Taller 2 – Informe Ejecutivo

John Vicente Moreno Triviño - 202210162
MINE-4101: Ciencia de Datos Aplicada

1. Introducción

El presente documento contiene el informe ejecutivo del Taller 2 de la asignatura. El notebook anexo contiene todos los detalles y comentarios sobre el entendimiento y preparación de los datos, así como sobre el modelo implementado. Por lo tanto, este informe presentará de manera muy resumida estos tópicos, y se enfocará más bien en las conclusiones y recomendaciones de negocio para el caso de estudio.

2. Entendimiento del negocio

El Banco Mundial tiene como uno de sus objetivos misionales prestar apoyo financiero a países en desarrollo para la ejecución de proyectos que permitan mejorar las condiciones de vida de la población y fortalecer el aparato socioeconómico. Sin embargo, una de sus mayores preocupaciones es hacer lo posible para que el dinero sea invertido de forma eficiente en los países receptores, al tiempo que se garantiza la devolución de estos préstamos.

Por otra parte, los países en desarrollo saben que mostrar estabilidad y capacidad de pago facilitará que el Banco Mundial les asigne fondos para invertir en sus proyectos, al tiempo que deben demostrar que estos fondos se usarán de manera sabía y eficiente. Adicional a esto, tanto el Banco Mundial como los países en desarrollo son conscientes de que el PIB Per cápita es uno de los mejores indicadores que predicen la estabilidad social y económica de un país. Por lo anterior, es necesario encontrar cuáles son las principales variables que tienen una mayor relación con el PIB Per Cápita, de modo que los gobiernos de estos países puedan enfocar sus políticas públicas en la mejora de las áreas que resulten claves.

Objetivos de negocio:

- Generar recomendaciones, basadas en datos, sobre las políticas públicas que los gobiernos de los países en desarrollo deberían implementar para demostrar que son economías estables, enfocadas en el desarrollo sostenible y con la capacidad de pago para recibir fondos suficientes por parte del Banco Mundial.
- Ayudar al Banco Mundial a entender cuáles son las variables que mejor se correlacionan con el PIB Per cápita de una nación, de modo que sea posible diseñar políticas de préstamos mejor informadas, enfocándose en sectores clave para cada país.

Objetivos del proceso de análisis de datos con ML:

- Utilizar algoritmos de regresión lineal multivariable para entender la relación entre el PIB Per cápita de un país y otras variables socioeconómicas
- Entrenar y validar un modelo de regresión lineal de Machine Learning, comprendiendo el proceso de ajuste del modelo y la interpretación de sus coeficientes.

3. Hallazgos durante el entendimiento de los datos

Para revisar los detalles y el paso a paso de los procesos de entendimiento y preparación de datos, se solicita dirigirse al notebook anexo en este repositorio. Entre tanto, se presenta aquí los principales hallazgos:

Sobre el dataset:

Se cuenta con un dataset de 178x16. La primera de las columnas representa el país (tipo de variable "String"), mientras que las restantes 15 columnas son variables numéricas (Float) con diferentes indicadores socioeconómicos de cada país. -Existen 12 registros duplicados, los cuales se eliminarán en el proceso de preparación de datos.

Top 6 variables:

