OBJETIVOS DE DESARROLLO SOSTENIBLE: PROYECTO DE MACHINE LEARNING

JOSÉ LUIS AGUILAR CHARFÉN

Resumen. Este reporte muestra la predicción del SDG Index Score en términos de los primeros dos Objetivos de Desarrollo Sostenible (SDGs). Se discute brevemente la construcción de tres modelos de regresión, uno de mínimos cuadrados ordinarios, y dos regularizados (Ridge y Lasso), así como el preprocesamiento para los datos. Se predicen adicionalmente los valores para aquellos países que no tienen un SDG Index Score [1], [2] con los modelos realizados en Python.

1. Introducción

Los Objetivos de Desarrollo Sustentable son 17 metas que se acordaron en la Organización de las Naciones Unidas (ONU) en 2015 [1] como una llamada a la acción para proteger al planeta y asegurar que para 2030 la población disfrute de paz y seguridad.

Estos objetivos varían desde reducir la pobreza, hasta temas de salud, equidad de género, higiene, crecimiento económico, energía y cuidado al medio ambiente. Cuando se consideran todos los diferentes objetivos, entonces se construye el SDG Index Score, que califica el progreso de un país en las 17 metas. Para este reporte, se enfocará en las primeras dos: *fin de la pobreza*, y *hambre cero*.

2. Metodología

Los datos utilizados son de las siguientes categorías, descritas brevemente:

1. *Poverty headcount ratio at $1.90/ day (%)*: Cuenta el porcentaje de gente que vive con menos de 1.90 dólares al día. Mientras mayor sea, indica un mayor nivel de pobreza.
2. *Poverty headcount ratio at $3.20/day (%)*: Cuenta el porcentaje de gente que vive con menos de 1.90 dólares al día. Se espera que esté fuertemente correlacionada con la variable anterior.
3. *Poverty rate after taxes and transfers (%)*: Cuenta la cantidad de gente que su ingreso cae por debajo de la línea de probreza, tomada como el 50 % del actual ingreso disponible después de impuestos y transferencias contadas como pagos a escuela o arreglos de alojamiento. Esta métrica solo está disponible para países de la OCDE. [3]
4. *Prevalence of undernourishment (%)*: Es el porcentaje de la población que su consumo alimenticio es insuficiente para proveer los niveles energéticos necesarios para mantener una vida activa y saludable.
5. *Prevalence of stunting in children under 5 years of age (%)*: Se refiere al porcentaje de la población infante que sufre de retraso en el crecimiento.
6. *Prevalence of wasting in children under 5 years of age (%)*: Se refiere al porcentaje de la población que es demasiado delgado para su altura y resulta por incapacidad de ganar peso o por bajar muy rápido.
7. *Prevalence of obesity, BMI $\geq$ 30 (% of adult population)*: Se refiere al porcentaje de gente que está por encima de un índice de masa corporal por arriba de 30[1].
8. *Human Trophic Level (best 2-3 worst)*: Nivel trófico del humano. Se refiere a la posición que ocuparía en una pirámide alimenticia. Un valor de 1 sería un productor primario,

Date: 3 de marzo de 2022.

[1]Pueden existir críticas por si es una métrica válida o útil, pero queda por fuera del alcance de este proyecto.

como una planta, y un nivel de 5 sería un depredador ápice, por ejemplo un oso polar o un esquimal. Se puede calcular con la siguiente ecuación[4]:

$$N_{T_i} = 1 + \sum_j \left(N_{T_j} y_{ij}\right) \tag{1}$$

Donde N_{T_i} es el nivel trófico de la población i, N_{T_j} el de la presa j, y y_{ij} es la fracción de la presa j en la especie i. Sin embargo, la mayoría de los países se encuentran con un nivel entre 2 y 3. No puede ser menor a 2 porque no somos productores primarios. Usualmente se interpreta como mejor que sea menor el nivel trófico porque significa que la energía y recursos usados para producir la comida – por efectos de la eficiencia de transferencia de biomasa, que es aproximadamente del 10 % [5] – son menores.

