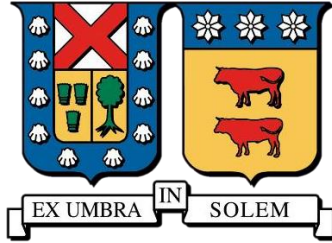


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INDUSTRIAS



**ANÁLISIS DE LA INTERACCIÓN ENTRE VARIABLES DE BIENESTAR SOCIAL
Y EL COMPORTAMIENTO DEL PIB PER CÁPITA MEDIANTE MODELOS DE
APRENDIZAJE SUPERVISADO**

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

AUTOR:	Srta: ZYANYA MONTENEGRO NOVOA
PROFESOR GUÍA:	Dr: PEDRO FERNÁNDEZ DE LA REGUERA B
PROFESOR CORREFERENTE:	Dra. MARIA FARIAS GORDON

SANTIAGO DE CHILE, ENERO 2023

Agradecimientos

Expreso mi más sincero agradecimiento a todos aquellos que me han acompañado durante este largo viaje llamado enseñanza universitaria, ha sido un desafío desde el día 1, pero he logrado aprender y ganar experiencia a medida que esta avanzaba.

Quiero dedicar un especial reconocimiento a mi familia y amigos, quienes me animaban en cada certamen, presentación o trabajo que hiciera, alegrándose por mis resultados y animándome en mis fracasos.

Agradecimientos especiales a la generación 2018 por el nivel de compañerismo en cada proyecto que realizamos o cada certamen que enfrentamos, es maravilloso cuando existen personas interesadas en compartir sus conocimientos y colaborar con otros.

A mi profesor guía que en este último año logro apoyarme en los diferentes hitos de esta investigación.

Y sobre todo gracias a ti Yesenia, por enseñarme lo bello que adquirir conocimiento y aprender sobre cosas nuevas. Tu enseñanza ha demostrado como el conocimiento nos empodera y enriquece, estoy agradecida por la inspiración y la sabiduría que has aportado a mi vida.

Contenido

Agradecimientos	2
1. PROBLEMA DE INVESTIGACIÓN	6
2. OBJETIVOS.....	8
2.1 Objetivo General	8
2.2 Objetivos Específicos.....	8
3. MARCO TEÓRICO	9
3.1 Bienestar Social en Chile	9
3.2 PIB como indicador económico y de Bienestar social	10
3.3 Bienestar social: Campos críticos del desarrollo socioeconómico.	12
3.3 Revisión de Estudios Previos: Estadísticas relevantes.	14
4. METODOLOGÍA.....	19
4.1 Regresión Lineal Múltiple	19
4.2 Datos de panel.....	24
4.3 Árboles de decisión.....	29
4.5 Variables de estudio.....	39
4.6 Descripción de la muestra	44
5. ANÁLISIS EXPLORATORIO DE DATOS (EDA).....	45
5.1 Histogramas	47

5.2	Relación entre PIB per cápita y variables de bienestar social.....	49
5.3	Boxplot PIB per cápita.....	60
5.4	Matriz de correlación: Todas las variables	60
5.6	Decisiones en eliminación de variables: Creación de casos para observar sensibilidad de decisiones.....	63
5.7	Caso 2: Eliminación de variables en base a altas correlaciones y multicolinealidad 64	
5.8	Caso 2: Transformaciones logarítmicas en variables seleccionadas	67
5.9	Caso 4: Tratamiento de datos atípicos.....	70
6.	RESULTADOS	76
6.1	Comparación entre metodologías y casos de sensibilidad en elección de variables 76	
6.2	Variables significativas	79
6.3	Regresión Lineal Múltiple	83
6.4	Árboles de decisión.....	91
6.5	Datos de panel.....	101
7.	DISCUSIÓN.....	106
8.	CONCLUSIÓN	112
9.	ANEXOS	115
	Anexo 1:	115
	Anexo 2	116

Anexo 3	124
Anexo 4:	126
Anexo 5:	133
Anexo 6	135
Bibliografía.....	139

1. PROBLEMA DE INVESTIGACIÓN

El bienestar social es un concepto complejo de definir. En general, se refiere al nivel de calidad de vida experimentado por los individuos. Según el Veenhoven (2000) la calidad de vida abarca tres aspectos: la calidad del entorno en el que vivimos, la calidad de nuestras acciones y el disfrute subjetivo de la vida

Muchos países miden el bienestar de las personas a través del crecimiento del Producto Interno Bruto (PIB) y otros indicadores económicos, asumiendo que un aumento de estos indicadores se traducirá en una mayor felicidad. No obstante, diversas investigaciones han demostrado que la felicidad no está garantizada por un mayor nivel de vida o ingreso. (Layard, 2005)

No obstante, si queremos comprender por qué los países presentan diferencias en sus niveles de vida, debemos tener en cuenta que experimentan divergencias en su tasa de crecimiento a largo plazo. Si estas diferencias se acumulan durante años, tienen consecuencias significativas en los niveles de vida. Una política de crecimiento a largo plazo puede contribuir a mejoras en los niveles de vida. (Barro & Sala, 2004)

Por lo tanto, al invertir en mejoras como el nivel educativo e investigación, se puede mejorar el nivel de vida y tener una influencia positiva en el incremento del PIB. El nivel educativo, por ejemplo, modera las tasas de natalidad y favorece el aumento del PIB, lo que se traduce en una mayor renta per cápita. El impacto económico de la educación se destaca aún más por su papel en mejorar la productividad, fomentar la innovación y aumentar la capacidad de adaptación de las personas a los cambios técnicos e institucionales. (Duran & Peña, 2020)

Sin embargo, la mayoría de los habitantes del mundo viven en países con niveles educativos muy bajos y no se han beneficiado de este efecto positivo en la calidad de vida.

En este contexto, factores como la salud y educación, la equidad en la distribución de ingresos, la seguridad social y la sostenibilidad ambiental emergen como elementos clave que, al potenciarse, no solo mejoran la calidad de vida de la población, sino que también influyen de manera significativa en el crecimiento económico y, por ende, en el PIB.

Lo que lleva a realizar preguntas tales como ¿Qué variables influyen en el Bienestar Social? ¿Como afecta el Bienestar Social al PIB y viceversa? ¿El PIB está relacionado con variables que describen el Bienestar Social? ‘¿Qué variables de bienestar social contribuyen al crecimiento o la disminución del PIB?

2. OBJETIVOS

2.1 Objetivo General

Desarrollar un modelo econométrico para analizar la relación entre variables de bienestar social y el PIB, con el objetivo de determinar si estas variables influyen en el comportamiento del PIB per cápita. Se emplearán modelos de aprendizaje supervisado para identificar el enfoque más eficaz en la comprensión del comportamiento de los datos.

2.2 Objetivos Específicos

- Identificar posibles indicadores que puedan relacionar el Bienestar Social y el Producto Interno Bruto (PIB), incorporándolos en el modelo y verificando su influencia en la relación.
- Analizar la influencia de indicadores sociales y económicos en el comportamiento del PIB utilizando bases de datos disponibles y realizando un análisis exhaustivo de los mismos.
- Determinar los indicadores más relevantes que describan el comportamiento del fenómeno estudiado, empleando preparación de datos, métricas de rendimiento y análisis de resultados para identificar aquellos con mayor poder explicativo.
- Analizar las diferencias entre metodologías de aprendizaje supervisado en la explicación del modelo, evalúan diversas métricas de rendimiento para comprender mejor su eficacia y aplicabilidad
- Realizar un análisis de sensibilidad para evaluar cómo cambios en la selección de indicadores afectan los resultados

3. MARCO TEÓRICO

3.1 Bienestar Social en Chile

Durante los años 1924 y 1973 existió un proceso de expansión de la influencia del Estado sobre las condiciones de vida de la población (Biblioteca Nacional de Chile, 2023). Este proceso se originó a partir de un diagnóstico de los graves problemas presentados por la cuestión social, implementando instituciones que buscaron cubrir las necesidades sanitarias, educativas, de vivienda y protección social, lo que derivó en una expansión de la cobertura de las distintas áreas de bienestar social, ampliando servicios de educación, salud, previsión y vivienda. Un ejemplo de esto son las organizaciones del Servicio Nacional de Salud (SNS) en 1952, la Junta Nacional de Auxilio Escolar y Becas (JUNAEB) en 1964 y la creación del Ministerio de la Vivienda y Urbanismo en 1965. Entre estos años se produjo un desarrollo en el que el común de los habitantes pudo acceder a una mejor salud, previsión, una mayor cobertura educativa y planes de vivienda, acompañado con un descenso marcado en la mortalidad infantil.

Posteriormente desde 1990 Chile ha experimentado una mejoría en indicadores relacionados con pobreza e ingresos. Siguiendo la línea de aumentar y establecer garantías en materia de previsión en salud, cobertura educativa y acceso a viviendas, sumado a un incremento en la cobertura de las pensiones y ampliando el acceso a bienes por parte de las familias. Una reducción significativa de la pobreza y pobreza extrema, tanto en zonas urbanas y rurales se ha visto acompañado por mejoras en el crecimiento del PIB per cápita.

En la actualidad Chile se encuentra en la categoría de desarrollo humano alto, con un índice de 0,847, siendo el más elevado en América Latina (Ministerio de Desarrollo Social y Familia, 2017).

3.2 PIB como indicador económico y de Bienestar social

El Producto Interno Bruto (PIB) es una medida cuantitativa del valor total de todos los bienes y servicios finales producidos dentro de las fronteras de un país durante un período en específico, generalmente un año. Indicando así el tamaño y actividad económica de un país.

Considerando que el Producto Interno Bruto (PIB) proporciona una medida del tamaño de la economía de un país, su crecimiento se asocia comúnmente con el aumento del ingreso personal. Este crecimiento se espera que facilite un mayor acceso a bienes y servicios, contribuyendo así al aumento del bienestar. No obstante, es importante destacar que esta perspectiva no tiene en cuenta la posible aparición de distorsiones e imperfecciones en el mercado, las cuales pueden afectar significativamente el funcionamiento económico de un país (Chacín, 2019).

Durante el año 2006, la Organización Gallup realizó una encuesta mundial en donde analizó los datos sobre la satisfacción con la vida y la salud, en donde se observaron relaciones con el ingreso nacional, la edad y la esperanza de vida. La satisfacción con la vida está fuertemente relacionada al ingreso nacional per cápita, en donde cada duplicación de los ingresos está asociada a un aumento de casi un punto en la satisfacción en escala de 0 a 10 (Deaton, 2009).

No obstante, los ingresos mejoraran la felicidad solo hasta que satisfacen las necesidades básicas, más allá los ingresos adicionales ya no aumentan el nivel de felicidad (Veenhoven., 1991)

El PIB es una variable de interés porque, si bien no mide directamente el bienestar de la población, es un indicador de la producción de bienes y servicios al interior de un país, por lo que eventualmente podría estar asociado al bienestar general de la población.

Lo que nos lleva a preguntarnos ¿Por qué es importante el crecimiento económico y con ello un aumento en el PIB? Esto es importante ya que mide el componente sustancial del progreso social, es decir, el bienestar económico o cuanto se beneficia la sociedad de su uso y asignación de los recursos. El crecimiento del PIB está estrechamente correlacionado con la disponibilidad de empleos e ingresos, los cuales son esenciales para mejorar el nivel de vida de los ciudadanos. Sin embargo, no mide de forma precisa fenómenos físicos. (Coyle, 2017)

Por otro lado, estudios vienen demostrando la interrelación entre el desarrollo humano y el crecimiento económico; la vinculación de la educación, la calificación de la fuerza laboral, con la competitividad del país y el impacto de la desigualdad sobre el crecimiento de la economía. En un sentido de retroalimentación mutua, los logros en materia de desarrollo humano pueden contribuir de manera decisiva al crecimiento económico (Requena, 2008)

.

3.3 Bienestar social: Campos críticos del desarrollo socioeconómico.

- 1. Educación:** La teoría del capital humano considera la educación como inversión con distintos tipos de rendimientos, relacionados con participación en el mercado laboral, la probabilidad de empleo, ingresos, condiciones de trabajo y salud. La educación presenta efectos positivos sobre la salud: las personas más instruidas suelen tener hábitos más saludables y acuden a controles médicos con la frecuencia adecuada. Asimismo, puede ayudar a conseguir trabajos más interesantes proporcionando mayor autonomía, en base a estos rendimientos se puede esperar que la educación incida de forma positiva en el bienestar individual. (Salinas & Salinas, 2008).
- 2. Desempleo:** El desempleo no solo conlleva problemas económicos, sino también patologías sociales, como tasas más altas de divorcio, suicidio y propensión al alcoholismo. Esta conexión va más allá de una simple correlación; existe una relación causal evidente. En repetidas ocasiones, observamos la instauración de un círculo vicioso: a medida que las tasas de desempleo se elevan, la cohesión social se resquebraja, dando lugar a manifestaciones de malestar en la sociedad, como la violencia urbana y los conflictos civiles. Estos fenómenos, a su vez, crean un entorno desfavorable para la inversión y la creación de empleo (Stiglitz, 2002).

3. **Salud:** Se define como perfiles de salud a las formas de enfermar y morir de poblaciones que comparten condiciones homogéneas de vida y trabajo. Durante la historia estos perfiles van presentando características particulares de acuerdo con la configuración de los procesos productivos y como los trabajadores se insertan y participan de este. La salud tiene relación con el bienestar y con las posibilidades de satisfacción de necesidades básicas (Escalona, 2006).
4. **Medio ambiente:** Algunos economistas y muchos científicos no economistas han argumentado que un incremento del Producto Interno Bruto (PIB) dañará el ambiente natural sin duda alguna. De hecho, durante las últimas décadas, el crecimiento económico se ha dado, gracias a un incremento en el uso de energía y a una mayor utilización de los recursos naturales. Otros economistas han sostenido que la economía puede crecer por siempre sin dañar la calidad del medio ambiente. Así, el progreso técnico es considerado como un factor crítico para la reconciliación entre crecimiento y medio ambiente (Correa, 2004).
5. **Pobreza:** El crecimiento económico es necesario, pero no suficiente para reducir la pobreza, en términos de tendencia estos indicadores muestran una correlación negativa, esto se refiere a que en episodios de expansiones económicas han estado relacionados con reducciones de la tasa de pobreza. El estudio de la relación de la pobreza y el crecimiento económico están relacionados con los cambios en la desigualdad de la distribución de ingresos, de este modo si la dinámica de la producción ocurre en actividades donde las poblaciones pobres participan más activamente, este tipo de crecimiento tendrá un efecto reductor de la pobreza como de la desigualdad de distribución del ingreso (García & Céspedes, 2011).

3.3 Revisión de Estudios Previos: Estadísticas relevantes.

1. Relación entre Índice de Gini, % de pobreza y % gasto social/PIB: El índice de Gini cuantifica la desigualdad en la distribución de ingresos dentro de una población, se expresa en escala de 0 a 1, donde 0 es igualdad perfecta y 1 desigualdad máxima. En el siguiente gráfico se observa el índice de Gini (ATI) y Gini (DTI) los que corresponden a índices de Gini antes y después de transferencias e impuestos respectivamente. También se consideran los 25 países de la OCDE y la cifra de la pobreza en Chile es según la Canasta Básica de Satisfacción de necesidades en el año 2006.

País	Población	Gini (ATI)	Gini (DTI)	%Pobreza	%Gasto Social/PIB	Ratio Q1/Q5
EE.UU	299 801.097	0.45	0.38	23.9	16.2	--
Francia	61 203.381	0.48	0.28	14.1	28.7	4.0
Suecia	9 073.522	0.43	0.23	11.4	31.3	3.5
España	43 696.229	0.33	0.31	21.0	20.3	5.3
Grecia	11 142.706	0.45	0.32	19.6	21.3	6.1
Polonia	38 049 168	0.56	0.37	20.8	22.9	5.6
México	104 874.282	--	.047	25.3	6.8	--
OCDE	----	0.44	0.30	17.4	20.7	4.8
Chile	16.730.470	0.54	0.52	13.7	14.9	13.1

Tabla 1: Indicadores Distributivos Chile y Selección de Países de la OCDE (Fuente: Política social e indicadores distributivos. Elementos para una caracterización de la política social en Chile, 2009)

La relación entre el gasto social, la desigualdad y los niveles de pobreza en distintos países es un fenómeno complejo. Países como Francia y Suecia, que destinan mayores porcentajes de su Producto Interno Bruto (PIB) al gasto social, exhiben menores tasas de pobreza.

Sin embargo, resulta intrigante observar que estos mismos países muestran índices de Gini más elevados en comparación con naciones como España, que tiene un menor porcentaje de gasto social/PIB.

El gasto social puede contribuir a reducir la pobreza al proporcionar servicios y apoyar a quienes lo necesitan, la presencia de un índice de Gini más alto podría indicar una mayor concentración de ingresos en ciertos segmentos de la población, a pesar de los esfuerzos de gasto social.

El caso de Chile, con altos niveles de desigualdad a pesar de un menor gasto social/PIB, y tasas relativamente bajas de pobreza, resalta la complejidad de estas relaciones. Este fenómeno podría estar influenciado por diversos factores, incluidos aspectos culturales, históricos y políticos específicos de cada país.

2. Comportamiento Nivel de Pobreza y PIB: La disminución de la tasa de pobreza por ingresos entre 1990 y 2017 se produce en un contexto de buen desempeño de la economía. Se observa que con el transcurso de los años la incidencia de la pobreza ha disminuido considerablemente y a la vez el PIB per cápita ha mostrado un aumento considerable. Si se mantiene constante la línea de pobreza en términos reales, la tasa de pobreza por ingresos solo puede cambiar por un crecimiento económico y con una redistribución de ingreso, los cuales responden a un conjunto de factores vinculados a la dinámica de los mercados y políticas públicas.