Figura 1. Mapa de calor con las correlaciones entre las variables

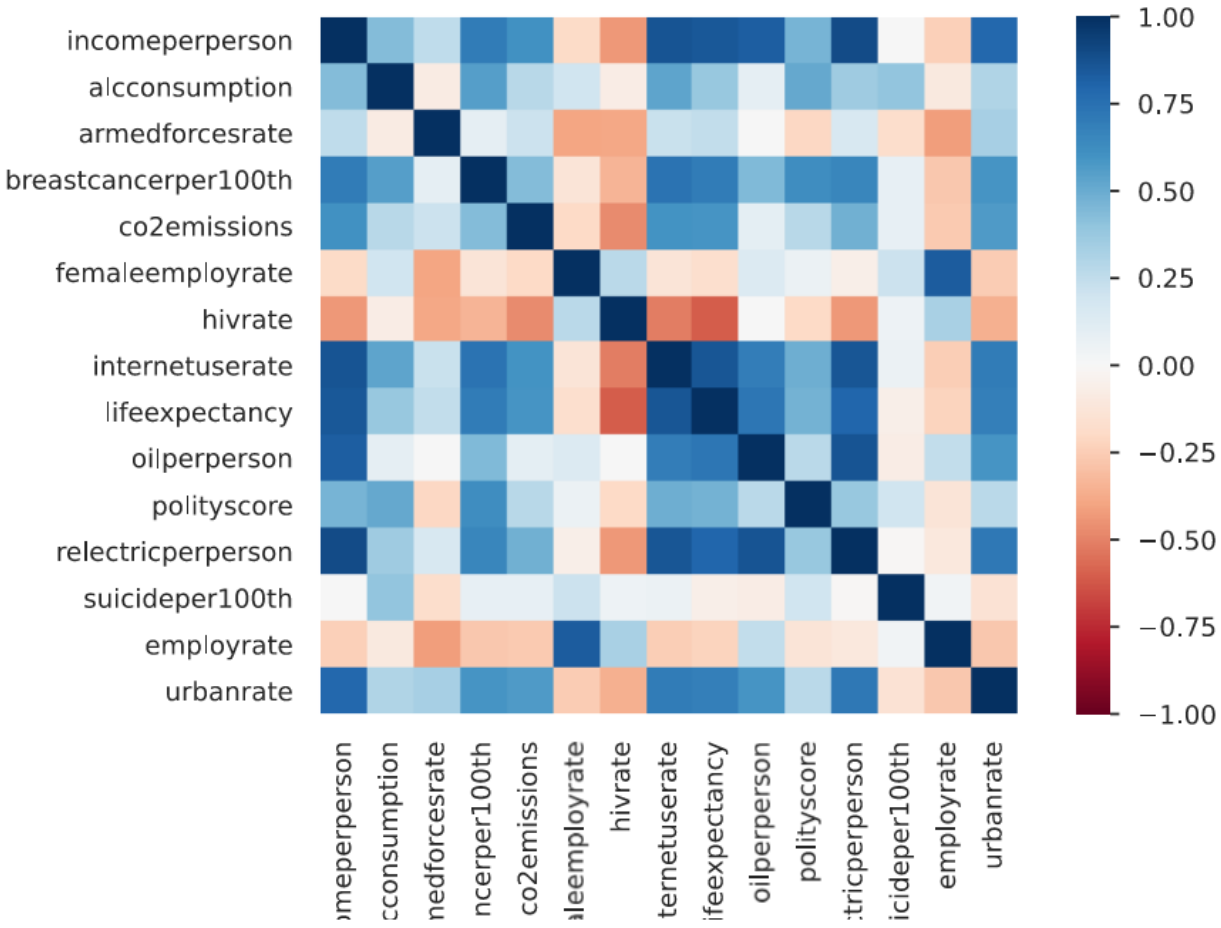
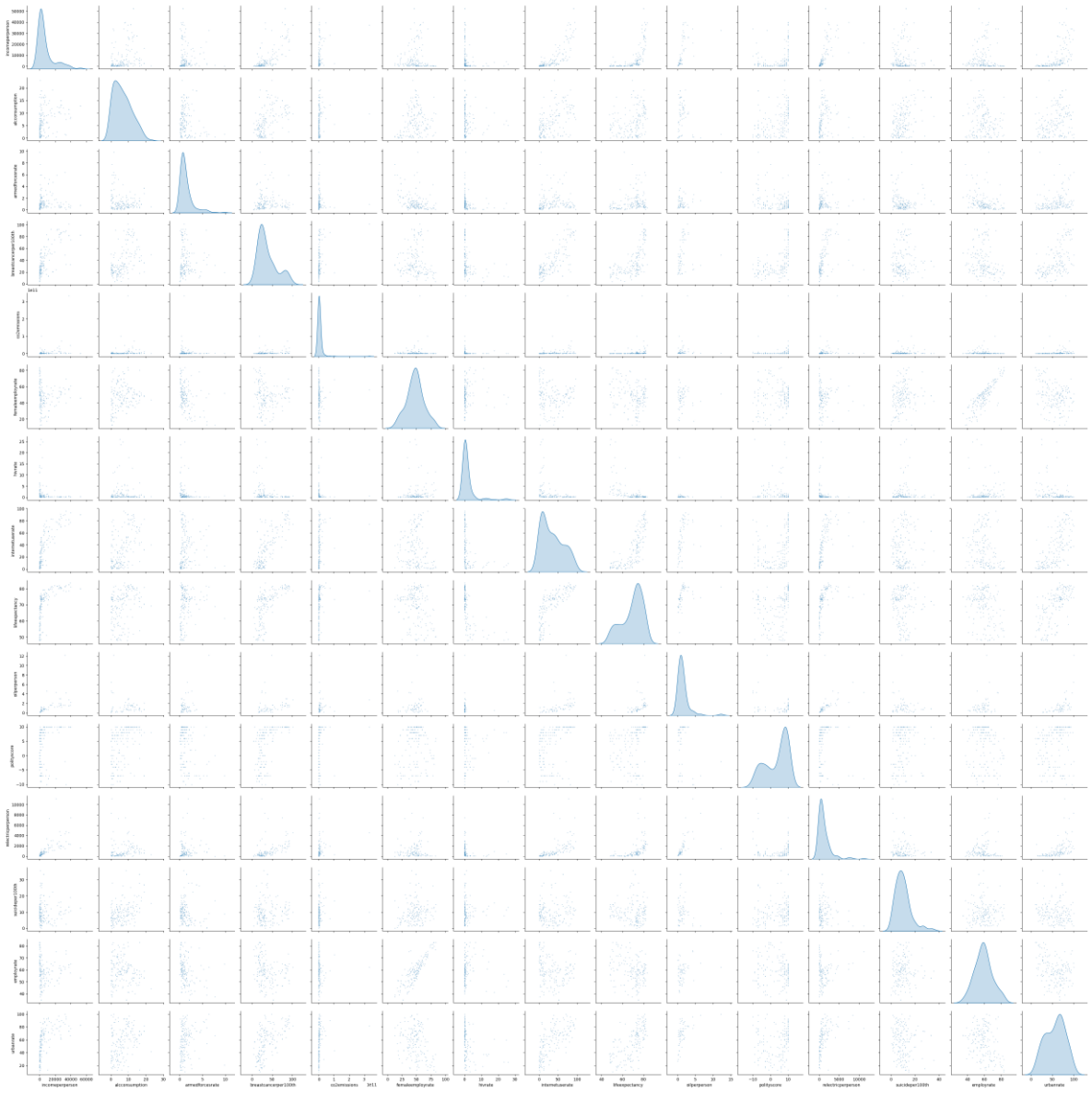