9. *Cereal yield (tonnes per hectare of harvested land)*: Es el rendimiento de producción de cereales. Usualmente, una mayor cantidad significa que la tierra es más fértil.
10. *Sustainable Nitrogen Management Index (best 0-1.41 worst)*: Esta métrica combina dos medidas de eficiencia en la producción agrícola: la eficiencia de uso de nitrógeno, y eficiencia de uso de tierras. La eficiencia de uso de nitrógeno se define como la fracción de entradas de nitrógeno que termina en los productos, y normalmente debe estar entre 0 y 1, con valores normalmente entre 0.5 y 0.9[6], indicando uso eficiente uso del nitrógeno, pero puede ser mayor a 1 si sale de la tierra, eliminando recursos. Combina además qué tanto se produce por hectárea, y la métrica construida por Zhang y Davidson idealmente alcanza 0, mostrando este número un alto rendimiento del uso de nitrógeno y de uso de tierras.
11. *Yield gap closure (% of potential yield)*: Indica en porcentaje el rendimiento potencial en los tres principales cultivos, ponderados por la importancia de cada cultivo en términos del área que ocupa. Solo está disponible para países de la OCDE. Mientras mayor sea su valor indica un uso más eficiente de tierra.
12. *Exports of hazardous pesticides (tonnes per million population)*: Toma el valor promedio de los últimos 5 años de las exportaciones de pesticidas vistos como materiales peligrosos; por ejemplo, el glifosato (o en general los pesticidas organofosfatados).

Debido a que tanto la tasa de pobreza medida con el ingreso disponible, como la cerradura de la brecha del rendimiento de plantas reportadas en el conjunto de datos utilizado solamente están disponibles para la OCDE, existen tres principales posibilidades para tratarlos. La primera consiste en modelar diferente para la OCDE e incluir estos valores, y otro modelo sin considerar estas variables para el resto del mundo[2]. Como segunda opción, pueden imputarse sus valores con alguna técnica, suponiendo que sus valores serán parecidos a los de la OCDE. Puede, por ejemplo, imputarse con la media, o la mediana. Como tercera opción, pueden removerse como variables, reconociendo que posiblemente, debido a que la situación sociopolítica y económica de los países de la OCDE es distinta a la del resto del mundo, entonces la información que se pueda extraer de los valores imputados de esta variable introduzcan un sesgo que no puede observarse fácilmente, y afectar las predicciones de maneras no deseables (por ejemplo, falsamente sugiriendo una relación más fuerte de la que existe realmente, o al revés, que no sea significativa).

Adicionalmente, se decide eliminar como valor para predecir a Eswatini, debido a que en el nivel trófico muestra un valor atípicamente alto (4), mucho mayor al siguiente país que es Islandia, con un valor de 2.583. Además, un nivel trófico de 4 supone que el consumo de alimentos es básicamente de animales carnívoros, o por su contra, que casi no existen plantas en la dieta de la población, que no coincide con la dieta usual de la mayoría de los países. Además de ser atípicamente alto, de incluirse en el modelo, puede provocar un apalancamiento muy grande, afectando fuertemente al valor del estimador.

[2]Alternativamente, puede hacerse un modelo por continente, pero se deja como propuesta meramente

Debido a que no todos los países tienen un SDG Index Score, no se consideran para entrenamiento ni validación ni prueba a estos países. Sin embargo, sí se predicen sus valores dadas las variables usadas para construir los primeros dos SDGs.

Para construir el modelo, se realiza un análisis de correlación de variables, y se observan tanto las distribuciones como las correlaciones con gráficos por pares de variables. Posteriormente, se transforman las variables con la transformación Yeo-Johnson [7], preferida sobre la de Box y Cox por la presencia de valores 0 en las variables. Posteriormente, se imputan los valores con metodología de vecinos más cercanos [8]. Después, se codifican las regiones usadas para el SDG Index por un método one-hot. Una vez transformadas las variables de estas maneras, se comparan un modelo de mínimos cuadrados ordinarios, regresión Ridge, y Lasso, ajustando sus regularizaciones por validación cruzada, y se evalúa si se encuentra sobreajustado comparando el error cuadrático medio del conjunto de datos de entrenamiento con uno de prueba por separado, corriendo 20 veces el experimento y promediando los valores.

Las decisiones de construcción del modelo y justificaciones se discuten a mayor profundidad en la sección de resultados y conclusión.

3. Resultados y discusión

Se observan en la figura 1 y 2 las correlaciones que existen entre las distintas variables de los datos. Puede observarse que aquellas asociadas con la pobreza tienden a estar negativamente correlacionadas con el índice SDG, lo cual hace sentido, y que están fuertemente correlacionadas entre sí. Al descomponer las correlaciones para que se observen por región, puede verse que existen ciertas diferencias en las correlaciones entre tipos de datos, siendo más fuertes por ejemplo en Oceanía, que en Latinoamérica.