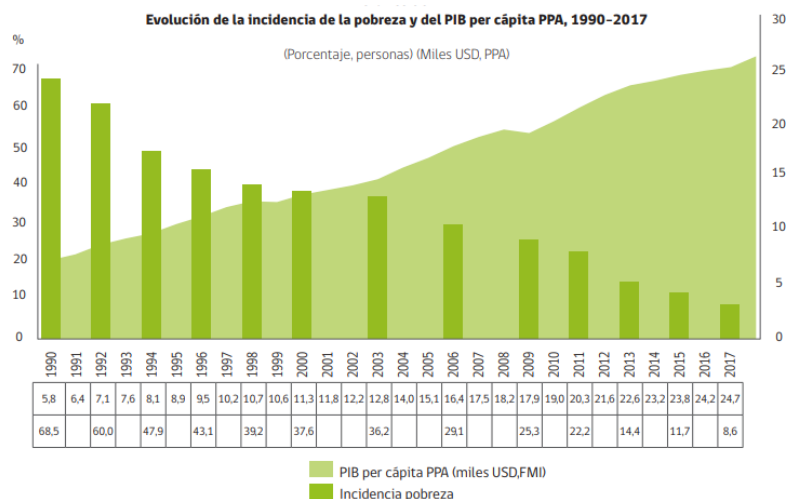


Gráfico 1: Evolución de la incidencia de la pobreza y el PIB, 1990-2017 (Fuente: Ministeriode Desarrollo Social y Familia Evolución de la pobreza 1990-2017, 2018)

En el estudio realizado por el Ministerio de Desarrollo Social y Familia (2018) se obtienen resultados que nos indican que la componente de crecimiento económico tiene un mayor peso relativo en la reducción de la pobreza que la distribución de ingresos, sobre todo en la década de los noventas donde su peso relativo fue sustantivamente mayor que desde el 2000 en adelante, esto sugiere que durante este periodo hubo un aumento en la actividad económica y una posible generación de empleos contribuyendo a reducir la proporción de personas en situación de pobreza.

3. Relación entre felicidad y PIB per cápita: Díaz, Maseda y Neira (2011) argumentan que el bienestar subjetivo enriquece la economía al ir más allá de la tradicional consideración. La integración del bienestar subjetivo en el núcleo económico amplía el enfoque tradicional, proporcionando una comprensión más precisa de cómo las políticas económicas impactan la calidad de vida y las dinámicas entre los elementos subjetivos del bienestar y las prácticas económicas. Observamos que a medida que aumenta el PIB per cápita aumentan los niveles de felicidad por lo tanto estos muestran una correlación positiva, lo que sugiere que, a nivel individual, el aumento del ingreso tiende a aumentar la felicidad. No obstante, esta relación es compleja y está influenciada por diversos factores contextuales y personales.

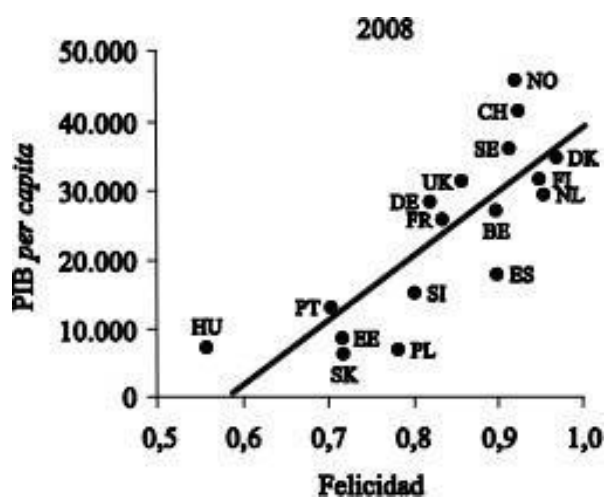


Gráfico 2: Relación entre la Felicidad y PIB per cápita 2008
(Fuente: Revista Galega de Economía, vol20, 2011)

4. Bienestar Social e Inversión en I+D: Cuando se invierte en I+D, se espera que se generen beneficios en términos de crecimiento económico, creación de empleo, avances en la calidad de vida, mejoras en atención médica, entre otros. Estos resultados positivos contribuirían al bienestar social general, esto observa en el siguiente gráfico el cual tiene una tendencia en donde va en aumento el Bienestar social a medida que aumenta la inversión en I+D, a excepción de 2 países que se salen de la tendencia, lo cuál sería interesante de abordar y analizar.

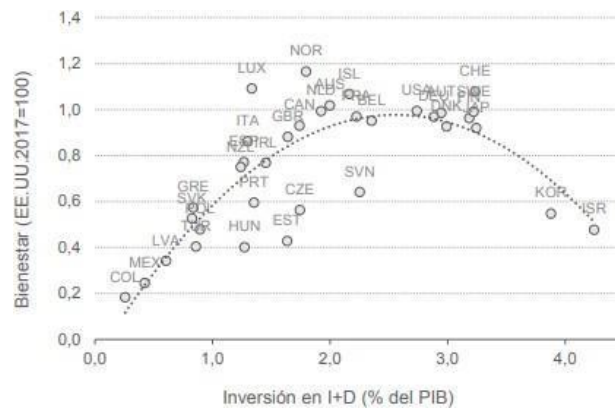


Gráfico 3: Bienestar Social e Inversión en I+D OECD 2010-2017
(Fuente: Innovación y Bienestar Social, 2018)

4. METODOLOGÍA

4.1 Regresión Lineal Múltiple

El análisis de regresión lineal es probablemente la forma más sencilla y utilizada de medir relaciones entre variables predictoras continuas y de respuesta. El método requiere un modelo que asocia una variable (resultado) a una función de variables causales. El modelo representa un promedio ponderado de las causas, donde los coeficientes de ponderación se entienden lineales; es decir, son únicos en cada término y están en potencias unitarias. Entonces, para cada combinación de las variables causales (escenario) se aplica la función y el resultado es la respuesta promedio a ese escenario

En la evaluación de modelos estadísticos, diversas métricas y pruebas ofrecen una panorámica completa. Estas herramientas proporcionan una comprensión sobre la eficacia de los modelos de regresión y análisis de varianza. Estas herramientas serán utilizadas para un mayor entendimiento del modelo y se pueden definir como:

- a. Coeficiente de determinación (R^2): Es una medida que indica la proporción de variabilidad en la variable dependiente que es explicada por el modelo de regresión. Es decir, es la proporción de varianza de las variables de respuesta que se puede predecir a partir de las variables predictoras. (Sitiobigdata, 2018)

- b. MSE (Error cuadrático medio) y RMSE (Raíz del error cuadrático medio): El MSE es una métrica comúnmente utilizada en la evaluación de modelos de regresión. Representa la media de los errores al cuadrado entre las predicciones del modelo y los valores reales de la variable de respuesta. La fórmula para calcularlo es la siguiente

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Siendo n el número de observaciones, y_i es el valor real de la variable de respuesta en la observación i y $\hat{y}_i$ corresponde a la predicción del modelo para la observación.

Cuando se utiliza MSE se elevan al cuadrado los errores antes de calcular la media, esto penaliza de manera más significativa los errores grandes. Por otro lado, el RMSE es la raíz cuadrada del MSE, la ventaja es que proporciona una medida de error en la misma escala que la variable de respuesta original, lo que facilita su interpretación. Ambas métricas, MSE y RMSE, son utilizadas para cuantificar la precisión del modelo (Sitiobigdata, 2018)

- c. P-values: En el contexto de los valores p , la comparación se realiza entre la hipótesis nula y la hipótesis alternativa (H_1), que sostiene la existencia de un efecto o relación significativa. Un valor p bajo indica que los resultados observados son poco probables bajo la suposición de que no hay efecto real (hipótesis nula). Por lo tanto, cuando se encuentra un valor p pequeño, se fortalece la evidencia en contra de la hipótesis nula, lo que sugiere que la variable en cuestión es significativa en el modelo o análisis. (Addlink, 2015).

d. AIC (Criterio de información Akaike) y BIC (Criterio de información Bayesiano):

Son medidas de calidad relativa de un modelo estadístico estos son usados para compararmodelos. Modelos con valores más bajos de AIC o BIC se consideran preferibles. El AIC mide el desajuste entre una distribución hipotética y una distribución teórica, es una medida global de bondad de ajuste del modelo, su cálculo se realiza desde un punto de vista predictivosuponiendo que los modelos identificados a partir de este criterio tienen un buen comportamiento respecto de su predicción.

Por otro lado, el BIC introduce una modificación bayesiana en la construcción del criterio AIC, este criterio penaliza más los modelos con un número mayor de parámetros estimados corrigiendo la tendencia a la sobrestimación observada en AIC. (Jiménez, 2018)

e. Durbin – Watson: El estadístico de Durbin – Watson se utiliza para evaluarla presencia de autocorrelación en los residuos del modelo de regresión. Toma valores entre 0 y 4, en donde valores cercanos a 2 indican ausencia de autocorrelación (Sandoval, 2011). La ausencia de autocorrelación en los residuos implica que los errores no están relacionados entre sí a travésdel tiempo o las observaciones.

f. Jarque – Bera: La prueba de Jarque - Bera es una prueba de bondad de ajuste que evalúa si una muestra tiene la asimetría y la kurtosis que se esperarían en una muestra normal. Es utilizado para verificar la normalidad de los residuos del modelo (Gascón, 2017). En el contexto de la regresión, la prueba de Jarque-Bera se aplica a los residuos del modelo para determinar si estos siguen una distribución normal. La asimetría se refiere a la falta de simetría en la distribución, mientras que la kurtosis mide la "colas" de la distribución, es decir, qué tan pesadas o ligeras son en comparación con una

distribución normal.

Modelos de regresión lineal múltiple utilizados: Los modelos utilizados fueron programados en Python utilizando librerías como numpy, pandas, matplotlib, seaborn, scikit-learn, entre otras.

1. **Regresión Lineal Múltiple SKlearn:** Sklearn es una biblioteca líder en aprendizaje automático, utiliza el método de mínimos cuadrados para regresión. Reconocida por su versatilidad, la librería ofrece una amplia variedad de algoritmos para tareas como clasificación, regresión, reducción de dimensionalidad y agrupamiento. Además de su robusto conjunto de modelos, Scikit-learn proporciona módulos esenciales para preprocesamiento de datos, extracción de características y optimización de hiperparámetros, simplificando todo el proceso de construcción y evaluación de modelos en el flujo de trabajo de aprendizaje automático (Hackeling, 2017).
2. **Regresión Lineal Múltiple Stepwise:** La regresión paso a paso (stepwise) es la construcción iterativa paso a paso de un modelo de regresión que implica la selección automática de variables independientes. El objetivo de la regresión stepwise, es a través de pruebas F y t encontrar un conjunto de variables independientes que influyan significativamente en la variable dependiente. (Mendez, 2019)

3. **Regresión Lineal Múltiple Ridge:** Como se mencionó anteriormente se utiliza el método de mínimos cuadrados para implementar la regresión lineal múltiple, esta técnica es ampliamente aceptada para resolver este problema. Sin embargo, en presencia de multicolinealidad, la técnica de mínimos cuadrados puede mostrar un desempeño deficiente como metodología de solución.

Aunque las estimaciones de mínimos cuadrados son insesgadas, sus variaciones son grandes y pueden estar lejos del valor real. Es por esto que la regresión Ridge es una técnica estándar para abordar este tipo de problemas. (Rajan, 2022)

El término α dentro de la regresión Ridge se refiere al parámetro de penalización o regularización. Un valor más alto de α implica una regularización más fuerte, esto conduce a coeficientes de modelos más pequeños y reduce la complejidad del modelo. Para este modelo se utilizará el valor de $\alpha=1$ lo que nos indica una regularización moderada que busca mitigar la multicolinealidad.

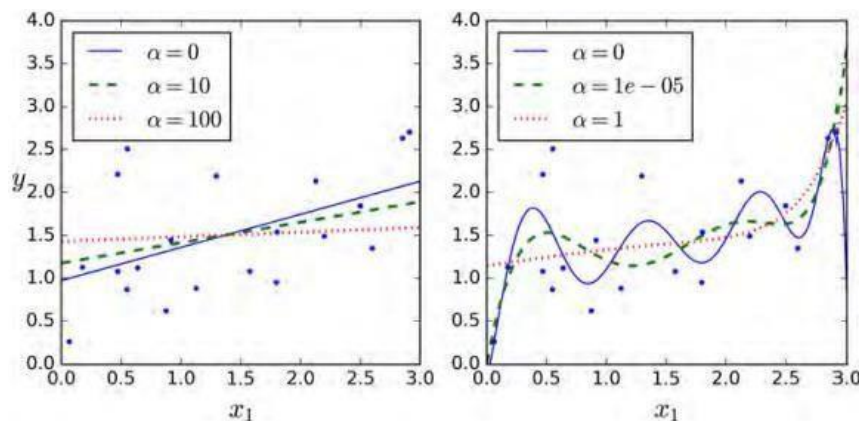


Figura 1: Ridge regression

(Fuente: Hands-on Machine Learning with Scikit -Learn, Keras& Tensorflow, 2017)

4.2 Datos de panel

Mayorga M y Muñoz S (2000) explica que para el análisis de la información pueden existir diferentes dimensiones sobre las cuales interesa obtener conclusiones según las estimaciones de los modelos que traten de extraer relaciones de causalidad o comportamiento entre diferentes tipos de variables a partir de datos disponibles. En este contexto, el modelo de datos de panel combina dos dimensiones esenciales: la dimensión longitudinal y dimensión temporal.

La dimensión longitudinal se refiere a la variedad de entidades o unidades observadas en el análisis. Pueden ser países, empresas, regiones, entre otros. Cada entidad constituye una unidad individual que aporta información única al modelo.

Por otra parte, la dimensión temporal aborda el aspecto del tiempo en el análisis. Se considera un conjunto de observaciones a lo largo de un periodo específico. Por ejemplo, datos recopilados anualmente, mensualmente o cualquier otra escala de tiempo.

Existen tres tipos de regresión para datos de panel, las cuales fueron detalladas por Brugger (2021) en su guía de regresiones de datos de panel para su implementación en Python, la información que viene a continuación es en base al artículo mencionado:

1. Pooled OLS: "Pooled OLS" o "modelo de efectos comunes", es una simplificación que trata los datos de panel mediante el método de Mínimos Cuadrados Ordinarios (OLS). Al omitir la consideración del tiempo y las características individuales, busca identificar patrones generales en las dependencias entre individuos.

La limitación clave radica en la necesidad de que no haya correlación entre variables independientes no observadas y las variables predictoras. Esto implica que las características no medidas individualmente no deben influir sistemáticamente en las variables explicativas, lo que puede ser una restricción significativa en casos donde existe esta correlación

2. Modelo de efectos fijos: Los modelos de efectos fijos son herramientas estadísticas que asumen que los efectos individuales de variables no observadas son constantes a lo largo del tiempo. Aunque permiten la presencia de heterogeneidad dentro del modelo, al considerar estos efectos como fijos, las dependencias solo pueden ser observadas dentro de los individuos. Este enfoque proporciona una comprensión detallada de las variaciones específicas a cada individuo, pero puede limitar la capacidad del modelo para generalizar patrones de comportamiento a lo largo del tiempo en el análisis de datos de panel.

3. Modelo de efectos aleatorios: Los modelos de efectos aleatorios son herramientas estadísticas que incorporan efectos individuales como variables aleatorias, permitiendo una flexibilidad única. Al considerar tanto la variabilidad entre individuos como a lo largo del tiempo, estos modelos ofrecen la capacidad de cambiar entre OLS y efectos fijos, lo que les permite abordar eficazmente tanto las dependencias entre individuos como las dependencias temporales.

Esta versatilidad hace que los modelos de efectos mixtos sean fundamentales para capturar la heterogeneidad no observada y proporcionar información precisa en el análisis de datos de panel, mejorando la capacidad de generalización y adaptación a complejas estructuras de variabilidad

Existen supuestos para los modelos de regresión lineal que deben cumplirse. Tres de ellos pueden ayudarnos a decidir entre estos modelos. Si los supuestos de exogeneidad, homocedasticidad y no autocorrelación no se cumplen, los modelos de efectos fijos y aleatorios podrían ser más adecuados que una regresión lineal.

Para elegir entre modelos de efectos fijos y aleatorios podemos implementar las siguientes pruebas detalladas por Brugger (2021):

a. **Gráfico de residuos para heterocedasticidad:** La representación de los valores previstos frente a los residuos, comúnmente conocido como gráfico de dispersión residual, es una herramienta valiosa para evaluar la presencia de heterocedasticidad en un modelo. Cuando los puntos trazados en el gráfico muestran una dispersión creciente o patrones no uniformes alrededor de la línea de cero, esto sugiere la presencia de heterocedasticidad.

b. **Prueba de Breusch-Pagan:** Es una herramienta estadística utilizada para evaluar la presencia de heterocedasticidad en un modelo de regresión. Este concepto se relaciona con la variabilidad no constante de la varianza de los errores a lo largo de los valores de las variables independientes. La prueba se centra específicamente en la varianza de los errores y se emplea para diagnosticar problemas de homocedasticidad.

La heterocedasticidad implica que la dispersión de los errores no es constante en todas las observaciones, lo que puede afectar la eficiencia y la validez de las estimaciones del modelo de regresión. La prueba de Breusch-Pagan busca detectar este fenómeno y proporcionar evidencia sobre la presencia o ausencia de heterocedasticidad en los residuos del modelo

La hipótesis nula de la prueba establece que no hay heterocedasticidad en los errores del modelo. Es decir, implica que la varianza de los errores es constante en todas las observaciones. La hipótesis alternativa, por otro lado, sugiere la presencia de heterocedasticidad.

La prueba produce un valor p asociado con la hipótesis nula. Si el valor p es inferior a un umbral predeterminado, comúnmente 0.05, se interpreta como evidencia significativa de heterocedasticidad. En este caso, se rechaza la hipótesis nula de homocedasticidad, indicando que la varianza de los errores no es constante.

c. **Prueba Durbin-Watson:** La prueba de Durbin-Watson, con resultados alrededor de 2 indicando ausencia de autocorrelación, valores entre 0-2 señalando autocorrelación positiva, y valores entre 2-4 indicando autocorrelación negativa, ofrece información sobre patrones en los residuos de un modelo de regresión. En casos de autocorrelación positiva, se sugiere que modelos de efectos fijos o aleatorios podrían ser más apropiados, mientras que, para autocorrelación negativa, estos modelos también son recomendados. El modelo de datos de panel, al considerar la variabilidad entre unidades y a lo largo del tiempo, ayudan a manejar patrones sistemáticos no capturados por modelos estándar de regresión. La elección entre efectos fijos y aleatorios dependerá de la naturaleza específica de la

autocorrelación observada en los residuos.

d. **Prueba de Hausmann:** La prueba de endogeneidad evalúa la covarianza entre las variables predictoras y el término alfa en un modelo. La hipótesis nula sugiere que esta covarianza es cero, indicando que no hay endogeneidad. Si la hipótesis nula se rechaza con un valor p pequeño, por lo tanto, se prefiere un modelo de efectos fijos, ya que sugiere la presencia de endogeneidad en el modelo. En caso de no rechazar la hipótesis nula, se opta por un modelo de efectos aleatorios.