Figura 2. Gráficas de dispersión bivariadas



Con base en el mapa de calor de correlaciones y las gráficas de dispersión de dos variables que se generaron en el notebook, además del conocimiento generado en el análisis univariado (el cual no se presenta aquí, pero sí en el notebook), se escogieron 6 variables que parecían ser las más importantes a priori:

- **breastcancerper100TH:** resulta muy sorpresivo que esta variable tenga una correlación positiva con el PIB per Cápita de un país. A primera vista, esto no tendría sentido. Sin embargo, haciendo una averiguación rápida en Google, resulta ser que en los países desarrollados están más extendidos los análisis de diagnóstico de cáncer de seno, lo que facilita su tratamiento posterior. Entonces, no es que las mujeres de países desarrollados se enfermen más, simplemente es que en los países pobres las mujeres no tienen acceso a diagnósticos eficaces, y muchas veces mueren antes de saber que tuvieron esta enfermedad.

- **HIVrate:** como era de esperarse, este indicador se relaciona negativamente con la riqueza de un país.

- **Internetuserate:** como era de esperarse, entre más rico un país, más acceso a internet tienen sus ciudadanos.

- **lifeexpectancy:** no es sorpresa que los países con mayores ingresos por persona tengan esperanzas de vida más altas.

- **relectricperperson:** tampoco es sorpresa que entre mayor desarrollo económico, haya más consumo energético a nivel residencial.

- **urbanrate:** para finalizar, y tal como se esperaba, la tasa de habitantes de zonas urbanas esta correlacionada con el PIB per Cápita.

4. Preparación de los datos

4.1. Análisis de calidad de datos:

Completitud:
Respecto a esta dimensión se tiene para cada una de las variables:

- **country:** 0% de los campos vacíos, sin problemas.
- **incomeperperson:** 3 campos vacíos (1,7%). Dado que esta es la variable que se quiere predecir, no es conveniente tener campos vacíos. Al ser tan pocos registros, estos se eliminarán. Además, dos de estos países son Afganistán y Myanmar, países sobre los que no existen datos confiables dada su complicada situación.
- **alcconsumption:** 0% de los campos vacíos, sin problemas.
- **armedforcesrate:** 4 campos vacíos (2.2%). Estos campos corresponden a Comoros, Solomon Islands, Bhutan y Swaziland, países que no tienen fuerzas militares activas, por lo que este campo se imputará como cero.
- **breastcancerper100th:** 1 campo vacío (0.6%). Este campo vacío corresponde a Timor-Leste. Este es un país muy pequeño y con muy poca relevancia a nivel internacional. Además, tiene otros campos relevantes vacíos también, por lo que se decide eliminarlo del dataset antes de la regresión.
- **co2emissions:** 2 campos vacíos (1,1%). Sin embargo, esta variable no se tendrá en cuenta en la regresión por sus problemas de consistencia: hace referencia a las emisiones totales, más no a las emisiones per cápita.
- **femaleemployrate:** 1 campo vacío (0,6%). Este campo pertenece a Djibouti, pequeño país de África. Además, este país contiene otros campos vacíos en el dataset y no es tan relevante a nivel mundial, por lo que el correspondiente registro se eliminará antes de la regresión.
- **hivrate:** 23 campos vacíos (12,9%). Esta es una de las variables más problemáticas aquí. Claramente, tener casi 13% de faltantes representa un problema importante de completitud de esta variable. Sin embargo, según lo visto en el entendimiento de datos, esta variable es importante para la regresión. Siendo así, se decide imputar esta variable según la región a la que pertenezca el respectivo país. En este caso, se identifican 4 regiones: Medio Oriente

(0,2% de tasa de VIH), Africa Subsahariana (8%), Los Balcanes (0.06%) y Sudamérica(0,5%). Para la imputación se tomará el promedio de tasa de infección por VIH para los demás países de la región que sí tienen datos.

- **internetuserate:** 4 campos vacíos (2,2%). Descontando un dato duplicado, son tres países los que carecen de datos sobre el acceso a Internet: Myanmar, el cual ya se había optado por eliminarlo de la regresión por su falta general de datos debido a sus problemas internos. Sudan, uno de los países más pobres del mundo, el cual se imputará con el promedio de países similares en su región: 2% de acceso a Internet. Y Sierra Leona, país africano pobre, pero no tan pobre como Sudan, el cual se imputará con los datos de otros países similares: 9% de acceso a Internet.
- **lifeexpectancy:** 0 campos vacíos, sin problema.
- **oilperperson:** 111 campos vacíos (62,4%). Esta variable simplemente tiene demasiados campos vacíos. Se decide dejarla por fuera de la regresión, puesto que no hay una forma coherente de imputar tantos datos.
- **polityscore:** 11 campos vacíos (6,2%). Aquí hay un dato duplicado para descontar. Aparte, hay 7 países con varios otros datos faltantes que se decide eliminar de la muestra; Bahamas, Barbados, Belice, Bosnia, Brunei, Cabo Verde y Surinam. Quedan entonces 3 países europeos (Islandia, Luxemburgo y Malta) conocidos por sus altos niveles de democracia, para los cuales se imputará un valor de 10, según el promedio de sus vecinos similares más cercanos.
- **relectricperperson:** 37 campos vacíos (20,8%). Lamentablemente, el nivel de campos vacíos de esta variable es muy grande como para imputarlo o ignorarlo. Así que no será posible usarla en la regresión. Sin embargo, durante el entendimiento de datos ya se comprobó que esta es una variable relevante para medir la riqueza de un país, lo cual ya genera insumos importantes a la hora de formular políticas públicas.
- **suicideper100th:** 0 campos vacíos, sin problemas.
- **employrate:** 1 campo vacío (0,6%). Este campo pertenece a Djibouti, país que igual se eliminará de la muestra por tener varios otros campos vacíos.
- **urbanrate:** 0 campos vacíos, sin problemas.