Figura 1. *Matriz de correlación de las variables numéricas del modelo.*

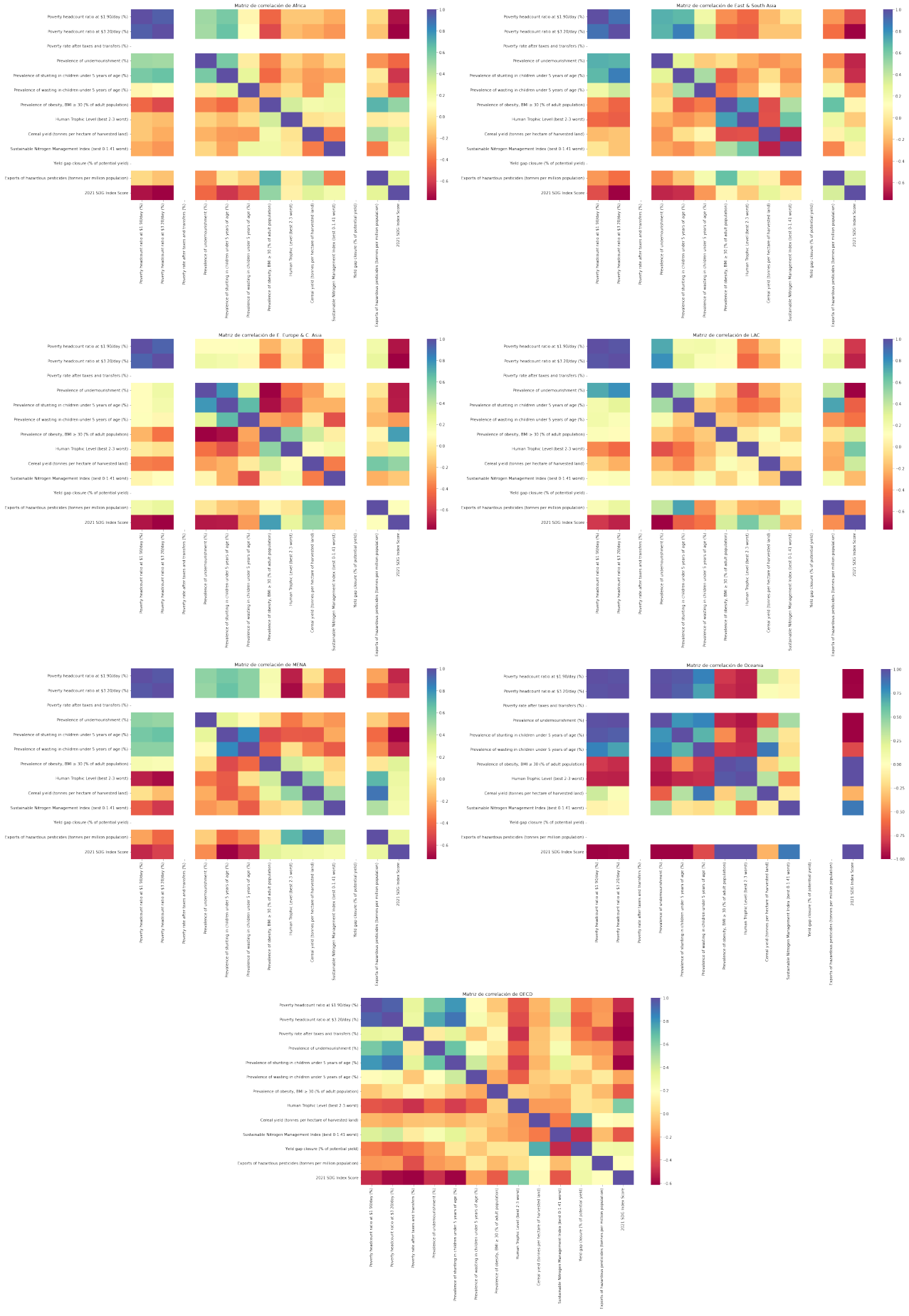


Figura 2. *Matrices de correlación de variables numéricas del modelo por región usada para calcular SDGs.*