4.3 Árboles de decisión

Un árbol de decisión es un modelo predictivo que separa el conjunto de predictores y agrupa observaciones con valores próximos para la variable dependiente. Un árbol se forma a través de variables binarias que se activan, si la variable es numérica se habla de árboles de decisión y en cambio, si es categórica corresponden a árboles de clasificación (Fernández, 2022).

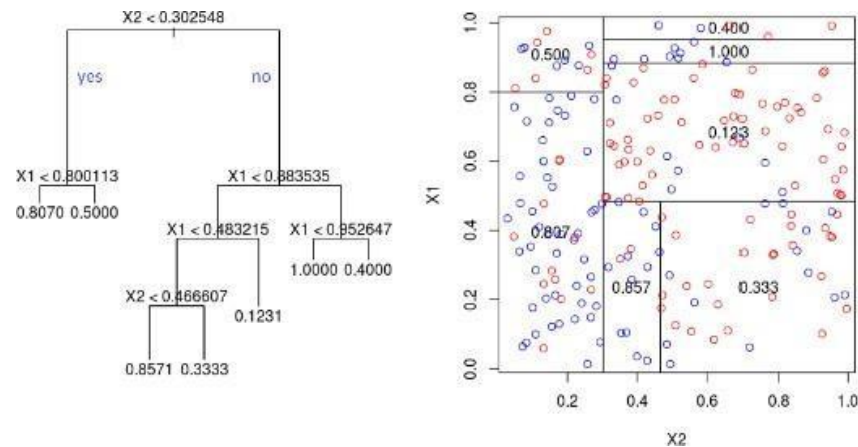


Figura 2: Estructura de Árbol de regresión y partición del espacio (Fuente: Modelospredictivos, 2023)

La metodología de árboles de decisión, según Ejea (2017), se basa en la división del espacio predictor en regiones simples mediante ramas que representan tomas de decisiones basadas en proposiciones lógicas. Cada hoja del árbol representa una región del espacio multidimensional de predictores, y la predicción se realiza utilizando la media de los datos de entrenamiento en esa región. Este enfoque toma decisiones localmente óptimas en cada paso, siendo eficiente al encontrar cortes

óptimos y reduciendo el gasto computacional.

Un árbol de regresión consiste en hacer preguntas tipo $\{x_k < c\}$ para las variables dependientes, de esta forma el espacio de las covariables es dividido en hiper-rectángulos, las observaciones que queden dentro de este tendrán el mismo valor estimado. El objetivo es realizar la partición del espacio de manera repetitiva para encontrar variables y valores de corte que minimicen la función de costos. (Hernández, 2023). Entendemos el parámetro c como un umbral o valor de corte, de esta forma dividimos el espacio de los valores de las variables independientes, creando subconjunto de datos.

El objetivo de realizar divisiones es que minimicen la función de costo global (MSE), proporcionando así un modelo de regresión que se ajuste bien a los datos de entrenamiento.

Las definiciones asociadas a los árboles son:

- Nodo: Unidad que construye el árbol
- Nodo hijo – Nodo padre: nodos resultantes de otros nodos, el nodo de origen será llamado nodo padre.
- Nodo raíz: Nodo sin antecesor.
- Nodos hermanos: Aquellos que compartan el mismo padre
- Nodos terminares: Aquellos que no tienen nodos hijos.
- Grado: número de descendientes directos que tiene un nodo
- Nivel: número de arcos que se recorren desde la raíz.

- Grado del árbol: Es el máximo de grados de sus nodos
- Altura: El máximo de los niveles

En su aplicación en Python estos criterios son detallados por Géron (2017)

los cuales se pueden definir como:

1. Criterion: Especifica la función para medir la cantidad de una división. El criterio usado para la regresión es el MSE, este mide la varianza de los valores objetivo en un nodo.
2. Max_depth: Controla la profundidad máxima del árbol. Limitar la profundidad puede ayudar a prevenir el sobreajuste.
3. Min_samples_leaf: El número mínimo de muestras requeridas para ser una hoja (nodo final). Ayuda a controlar el tamaño mínimo de las hojas.
4. Random_state: Semilla para la generación de números aleatorios, esta es útil para poder reproducir el resultado.

Para evitar el sobreajuste (*Overfitting*), que puede ocurrir al tomar muestras muy pequeñas y hacer que el modelo se ajuste demasiado a los datos de entrenamiento, se implementan criterios de parada. Estos criterios, como la profundidad máxima del árbol limitan la profundidad del árbol, ayuda a evitar que el árbol se vuelva demasiado complejo y se ajuste demasiado a los datos de entrenamiento.

También se destaca la posibilidad de pre-podar el árbol especificando criterios de parada antes de que se cultive completamente. En esencia, estos criterios actúan como precaución para garantizar que el árbol no capture detalles irrelevantes o ruido en los datos de entrenamiento, lo que podría conducir a un rendimiento deficiente en nuevos datos no vistos. Es decir, se busca un equilibrio para que el modelo aprenda patrones generales sin ajustarse en exceso a las particularidades del conjunto de entrenamiento.

Se implementarán las siguientes metodologías para evitar el sobreajuste:

1. *Cost complexity Prunning*: En este caso, vamos a realizar un tipo de poda llamado "poda de coste de complejidad" o "poda del eslabón más débil". Esta es utilizada en árboles de decisión para evitar el sobreajuste y se basa en la incorporación de un término de penalización en la función de costo del árbol. Esta se expresa mediante la formula obtenida desde *PennState Eberly College of Science* (2023)

$$R\alpha(T)=R(T)+\alpha\cdot|T|$$

Donde:

$R\alpha(T)$: Costo total ajustado incluyendo término de penalización.

$R(T)$: Costo original del árbol

α : Parámetro de penalización

$|T|$: Número de nodos terminales en el árbol.

El objetivo principal es minimizar el Costo total ajustado de manera que sea lo suficientemente complejo para adaptarse a los datos de entrenamiento, pero no tan complejo para sobreajustarse. El parámetro α controla la fuerza de penalización, a medida que esta aumenta, la penalización por tener más nodos en el árbol se vuelve más fuerte.

La metodología para llevar a cabo esta poda es detallada por Ejea (2017) que implica estimar el error de un subárbol específico utilizando validación cruzada. Esta técnica nos ayuda a evaluar qué tan preciso es nuestro modelo.

Para llevar a cabo esto, utilizaremos la técnica de "k bloques". Imaginemos que divides tu conjunto de datos en k partes. Luego, k veces, seleccionas una de esas partes como conjunto de prueba y el resto como conjunto de entrenamiento. Esto se conoce como "Método de *K-fold cross validation*".

En cada iteración, entrenas el modelo en el conjunto de entrenamiento y lo evalúas en el conjunto de prueba. Repites este proceso k veces, cambiando qué parte es el conjunto de prueba en cada iteración.

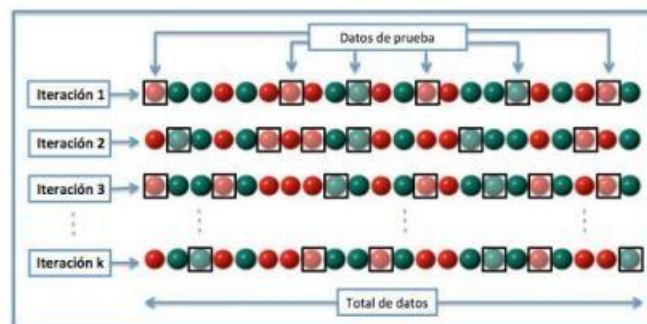


Figura 3: Método de *K-fold cross validation*

(Fuente: Análisis de Sensibilidad mediante Random Forest, 2018)

Después de estas k iteraciones, promediamos los errores obtenidos en cada prueba. Este promedio nos da una buena estimación del rendimiento del modelo en diferentes conjuntos de datos. En términos más simples, estamos tratando de entender qué tan bien nuestro modelo generaliza a datos que no ha visto durante el entrenamiento.

Este enfoque de poda se basa en encontrar el equilibrio adecuado entre la complejidad del modelo y su capacidad para generalizar a nuevos datos.

2. Mejores hiperparámetros: El objetivo aquí es encontrar la combinación óptima de hiperparámetros para el árbol de decisión, especialmente la profundidad máxima y el número mínimo de muestras en una hoja. Este enfoque sigue la metodología descrita por Kohrsen (2018).

En lugar de muestrear aleatoriamente distribuciones de hiperparámetros, se opta por evaluar explícitamente cada combinación predefinida de configuraciones. Es decir, se establecen valores específicos para la profundidad máxima y el número mínimo de muestras en una hoja y se evalúa cómo el modelo se comporta con estas configuraciones.

Se inicia probando diferentes combinaciones de hiperparámetros y registrando el rendimiento del modelo para cada una. A medida que se avanza y se identifican las combinaciones que conducen a un rendimiento óptimo, se puede refinar la búsqueda utilizando una cuadrícula más detallada alrededor de las configuraciones previamente exitosas.

Esto implica ajustar más finamente los valores de los hiperparámetros para encontrarla configuración que maximice el rendimiento del modelo.

En lugar de depender de la aleatoriedad, este enfoque sigue un proceso más sistemático, evaluando de manera explícita combinaciones predefinidas y ajustándolas de manera refinada a medida que se obtienen mejores resultados.

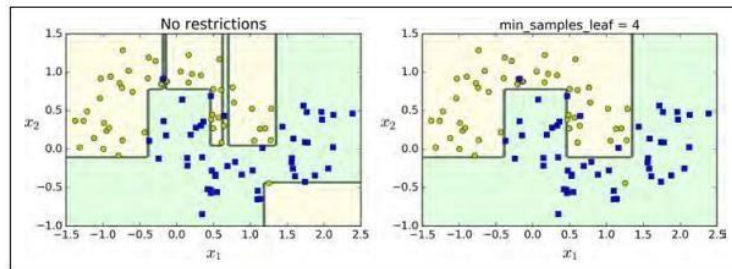


Figura 4: Uso de *min_samples_leaf* como hiperparámetro (Fuente: Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow, 2017)

4.4 Random forest tree: Importancia de las variables.

Un árbol de bosque aleatorio o *Random forest tree* es un algoritmo de aprendizaje automático supervisado utilizado para clasificación y regresión. Se basa en la construcción de múltiples árboles de decisión, donde cada árbol se entrena con una muestra aleatoria independiente. Estos árboles se combinan para formar un bosque, y al contar con varios modelos, se busca mejorar la precisión en comparación con un solo árbol. Este algoritmo evita el sobreajuste al considerar diversas muestras y, en problemas de regresión, proporciona como resultado final el promedio de las predicciones de todos los árboles en el conjunto (Talari, 2022).

En cada nodo de decisión de cada árbol, se elige una muestra diferente de características para llevar a cabo la división. Esto garantiza que cada árbol tenga en cuenta diferentes perspectivas y características en su proceso de toma de decisiones. Según Mwiti (2023) cada árbol realiza su propia predicción individual basa en la muestras y características seleccionadas, luego obtiene una predicción final, las salidas de cada árbol se promedian. Este proceso de promedio contribuye a reducir el impacto de las decisiones individuales sesgadas, mejorando así la robustez y la capacidad de generalización del modelo.

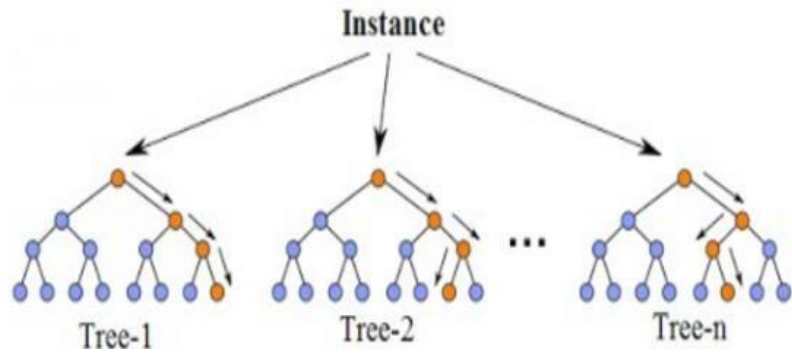


Figura 5: Estructura de árbol de bosque aleatorio (Fuente:Random Forest Regression, 2023)

Para evaluar la importancia de las características, podemos utilizar la profundidad de las características utilizadas como nodo de decisión en un árbol. Este enfoque nos ayuda a entender la importancia relativa de cada característica en relación con la capacidad del modelo para predecir la variable dependiente.

En el contexto de importancia relativa y su implementación en SKlearn, Pedregosa (2011) indica que las características que se utilizan en los nodos superiores del árbol contribuyen a una mayor proporción de las muestras de entrada, lo que significa que influyen más en las predicciones finales. Según sus definiciones, la fracción esperada de muestras a las que una característica contribuye se utiliza como una estimación de su importancia relativa.

La clave es comprender cómo las variables de entrada impactan en la salida del modelo. Aquellas que causen más variabilidad en las predicciones serán cruciales para entender el comportamiento del modelo. Según García (2018), la importancia se mide mediante el aumento en el Error Cuadrático Medio (MSE). Cada árbol en el *Random Forest* calcula su propio MSE. Aquí se expone el proceso:

1. Cálculo del MSE Inicial: Iniciamos calculando el MSE con las variables que no se utilizaron para entrenar cada árbol. Este MSE inicial nos indica la discrepancia en las predicciones del modelo.
2. Permutación Aleatoria: Luego, elegimos una variable y la permutamos aleatoriamente, alterando así las predicciones del modelo.
3. Cálculo del Nuevo MSE: Se recalcula el MSE con la variable permutada. Un aumento sustancial en el MSE sugiere que esa variable tiene un impacto significativo en las predicciones del modelo.

Este proceso se repite para cada variable, identificando aquellas que generan mayores cambios en las predicciones (mayor aumento del MSE), las cuales se consideran más importantes en la explicación del modelo

4.5 Variables de estudio

Las variables por estudiar abarcan aspectos sociales y de calidad de vida en torno a empleo, salud, educación, medioambiente y pobreza. Los datos corresponden a registros anuales, entre los años 2000 y 2020 ambas fechas incluidas.

Variable dependiente.

PIB per cápita: El PIB per cápita es el producto interno bruto dividido por la población a mitad de año. El PIB es la suma del valor agregado bruto de todos los productores residentes en la economía más todo impuesto a los productos, menos todo subsidio no incluido en el valor de los productos. Se calcula sin hacer educciones por depreciación de bienes manufacturados o por agotamiento y degradación de recursos naturales. Datos en US\$ a precios actuales. (World Bank ,2023)

Variables independientes.

1. *Tasa de finalización nivel primario*: Corresponde al número total de estudiantes que ingresan al último año de educación primaria, independientemente de su edad, expresado como porcentaje de la población total en edad oficial de ingresar a dicho grado. Este indicador también se denomina “tasa bruta de ingreso al último grado de primaria”. Esta tasa puede ser superior a 100% debido a niños mayores y menores de la edad oficial de ingreso que entran a la educación primaria de forma precoz o tardía y (o) que han repetido grado (World Bank,2023). Este indicador no solo refleja la matrícula de último año, sino también los posibles desafíos en la progresión de los estudiantes a través de la educación primaria

Si consideramos la repetición de grado esta puede influir en la tasa, y la identificación de estas variaciones proporciona información valiosa para entender el panorama educativo. Además, el contexto demográfico y las políticas educativas pueden influir en el indicador.

2. *Tasa de desempleo*: El desempleo es la proporción de la población activa que no tiene trabajo pero que busca trabajo y está disponible para realizarlo (World Bank, 2023). Es crucial para evaluar el rendimiento del mercado laboral y su capacidad de absorber la mano de obra disponible, una tasa de desempleo alta puede indicar desafíos económicos y dificultades en la inserción laboral de la población activa. Debemos considerar que ocurren fluctuaciones a lo largo de los años, lo cual puede estar influido por políticas gubernamentales, inversión en educación u otros indicadores macroeconómicos.

3. *Tasa de mortalidad en menores de 5 años*: Es la probabilidad por cada 1.000 que un recién nacido muera antes de cumplir cinco años, está sujeto a las tasas de mortalidad específicas por edad del año especificado (World Bank, 2023). Esta puede indicar un mejor acceso a la atención médica como mejores condiciones sanitarias y sociales. Por otro lado, si la tasa tiene un valor elevado puede señalar deficiencias en la atención médica.

4. *Esperanza de vida al nacer*: Indica la cantidad de años que viviría un recién nacido si los patrones de mortalidad vigentes al momento de su nacimiento no cambian a lo largo de la vida del infante. (World Bank, 2023). Es fundamental para evaluar la calidad de vida y eficacia de los sistemas de salud, esta sugiere un acceso efectivo a la atención médica y una población relativamente saludable

Por otro lado, si la esperanza de vida es baja puede indicar desigualdades en el acceso a servicios de salud o condiciones sociales complejas. Hacer un seguimiento de la esperanza de vida al nacer puede proporcionar información sobre el proceso de mejora de la salud pública y la calidad de vida lo cual puede estar relacionado con PIB per cápita

5. *Tasa de mortalidad materna*: Es la cantidad de mujeres que mueren durante el embarazo y el parto por cada 100.000 nacidos vivos. Los datos se estiman según un modelo de regresión que usa información sobre fertilidad, asistentes al parto y prevalencia del VIH (World Bank, 2023). Este indicador puede ayudarnos a evaluar la eficacia de los sistemas de atención materna y salud reproductiva en una sociedad. Una tasa de mortalidad materna baja puede indicar un acceso efectivo a la atención prenatal, durante el parto y post parto.

6. *Incidencia de tuberculosis*: Representa la cantidad estimada de nuevos casos de tuberculosis pulmonar, de frotis positivo y extrapulmonar por cada 100.000 personas. (World Bank, 2023) Este indicador proporciona información valiosa sobre la carga de la enfermedad y la eficacia de los programas de prevención y control de tuberculosis. Una incidencia baja de tuberculosis podría indicar un control efectivo

de la enfermedad, lo cual se puede originar debido a un diagnóstico temprano y tratamiento adecuado.

7. *Prevalencia de desnutrición*: Población ubicada por debajo del nivel mínimo de consumo alimenticio de energía, muestra el porcentaje de la población cuya ingesta de alimentos no alcanza a satisfacer sus requisitos alimenticios de energía de manera continua (World Bank, 2023).

Una prevalencia baja indica que la mayoría de la población tiene acceso adecuado a alimentos nutritivos, al contrario, si existe una alta prevalencia de desnutrición sugiere dificultades al momento de acceder a alimentos adecuados, lo que puede conducir a problemas en la salud.