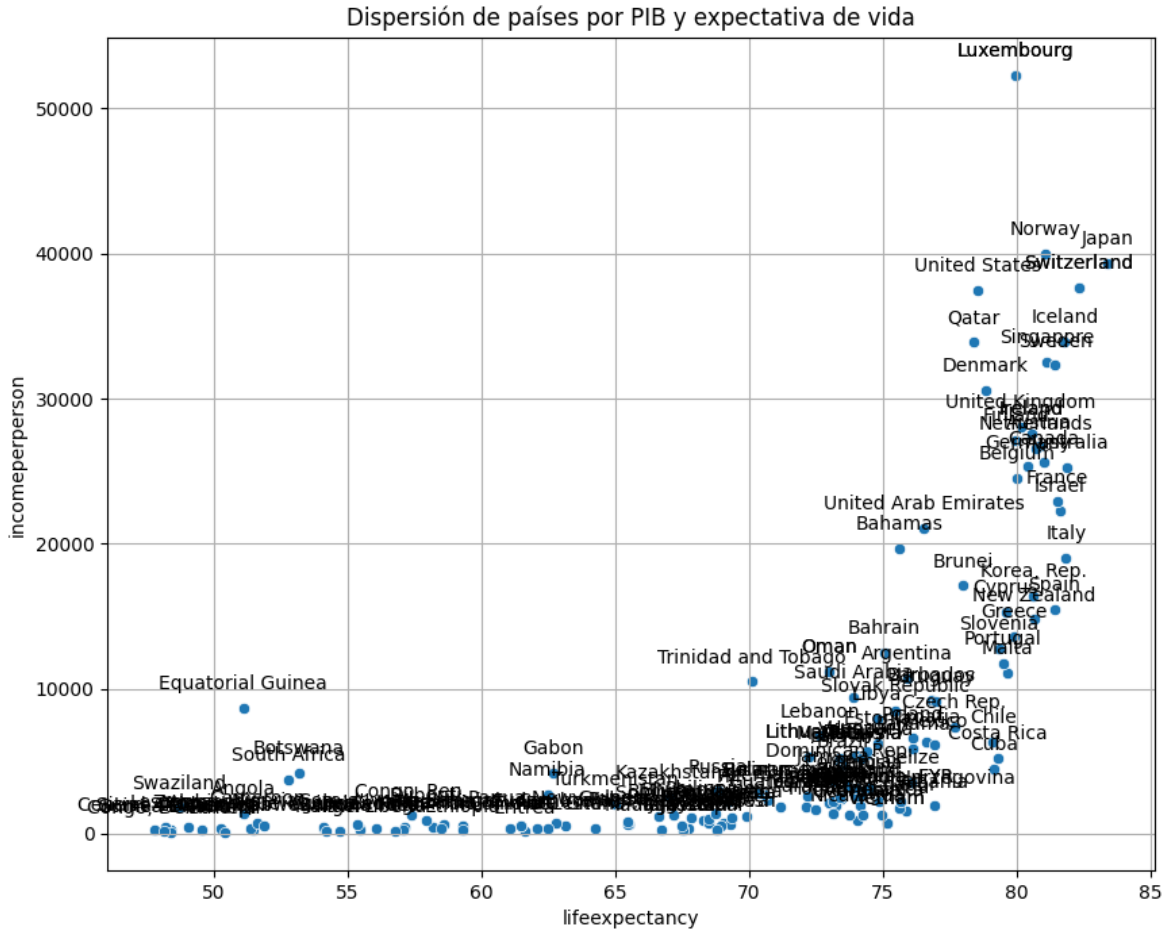
Consistencia:

Como ya se hizo en las líneas anteriores, la variable de emisiones acumuladas de CO2 se ha eliminado por falta de consistencia. Esta variable correspondía al estimado total por país, más no a un estimado per cápita como el resto de las variables. Por esta razón, las emisiones totales de países como China y USA no son comparables, teniendo en cuenta que China tiene una población mucho mayor. Por otra parte, las demás variables resultan ser consistentes con el objetivo del ejercicio. Solamente, se ha evidenciado que existen 12 registros duplicados, los cuales se procede a eliminar.

Precisión:

En lo relativo a esta dimensión, se ha verificado que la fuente de los datos "Gapminder" realizar compilados serios y rigurosos de la información socioeconómica de diferentes países. Por lo tanto, Gapminder constituye una fuente seria en la que se puede confiar. Además, el entendimiento de datos que se acaba de realizar muestra que, a nivel general, existe coherencia entre los datos y lo que de antemano se sabe del nivel de desarrollo de cada país. La siguiente gráfica de dispersión es una muestra de ello:

Figura 3. Dispersión de países por PIB y expectativa de vida



Oportunidad (A tiempo):
Existe aquí un aspecto importante sobre la oportunidad temporal de los datos. Y este se trata de que una buena parte de las variables no son del mismo año. Concretamente, se desea predecir el PIB per Cápita de un país en 2010, con base en variables que fueron tomadas en diferentes años entre el 2006 y el 2011.

Sin embargo, se considera que en este caso es válido pasar por alto esto, puesto que en todos los casos se trata de datos de la misma década, que da una u otra forma reflejan la realidad del cada país hacia 2010. Además, la verdad es que los cambios socioeconómicos no suelen suceder tan rápido, como para que un país cambie demasiado en unos 5 años. Por lo tanto, para efectos de este ejercicio se considerará como válida la mezcla entre diferentes variables que difieren levemente en su año de muestreo, pero que de todas formas no están a más de 5 años entre sí.

Conformidad y validez:
Se ha revisado el dataset, y todas las columnas tienen valores que son conformes a su especificación, en los formatos adecuados. Además, los datos resultan ser válidos dentro de su domino del conocimiento.

Relevancia:
Durante el proceso de entendimiento de datos se ha demostrado que las variables del dataset son relevantes para la regresión propuesta, particularmente aquellas escogidas dentro del TOP 6.

Procesabilidad:
Aquí se está tratando con un dataset relativamente sencillo, que después de depurarse contiene 154 filas x 13 columnas. La primera columna contiene nombres de países, mientras que las otras 12 contienen datos númericos (FLoat) fáciles de tratar. Por lo tanto, no hay ningún problema para procesar esta dataset.

Disponibilidad:
Por defecto, al haber descargado el dataset desde el repositorio enunciado, los datos necesarios para este ejercicio ya se encuentran disponibles.

4.2. Complementación de los datos con fuentes externas:

De manera adicional, y con el fin de enriquecer el dataset, se ha buscado información que permita complementar los datos. Concretamente, se utilizaron datos extraídos de Gapminder con la región geográfica de cada país (esquema de 8 regiones) y el nivel de ingresos con el cual el Banco Mundial clasifica a cada país (Ingresos altos, medio altos, medio bajos, bajos).

Dado que las dos nuevas variables son categóricas, fue necesario transformarlas para incorporarlas en la regresión. En el caso de la variable que indica una de las 8 regiones geográficas, se utilizó un one-hot encoder. Para el nivel de ingresos según el Banco Mundial, se utilizó una categorización ordinal.