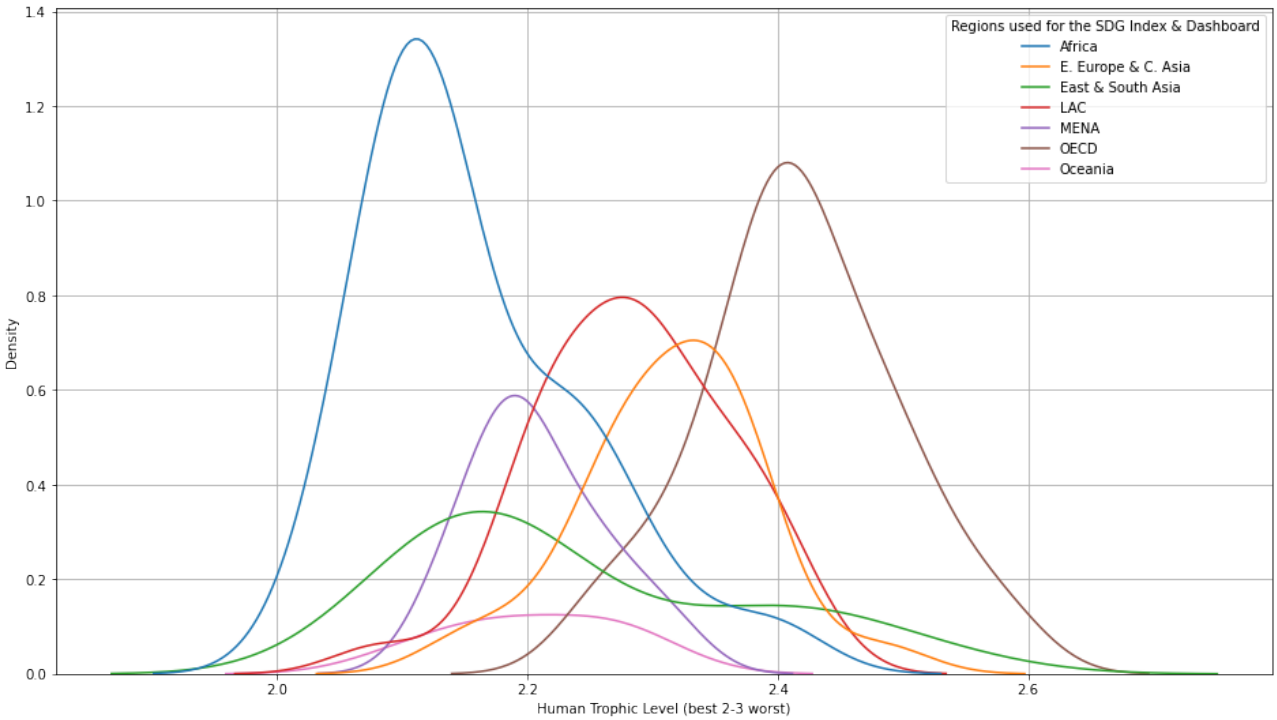


Figura 3. *Distribuciones del índice trófico humano por región.*

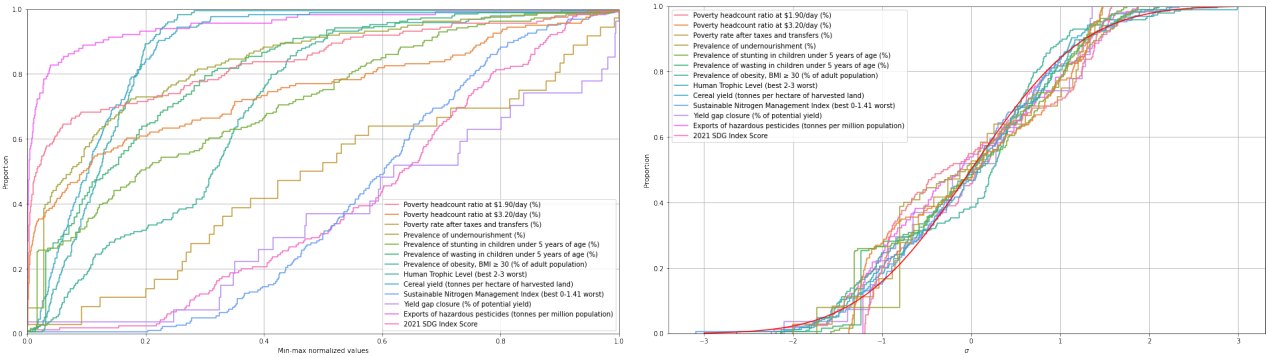
Es notable mencionar que no necesariamente se cumple el que las distribuciones estén idénticamente distribuidas por región. Debido a esta dependencia, se decide incluir en el modelo la región de la que viene como variable predictora; para ejemplificar, se ve claramente la diferencia en distribuciones en la figura 3, lo que sugiere que es una buena idea incluirse. Debe reconocerse, sin embargo, que aunque exista una relación, no implica causalidad, y el modelo meramente indica su presencia. Debe tenerse en cuenta que es un modelo reducido, y que en uno completo en principio no debería usarse, bajo la interpretación de que el pertenecer a una región favoreciera tu puntuación[3].