8. *Emisiones de CO₂*: Se refiere a la liberación de dióxido de carbono (CO₂) a la atmósfera como resultado de actividades humanas, como la quema de combustibles fósiles, deforestación y procesos industriales (World Bank, 2023). El CO₂ es uno de los gases de efecto invernadero responsables del cambio climático, ya que contribuye al atrapar el calor en la atmósfera y aumentar la temperatura global. Este indicador puede servir para evaluar el impacto ambiental de las actividades humanas, un aumento de CO₂ puede traer consecuencias significativas en el ambiente afectando patrones climáticos.

9. *Consumo de energía renovable*: El consumo de energía renovable es la proporción de energía renovable en el consumo total de energía final (World Bank, 2023). Las fuentes renovables incluyen recursos naturales como energía solar, eólica, hidroeléctrica, geotérmica y biomasa, estas son sostenibles y generalmente tienen un impacto ambiental menor en comparación a combustibles fósiles.

10. Población total: Se basa en la definición de facto de la población, que cuenta a todos los residentes independientemente de su estatus legal o ciudadanía. (World Bank, 2023)

11. Población mayor de 65 años: Se basa en la definición establecida de dicho término, que contabiliza a todos los residentes independientemente de su situación legal o de ciudadanía, excluyendo a los refugiados no asentados permanentemente en el país de acogida, a quienes se les considera usualmente.

12. Tasa de incidencia de pobreza por debajo del 40%: Se refiere a la proporción de la población que vive en condiciones de pobreza, donde la línea de pobreza está definida como un ingreso per cápita que es inferior al 40% del ingreso mediano de la población. En otras palabras, mide la prevalencia de la pobreza extrema al considerar a aquellos cuyos ingresos caen significativamente por debajo de un umbral crítico (Our World in Data, 2023). Esta tasa nos proporciona la prevalencia de la pobreza extrema, ya que se centra en aquellos cuyos ingresos están significativamente por debajo del umbral crítico, lo que puede servir para definir políticas y programas diseñados para elevar o equilibrar los ingresos de los grupos más desfavorecidos lo que puede llegar a derivar en una mejor calidad de vida de la población.

13. Participación de ingreso del 20% peor remunerado: Es la participación porcentual en el ingreso o en el consumo, es la participación que se devenga a subgrupos de población representados en deciles o quintiles (World Bank, 2023). Este enfoque se puede utilizar para evaluar la distribución del ingreso, nos lleva a comprender como se comparten los recursos económicos.

4.6 Descripción de la muestra

En esta investigación se tomó como muestra 33 países (véase en anexo 1) los cuales se distribuyen de la siguiente forma:

- Europa: 22 países (66.67%)
- América: 9 países (27.27%)
- Asia: 2 países (6.06%)

Para la obtención de los indicadores se tuvo como criterio un lapso de 21 años (2000-2020) en donde cada país debe tener al menos 15 datos por indicador, para los datos faltantes se rellenó con el promedio de las observaciones por año para cada país, esto se realizó para obtener una muestra más representativa, en total se tienen 693 observaciones para 14 indicadores (9.702 observaciones en total).

La muestra se puede clasificar en tres categorías principales según sus niveles de ingreso:

Niveles de ingreso	Cantidad de países
Ingreso alto	18
Ingreso mediano alto	13
Países de ingreso mediano bajo	2
Total general	33

Tabla 2: Cantidad de países según niveles de ingresos (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

5. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

El análisis exploratorio de datos es una fase crítica de la investigación, nos proporciona una comprensión más profunda de la estructura y las tendencias presentes en la muestra de países. En esta sección, utilizaremos una variedad de gráficos y técnicas exploratorias para visualizar y analizar las relaciones preliminares entre las variables clave.

Para facilitar la manipulación de variables se realizó un cambio de nombre para una, los cambios realizados son:

Indicador	Código
PIB per cápita	PIBPERCAP
Tasa de finalización nivel primario	FISPRIM
Tasa de desempleo	DESEMPLEO
Tasa de mortalidad, menores de 5 años	MORT5
Participación de ingreso del 20% peor remunerado	PART20
Tasa de incidencia de pobreza por debajo del 40%	POBRE40
Esperanza de vida al nacer	ESPERANZANACER
Tasa de mortalidad materna	MORTMATERNA
Incidencia de tuberculosis	TUBERCULOSIS
Consumo de energía renovable	ENERGIARENOVABLE
Población total	POBLACIONTOTAL
Emisiones de CO2	EMISIONCO2
Población mayor de 65 años	POBLACION65
Prevalencia de desnutrición	DESNUTRICION

Tabla 3: Códigos de variables (Fuente: Elaboración propia, 2023)

5.1 Histogramas

Los histogramas son herramientas visuales fundamentales que nos permiten analizar la distribución de un conjunto de datos. Examinar la forma de esta distribución es esencial, ya que la simetría indica una distribución equitativa, sugiriendo la presencia de una distribución normal. Por otro lado, la asimetría en la distribución puede revelar información valiosa sobre posibles sesgos en los datos y la existencia de valores atípicos. A continuación, se presentan los histogramas correspondientes a las variables de estudio, ofreciendo una representación gráfica que facilita la identificación de patrones y características en los datos

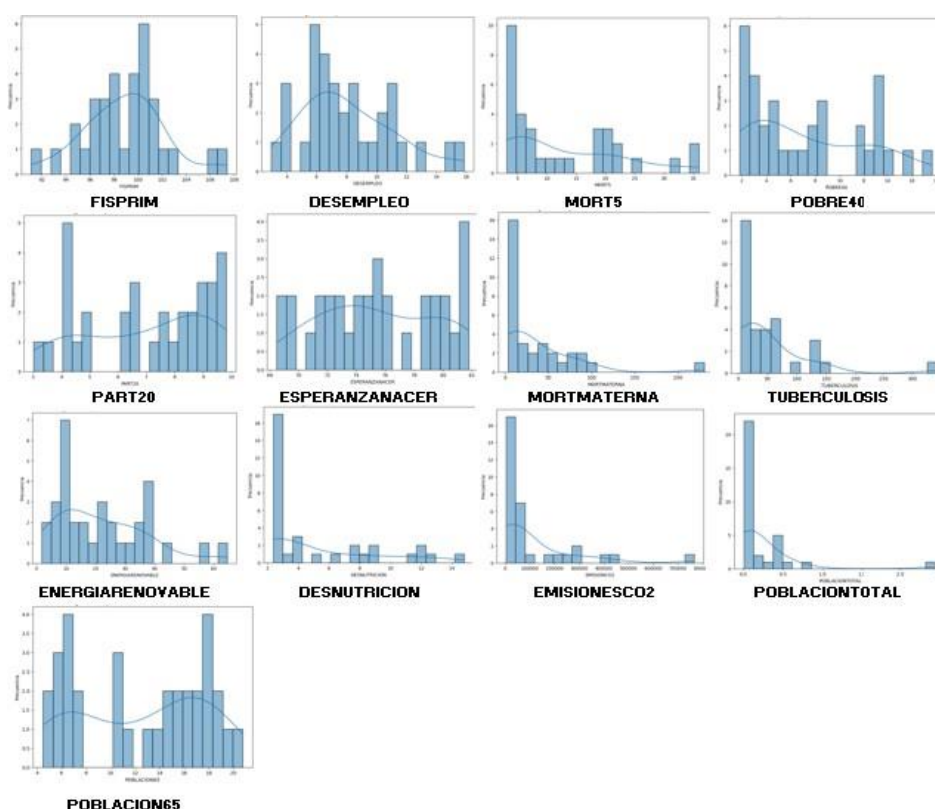


Gráfico 4: Compilado de histogramas de variables en estudio (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020)).

Se puede observar que se presenta una distribución asimétrica hacia la izquierda en múltiples variables, es decir se extiende hacia los valores más bajos, lo que indica que algunos de los valores más pequeños son atípicamente bajos en comparación con mayoría de los datos. Esta distribución nos indica que la media es menor a la mediana, ya que valores bajos disminuirán la media, por otro lado, la moda está en la parte derecha del histograma, ya que es la región con mayor concentración de datos. Es importante destacar que, para la variable “Población total” se aprecia una mezcla de al menos dos grupos de países los cuales pueden concordar con el detalle de la muestra

5.2 Relación entre PIB per cápita y variables de bienestar social

A continuación, examinamos cómo se relaciona la variable PIB per cápita con otras variables de interés. En caso de que se necesite ver el detalle por año de cada variable se puede ver en el anexo 2. Se consideró el promedio por país desde 2000-2020 incluidos.

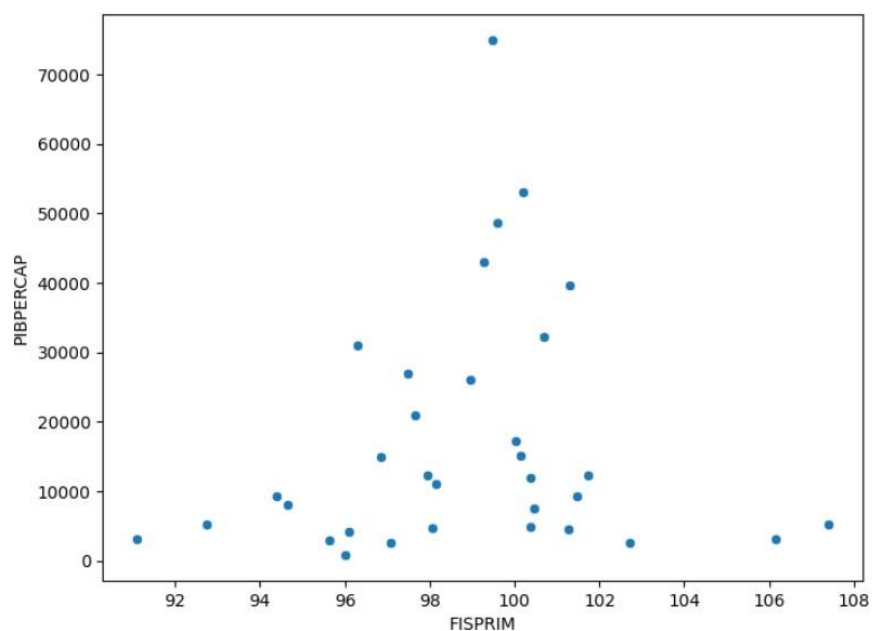


Gráfico 5: PIB per cápita en función de Tasa de finalización nivel primario (Fuente:Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Se observa que la gráfica tiene una tendencia que parece ser exponencial hasta el 100%, lo que quiere decir que a mayor PIB per cápita se tiene una mayor tasa de finalización nivel primario, no obstante, considerando que existen estudiantes que ingresaron de forma adelantada, tardía o bien hayan repetido de grado, lo que explica la forma simétrica del gráfico, también cabe destacar que valores mayores a 110% destacan por tener bajos valores de PIB per cápita, lo que podría indicar una alta tasa de repetición o ingreso de forma tardía.

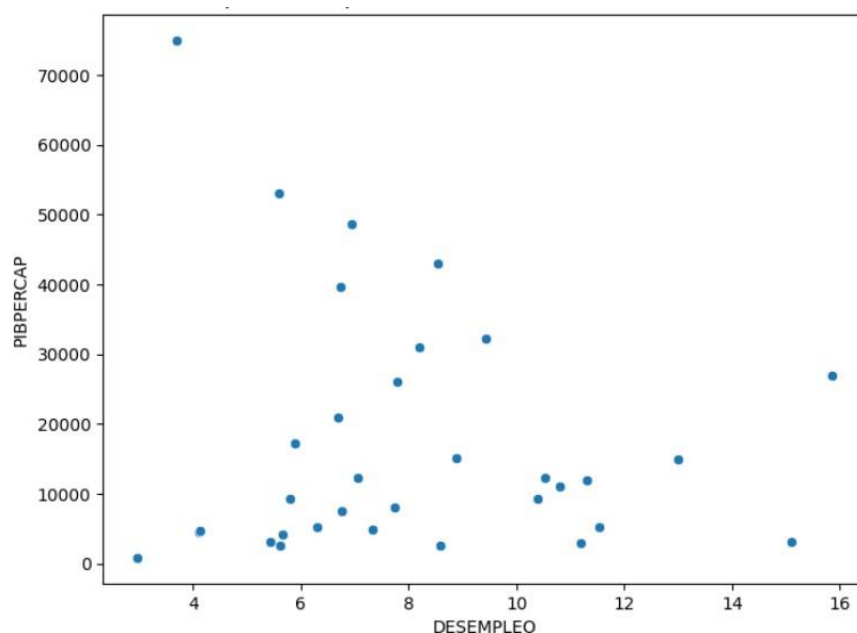


Gráfico 6: PIB per cápita en función de Tasa de desempleo (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Observamos que a mayores niveles de PIB per cápita la tasa de desempleo es menor llegando hasta un 10%, no obstante, para observaciones de PIB per cápita más bajos se pueden dividir en países que muestran valores de Tasa de desempleo menores a 10% y un grupo de países con tasas entre 10% y 16%. Cabe destacar que esta gráfica

nos muestra un promedio de los valores anuales por país, por lo tanto, puede influir ciertos periodos en donde la tasa de desempleo sufrió variaciones para cada país, lo que puede generar que no sea tan clara la tendencia de los datos.

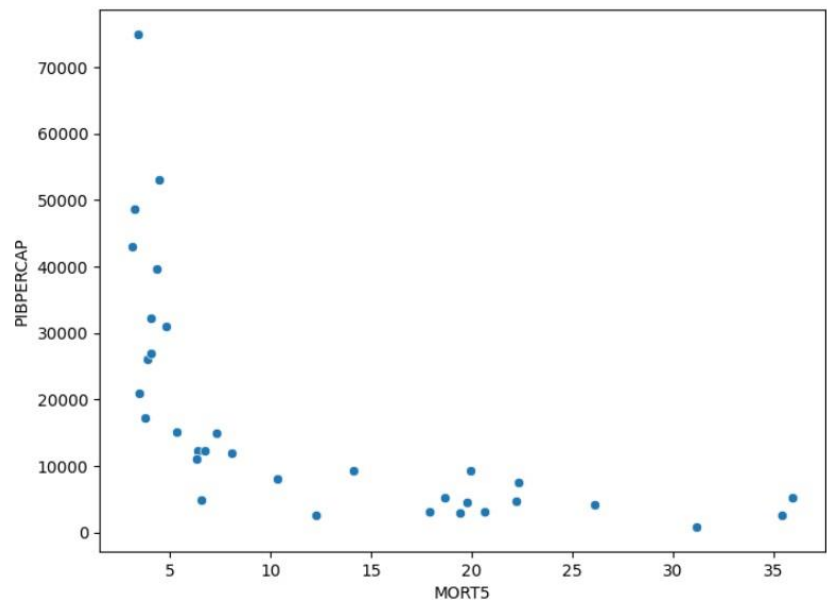


Gráfico 7: PIB per cápita en función de Tasa de Mortalidad en menores de 5 años
(Fuente:Elaboración propia en base a datos obtenidos desde World Bank
(2000-2020))

Observamos que para bajos valores de PIB per cápita la probabilidad por cada 1.000 que un recién nacido muera antes de cumplir cinco años aumenta sobre un 10% llegando a valores de 50% para economías con valor más bajos de PIB per cápita. Teniendo una tendencia exponencial.

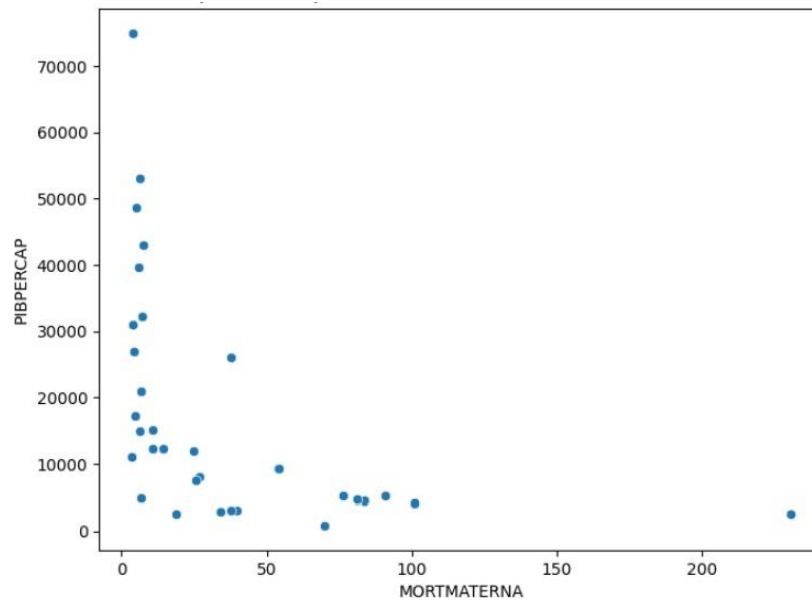


Gráfico 8: PIB per cápita en función de Tasa de mortalidad materna (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

La tasa de mortalidad materna se comporta de manera similar a la tasa de mortalidad en menores de 5 años lo que nos puede indicar que con una mejor economía en términos de PIB per cápita se reduce la probabilidad de fallecimiento tanto de la madre como en menores de 5 años y queda en evidencia que en economías con valores bajos de PIB per cápita ambos indicadores aumentan.

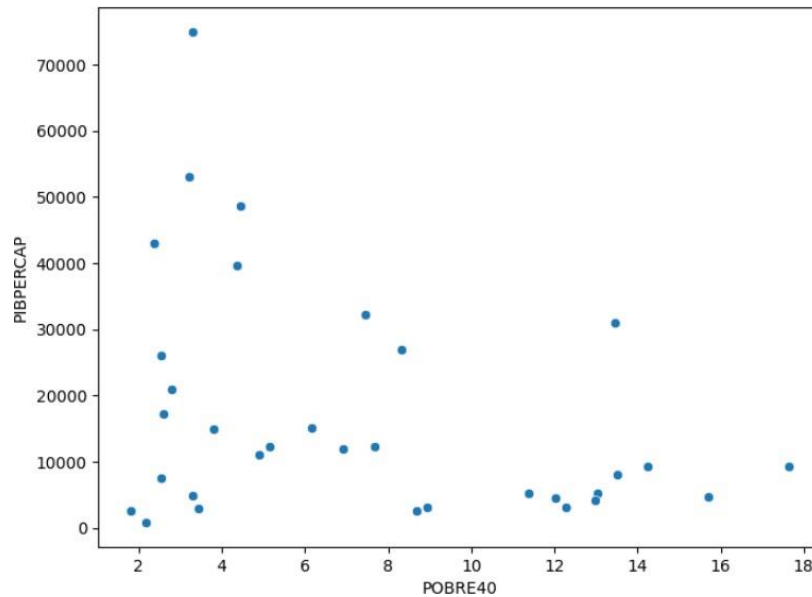


Gráfico 9: PIB per cápita en función de Tasa de incidencia de pobreza por debajo del 40%(Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Para economías con altos niveles de PIB per cápita se muestran bajos niveles de Tasade incidencia de pobreza por debajo del 40%, al contrario, a medida que estos niveles disminuyen se observan un porcentaje mayor al 10% en la tasa de incidencia de pobreza. Existiendo observaciones que si bien tienen valores intermedios de PIB per cápita tienen valores de tasa de incidencia de pobreza menores a 10%.