5. Interpretación del modelo y sus métricas de error

Se han implementado dos modelos para evaluar cuál podría ser más conveniente para los objetivos del ejercicio: una regresión lineal y una regresión regularizada Lasso. Se presentan los resultados:

Figura 4. Resultados de las modelaciones

Regresión lineal:

```
Mean Squared Error (MSE): 20722999.279324375
R-squared (R2): 0.6485335322488124

Coeficientes de la regresión lineal:
alcconsumption: -943.6320579898855
armedforcesrate: -245.34128827349105
breastcancerper100th: 2437.745096797727
femaleemployrate: 616.5993755513317
hivrate: 1118.9893150961825
internetuserate: 7280.020379280859
lifeexpectancy: 2326.878927880945
polityscore: -284.4289529280069
suicideper100th: 631.0021529494929
employrate: 1091.8947535976783
urbanrate: 1059.7854895977134
World bank, 4 income groups 2017: 265.9664740493914
eight_regions_africa_north: 37.726577131974125
eight_regions_africa_sub_saharan: 1690.4488766349468
eight_regions_america_north: 156.20458200018567
eight_regions_america_south: -1291.1545823593037
eight_regions_asia_west: 139.55507731311786
eight_regions_east_asia_pacific: 280.33246848382817
eight_regions_europe_east: -2319.8341689548515
eight_regions_europe_west: 544.0443040587661
intercept: 7545.424951708311
```

Regresión Lasso:

```
Mean Squared Error (MSE): 15885429.22747667
R-squared (R2): 0.7305797474565792

Coeficientes de la regresión Lasso:
alcconsumption: -0.0
armedforcesrate: -0.0
breastcancerper100th: 2205.375886346924
femaleemployrate: 0.0
hivrate: 476.98851062744274
internetuserate: 7945.076770828324
lifeexpectancy: 0.0
polityscore: -0.0
suicideper100th: 0.0
employrate: 1026.254473851898
urbanrate: 211.86508278003666
World bank, 4 income groups 2017: 380.6431230360768
eight_regions_africa_north: -0.0
eight_regions_africa_sub_saharan: 0.0
eight_regions_america_north: -0.0
eight_regions_america_south: -1069.7753863207788
eight_regions_asia_west: -0.0
eight_regions_east_asia_pacific: 0.0
eight_regions_europe_east: -2791.993527142519
eight_regions_europe_west: 197.82014654650618
intercept: 7476.800795637643
```

En lo relativo a las métricas, se escogieron el MSE y el R2 para la evaluación de los modelos. La regresión Lasso ha mostrado tener un mejor comportamiento a la hora de predecir el PIB per Cápita de los países. Particularmente, si bien el R2 = 0,73 de la regresión Lasso no puede parecer el mejor para una regresión, es más que suficiente para el objetivo de esta modelación: identificar las variables que mejor se relacionan con el PIB.

Siguiendo esta línea, la regresión Lasso tiene como una de sus principales características ser capaz de llevar a cero los coeficientes de las variables que menos influyen en un modelo. Por esta razón, los modelos Lasso son especialmente buenos para identificar variables de interés en una regresión, siempre y cuando el coeficiente "Alpha" tenga un valor adecuado. En este caso, se probó manualmente (en el notebook) con varios coeficientes Alpha, hasta llegar al valor de 300 como el que mejor R2 brindaba.

6. Conclusiones y recomendaciones

En primer lugar se contestará la pregunta relacionada con el segundo objetivo de negocio:

¿Cuáles son las variables que mejor se correlacionan con el PIB Per cápita de una nación?

De acuerdo con la regresión Lasso, y tomando aquellas variables cuyos coeficientes de regresión son diferentes a cero, se tiene:

- **breastcancerper100th:** nivel alto de correlación positiva. Aquí lo que debe entenderse es que los países con mayores ingresos tienen mejores esquemas para la detección temprana del cáncer de seno. Mientras que en los países pobres muchas veces las mujeres mueren sin haberse enterado de que tenían la enfermedad por falta de diagnóstico.

- **hivrate:** nivel medio de correlación positiva. La realidad es que se esperaba que la correlación con esta variable fuera negativa. En este caso, en el entendimiento de datos fue evidente que los países más pobres son los que tienen tasas más altas de contagio de VIH. Por lo tanto, se mantendrá esta última filosofía.

