Universidad de los Andes.
Departamento de Ingeniería de Sistemas y Computación.
Ciencia de Datos Aplicada.

TALLER 2:

Objetivo: utilizando la información del conjunto de datos Gapminder del Banco Mundial y aplicando un algoritmo de Machine Learning, se propondrán políticas para 166 países buscando mejorar el bienestar económico y social de su población.

Caracterización del dataset.

Dimensiones: el dataset está compuesto de 178 registros y 16 variables pertenecientes a diferentes indicadores entre los que se encuentran: prevalencia de VIH, uso de Internet, esperanza de vida, consumo de petróleo per capital, puntaje de democracia, tasa de suicidio, consumo residencial de electricidad por persona, etc.

Duplicidad: el dataset presenta 12 registros duplicados, correspondiente a los países de Switzerland, Oman, Macedonia FYR, Sudan, Malaysia, Iran, Lithuania, Belarus, Vietnam, Ireland, Tajikistan, Luxembourg, los cuales fueron eliminados para una mejor estimación.

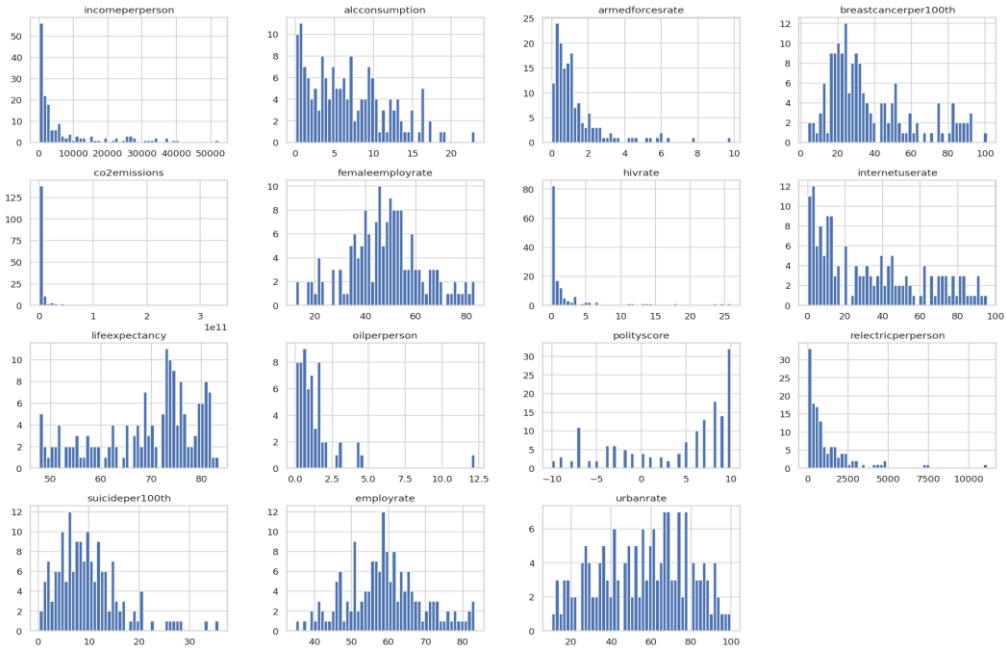
Faltantes: Para la gran mayoría de variables hay información faltante, siendo las variables hivrate 13%, oilperperson 62%, polityscore 6%, relectricperperson 21% las variables con las tasas más altas para imputar datos.

Imputación: Se utiliza la media por continente para la imputación de los valores faltantes a las variables cuya tasa fuera menor al 5%.

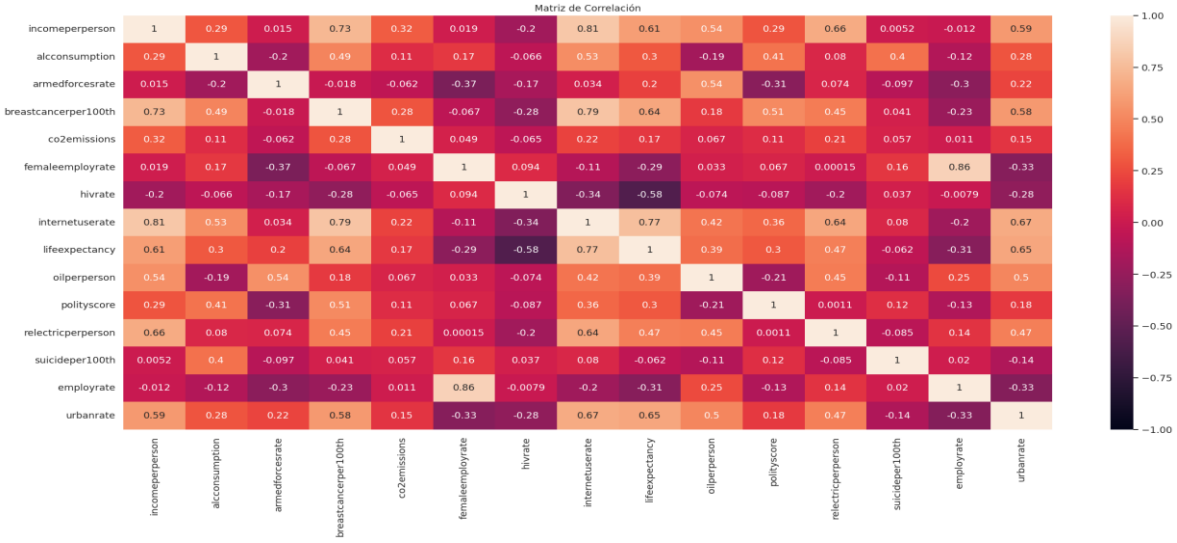
Análisis exploratorio de datos.