En la figura 4 se observan las distribuciones cumulativas empíricas de los datos utilizados para el análisis. En los datos no transformados, puede observarse que el común supuesto de normalidad no se respeta, por lo que existe que los puntos que se encuentren en la cola de la distribución sean muy influyentes en los valores de los parámetros usados para estimar, en particular si son datos atípicos. Debido a esto, se propone una transformación de los datos con la técnica de Yeo-Johnson [7], favorecida sobre la de Box y Cox por la presencia de valores 0 en los datos. Puede notarse que con esta transformación se ven mucho más cercanos a una distribución normal.

En la figura 5 se observan las porporciones de descomposición de varianzas para las variables numéricas. Puede notarse que en los datos originales existe colinealidad entre las dos primeras variables, que son las que indican la cantidad de gente que vive debajo de cierto nivel de pobreza, pero son las únicas que presentan una colinealidad relativamente fuerte. Existen otras más débiles, pero resultan poco preocupantes.

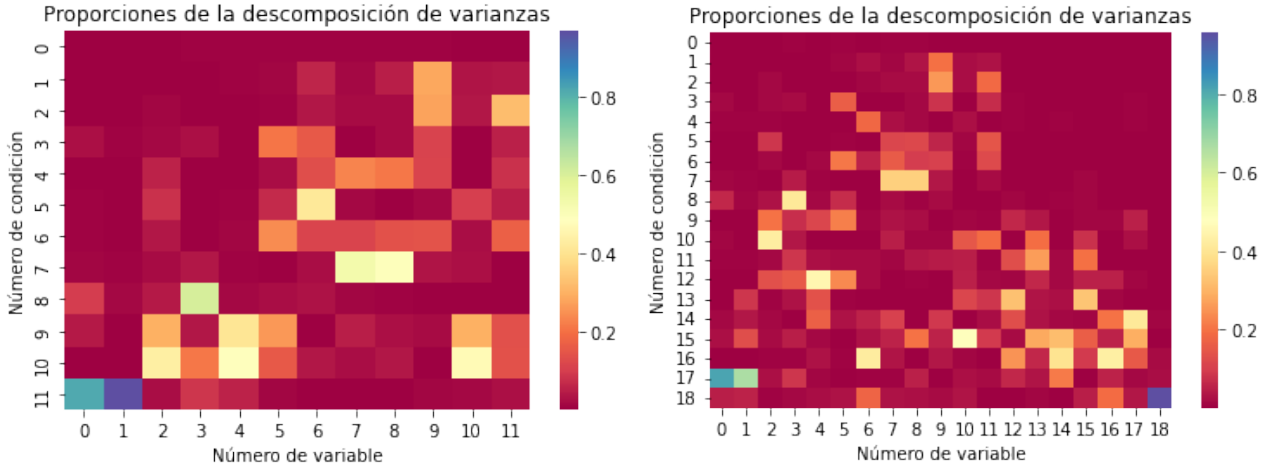
Una vez hecho este análisis, entonces puede construirse el modelo. Para poder imputar los datos faltantes, se propone usar una técnica de k-vecinos más cercanos, que aunque bien no se conocen los efectos de esta técnica en la estructura inherente de los datos [8], tienden a mejorar las propiedades inferenciales de los datos. Después de este preprocesamiento, se construyen los siguientes modelos:

[3]Vale la pena recordar que la metodología usada para calcular el score es promediar los puntajes de cada uno de los SDGs[2]



(a) *Distribuciones de datos no transformados.* **(b)** *Distribuciones de datos transformados. En rojo: distribución normal acumulada con media de 0 y varianza de 1*

Figura 4. *Distribuciones de datos antes y después de transformar.*



(a) *Proporción de descomposición de varianzas solo para variables numéricas.* **(b)** *Proporción de descomposición de varianzas incluyendo también la región que se encuentran.*

Figura 5. *Análisis de descomposición de las varianzas en términos de los números de condición.*

1. Regresión Ridge: se hace validación cruzada para ajustar el parámetro de regularización λ.
2. Regresión lineal: se hace de manera ordinaria.
3. Regresión Lasso: se hace validación cruzada para ajustar el parámetro de regularización λ.

Después de construidos, se corren 20 veces los experimentos y se calculan sus errores cuadráticos medios para poder compararlos. Adicionalmente, se observan las distribuciones de los residuales en la figura 6.

