**Indice**

* Objetivo
* Desarrollo del trabajo
  + Limpieza de datos y preparación del dataset
  + Reducción de dimensionalidad
  + Aprendizaje supervisado: regresión multiple
  + Aprendizaje supervisado: árbol, KNN y regresión logistica
  + Análisis estadístico
  + Aprendizaje no supervisado
* Métricas y resultados
* Conclusión

**Objetivo**

Se trabajó sobre el CSV Global\_Economy que reúne los datos macroeconómicos de todos los países del mundo desde 1970 hasta 2021.

**Identificar las variables que pueden permitir que Argentina pase a ser un país rico.**

**Desarrollo del trabajo**

**Limpieza da datos y preparación del dataset**

* Se estandariza el Df seleccionado y se eliminan nulos
* Se corre prueba dtype “antes” y “después” de encodear datos tipo “object” a través del label encoder. En el caso particular de los datos “AMAexchangerate” y “IMFbasedexchangerate”, se detectan strings mezclados, por lo que se aplica el método regex para dejar nulo todo lo que no sea número.
* Se verifican nulos a través de los métodos suma de nulos, df.info y a través de la librería missingo.
* Se completan nulos a través del metido ffill

**Reducción de dimensionalidad**

* Se utiliza PCA para reducción de dimensionalidad y se grafican los resultados en un gráfico de barras
* Se corre método “Forward Selection” como ejercicio complementario al PCA para establecer una adecuada “featured selection”
* Como ejercicio didáctico, se escalan variables clave a través del método MinMax Scaler

**Aprendizaje supervisado: regresión múltiple**

* Se corre regresión múltiple para el “Percapita GNI” de todos los países con múltiples variables del Df
* Se corre regresión múltiple para el ‘Percapita GNI’ del país Argentina utilizando como “x” solamente las variables ‘Exportaciones y ‘Desarrollo Industrial’
* Se realiza gráfico de puntos para la recta de regresión de ‘Exportaciones y ‘Percapita GNI’
* Se realiza gráfico de puntos en 3 D para la correlación entre ‘GNI en USD’, ‘Exportaciones y ‘Desarrollo industrial’
* Se corre regresión múltiple para el ‘Percapita GNI’ del país Argentina utilizando como “x” solamente las variables ‘Población’ y ‘GNI en USD’
* Se realiza gráfico de puntos para la recta de regresión de ‘Población’ y ‘Percapita GNI’
* Se realiza gráfico de puntos para la recta de regresión de ‘GNI en Usd’ y ‘Percapita GNI’
* Se realiza gráfico de puntos en 3 D para la correlación entre ‘GNI en USD’, ‘Población’ y ‘Percapita GNI’

**Aprendizaje supervisado: árbol, KNN y regresión logística**

* Se crea una variable “tipo”, tipo de dato str, que clasifica a los países en pobre, medio pobre, medio y rico, según los valores de su medida de posición estadística en el set de datos (cuartil 1,2,3 y 4). Se realiza esta clasificación a titulo didáctico para poder aplicar el algoritmo “árbol de decisión"
* Luego de entrenar los datos con “árbol de decisión”, se encodean los datos de la variable “tipo” en forma manual (a través de un replace) y se aplica “KNN”.
* Luego se vuelven a preparar los datos para entrenar el sistema con “Regresión Logística”. Se aplica “matriz de confusión” para evaluar la predicción de este último.
* Se evalúa “accuracy” de la “Regresión Logística” para “Exportaciones”, “Desarrollo Industrial” y “Construcción” como variables independientes para predecir el “tipo” de país.

**Análisis estadístico**

* Se revisan valores típicos y atípicos utilizando Zcore para todos los países del dataset y para los datos del país Argentina.
* Se grafican valores típicos y atípicos del Zcore para el mundo.