medida	incomeperperson	alcconsumption	armedforcesrate	breastcancerper100th	co2emissions	femaleemployrate	hivrate	internetuserate
count	163.000.000	163.000.000	159.000.000	162.000.000	1,61E+08	162.000.000	143.000.000	161.000.000
mean	7.202.019.107	6.782.086	1.359.446	37.656.790	6,16E+15	48.001.234	1.979.371	33.809.838
std	10.469.295.971	4.995.312	1.528.647	23.189.042	2,86E+16	14.731.864	4.429.626	28.004.156
min	103.775.857	0.050000	0.000000	3.900.000	8,51E+11	12.400.000	0.060000	0.210066
25%	602.502.870	2.625.000	0.468250	20.600.000	7,89E+13	39.250.001	0.100000	9.196.775
50%	2.344.896.916	6.080.000	0.904025	29.900.000	2,77E+14	48.450.001	0.400000	28.731.883
75%	8.550.031.767	9.870.000	1.544.014	50.250.000	2,41E+15	56.150.001	1.350.000	53.024.745
max	52.301.587.180	23.010.000	9.820.127	101.100.000	3,34E+17	83.300.003	25.900.000	95.638.113

medida	lifeexpectancy	oilperperson	polityscore	relectricperperson	suicideper100th	employrate	urbanrate
count	163.000.000	60.000.000	153.000.000	128.000.000	163.000.000	162.000.000	163.000.000
mean	69.219.650	1.379.010	3.849.673	1.144.245.457	9.894.519	59.074.691	56.015.215
std	9.924.945	1.747.206	6.226.821	1.596.990.968	6.322.557	10.364.735	22.600.016
min	47.794.000	0.032281	-10.000.000	0.000000	0.201449	34.900.002	10.400.000
25%	62.470.000	0.490364	-2.000.000	219.736.499	5.700.370	51.575.001	37.090.000
50%	72.974.000	0.883796	7.000.000	597.136.436	8.973.104	58.850.000	57.940.000
75%	76.337.000	1.583.022	9.000.000	1.491.145.249	12.640.498	65.000.000	73.470.000
max	83.394.000	12.228.645	10.000.000	11.154.755.030	3,58E+07	83.199.997	100.000.000



Matriz de correlación: se envician correlaciones positiva y negativas, así como correlaciones bajas, algunas de mayor interés entre las variables y el PIB



Modelo regresión lineal: se ajustó teniendo en cuenta las variables: 'femaleemployrate', 'lifeexpectancy', 'employrate', 'urbanrate', 'suicideper100th', y el uso de las variables dummies perteneciente a la región (continente al que pertenecen los países 'region_Africa', 'region_America', 'region_Asia', 'region_Europa','region_Oceania')

Encontrando buenas métricas en el de regresión logarítmica, con lo cual se generan las siguientes recomendaciones de políticas públicas para mejora del PIB en los países.

Los coeficientes del modelo son:

	columns	coef
0	femaleemployrate	-0.015546
1	lifeexpectancy	0.064046
2	employrate	0.014862
3	urbanrate	0.024663
4	suicideper100th	0.020027
5	region_Africa	-0.186537
6	region_America	0.010768
7	region_Asia	-0.309288
8	region_Europa	0.170280
9	region_Oceania	0.314778

Las métricas del modelo son:

```
MAE Logaritmical regresion:

Train: 0.5528606398121843
Test: 0.4889678055342581

RMSE Logaritmical regresion:
Train: 0.7555345990403958
Test: 0.6548747411873668
```

Para la interpretación de los coeficientes es:

Un aumento de una unidad en la variable X incrementa en % del ingreso por persona, manteniendo constante las demás variables.

Políticas:

1. Fomentar la educación de las niñas y mujeres en áreas de alta demanda laboral, como la tecnología, la ciencia, política, etc. para aumentar su empleabilidad y reducir la brecha de género en estos campos, además de Promover la igualdad salarial entre hombres y mujeres.
2. Política que garantice que los niños y jóvenes tengan acceso a una alimentación saludable y equilibrada, fortalecer el sistema de salud pública, incentivar la práctica de deportes y actividades físicas, y promover la educación en valores y habilidades socioemocionales para mejorar la calidad de vida de la población y fomentar la convivencia pacífica.
3. Fomentar el crecimiento y desarrollo de las pequeñas y medianas

empresas, promover la formalización del empleo, implementar políticas de igualdad de género en el empleo y fomentar la creación de empleos verdes.

4. Promover el crecimiento económico y la competitividad de las ciudades intermedias, a través de la implementación de políticas de desarrollo económico local, fomento de empleos en sectores estratégicos, inversión en infraestructura, y políticas de desarrollo sostenible.
5. Garantizar el acceso a servicios de salud de calidad, promover estilos de vida saludable, aumentar la cobertura de acceso a los servicios de salud, especialmente en zonas rurales y marginadas, implementar programas de prevención y tratamiento de enfermedades mentales y de detección temprana de enfermedades mentales y de riesgo de suicidio.

Jenniffer Escudero.