Informe de resultados

Taller 2

Juan David Ayala

1. **Revisión del dataset.**
   Se cuenta con un dataset de países con información general de los mismos. La idea es determinar mediante una regresión lineal los principales componentes para su predicción.

   a. Limpieza de datos:

      Los datos se someten a una limpieza donde se buscan eliminar valores duplicados y y validar que los valores que tienen cada una de las columnas son los esperados.

      En este caso se encontraron 15 registros donde no se contaba con la variable objetivo o esta era nula.

      Los registros parecen tener valores y tipos de datos coherentes que no afectan su medición (numéricos, cadenas, etc.)

   b. Se identifica la variable **incomeperperson** como la variable objetivo.

2. **Inclusión de información externa.**
   Se agregó a la información existente el continente de los países y si es un país desarrollado o no. Para hacer la unión de estos datsets fue necesario hacer ajustes a la variable "country" como la capitalización del nombre y eliminación de ',' ya que en algunos casos no era posible cruzar los países.

   Una vez se incluyó la información externa se realizó la separación entre los sets de **entrenamiento** y **prueba**.

3. **Análisis univariado:**
   Al tener información numérica se realiza un análisis univariado con ayuda de y data profiling. Se encuentra algunas columnas con un número alto de valores nulos y una indicación de alta correlación entre variables. Esto se evaluará más adelante.

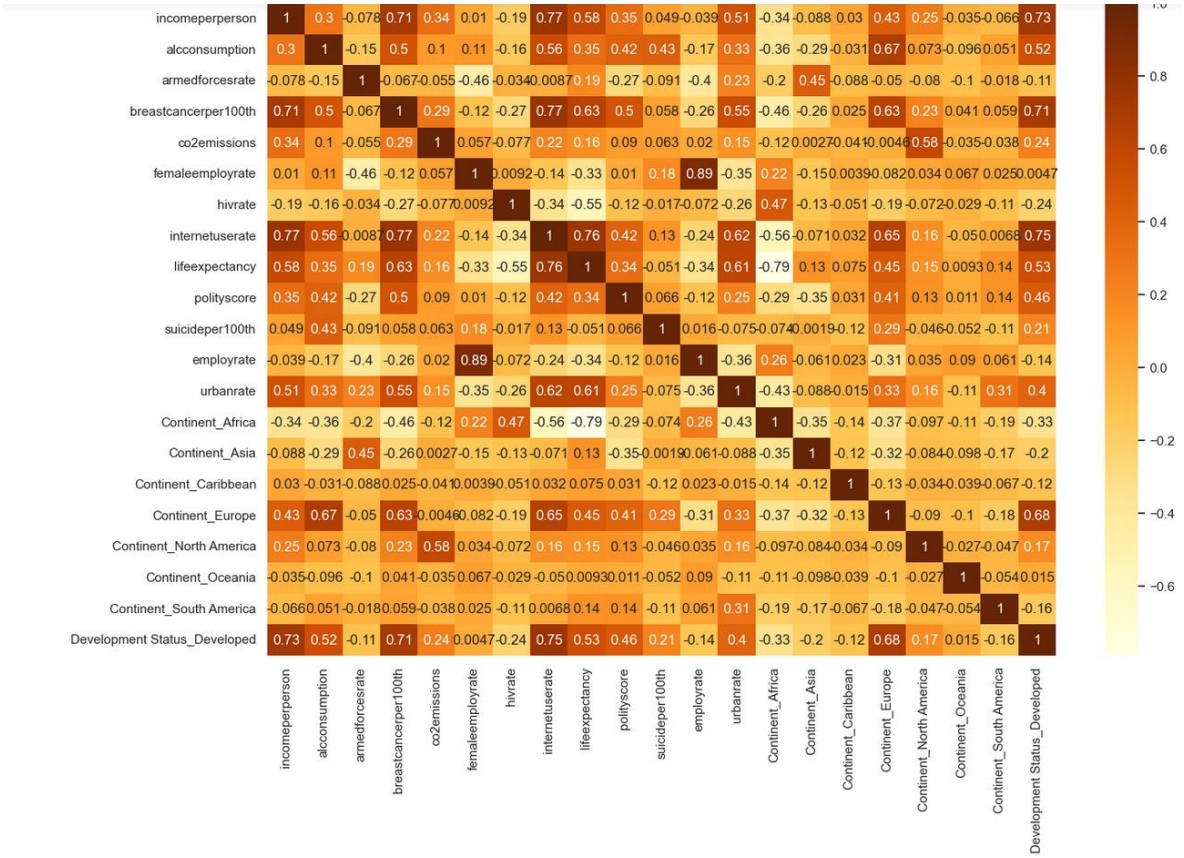
4. **Imputación de valores nulos.**
   En este caso se realizó una imputación por K-vecinos cercanos. La decisión se tomó teniendo en cuenta que usualmente los países son similares por regiones y comparten características. De esta manera se obtiene un modelo mejor ajustado que si se hace la imputación por medias o medianas.}
   2 columnas (*oilperperson* y *relectricperperson*) fueron eliminadas del dataset por su gran cantidad de valores nulos (63.4% y 20.8% respectivamente.)