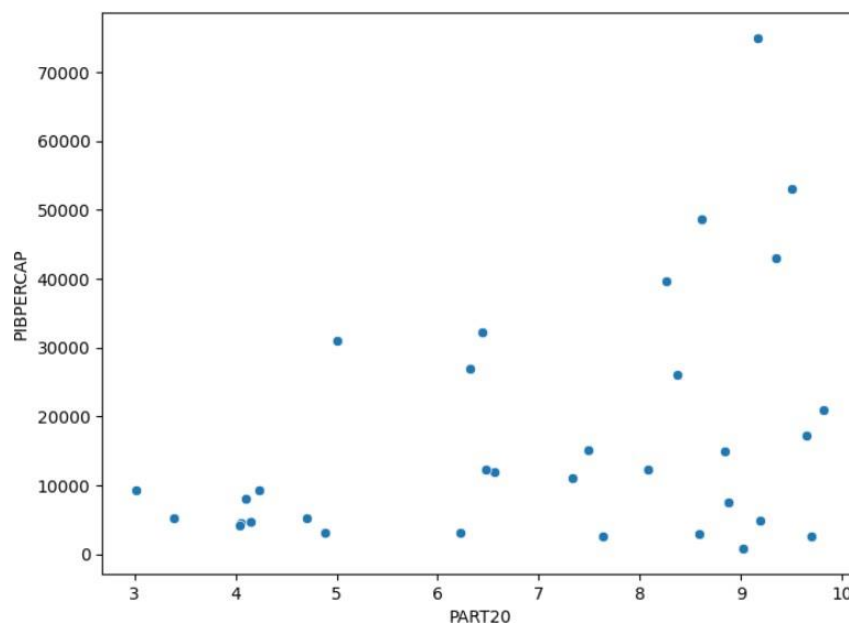


Gráfico 10: PIB per cápita en función de Participación de ingreso del 20% peor remunerado(Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

A medida que las economías perciben mayores niveles de PIB per cápita la participación de ingreso del 20% peor remunerado aumenta, esto puede significar que, a medida que una economía progresa, se destina un porcentaje mayor del ingreso total a la parte más baja de la distribución de ingresos, lo que podría interpretarse como una mayor inclusión o beneficio para la porción de la población con ingresos más bajos.

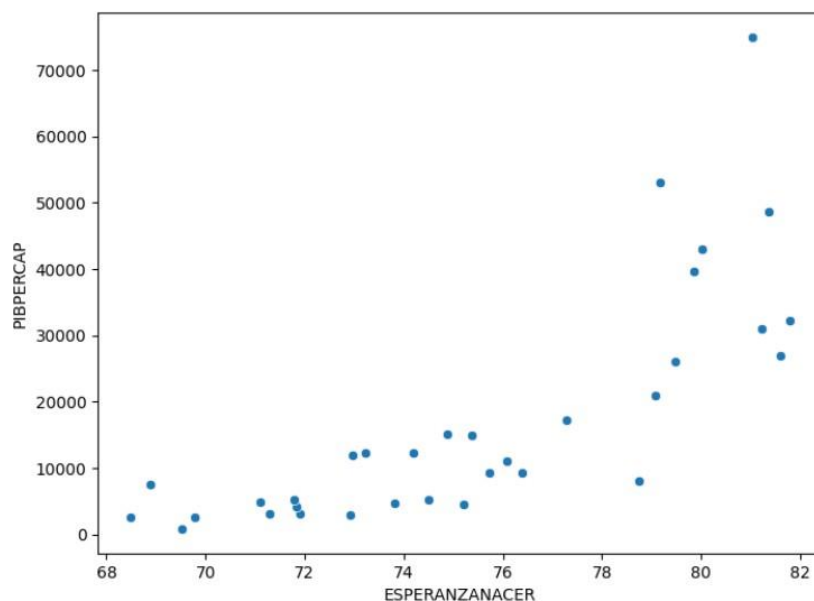


Gráfico 11: PIB per cápita en función de Esperanza de vida al nacer (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Al igual que el gráfico anterior podemos ver una relación positiva entre el PIB per cápita y la esperanza de vida al nacer, esto se puede interpretar como una mayor capacidad financiera la cual se puede invertir en un sistema de salud más avanzado, junto con un mayor acceso a alimentos de mejor calidad, considerando que la nutrición es fundamental para el desarrollo y el mantenimiento de la salud, esto puede influir positivamente en la esperanza de vida.

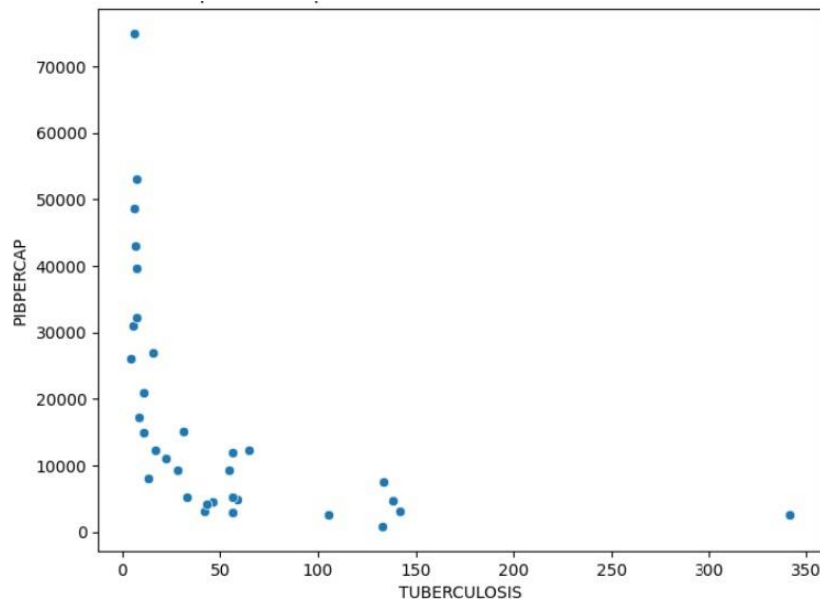


Gráfico 12: PIB per cápita en función de Incidencia de tuberculosis (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Al igual que la Tasa de mortalidad en menores de 5 años y materna, podemos observar que la Incidencia de tuberculosis a medida que la economía percibe menos valores de PIB per cápita aumenta la incidencia de tuberculosis, esto nos puede hablar del sistema de salud presente en cada país y el total disponible para destinarlo a la salud de la población. Los valores más altos de incidencia de tuberculosis corresponden a Indonesia el cual ocupa el tercer lugar a nivel mundial en casos de tuberculosis después de India y China, el mayor número de casos se registra en el grupo de edad productiva, especialmente el de 45 a 54 años (Fundación IO, 2023).

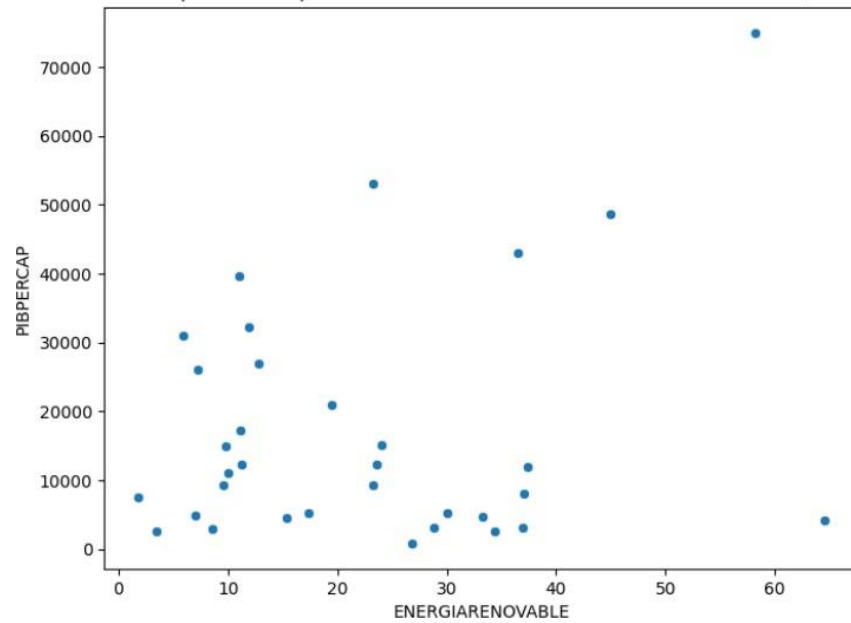


Gráfico 13: PIB per cápita en función de Consumo de energía renovable
(Fuente:Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

No se logra evidenciar una tendencia clara en la gráfica, para países con diferentes niveles de PIB per cápita muestran usar porcentajes de consumo de energía renovable mayores a un 50%, también cabe destacar que países con altos valores de PIB per cápita muestran tener un consumo de energía renovable entre 55-65%.

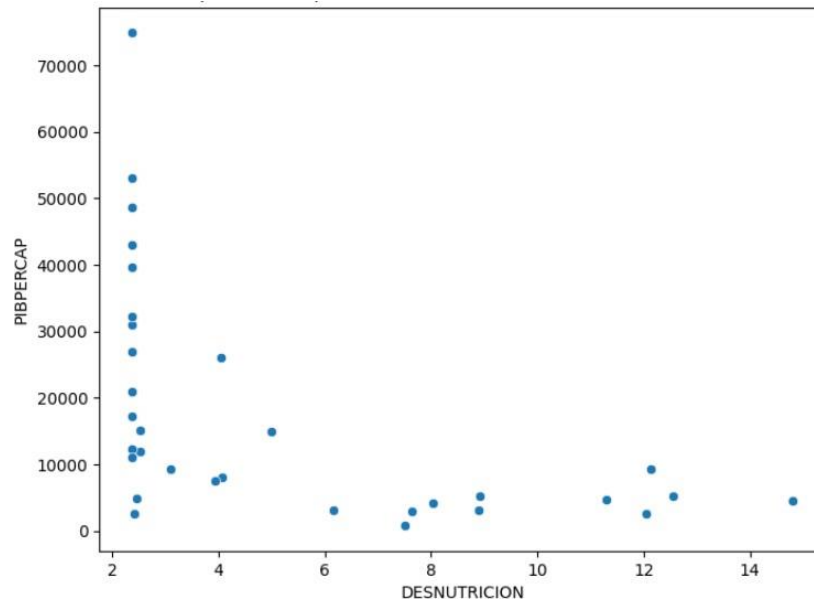


Gráfico 14: PIB per cápita en función de Prevalencia de desnutrición (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Se puede observar la presencia de valores 0 los cuales pueden evidenciar valores atípicos o bien datos faltantes, los cuales serán posteriormente eliminados del modelo. Por otro lado, se ve que para valores más altos de PIB per cápita se mantiene una prevalencia de desnutrición menor a países con valores más bajos, esto se puede relacionar con el acceso a alimentos de mejor calidad y nutrición como se mencionó anteriormente para el indicador Esperanza de vida al nacer.

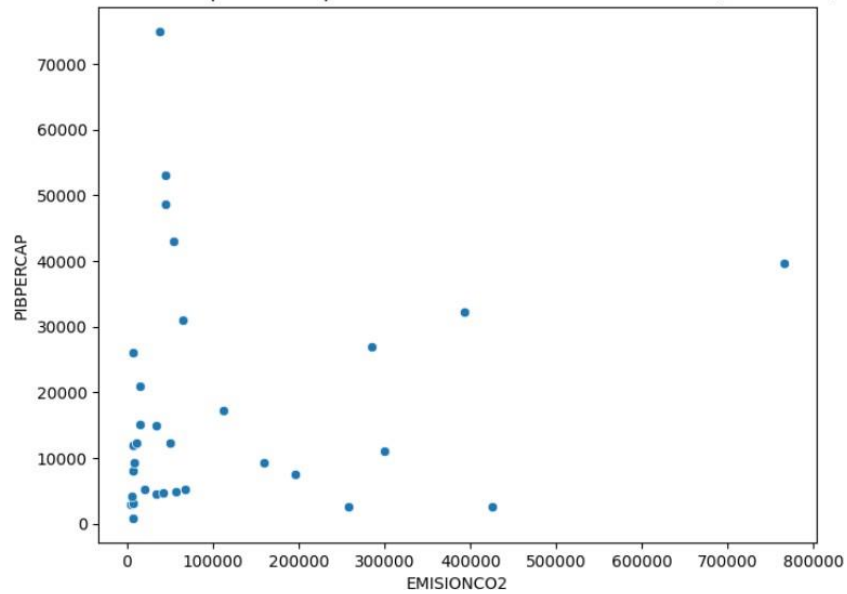


Gráfico 15: PIB per cápita en función de Emisiones de CO2 (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Si bien la tendencia no es del todo exponencial se puede observar que para economías con mayores valores de PIB per cápita se registran menores valores de emisiones de CO2, en el caso contrario se ve un aumento en las emisiones de CO2 desde 200.000 [kt] para economías con menores valores de PIB per cápita, llegando a valores de más de 600.000 [kt] en economías con niveles intermedios.

El tratamiento de datos atípicos de llevar a cabo en el desarrollo del Caso 4, en donde se definirán los países que tienen observaciones atípicas en este caso al realizar el promedio de los 21 años. También se llevará a cabo un procedimiento de eliminación de datos atípicos que considera eliminar datos atípicos bajo reglas cuantitativas, considerando observaciones puntuales.

5.3 Boxplot PIB per cápita

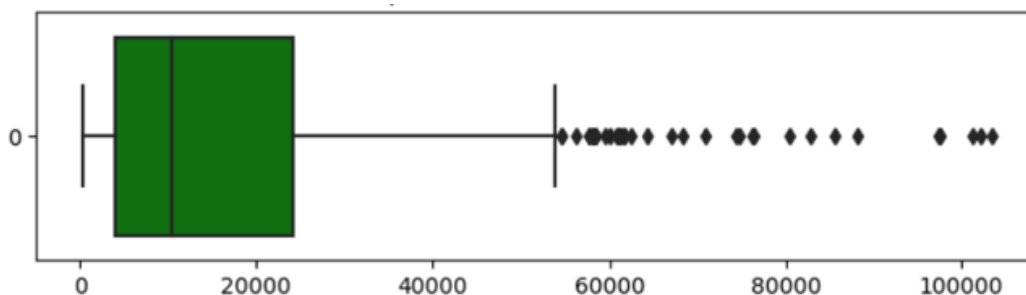


Gráfico 16: Boxplot PIB per cápita (Fuente: Elaboración propia, 2023)

Observando la simetría del boxplot la distribución de la variable de estudio es asimétrica, también cuenta con valores atípicos los cuales pueden corresponder a valores más altos de PIB per cápita. Por otro lado, la longitud de la caja es más larga y los bigotes son extendidos lo que podría sugerir mayor variabilidad en los datos.

5.4 Matriz de correlación: Todas las variables

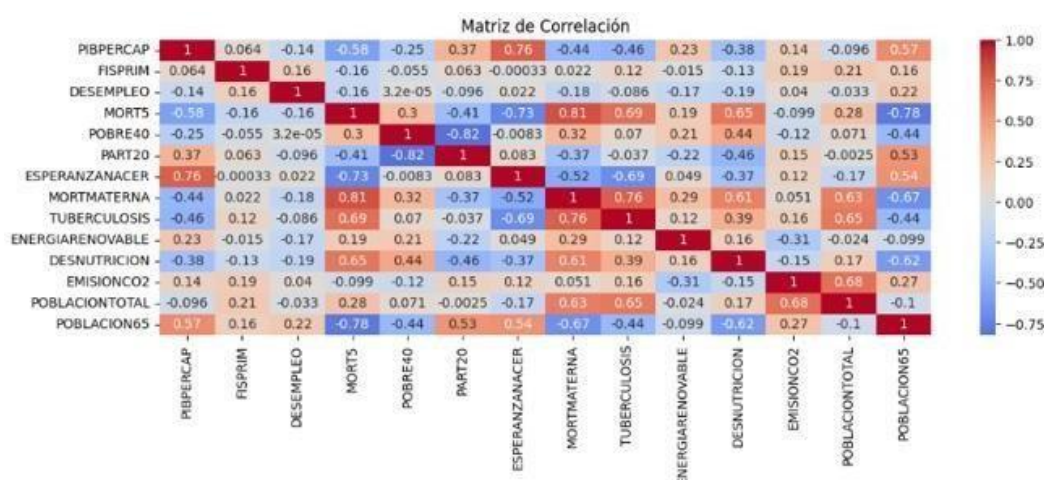


Figura 6: Matriz de correlación para todas las variables (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

En torno a la correlación con la variable de estudio (PIB per cápita) se observa que tiene alta correlación con la “Tasa de mortalidad en menores de 5 años”, “Esperanza de vida al nacer”, “Tasa de mortalidad materna”, “Incidencia de tuberculosis” y “Población mayor a 65 años”. Esto nos da una noción de que la salud es una influencia fundamental en el comportamiento del PIB. Por otro lado, una alta correlación entre las variables exógenas podría indicar multicolinealidad y afectar en el modelo, dentro de ellas se encuentra “Tasa de mortalidad en menores de 5 años”, “Tasa de mortalidad materna”, “Incidencia de tuberculosis”, lo que podría dar indicios de cómo afecta el sistema de salud en la calidad de vida de las personas. También existe alta correlación entre las variables “Tasa de incidencia de pobreza” y “Participación de ingreso del 20% peor remunerado”

5.5 Variance Inflation Factor (VIF)

Debido a que existen variables independientes que parecen estar altamente correlacionadas entre sí, es por esto que decidió hacer una estimación directa del modelo de regresión con el objetivo de encontrar los valores del VIF antes de formular el modelo definitivo. El criterio para decidir que una variable está excesivamente correlacionada a las restantes se fijó en un R^2 del 80%, es decir, que los VIF mayores a 5 representan una colinealidad inadmisible. Los resultados se muestran en la Tabla 3.