**Aprendizaje no supervisado**

* Se aplica algoritmo Agglomerative clustering en 4 clusters para visualizar segmentaciones de la variable “tipo” en relación al “PercapitaGNI”

**Métricas y resultados**

* La PCA identifica que el 91% de la varianza del dataset se explica por las variables ‘Población’ y ‘Percapita GNI’.
* Además, el ‘Percapita GNI’ es la relación entre las variables ‘Población’ y ‘GNI en USD’
* La aplicación del método “forward selection” identifica las siguientes características como relevantes:

['Exportsofgoodsandservices','ManufacturingISICD',

'Population', 'ConstructionISICF', 'Importsofgoodsandservices', 'Wholesale,retailtrade,restaurantsandhotelsISICGH',

'HouseholdconsumptionexpenditureincludingNonprofitinstitutionsservinghouseholds', 'OtherActivitiesISICJP',

'GrossNationalIncomeGNIinUSD', Finalconsumptionexpenditure',

'Mining,Manufacturing,UtilitiesISICCE', 'TotalValueAdded']

* La regresión múltiple para predecir el comportamiento de ‘Percapita GNI’ con todas las variables arroja un R2 de 0,165, un F-Statistic de 258, un skew de 5 y una curtosis de 47
* La regresión aumenta la correlación entre las variables y el carácter normal de la distribución cuando se utilizan las variables ‘Población’ y ‘GNI en USD’ para predecir el ‘Percapita GNI’ y se aíslan los países para su análisis. Para Argentina el OLS arroja un R2 de 0,987, un F-Statistic de 1937 un skew de 0,47 y una curtosis de 2,3. Pero estas métricas son indicio de que el modelo está “sobreajustado”, precisamente por su simpleza. La relación entre “Población” y “GNI” es igual al “PercapitaGNI” por definición.
* Utilizando la característica Exportsofgoodsandservices el R2 del OLS para la predicción del ‘Percapita GNI’ de Argentina queda en 0,7
* Sumando la característica ManufacturingISICD como variable independiente el OLS muestra métricas positivas: R2 de 0,96, F-Statistic de 634, Skew de 0,7 y curtosis de 2,5

Por otro lado, a fines didácticos se trabajan los siguientes algoritmos:

* El árbol de decisión se entrena con el 30% de los datos y expone los 4 tipos de observaciones:
  + 1861 registros con ‘Percapita GNI’ mayor o igual a 8.906 (P75, o “piso del Q4”),
  + 1837 registros con ‘Percapita GNI’ mayor o igual a 2.359 y menor a 8.906 (piso del Q3 -P50- a “techo del Q3/piso del Q4” – P75-),
  + 1840 registros con ‘Percapita GNI’ mayor o igual a 729 y menor a 2.359 (piso del Q2 -P25- a “techo del Q2/ piso del Q3” – P50-)
  + 1820 registros con ‘Percapita GNI’ menor a 729 (hasta “piso del Q2/techo del Q1” -P25)
* El primer grupo responde al tipo “rico”, el siguiente al “medio”, el siguiente al “medio pobre” y el ultimo al “pobre”
* El accuracy score es 1, justamente por la simpleza del modelo, lo que es indicio de “overfitting” (al segmentar la característica tipo de país por la posición estadística del ‘PercapitaGNI’)
* El accuracy score del KNN para k=1 también es de 1, por la misma razón.
* Lo propio sucede con “Regresión Logística”, según puede verse en la Matriz de confusión, cuando se utiliza la característica “Percapita GNI” para predecir “tipo”.
* Probada la “Regresión Logística” con las características “Exportaciones”, “Desarrollo industrial” y “Construcción”, cada uno arroja un accuracy de 0,25 por tratarse de valores de múltiples países con distintas estructuras productivas, conteniendo además paises con valores “outliers” (top 30 mega ricos, como evaluado en la primera presentación)
* El Z\_score para la distribución del ‘Percapita GNI’ de Argentina arroja valores típicos en un 100%
* Reforzando las conclusiones obtenidas en el OLS y en los gráficos de Boxplot (primera entrega) sobre el dataset, el Z\_score para el total de los países del mundo muestra valores atípicos

**Conclusión**

**Para la Argentina el aumento de las exportaciones y de la inversión en capacidad instalada para el desarrollo industrial correlacionan positivamente con el aumento del Producto interno Bruto per cápita (‘PercapitaGNI’). El OLS para predecir el ‘PercapitaGNI’ a partir de las exportaciones y el crecimiento industrial, es un algoritmo adecuado para evaluar la situación mencionada.**