Cuadro 1. *Resultados de regresiones*

Valor	Ridge	OLS	Lasso
λ	2.637	–	0.039
R^2 en todos los datos	0.879	0.881	0.879
MSE entrenamiento	13.45	12.99	13.34
MSE prueba	18.85	19.34	18.40

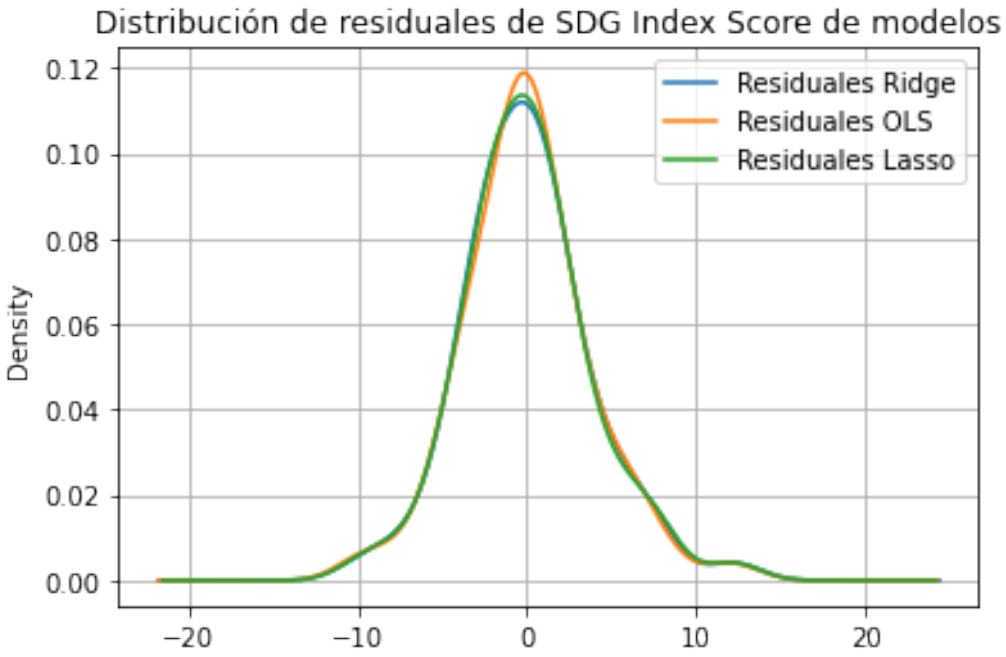


Figura 6. *Distribución de residuales de cada uno de los modelos. Puede observarse son muy similares, y son relativamente normales.*

Se puede observar que todos los modelos tienen un rendimiento relativamente cercano, siendo muy ligeramente menores los valores del error cuadrático medio en los datos de prueba para la regresión Ridge y Lasso. Observando la figura 6 puede notarse que los residuales son aproximadamente normales. Observando el resumen de los residuales en el cuadro 2 se verifica esta idea. Resulta importante debido a que los residuales pueden considerarse como un estimador del error del proceso de generación de datos, y la suposición de que son normales es respetada[4].

Cuadro 2. *Resumen estadístico de residuales de los modelos*

Valor	Ridge	OLS	Lasso
Media	$5{,}6 \cdot 10^{-15}$	$1{,}5 \cdot 10^{-14}$	$3{,}7 \cdot 10^{15}$
Error estándar	0,03	0,03	0,03
Asimetría	0.76	0.54	0.71
Kurtosis	4.34	4.06	4.17
Jarque-Bera	21.22	11.74	17.44

Debido a que los modelos con regularización son muy similares a los mínimos cuadrados, se puede asumir que entonces la varianza de los estimadores no es muy elevada. Esto es de esperarse pues aunque se observa algo de colinealidad, notable en la figura 5, es menor esta colinealidad y solamente es entre dos variables, por lo que no le afecta mucho este tipo de problemas numéricos. Por el otro lado, el que se vea poco reducida la varianza por medio de la regularización también sugiere que, aunque el rendimiento de los modelos es muy similar, no existe problemas con utilizar una regresión lineal para predecir el SDG Index Score con este método.

En la figura 7 se observan comparativamente los valores reales contra los predichos por el modelo. Puede observarse que en el rango menor de valores, tiende a ligeramente sobreestimar los puntajes, lo que sugiere que probablemente términos de interacción o de mayor orden puedan mejorar la regresión. Dicho esto, solamente en los menores valores es en los que se encuentran

[4]Se entiende como el cuarto momento central, no como kurtosis en exceso.