5. **One hot encoding de variables categóricas**

   Se realiza un one hot encoding de las dos columnas que se agregaron: continente y el estado del país (desarrollado, en desarrollo). La columna Country fue eliminada porque es única de cada registro y no nos ofrece mayor información adicional.

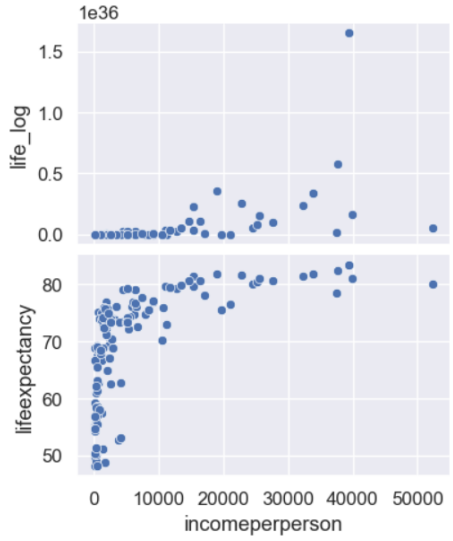
6. **Análisis multivariado.**

   Entre los análisis se realiza una correlación de los datos. Algunas variables tienen una alta correlación como employrate y femaleemployrate. En estos casos se eliminará alguna de las variables.



Adicionalmente se realiza un grafica de dispersión para determinar posibles correlaciones o transformaciones que se podrían realizar a laos datos y que no se observan con las correlaciones.

Se realiza una transformación al campo “lifeexpectancy” (exp(lifeexpectancy)) y se renombra life_log para recordar la transformación inversa que hay que realizar. Este campo se incluye para ver su comportamiento. Es importante tener en cuenta que es necesario hacer la transformación inversa a el coeficiente resultante.



7. **Posibles columnas importantes.**

   En este punto se definen las posibles columnas importantes:

   - internetusagerate: mayor uso de internet puede indicar una mayor capacidad adquisitiva que estaría correlacionada con el PIB
   - lifeexpectancy: una mayor esperanza de vida indica que hay condiciones donde la gente puede llegar a vivir más (buena salud, bajo crimen, etc.).
   - Development Status_Developed: la forma en la que se calcula si un país es desarrollado es mediante algunas medidas entre las que se encuentra si es desarrollado o no. Esto se ve en la tabla de correlaciones.
   - polityscore: se espera que si hay una fuerte democracia en un país
   - suicideper100th: una baja tasa de suicidios podría indicar una mejor calidad de vida.

8. **Entrenamiento:**

   Se entrena un modelo de regresión lineal y un modelo lasso para identificar si este logra eliminar algunas variables no importantes.

   También se realiza una normalización de los datos para poder determinar las variables principales.

   Se valida vs el dataset de test y se comparan los resultados:

	Algoritmo	MSE_train	RMSE_train	R2_train	MSE_test	RMSE_test	R2_test
0	Reg. lineal	20579073.12	4536.416330	0.791115	25561320.94	5055.820501	0.823698
1	Lasso	20579226.49	4536.433234	0.791113	25555499.99	5055.244801	0.823739

9. **Resultados:**

   Tanto la regresión lineal como la regresión lasso tuvieron un desempeño similar tanto en el se de entrenamiento como en el de prueba, sin embargo, la regresión lasso presenta una ligera mejoría en el RMSE del grupo de prueba. Las 5 características principales se presentan a continuación en orden.

variable	coeficiente
internetuserate	4698.898664
Development Status_Developed	2778.576253
alcconsumption	-1917.934565
employrate	1755.124799
urbanrate	1730.599211

Internetusagerate es la característica más importante seguido de el indicador de si el país es desarrollado o no.
El consumo de alcohol es importante, pero tiene una connotación negativa: entre más alto sea el consumo del alcohol, mayor el impacto en el PIB de dicho país.
Las otras 2 características son employrate y urban rate.

Urbanrate es una característica sobre la cual no se pueden realizar recomendaciones fácilmente y es necesario obtener más información, ya que puede estar ligada a otras características importantes que permitan identificar si hay alguna causa subyacente. Por ejemplo, existen países donde la migración de personas a la ciudad puede deberse a guerras o falta de oportunidades en el campo y simplemente recomendar promover esta migración sería erróneo.

Teniendo en cuenta estas características se realizarán las recomendaciones.

10. **Políticas:**
    Las siguientes políticas se recomiendan a los países:

    1. Identificar si el país tiene un alto consumo de alcohol, identificar sus causas e intentar reducirlo mediante campañas de educación o de salud.
       Esto puede influir en algunas otras métricas como la expectativa de vida e indicadores de salud.
    2. Incentivar la creación de puestos de trabajo formal con el fin de incrementar la tasa de empleo.
    3. Impulsar la educación digital de los ciudadanos (manejo de internet, ventajas, desventajas, **riesgos**) e impulsar carreras en el ámbito tecnológico. Esto podría abrir posibilidades económicas para los ciudadanos mediante empleos virtuales.