

![](figuras\_y\_tablas/Figura\_7.png)

**\*\*Figura 7\*\***

:arrow\_right:Como último paso antes de entrenar modelos, se deben imputar las columnas 'Exportaciones' e 'Importaciones'; en este caso usaremos los valores del promedio de la columna y la función **\*\*SimpleImputer()\*\*** de **\*\*sklearn\*\***, ver siguiente figura.

![](figuras\_y\_tablas/Figura\_8.png)

**\*\*Figura 8\*\***

**## :brain:Entrenamiento y evaluación de modelos:robot:**

:arrow\_right:Se decidió, en esta fase, entrenar y evaluar el desempeño de tres modelos de regresión, los cuales son:

    :bulb:RandomForestRegressor()

    :bulb:LinearRegression()

    :bulb:DecisionTreeRegressor()

:arrow\_right:El desempeño de cada modelo se calculó usando el **\*\*Error Medio Absoluto\*\***, el cual indica **\*\*En promedio, las predicciones del modelo difieren 'tantas' veces del valor real\*\***. A continuación se muestra un ciclo for, el cual itera sobre los 3 modelos y genera el error que cada uno produce evaluado en los datos de 'testeo'

![](figuras\_y\_tablas/Figura\_9.png)

**\*\*Figura 9\*\***

**## :notebook:INTERPRETACIÓN COEFICIENTES DEL MODELO DE REGRESIÓN LINEAL:mag\_right:**

:arrow\_right:Pese a que el modelo de regresión lineal **\*\*no fue el modelo que mejor capacidad predictiva tuvo\*\***, éste sin duda, es el modelo que mayor información nos entrega acerca del problema que queremos solucionar, puesto que **\*\*el valor de los coeficientes indica el grado de importancia que tiene cada variable a la hora de predecir la variable objetivo\*\***, en este caso el icv. A continuación se muestra el desarrollo en código del modelo de regresión lineal, junto con los valores de los coeficientes de cada variable.

![](figuras\_y\_tablas/Figura\_Regresion\_lineal.PNG)

**\*\*Figura regresión lineal\*\***

:arrow\_right:De la figura anterior, se puede inferir que la variable que más influencia tiene a la hora de predecir el icv es **\*\*Desempleo\*\***; inclusive, desde un punto de vista simplificado, se podría decir que **\*\*es la única variable que tiene un impacto considerable a la hora de predecir el icv\*\***, ya que su coeficiente es **\*\*varios ordenes de magnitud\*\*** mayor que los coeficientes de las demás variables.

**## Valor proyectado del icv a un año :chart\_with\_downwards\_trend:**

**\*\*NOTA:exclamation::\*\*** El código para la predicción del icv se encuentra en el archivo **\*\*icv\_anual.ipynb\*\***

:arrow\_right:Para realizar la predicción del icv a uno año, se debe realizar un 'resampling' similar al realizado previamente, pero esta vez no trimestral sino anual; adicionalmente este cambio de frecuencia temporal se debe realizar sobre las dos series de tiempo 'icv' y 'variables'.

:arrow\_right:El procedimiento de predicciones anuales no contiene nada diferente al procedimiento de predicciones trimestrales, la decisión de haber mostrado el procedimiento de desarrollo de modelos para frecuencias trimestrales es porque **\*\*con frecuencias trimestrales disponemos de más datos y por ello los modelos alcanzan mejores desempeños\*\***.

:arrow\_right:Para la predicción anual se eligió el modelo **\*\*RandomForestRegressor()\*\*** porque éste fue el que mejor desempeño tuvo en las predicciones trimestrales; el error encontrado en las predicciones anuales fue **\*\*0.022005\*\***, valor mayor al error trimestral, **\*\*puesto que se disponen de menos datos para entrenar el modelo\*\***. La construcción del modelo se enseña a continuación.

![](figuras\_y\_tablas/Figura\_10.png)

**\*\*Figura 10\*\***

:green\_book:Para realizar la prediccion del icv anual del 2021, se utilizó el siguiente fragmento de código, y el resultado de esta predicción fue de **\*\*0.04638 = 4.638%\*\***

![](figuras\_y\_tablas/Figura\_11.png)

**\*\*Figura 11\*\***

:green\_book:Contrastando el valor predicho del icv con los valores reales, se encuentra poca diferencia, hecho que aumenta la credibilidad de los modelos acá entrenados; el artículo de **\*\*la república\*\***, titulado **\*\*Estos son los bancos con mayores y menores índices de cartera vencida a julio de 2021\*\*** indica que para Julio del 2021 se tenía para Bancolombia un icv de 4,6%, valor similar al predicho por el modelo; pese a que sólo se tiene el valor de los primero 6 meses del icv para el 2021, es un estimativo significativo sobre la efectividad del modelo acá desarrollado.

**\*\*NOTA:exclamation:: la fuente del artículo de la república, se encuentra en el archivo de texto plano adjunto al proyecto\*\***

**### :books:Variables adicionales en la predicción del icv:label:**

:green\_book:A la hora de considerar otras variables con capacidad predictiva para el icv, habrá que tener en cuenta que el icv es una **\*\*serie de tiempo\*\*** y por ende, se puede aprovechar la **\*\*teoría de modelación de series de tiempo\*\*** **\*\*(:bangbang: ARIMA & SARIMA :bangbang:)\*\***, la cual emplea fuertemente los valores pasados o históricos de la serie de tiempo para predecir los futuros valores. por ejemplo, unas de las herramientas más usadas en series de tiempo son **\*\*medias móviles simples\*\*** y también **\*\*medias móviles exponenciales\*\***.

:green\_book:Por lo anterior, yo recomendaría inferir nuevas variables a partir de los valores pasados del icv, variables como:

:herb: media móvil simple

:herb: media móvil exponencial

:herb: indice anterior

:herb: desviación estandard de la serie de tiempo 'n' periodos atrás

:maple\_leaf:**\*\*Se pueden inferir ilimitados posibles predictores en base a los mismos datos históricos\*\***, un ejemplo de uso extensivo de la teoría de predicción de series de tiempo son los modelos de trading financiero, los cuales analizan series de precios y deciden en base a ello tomar alguna decisión en el mercado.