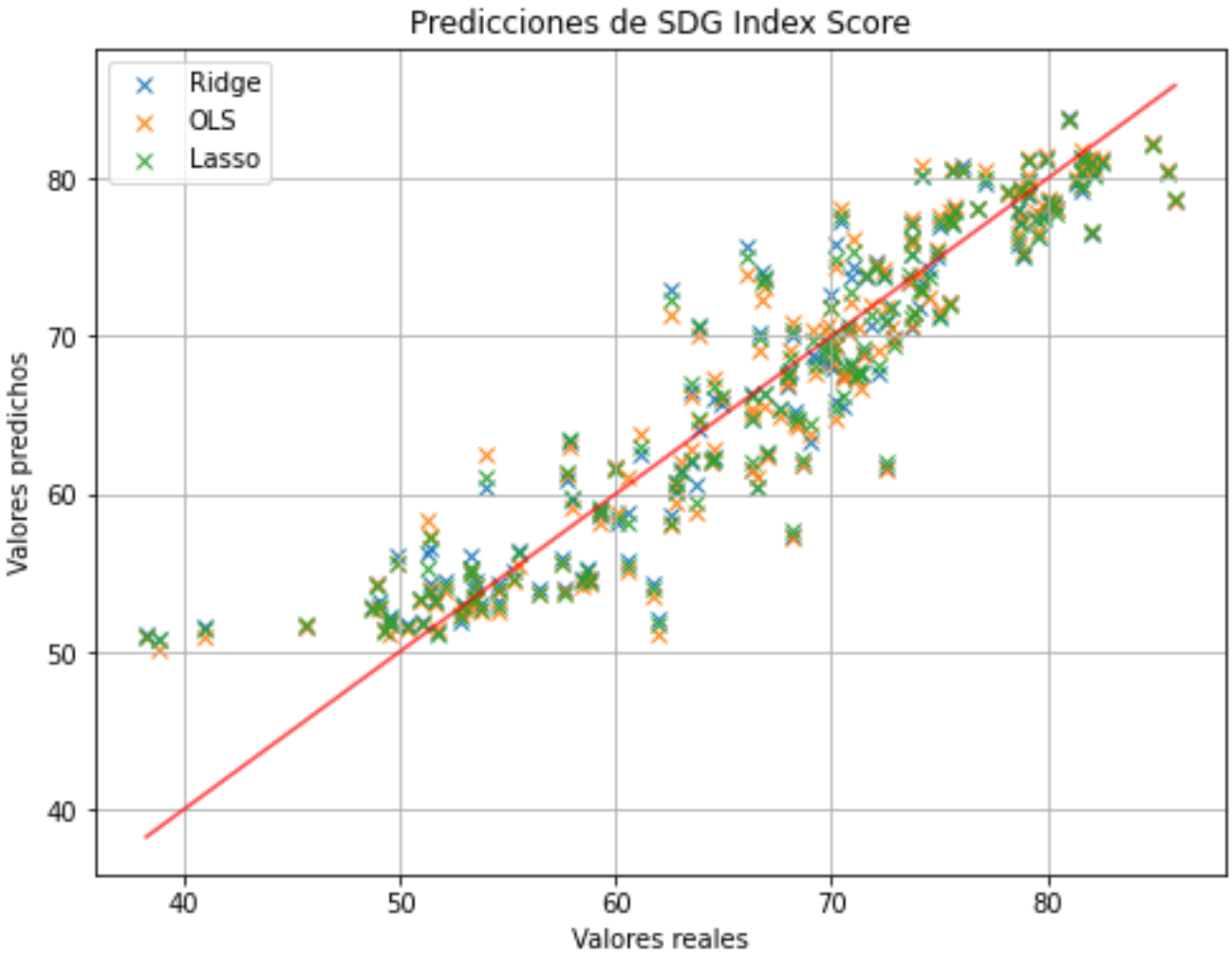


Figura 7. *Predicciones contra valores reales del SDG Index*

las sobreeestimaciones, y el resto de los datos tienden a acercarse a la línea identidad, lo que sugiere que son relativamente cercanos los valores predichos a los valores reales.

Habiendo hecho esto, entonces se predicen los valores de aquellos países para los cuales no estaba documentado el SDG Index Score en los datos de las Naciones Unidas [2]. Estos valores se encuentran reportados en el cuadro 3 para los tres modelos (en todos muy similares).