Indicador	VIF	COLINEALIDAD
Tasa de finalización nivel primario	1.26	Baja
Tasa de desempleo	1.37	Baja
Tasa de mortalidad, menores de 5 años	10.14	Alta
Participación de ingreso del 20% peor remunerado	4.67	Regular
Tasa de incidencia de pobreza por debajo del 40%	3.58	Baja
Esperanza de vida al nacer	3.76	Baja
Tasa de mortalidad materna	13.39	Alta
Incidencia de tuberculosis	6.57	Alta
Consumo de energía renovable	1.65	Baja
Población total	13.62	Alta
Emisiones de CO2	5.23	Regular
Población mayor de 65 años	4.1	Baja
Prevalencia de desnutrición	2.16	Baja

Tabla 4: Valores VIF de variables independientes (Fuente: Elaboración propia en base adatos obtenidos desde World Bank (2000-2020))

5.6 Decisiones en eliminación de variables: Creación de casos para observar sensibilidad de decisiones.

En base al análisis exploratorio de datos podemos tener una noción de cuáles son las posibles variables candidatas para eliminar, intentando disminuir la multicolinealidad presente en el modelo, con el objetivo de realizar un análisis en la sensibilidad de la toma de decisiones en la elección de variables se crean 4 casos

- I. Caso 1: Se consideran todas las variables, se busca ver el comportamiento de las variables al estar todas incluidas a pesar de la multicolinealidad y tratando de obtener información sobre la significancia de las variables en el modelo.
- II. Caso 2: Se realiza eliminación de variables con valores altos en la matriz de correlación y se utilizará el criterio VIF para medir la multicolinealidad, eliminando variables para mantener los valores de VIF bajo 5.
- III. Caso 3: Se utilizan los mismos criterios anteriores pero esta vez se realizan transformaciones logarítmicas en variables seleccionadas.
- IV. Caso 4: Se eliminan datos atípicos, se consideran atípicos a los datos que están sobre el percentil 95%. Se realizan la misma eliminación de variables que en Caso 2.

A continuación, se verá el detalle de decisiones tomadas en cada caso, las decisiones reales fueron en base a resultados de valores VIF y correlaciones entre variables con el objetivo de disminuir la multicolinealidad del modelo buscando tener certeza del efecto individual de las variables.

5.7 Caso 2: Eliminación de variables en base a altas correlaciones y multicolinealidad

En base a la matriz de correlación inicial se puede observar de las siguientes variables:

- Tasa de mortalidad en menores de 5 años: Tiene alta correlación con “Esperanza de vida al nacer”, “Tasa de mortalidad materna”, “Incidencia de tuberculosis” y “Población mayor a 65 años”.
- Tasa de mortalidad materna: Tiene alta correlación con “Tasa de mortalidad en menores de 5 años” e “Incidencia de tuberculosis”.
- Tasa de incidencia a la pobreza por debajo del 40%: Tiene alta correlación con “Participación del ingreso del 20% peor remunerado”.

Debido a que, si bien estas variables pueden influir en el modelo, buscamos representar áreas del bienestar social, en este caso tenemos variables en el ámbito de la salud y pobreza. Por lo tanto, se decide eliminar las variables mencionadas anteriormente ya que hay variables que explican el área que queremos estudiar. Obteniendo la siguiente matriz de correlación.

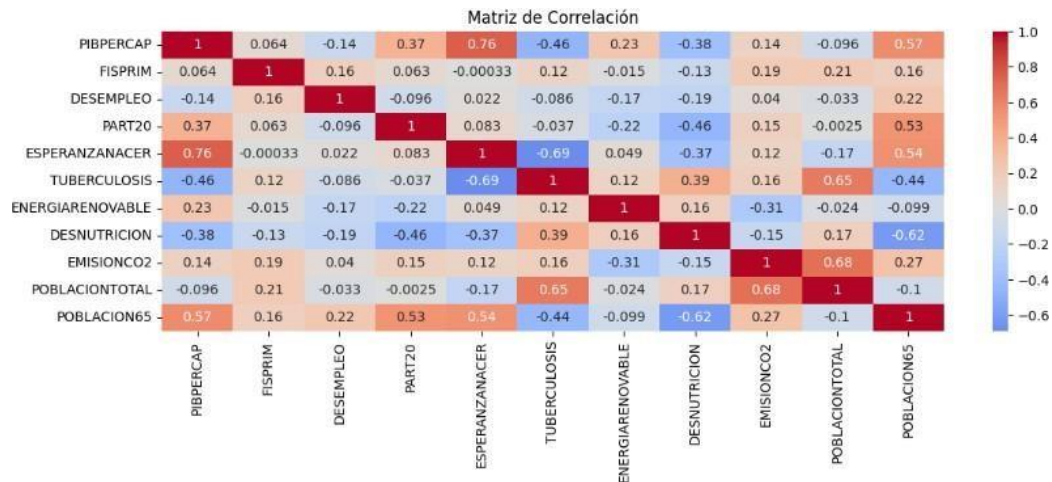


Figura 7: Matriz de correlación con variables seleccionadas en caso 2 (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Para calcular los VIF, y solamente con este objetivo, se estimó el modelo de regresión lineal múltiple de la estructura:

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \dots + \beta_{15} X_{15ij}$$

donde $i = 1, 2, \dots, 33 \text{ ó } 34 = \text{número de países}$
 $j = 1, 2, \dots, 15 = \text{año del registro.}$

De los resultados se desprende que los valores de VIF de cada variable se encuentran altos para “Población total” e “Incidencia de tuberculosis”, por lo tanto, se plantea la eliminación de una de estas variables para disminuir la multicolinealidad. En este caso la variable eliminada será “Población total” ya que nos interesa examinar de mayor forma el efecto que tiene la variable “Incidencia de tuberculosis” en el modelo.

Luego de la eliminación de la variable población total se obtienen los siguientes valores de VIF

Indicador	VIF Inicial	VIF luego de eliminación
Tasa de finalización nivel primario	1.12	1.11
Tasa de desempleo	1.28	1.28
Participación de ingreso del 20% peor remunerado	1.98	1.98
Esperanza de vida al nacer	3.14	2.63
Incidencia de tuberculosis	5.61	2.85
Consumo de energía renovable	1.42	1.41
Población total	5.38	-
Emisiones de CO2	3.26	1.56
Población mayor de 65 años	3.03	3
Prevalencia de desnutrición	1.84	1.84

Tabla 5: Valores VIF de variables independientes caso 2 posterior a eliminación de variables con altos valores VIF (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

5.8 Caso 2: Transformaciones logarítmicas en variables seleccionadas

Del análisis exploratorio de datos podemos observar que existen variables con líneas de tendencias más marcadas y de carácter exponencial, por lo tanto, se considera realizar una transformación logarítmica tratando de capturar la tendencia existente. Para esto consideramos las variables “Tasa de desempleo”, “Tasa de mortalidad en menores de 5 años”, “Participación del ingreso del 20% peor remunerado”, “Esperanza de vida al nacer”, “Tasa de mortalidad materna” e “Incidencia de tuberculosis”. (véase en anexo 3)

Luego de estos cambios realizados se procede a generar la nueva matriz de correlación incluyendo los cambios en las variables mencionadas y su variable original, se puede observar que a las variables que se le aplicó el ajuste logarítmico, la “Tasa de mortalidad materna” e “Incidencia de tuberculosis” tienen un aumento en la correlación con la variable dependiente. No obstante, se eliminan las variables: “Tasa de mortalidad materna”, “Tasa de incidencia de pobreza bajo 40%”, “Tasa de mortalidad en menores de 5 años” e “Incidencia de tuberculosis”. Por alta correlación entre variables independientes.

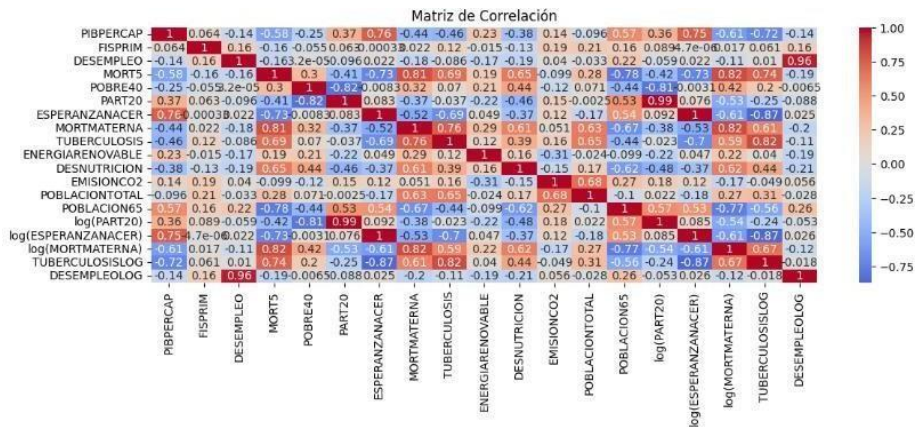


Figura 8: Matriz de correlación con variables con transformaciones logarítmicas (Fuente:Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

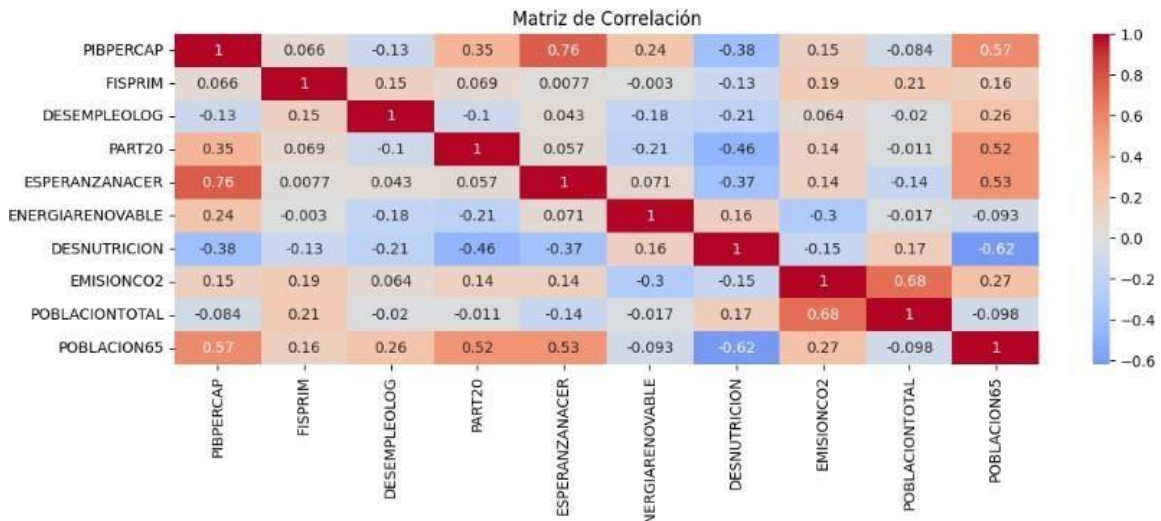


Figura 9: Matriz de correlación con variables seleccionadas caso 3 (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Para verificar la existencia de multicolinealidad, el modelo se vuelve a estimar y calcula nuevamente el VIF para las variables seleccionadas obteniendo los siguientes resultados. En este caso no se ven valores superiores o cercanos a 5 por lo tanto se opta por mantener las variables.

Indicador	VIF
Tasa de finalización nivel primario	1.11
Tasa de desempleo	1.35
Participación de ingreso del 20% peor remunerado	1.95
Esperanza de vida al nacer	1.7
Consumo de energía renovable	1.38
Población total	2.73
Emisiones de CO2	3.02
Población mayor de 65 años	3.12
Prevalencia de desnutrición	1.83

Tabla 6: Valores VIF de variables independientes seleccionadas para caso 3
(Fuente:Elaboración propia en base a datos obtenidos desde World Bank
(2000-2020))

5.9 Caso 4: Tratamiento de datos atípicos

En las gráficas y gráficos Boxplot se pudo observar la presencia de datos atípicos, es por esto se consideró como dato atípico al que estuviera sobre el 95%, de esta forma se logró obtener que en la muestra total existían 304 datos atípicos. Por lo tanto, se generó un código que eliminaba un porcentaje de los datos atípicos. Se realizó una comparación tratando de no reducir en gran magnitud el tamaño de la muestra, por lo tanto, se eliminó el 50% de los datos atípicos quedando 152 datos. De esta forma la muestra pasa de tener 693 observaciones por indicador a un total de 541 observaciones por indicador.

De los gráficos de dispersión podemos observar datos atípicos, en la siguiente tabla se pueden ver los datos atípicos por cada indicador, cabe destacar que los gráficos de dispersión se realizaron con el promedio del indicador por país durante los 21 años de observaciones.

Variables	Max	Promedio	Min
Tasa de finalización nivel primario	107.4	98.8	91.1
	Colombia		El Salvador
Prevalencia de desnutrición	14.8	5.2	2.4
	Ecuador		Alemania
Tasa de desempleo	15.9	8.0	3.0
	España		Kirguistán
Emisiones de CO2	765601.2	107685.5	4871.3
	Alemania		Armenia
Consumo de energía renovable	64.5	22.3	1.8
	Paraguay		Kazajstán
Tasa de mortalidad materna	230.7	36.2	3.6
	Indonesia		Polonia
Tasa de mortalidad en menores de 5 años	35.9	12.6	3.1
	República Dominicana		Finlandia
Tasa de incidencia de pobreza por debajo del 40%	17.6	7.4	1.8
	Panamá		Ucrania
Incidencia de Tuberculosis	341.4	52.0	4.6
	Indonesia		Chipre
Participación del 20% peor remunerado	9.8	7.0	3.0
	Eslovenia		Panamá
Esperanza de vida al nacer	81.8	75.5	68.5
	Italia		Indonesia

Tabla 7: Valores atípicos de promedio de observaciones por país (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Para no comprometer el tamaño de la muestra considerando que son múltiples países que marcan promedios atípicos, se realizará un procedimiento para eliminar observaciones puntuales atípicas. Se establece un percentil umbral de 95. Podemos definir los siguientes pasos como:

Cálculo del umbral para cada columna: se calcula este umbral para cada columna basándose en el percentil 95

- Aplicación de criterios para identificar observaciones atípicas: Si se cumple la condición de que los datos superan el umbral establecido esta será observación atípica.
- Se calcula el total de observaciones atípicas resultantes.
- Definición de porcentaje de observaciones atípicas que se eliminará
- Se itera buscando un punto en el que el modelo tenga mejores métricas de rendimiento en una regresión lineal múltiple.

De esta forma se llega a los siguientes resultados:

	% de eliminación de atípicos			
	25%	50%	75%	100%
R-squared	73.7%	75.1%	74.5%	70.7%
MSE	7.50E+08	6.70E+08	4.10E+07	3.90E+07
AIC	9698	8423	7211	5964

Tabla 8: Métricas de rendimiento luego de eliminación de datos atípicos
(Fuente:Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

	693 observaciones	465 observaciones
	304 observaciones Atípicas	Luego de eliminación 75%
País	N° observaciones Atípicas	N° observaciones atípicas
Alemania	21	5
Paraguay	21	7
Indonesia	21	2
Noruega	20	5
Italia	17	3
Georgia	16	6
Dinamarca	14	4
Panamá	14	3
Ecuador	13	2
Colombia	12	3
España	12	3
República Dominicana	11	3
Kirguistán	10	4
Suecia	10	5
Perú	10	3
Ucrania	10	1

Tabla 9: Valores atípicos por país antes y después de eliminación del 75% de las observaciones atípicas (Fuente:Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Se realiza la matriz de correlación para este caso obteniendo los siguientes resultados:

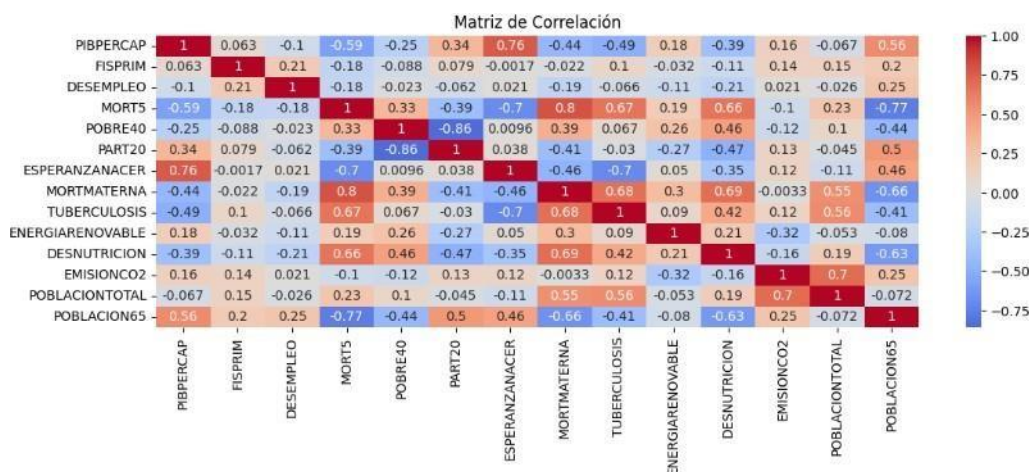


Figura 10: Matriz de correlación caso 4 (Fuente: Elaboración propia en base adatos obtenidos desde World Bank (2000-2020))

En este punto de la investigación el caso 4 tiene resultados bastantes similares entornoa la matriz de correlación y prueba VIF al caso 2, por lo tanto, se realiza la misma eliminación de variables. Teniendo los siguientes resultados:

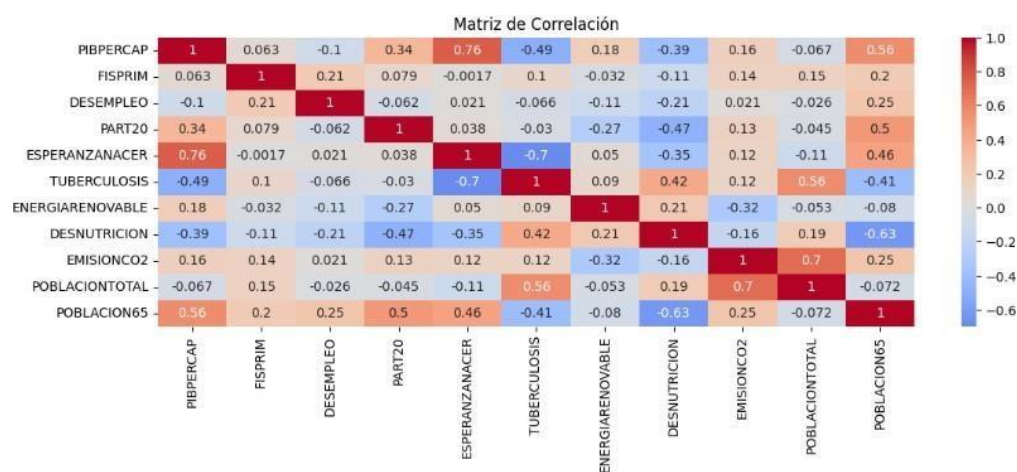


Figura 11: Matriz de correlación con variables seleccionadas caso 4 (Fuente Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Indicador	VIF Inicial	VIF luego de eliminación
Tasa de finalización nivel primario	1.12	1.12
Tasa de desempleo	1.26	1.26
Participación de ingreso del 20% peor remunerado	1.92	1.92
Esperanza de vida al nacer	2.85	2.42
Incidencia de tuberculosis	4.67	2.77
Consumo de energía renovable	1.42	1.42
Población total	4.49	-
Emisiones de CO2	3.2	1.47
Población mayor de 65 años	2.69	2.68
Prevalencia de desnutrición	2.04	1.47