- **internetuserate:** nivel alto de correlación positiva. Esta es una de las variables que desde el principio se pensaba iba a ser determinante. El tema es simple, entre más rico un país, mayor acceso a Internet tienen sus habitantes.

- **employrate:** nivel medio de correlación positiva. Esta variable no estaba dentro del top 6 original, pero tampoco sorprende. No resulta sorprendente que los países más ricos tengan mejores tasas de empleo.

- **urbanrate:** nivel medio de correlación positiva. Por último, esta también era una de las variables del top 6 original. No es una sorpresa para nadie que los países con mayor porcentaje de población urbana sean los más ricos: urbanizarse conlleva mayores niveles de industrialización e incentiva el desarrollo de economías basadas en servicios.

Se contesta ahora la pregunta del segundo objetivo de negocio:

¿Qué conjunto de políticas públicas recomendaría implementar, a partir de la premisa de que la mejora en estas áreas indicaría al Banco Mundial que el país es estable, está en una trayectoria de desarrollo sostenible y tiene la capacidad de administrar y reembolsar préstamos de manera efectiva?

Según las variables identificadas con el modelo Lasso, estas son las recomendaciones de política pública:

- Fortalecer y ampliar **los programas de detección de cáncer de seno en las mujeres**, con el fin de facilitar el tratamiento temprano de esta dolencia y mejorar las tasas de supervivencia a esta enfermedad. A priori, esto tendrá como consecuencia incrementar las tasas de cáncer de seno en las mediciones oficiales, pero a largo plazo contribuirá con la disminución de las tasas de mortalidad y el incremento de la justicia social. Estos programas deben estar dirigidos a todas las mujeres, sin importar sus condiciones socioeconómicas o regiones de vivienda. Además, será muy importante realizar campañas de concientización sobre esta problemática, enfocándose en el autoexamen para detectar señales tempranas de alerta.
- Fortalecer y ampliar los **programas de educación sexual, particularmente aquellos dirigidos a incentivar el uso del preservativo**, como medio más eficaz para **prevenir** las enfermedades de transmisión sexual, entre ellas el **VIH**. Estos programas deben ser dirigidos a toda la población, sin importar su edad, orientación sexual, condición socioeconómica o lugar de residencia.
- **Fortalecer la red de comunicaciones basada en Internet**, de modo que las personas en general puedan tener un mejor acceso a la red, bien sea mediante conexiones en casa o por celular. Aquí es necesario que los gobiernos evalúen formas para **que los ciudadanos más vulnerables puedan tener acceso a un servicio de Internet asequible** (una opción es el subsidio del servicio), al tiempo que se evalúan los mejores mecanismos para que **puedan adquirir dispositivos celulares o computadores** que sean su puerta de entrada a la red mundial.
- Con relación a las tasas de empleo, resulta obvio en cualquier contexto que entre más baje el desempleo, mejor. Sin embargo, generar **una política pública para esto depende del contexto de cada país y de su economía**. En el caso concreto de un país como Colombia, se podría pensar que **facilitar la conformación de empresas, reducir los impuestos sobre las contrataciones (parafiscales) y alinear la carga tributaria de las empresas a su tamaño y sector** son estrategias que podrían contribuir a aumentar la tasa de empleo y de formación de empresas en general.
- Para finalizar, los datos muestran que entre mayor tasa de urbanización, mayor es el PIB per Cápita de un país. Sin embargo, esta conclusión podría ser engañosa: **la llegada masiva y desordenada de población rural a las grandes ciudades puede crear problemas** para el desarrollo de los países y generar condiciones de vida paupérrimas para los nuevos habitantes de los centros urbanos. Por lo tanto, aquí se deberá pensar en como **nivelar el nivel de vida de las personas en el campo** con el de las personas de la ciudad, de modo que las primeras no se vean obligadas a migrar. Además, el campo es la fuente de los alimentos de un país, y es un sector que no debe descuidarse. Por último, el crecimiento de las ciudades, especialmente en el contexto latinoamericano, es inevitable. Pero es urgente que los gobiernos diseñen estrategias para que **este crecimiento se dé de una manera ordenada, sostenible y alineada con las metas ambientales** de la actualidad.