4. Conclusiones

Se construyeron tres modelos de regresión para predecir el SDG Index Score utilizando como modelo reducido solo los indicadores de las primeras dos metas de desarrollo sustentable. Se encontró que la regularización disminuye muy poco al error cuadrático medio; teniendo resultados similares en algunas métricas. Los residuales tienden a comportarse de manera normal, pero sí parece existir una ligera sobreestimación cuando los valores reales son de SDG Index Scores bajos. El modelo se acerca bastante a los puntajes de las metas de sustentabilidad, lo que sugiere que con un modelo reducido pueden aproximarse los puntajes para países que no se tenga este valor. Se pudieron predecir los valores para los países que no tienen un SDG Index Score con los tres modelos.

Cuadro 3. *Predicciones de valores para variables sin SDG Index*

Country	Ridge	OLS	Lasso
Andorra	78.44	78.28	78.04
Antigua and Barbuda	72.34	71.78	72.53
Comoros	53.99	53.42	53.70
Dominica	67.44	67.73	68.21
Eritrea	54.26	53.92	54.00
Guinea-Bissau	60.97	61.10	61.54
Equatorial Guinea	52.93	52.72	52.77
Grenada	69.87	71.17	68.95
Kiribati	63.23	63.91	63.30
St. Kitts and Nevis	56.36	56.07	56.56
Libya	79.54	79.22	79.02
St. Lucia	60.80	62.58	59.99
Liechtenstein	64.38	66.25	63.82
Monaco	79.54	79.22	79.02
Marshall Islands	70.99	71.80	69.59
Nauru	77.88	77.81	76.05
Palau	69.35	71.37	69.08
Solomon Islands	65.29	63.42	64.47
San Marino	55.86	58.74	55.28
Seychelles	72.98	71.99	72.89
Timor-Leste	65.88	65.70	66.22
Tonga	72.76	71.93	72.48
Tuvalu	55.35	56.19	55.42
St. Vincent and the Grenadines	69.26	71.71	69.20
Samoa	66.83	68.58	66.13

Referencias

[1] United Nations Development Programme. «Sustainable Development Goals — United Nations Development Programme.» (), dirección: https://www.undp.org/sustainable-development-goals?utm_source=EN&utm_medium=GSR&utm_content=US_UNDP_PaidSearch_Brand_English&utm_campaign=CENTRAL&c_src=CENTRAL&c_src2=GSR&gclid=EAIaIQobChMIxePX7YGe9gIVcyCtBh3XCgZREAAYAiAAEgJQV_D_BwE (visitado 26-02-2022).

[2] ——, «Sustainable Development Report 2021,» Sustainable Development Report 2021: Downloads. (), dirección: https://dashboards.sdgindex.org/ (visitado 26-02-2022).

[3] «Income Distribution Database : By Country - POVERTY.» (), dirección: https://stats.oecd.org/index.aspx?queryid=66598 (visitado 28-02-2022).

[4] S. Bonhommeau, L. Dubroca, O. L. Pape y col., «Eating up the world's food web and the human trophic level,» *Proceedings of the National Academy of Sciences*, vol. 110, n.º 51, págs. 20 617-20 620, 17 de dic. de 2013, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1305827110. pmid: 24297882. dirección: https://www.pnas.org/content/110/51/20617 (visitado 28-02-2022).

[5] «Calculating efficiency of biomass transfers - Trophic levels in an ecosystem - AQA - GCSE Biology (Single Science) Revision - AQA,» BBC Bitesize. (), dirección: https://www.bbc.co.uk/bitesize/guides/zs7gw6f/revision/4 (visitado 26-02-2022).

[6] X. Zhang y E. Davidson, «Sustainable Nitrogen Management Index,» Soil Science, preprint, 24 de nov. de 2019. DOI: 10.1002/essoar.10501111.1. dirección: http://www.essoar.org/doi/10.1002/essoar.10501111.1 (visitado 26-02-2022).

[7] I.-K. Yeo y R. A. Johnson, «A New Family of Power Transformations to Improve Normality or Symmetry,» *Biometrika*, vol. 87, n.º 4, págs. 954-959, 1 de dic. de 2000, ISSN: 0006-3444. DOI: 10.1093/biomet/87.4.954. dirección: https://doi.org/10.1093/biomet/87.4.954 (visitado 28-02-2022).

[8] L. Beretta y A. Santaniello, «Nearest Neighbor Imputation Algorithms: A Critical Evaluation,» *BMC Medical Informatics and Decision Making*, vol. 16, n.º 3, pág. 74, 25 de jul. de 2016, ISSN: 1472-6947. DOI: 10.1186/s12911-016-0318-z. dirección: https://doi.org/10.1186/s12911-016-0318-z (visitado 26-02-2022).