Tabla 10: Valores VIF de variables independientes seleccionadas para caso 4 (Fuente:Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

6. RESULTADOS

Este estudio propone explorar en detalle como diversas variables sociales que abarcan aspectos como empleo, salud, educación, medioambiente y pobreza, influyen en el rendimiento económico medido a través de PIB per cápita. En primera instancia se realizará una comparativa entre metodologías para posteriormente analizar cada una de estas:

6.1 Comparación entre metodologías y casos de sensibilidad en elección de variables

Modelo / Caso	R – squared				MSE			
	C1	C2	C3	C4	C1	C2	C3	C4
Linear SKLearn	75.6%	74.4%	74.6%	74.4%	7.93E+07	8.60E+07	8.46E+07	4.08E+07
Linear Stepwise	75.7%	74.4%	74.6%	74.6%	7.94E+05	8.65E+07	8.44E+07	4.11E+07
Linear Ridge	79.1%	77.2%	77.7%	75.7%	7.95E+07	8.65E+07	9.47E+07	4.07E+07
Datos de panel	42.5%	41.8%	42.2%	41.8%	2.14E+07	2.10E+07	2.16E+07	2.17E+07
Árboles de regresión CCP	94.9%	92.3%	85.0%	89.6%	1.83E+08	2.77E+07	5.40E+08	2.88E+07

Tabla 11: Comparación entre métricas de rendimiento de modelos implementados (Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Las relaciones lineales múltiples tienen métricas de rendimientos suficientes para explicar parte del comportamiento de la variable dependiente, pero no logran capturar tendencias no lineales del estudio, entre ellas destaca Regresión lineal múltiple Ridge la cual mitiga la multicolinealidad del modelo obteniendo mejores métricas de rendimiento y disminuyendo el número de condición.

En el caso de datos de panel tenemos rendimientos deficientes, esto puede ocurrir debido a que no se cumple uno de los supuestos que nos indica si es mejor usar la metodología de datos de panel, este supuesto se verifica con el estadístico de Durbin- Watson el cual nos indica la autocorrelación de las variables. Si bien el código implementado de datos de panel indica que existe autocorrelación positiva, los resultados de las regresiones lineales indican lo contrario.

La metodología de árboles de decisión obtiene mejores métricas de rendimiento, esto se origina debido a que logra capturar relaciones no lineales. Se opta por realizar una comparativa entre decisiones de poda y mejor hiperparámetros obteniendo resultados favorables y reduciendo el sobreajuste del modelo.

Sobre los casos presentados podemos obtener mejores rendimientos de los casos que evitan las altas correlaciones y valores de VIF (Caso 2) y el tratamiento de datos atípicos (Caso 4). Si consideramos la combinación caso-modelo la elección de variables influye de manera significativa para los modelos de regresiones lineales múltiples y árboles de decisión.

Ocurre un fenómeno interesante en el caso 3 al realizar transformaciones logarítmicas, se presencia una tendencia logarítmica para “Incidencia de tuberculosis”, lo que puede sugerir que esta variable podría tener un comportamiento más complejo o no lineal, sin embargo, esta presenta una alta correlación con otras variables siendo eliminada del modelo.

A pesar de ser eliminada debido a la correlación, al analizarla con árboles de decisión, observamos que parece ser influyente en el rendimiento del modelo. Al ser excluida, los árboles de decisión obtienen menores valores en las métricas de rendimiento, indicando que la eliminación de esta variable afectaría negativamente en la capacidad del modelo para ajustarse a los datos. Esto puede explicarse por la naturaleza más flexible de los árboles de decisión en comparación con los modelos lineales que asumen relaciones lineales, los árboles de decisión pueden adaptarse mejor a patrones no lineales o de mayor complejidad. Por lo tanto, la eliminación de una variable influyente para árboles de decisión puede resultar en la pérdida de información.

Esto destaca la importancia de considerar la naturaleza de las variables y como interactúan con diferentes modelos.

En el caso 1, se evidencia un alta multicolinealidad entre las variables, a pesar de obtener buenas métricas de rendimiento. Este fenómeno sugiere un posible sobreajuste, ya que el modelo puede estar capturando la complejidad de la multicolinealidad en lugar de las relaciones individuales de cada variable.

6.2 Variables significativas

Los modelos que permiten conocer esta influencia o peso son los de regresión lineal y *Random forest tree*. Obteniendo la siguiente tabla:

	Linear SKLearn				Linear Stepwise				Linear Ridge				Random forest tree				Frecuencia Relativa
Variables	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	
Tasa de desempleo	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100.0%
Esperanza de vida al nacer	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100.0%
Consumo de energíarenovable	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100.0%
Participación de ingreso del 20% peor remunerado	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		87.5%
Emisiones de CO2		✓	✓	✓		✓	✓	✓		✓	✓	✓		✓		✓	68.8%
Población mayor de 65 años	✓			✓	✓	✓		✓	✓			✓		✓	✓	✓	62.5%
Tasa de finalización nivel Primario	✓		✓		✓				✓				✓	✓	✓	✓	50.0%
Tasa de mortalidad, menores de 5 años	✓				✓				✓				✓				25.0%
Población total	✓		✓				✓				✓						25.0%
Incidencia de tuberculosis													✓	✓		✓	18.8%
Tasa de mortalidad materna	✓				✓				✓								18.8%
Total	9	5	7	6	8	6	6	6	8	5	6	6	6	8	6	7	

Tabla 12: Resumen de variables con significancia e importancia en los modelos realizados (Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Las variables marcadas con un “ticket” en cada combinación caso-modelo corresponden a variables con $p < 0.5$ indica que hay evidencia estadística suficiente para rechazar la hipótesis nula de que el coeficiente de la variable es igual a cero. En otras palabras, sugiere que la variable tiene un efecto significativo en la variable dependiente, lo cual es fundamental para entender la relación causal o explicativa entre las variables.

Si consideramos *Random forest tree* corresponden a variables que tienen importancia y aportan en la predicción del modelo, esta se calcula evaluando cómo cada variable contribuye a la reducción de la impureza en los nodos del árbol de decisión, esta impureza se mide con criterio como la ganancia de información o la reducción de la varianza. Si la impureza disminuye debido a una variable considerada en el nodo, mayor será la importancia, las variables que más contribuyen a la capacidad de predecir valores o discriminar entre clases son consideradas más importantes.

Las variables más relevantes deberían aparecer en los nodos superiores de los árboles de decisión afectando más el resultado al tener una mayor contribución al modelo. En el proceso de construcción del árbol y en todos los árboles del bosque se promedian las contribuciones para obtener la importancia final de cada variable.

La frecuencia relativa revela que **"Tasa de desempleo"**, **"Esperanza de vida al nacer"** y **"Consumo de energía renovable"** son críticas en todos los modelos con un 100% de frecuencia relativa. Le siguen de cerca **"Participación de ingreso del 20% peor remunerado"**, **"Emisiones de CO2"** y **"Población mayor a 65 años"** con frecuencias superiores al 60%. Para un análisis integral, se compararán la frecuencia relativa, los coeficientes de regresión y la importancia de estas variables con una frecuencia relativa superior al 50%, proporcionando una visión completa de su relevancia en todos los modelos.

Variables	Frecuencia relativa	Coeficientes de regresión			Importancia
		Linear SKLearn	Linear Stepwise	Linear Ridge	Random forest tree
Esperanza de vida al nacer	100%	2,812.1	2,854.9	11,907.0	22.7%
Consumo de energía renovable	100%	302.0	297.2	4,699.8	15.2%
Tasa de desempleo	100%	-1,307.9	-1,334.3	-824.8	4.1%
Tasa de finalización nivel primario	50%	187.9	178.7	925.0	2.8%
Población mayor de 65 años	63%	317.0	315.5	1,664.7	2,2%
Participación de ingreso del 20% peor remunerado	88%	2,500.1	2,474.7	5,416.4	1,39%
Emisiones de CO2	69%	9.E-03	9.E-03	1,528.1	2,03%
Incidencia de tuberculosis	19%	-9.E+00	0.E+00	-667.3	48,1%

Tabla 13: Coeficientes e importancia de variables con mayor frecuencia relativa en la significancia e importancia del modelo (Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Nuestras variables de interés se centran inicialmente en aquellas con una frecuencia del 100%, siendo "Esperanza de vida al nacer" la que destaca como la más significativa en los modelos. La magnitud de su importancia coincide con la frecuencia relativa. No obstante, algunas variables como "Consumo de energía renovable", "Tasa de desempleo" y "Participación de ingreso del 20% peor remunerado" muestran discrepancias entre los coeficientes de regresión y su nivel de importancia.

Es esencial considerar que los modelos de regresión lineal, al asumir relaciones lineales, pueden no capturar patrones no lineales o interacciones complejas entre los datos, limitando su eficiencia. En contraste, *Random Forest tree* combina múltiples árboles de decisión, lo que le permite capturar relaciones más complejas y no lineales de manera más efectiva. Destaca que la importancia de "Incidencia de tuberculosis" supera a la de "Esperanza de vida al nacer", pero su frecuencia relativa es menor, y muestra poca significancia en los modelos de regresión lineal. Esto se debe a que "Incidencia de tuberculosis" sigue una relación logarítmica en lugar de lineal. La capacidad de Random Forest para capturar esta complejidad revela el verdadero impacto de la variable.

6.3 Regresión Lineal Múltiple

Métricas de rendimiento

	R –squared				MSE			
Modelo / Caso	C1	C2	C3	C4	C1	C2	C3	C4
Linear SKLearn	75.6%	74.4%	74.6%	74.5%	7.93E+07	8.60E+07	8.46E+07	4.08E+07
Linear Stepwise	75.7%	74.4%	74.6%	74.6%	7.94E+05	8.65E+07	8.44E+07	4.11E+07
Linear Ridge	79.1%	77.2%	77.7%	75.7%	7.95E+07	8.65E+07	9.47E+07	4.07E+07
	RMSE				n° Condición			
Modelo / Caso	C1	C2	C3	C4	C1	C2	C3	C4
Linear SKLearn	8.90E+03	9.31E+03	9.20E+03	6.39E+03	2.35E+09	6.97E+06	1.67E+09	5.69E+06
Linear Stepwise	8.91E+03	9.30E+03	9.19E+03	6.14E+03	2.27E+09	6.23E+06	1.62E+09	4.81E+06
Linear Ridge	8.91E+03	9.31E+03	9.20E+03	6.38E+03	12.80	4.08	4.12	3.64

Tabla 14: métricas de rendimiento en regresiones lineales múltiples (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Se evidencia que la regresión lineal Ridge exhibe un rendimiento superior en términos de R^2 , indicando una mayor capacidad para explicar la variabilidad de los datos. Esta tendencia se refleja de manera consistente en el MSE, donde la regresión Ridge muestra una menor dispersión de errores en las predicciones en la mayoría de los casos. Además, tanto el RMSE como el número de condición mantienen valores más bajos para la Regresión de Ridge, sugiriendo así una mayor estabilidad numérica en los cálculos.

En los casos de Linear SkLearn y Linear Stepwise, donde los números de condición son del orden de 10^6 y 10^9 , indican multicolinealidad entre las variables predictoras que están altamente correlacionadas. La introducción de penalización en los coeficientes por parte de Ridge mitiga eficazmente la multicolinealidad y mejora la estabilidad de los resultados. Esta estrategia también contribuye a prevenir el sobreajuste al agregar penalización a los coeficientes, lo cual es especialmente útil al trabajar con un conjunto de datos pequeño y múltiples variables.

Al enfocarnos exclusivamente en la regresión lineal múltiple Ridge, se presentan los resultados específicos por caso

	R –squared				MSE			
Modelo / Caso	C1	C2	C3	C4	C1	C2	C3	C4
Linear Ridge	79.1%	77.2%	77.7%	75.7%	7.95E+07	8.65E+07	9.47E+07	4.07E+07
	RMSE				n° Condición			
Modelo / Caso	C1	C2	C3	C4	C1	C2	C3	C4
Linear Ridge	8.91E+03	9.31E+03	9.20E+03	6.38E+03	12.80	4.08	4.12	3.64

Tabla 15: Métricas de rendimiento en regresiones lineales múltiples Ridge
(Fuente:Elaboración propia en base a datos obtenidos desde World Bank
(2000-2020))

Al analizar los resultados, el caso 1, que considera todas las variables, destaca por tener el mayor coeficiente de determinación. Sin embargo, las otras métricas revelan una mayor dispersión y menor estabilidad numérica en los cálculos. Específicamente, el número de condición es tres veces más alto en comparación con los casos en los que se eliminan variables para reducir la multicolinealidad.

Los casos más prometedores son el 2 y el 4. En el caso 2, se tomaron decisiones estratégicas para eliminar variables con alta correlación y con $VIF > 5$. Aunque su rendimiento es similar al caso 3, donde se aplicaron transformaciones logarítmicas antes de los mismos criterios, se destacan diferencias en MSE y el número de condición, con valores más bajos que en el caso 1 y 3. Esto sugiere una efectiva reducción de la multicolinealidad, aunque no tan destacada como en el caso 4.

En cuanto al caso 4, las métricas de rendimiento, incluyendo MSE, RMSE y el número de condición, muestran valores más bajos en comparación con otras variables. Esto indica que la decisión de eliminar datos atípicos y aplicar criterios de alta correlación y VIF contribuye a disminuir la multicolinealidad.

Autocorrelación y Normalidad de los datos

Modelo / Caso	Durbin – Watson				Jarque Bera			
	C1	C2	C3	C4	C1	C2	C3	C4
Linear SKLearn	1.90E+00	1.89E+00	1.89E+00	1.98E+00	6.70E+02	7.15E+02	6.18E+02	1.80E+02
Linear Stepwise	1.88E+00	1.89E+00	1.89E+03	1.98E+00	6.77E+02	7.09E+02	6.08E+02	1.82E+02

Tabla 16: Pruebas estadísticas de autocorrelación y desviaciones de la normalidad de los datos (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000- 2020))

El estadístico Durbin-Watson, con valores entre 0 y 4, cerca de 2 indica ausencia de correlación. En todos los modelos, se sitúa en un rango cercano a esta ausencia, sugiriendo independencia lineal entre las variables. Sin embargo, los gráficos de dispersión revelan relaciones no lineales entre las variables independientes, aspecto que puede escapar a la captura del modelo y generar una falta de relación general.

Por otro lado, la prueba de Jarque Bera arroja valores altos, señalando que los residuos no siguen una distribución normal. Esta falta de normalidad puede sesgar la estimación de los parámetros del modelo.

AIC y BIC

Modelo / Caso	AIC				BIC			
	C1	C2	C3	C4	C1	C2	C3	C4
Linear SKLearn	1.09E+04	1.10E+04	1.10E+04	8.42E+03	1.10E+04	1.10E+04	1.10E+04	8.46E+03
Linear Stepwise	1.09E+04	1.10E+04	1.10E+04	8.42E+03	1.10E+04	1.10E+04	1.10E+04	8.45E+03

Tabla 17: Medidas estadísticas para evaluación de bondad de ajuste (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Estas medidas son utilizadas para la comparación de modelos. Se busca el modelo que tenga menores valores de AIC y BIC. Frente a esto se puede decir que el caso 4 posee mejor calidad relativa del modelo, esto puede ocurrir ya que el caso 4 no considera valores atípicos.

P-values

Se presentan los siguientes resultados de variables significativas, esto se hizo en base a la gráfica de p-values de los estimadores, por efecto de la cantidad de casos y modelos se adjuntarán las gráficas del caso con mejor rendimiento. El resto de los gráficos pueden ser revisados en anexo 3. En este caso se consideró como variable significativa a aquella que tuviera un valor p menor a 0.05

Variables	Linear SKLearn				Linear Stepwise				Linear Ridge				Árboles de decisión				Frecuencia Relativa
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4	
Tasa de desempleo	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100.0%
Esperanza de vida al nacer	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100.0%
Consumo de energía renovable	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100.0%
Participación de ingreso del 20% peor remunerado	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		87.5%
Emisiones de CO2		✓	✓	✓		✓	✓	✓		✓	✓	✓		✓		✓	68.8%
Población mayor de 65 años	✓			✓	✓	✓		✓	✓			✓		✓	✓	✓	62.5%
Tasa de finalización nivel primario	✓		✓		✓				✓				✓	✓	✓	✓	50.0%
Tasa de mortalidad, menores de 5 años	✓				✓				✓				✓				25.0%
Población total	✓		✓				✓				✓						25.0%
Incidencia de tuberculosis													✓	✓		✓	18.8%
Tasa de mortalidad materna	✓				✓				✓								18.8%
Tasa de incidencia de pobreza por debajo del 40%																	0.0%
Prevalencia de desnutrición																	0.0%
Total	9	5	7	6	8	6	6	6	8	5	6	6	6	8	6	7	

Tabla 18: Resumen de significancia de variables exógenas (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Observando la significancia de las variables en los gráficos p-values se puede destacar la frecuencia de las variables “Tasa de desempleo”, “Esperanza de vida al nacer”, “Consumo de energía renovable” y “Participación de ingreso del 20% peor remunerado”, siendo variables significativas en todos los modelos en los diferentes casos. También “emisiones de CO2” que no es significativa solo en el caso 1, esto puede suceder por el alta multicolinealidad de este caso ya que esto implica que para el modelo es difícil reconocer la influencia individual de las variables predictoras y considerando que se tienen 13 variables predictoras esto influye de mayor forma en el modelo.

Como identificamos como mejor modelo Regresión Lineal Ridge a continuación se realizará el análisis de las variables significativas y sus coeficientes de regresión:

Variables	Linear Ridge							
	Significancia				Coeficientes			
	C1	C2	C3	C4	C1	C2	C3	C4
Tasa de desempleo	✓	✓	✓	✓	-1,903	-1,775	1,885	-1,507
Esperanza de vida al nacer	✓	✓	✓	✓	13,379	11,644	11,957	10,648
Consumo de energía renovable	✓	✓	✓	✓	4,938	4,995	4,978	3,889
Participación de ingreso del 20% peor remunerado	✓	✓	✓	✓	5,652	5,746	5,606	4,661
Emisiones de CO2		✓	✓	✓	250	1,866	2,618	1,378
Población mayor de 65 años	✓			✓	1,990	1,274	1,339	2,055
Tasa de finalización nivel primario	✓				1,559	772	815	554
Población total			✓		2,940	-	-1,375	-
Tasa de mortalidad, menores de 5 años	✓				6,106	-	-	-
Tasa de mortalidad materna	✓				-5,256	-	-	-
Incidencia de tuberculosis					-1,313	-965	-	276
Tasa de incidencia de pobreza por debajo del 40%					-371	-	-	-
Prevalencia de desnutrición					229	337	372	-114
Total	8	5	6	6				

Tabla 19: Resumen de significancia de variables exógenas (Fuente: Creación propia en base a datos obtenidos desde World Bank (2000-2020))

Se observan exactamente en el caso 2 las variables con frecuencia en todos los casos y modelos, lo que nos indica una congruencia entre el caso 2 y la predicción total. La magnitud de los coeficientes de las variables seleccionadas nos da noción de como aportan las variables al modelo siendo la “Esperanza de vida al nacer” la que tiene un mayor coeficiente de regresión para todos los casos, siguiendo de “Participación de ingreso del 20% peor remunerado” y “Consumo de energía renovable” se puede observar que todas estas variables tienen una dirección positiva con la variable dependiente, al contrario la “Tasa de desempleo” tiene una dirección negativa, por lo tanto a medida que disminuye el PIB aumentaría la tasa de desempleo.

Si bien “Emisiones de CO2” es una variable significativa su coeficiente de regresión es bajo en comparación con las otras variables, lo que nos podría indicar que puede tener un efecto marginal pequeño con PIB per cápita, el hecho de que su magnitud sea modesta indica que no es tan sustancial para explicar la variabilidad de la variable dependiente.

Por último, se muestra la gráfica de p-values del caso 2 podemos obtener información sobre que variables tienen menos significancia en el modelo como es la 2, “Prevalencia de desnutrición e “Incidencia de tuberculosis” y en una menor medida “Tasa de finalización nivel primario” y “Población mayor a 65 años”. Para mayor detalle de los casos restantes véase en anexo 4 y 5.

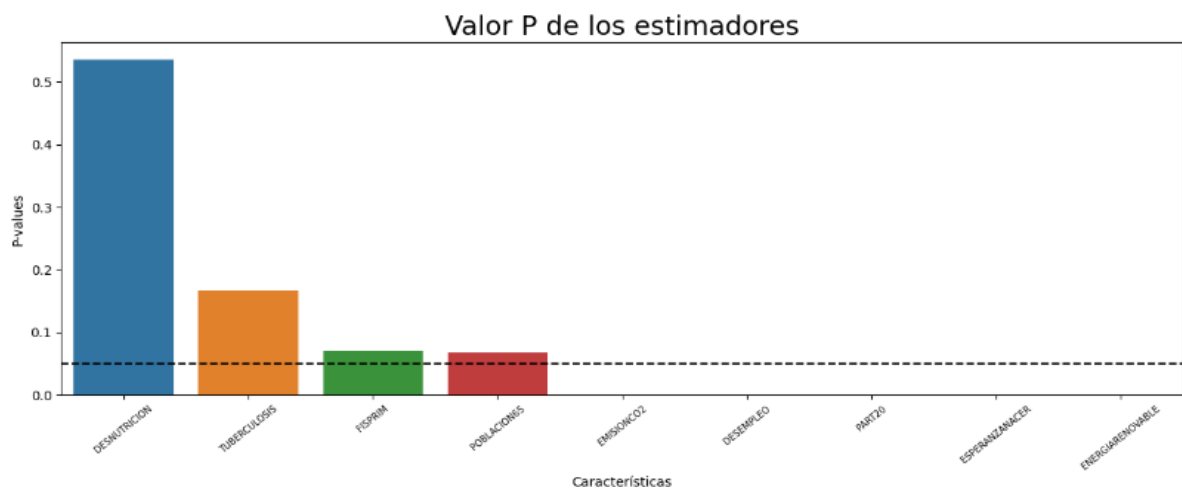


Gráfico 17: Gráfica de p-values para C2 Regresión Ridge (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

6.4 Árboles de decisión

En esta sección, exploraremos diversos indicadores clave que proporcionan una visión completa del rendimiento del modelo utilizando árboles de decisión en su variable de regresión, también se contempla la utilización de 2 estrategias utilizadas para optimizar su desempeño. La primera técnica de *Cost complexity pruning* (Poda), que busca encontrar el equilibrio adecuado entre la complejidad del árbol y su capacidad de generalizar nuevos datos. En cada división se evalúa el beneficio en la reducción de la impureza versus el costo asociado con la complejidad adicional, se realiza la poda, eliminando las ramas menos relevantes.

La segunda estrategia implica la elección de “Mejores hiperparámetros” a través de iteración, especificando configuraciones para mejorar la capacidad predictiva del modelo.

A continuación, se presentan las métricas de rendimiento R^2 y MSE para los escenarios presentes:

Metodología	R-squared				MSE			
	C1	C2	C3	C4	C1	C2	C3	C4
Sin hiperparámetros	94.8%	92.7%	81.4%	90.2%	1.80E+07	2.60E+07	6.70E+07	2.70E+07
Cost complexity pruning	88.1%	92.9%	85.8%	91.4%	1.85E+07	2.56E+07	5.10E+07	2.39E+07
Mejor hiperparámetros	95.1%	92.3%	85.1%	89.6%	1.70E+07	2.76E+07	5.30E+07	2.89E+07

Tabla 20: Métricas de rendimiento para diferentes metodologías de Árbol de regresión
(Fuente:Elaboración propia en base a datos obtenidos desde World Bank(2000-2020))

Los resultados son bastante similares entre ellos, obteniendo mejores resultados para la metodología de definir “mejores hiperparámetros” en términos de R^2 , no obstante, se ven rendimientos bastante similares para la metodología “sin parámetros” y “*Cost complexity pruning*” exceptuando el caso 1, este último tiene MSE más neutrales que la metodología “mejores hiperparámetros”.

Cabe destacar que, si bien “Sin hiperparámetros” tiene buenos rendimientos y menores valores de MSE, hay que evitar el sobreajuste u *Overffiting* al que tiende el modelo de árbol de decisión.

La técnica “*Cost complexity pruning*” se refiere al costo total, que consiste en la suma del error de ajuste y un término de penalización proporcional al tamaño del árbol, si el árbol más grande, el costo es mayor. Por lo que se realiza el proceso de poda con un árbol completo (Sin hiperparámetros), se busca minimizar el costo total del árbol eliminando subárboles menos relevantes y simplificando la estructura general. Por otro lado, la técnica de “mejores hiperparámetros” se hace en base a hiperparámetros claves como la profundidad máxima del árbol, número mínimo de muestras requeridas para dividir un nodo y el número mínimo de hojas, en este caso se utiliza la búsqueda de una cuadrícula en donde se especifica el conjunto de posibles hiperparámetros y evalúan las combinaciones posibles que optimizan el rendimiento, la métrica de evaluación es el MSE.

Dicho esto, se analizará la gráfica de árboles de regresión en las variantes implementadas:

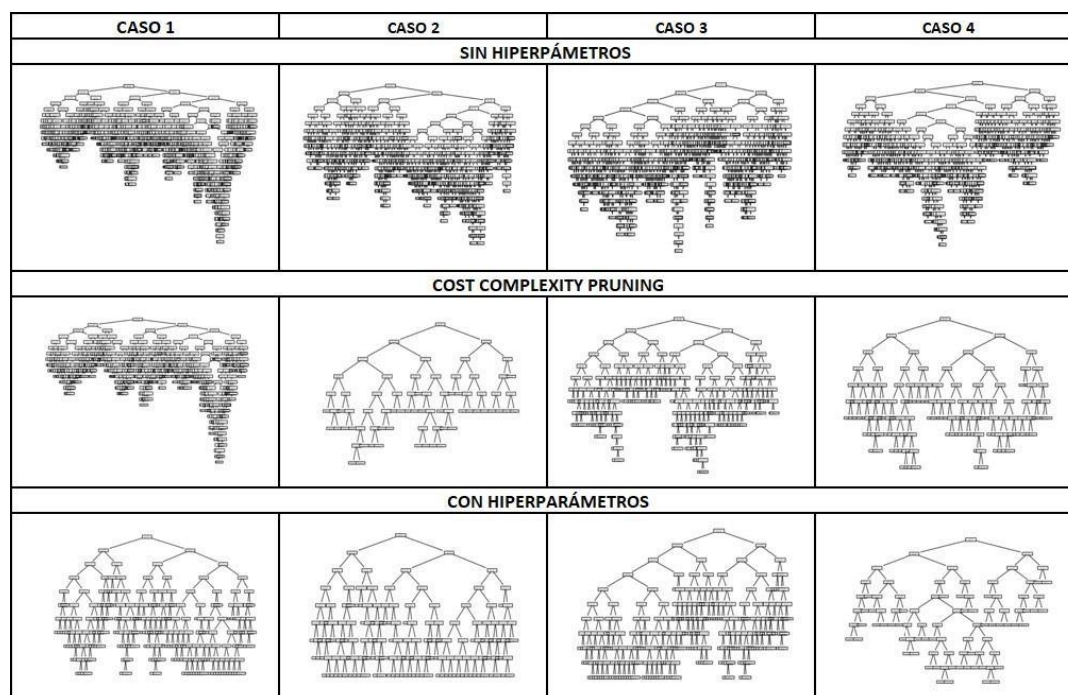


Figura 12: Compilado de resultados Árboles de regresión – Comparación todos los casos(Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Un árbol más grande puede adaptarse mejor a los datos de entrenamiento, tratando de explicar la complejidad de las variables, pero también se puede volver demasiado específico, no generalizando bien los datos nuevos, por lo tanto, el modelo se sobreajusta. Debido a las técnicas implementadas poseen valores similares en las métricas de rendimiento, usaremos como guía la representación visual para identificar un posible sobreajuste. Si notamos la primera línea de compilado de diagramas de árbol de regresión corresponden a árboles de mayor tamaño en comparación con las demás variantes lo que nos podría indicar un sobreajuste al no podar ni mejorar los hiperparámetros del modelo.

También ocurre esto en el caso 1 con “*Cost complexity pruning*”, el resto de los árboles se ven más equilibrados, podemos notar que para el caso 3 los árboles tienen cierto grado de complejidad mayor al resto de los árboles y por último el árbol del caso 4 con técnica de mejorar los hiperparámetros se muestra con la estructura más simple.

Ahora que tenemos una noción sobre que árboles pueden estar sobreajustado, por lo que se hace una prueba adicional, en donde se compararan las métricas de rendimiento de los datos de entrenamiento con los datos de prueba. Lo que nos entrega la siguiente tabla:

Técnica	Data	R-squared				MSE			
		C1	C2	C3	C4	C1	C2	C3	C4
Sin hiperparámetros	Test	94.9%	91.5%	82.7%	90.2%	1.80E+07	3.07E+07	6.20E+07	2.73E+07
	Train	100.0%	100.0%	100.0%	100.0%	0	0	0	0
Cost complexity Pruning	Test	94.9%	92.9%	85,82%	91.4%	1.79E+07	2.56E+07	5.11E+07	2.30E+07
	Train	99.9%	97.7%	99.8%	99.3%	3.66E+03	7.70E+06	5.32E+05	1.70E+06
Mejores hiperparámetros	Test	95.0%	92.3%	84.7%	89.6%	1.85E+07	2.70E+07	5.53E+07	2.80E+07
	Train	98.8%	98.0%	98.7%	94.1%	3.88E+06	6.79E+06	4.49E+06	1.43E+07

Tabla 20: Métricas de rendimiento para diferentes metodologías de Árbol de regresión para conjuntos de prueba y entrenamiento (Fuente: Creación propia en base a datos obtenidos desde World Bank (2000-2020))

Para comprender estos resultados, es esencial considerar la discrepancia entre los valores de R^2 y MSE en los conjuntos de entrenamiento y prueba. Un R^2 significativamente más alto en el conjunto de entrenamiento que en el de prueba

sugiere que el modelo se ajustaba a los datos utilizados para su entrenamiento, explicando una gran proporción de la variabilidad presente en esos datos

Esta situación es indicativa de una buena capacidad del modelo para adaptarse a patrones específicos en los datos de entrenamiento.

Sin embargo, si existe una brecha considerable entre los valores de R^2 en los conjuntos de entrenamiento y prueba, esto plantea un indicio de que el modelo podría tener dificultades para generalizar efectivamente a nuevos datos, ya que no logra explicar la variabilidad en el conjunto de prueba de manera tan eficiente como en el de entrenamiento. En otras palabras, el modelo podría estar demasiado ajustado a las particularidades de los datos de entrenamiento y no capturar de manera efectiva la estructura subyacente que se espera generalizar.

Es importante destacar que la discrepancia entre el rendimiento del modelo en los conjuntos de entrenamiento y prueba nos lleva a equilibrar la complejidad del modelo. Un modelo demasiado complejo puede memorizar los datos de entrenamiento teniendo buenas métricas de rendimiento, pero podría fallar al generalizar nuevos datos, esta brecha se evidencia en el conjunto de prueba.

En el contexto del MSE, una marcada discrepancia entre los conjuntos de entrenamiento y prueba, con un MSE más bajo en el conjunto de entrenamiento, sugiere que las predicciones del modelo se adaptan eficientemente a los datos de entrenamiento. Sin embargo, una diferencia significativa indica un posible sobreajuste, donde el modelo podría tener dificultades para proporcionar predicciones precisas en nuevos datos del conjunto de prueba.

En la Tabla 20, observamos que el modelo "Sin hiperparámetros" muestra altos valores de R^2 del 100% en el conjunto de entrenamiento, acompañados de un MSE de 0, considerablemente menor que los resultados en el conjunto de prueba. Este patrón confirma la presencia de sobreajuste en el modelo "Sin hiperparámetros". Resultados similares se obtienen para "*Cost Complexity Pruning*" en el caso 1, indicando cierto grado de sobreajuste. En el caso 3, los rendimientos son inferiores al resto, con una diferencia porcentual de 14 puntos en el R^2 entre el conjunto de entrenamiento y prueba. Esto podría sugerir sobreajuste, especialmente considerando que presenta los mayores valores de MSE, indicando posiblemente una limitación del modelo para capturar la relación entre variables independientes y dependiente. Al examinar el error de "*Cost Complexity Pruning*" en este caso, los valores de MSE en el conjunto de entrenamiento y prueba difieren en dos órdenes de magnitud, confirmando la presencia de sobreajuste.

Por último, los casos 2 y 4 exhiben métricas superiores y una menor brecha entre el R^2 del conjunto de entrenamiento y prueba. Al comparar los modelos y considerar las diferencias entre R^2 y MSE en ambos conjuntos, se puede concluir que el caso 2 en la técnica "*Cost Complexity Pruning*" muestra mejores métricas de rendimiento y una menor presencia de sobreajuste.

Además, se puede realizar un análisis de los residuos en el caso seleccionado, cuyos resultados y gráficas correspondientes se encuentran detallados en el Anexo 6. Para explorar estos mismos resultados en los otros casos, se puede consultar el mencionado anexo.

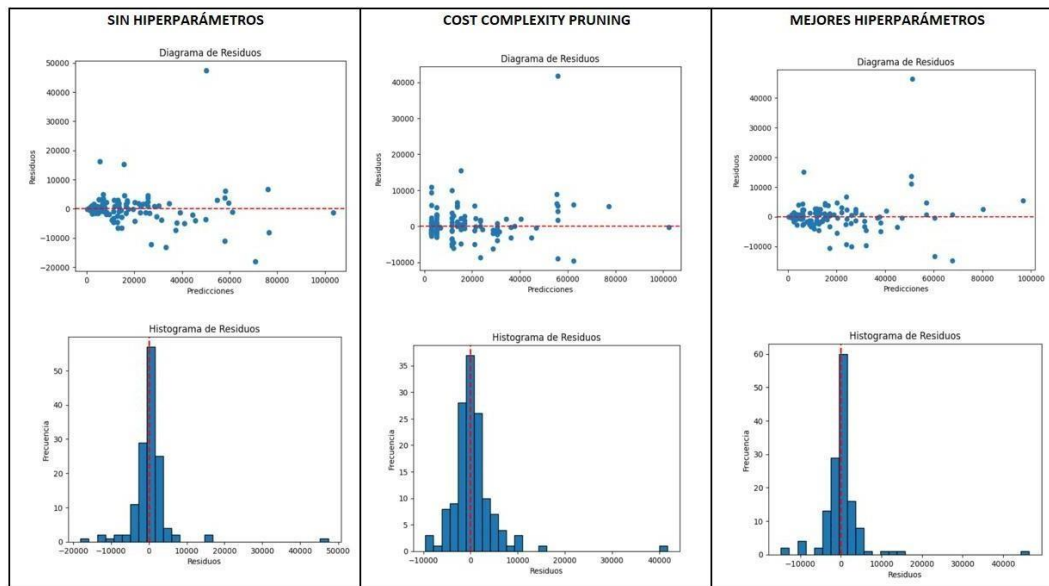


Figura 13: Diagrama e Histograma de Residuos – Caso 2 (Elaboración propia en base adatos obtenidos desde World Bank (2000-2020))

Se observa que los residuales se distribuyen de un alrededor del 0 y no se encuentran tendencias presentes, por lo tanto, se está en presencia de homocedasticidad y una normalidad en los residuales.

Variables importantes para cada caso.

En el análisis de importancias de características, exploraremos la relevancia de cada variable en la capacidad de nuestro modelo para predecir la variable dependiente. Un enfoque clave es examinar la profundidad de las características utilizadas como nodos de decisión de un árbol. Las características ubicadas en los nodos superiores del árbol desempeñan un papel más influyente en las predicciones finales al afectar a una mayor proporción de muestras de entradas. A continuación, se muestra un compilado de las importancias de las variables para cada caso:

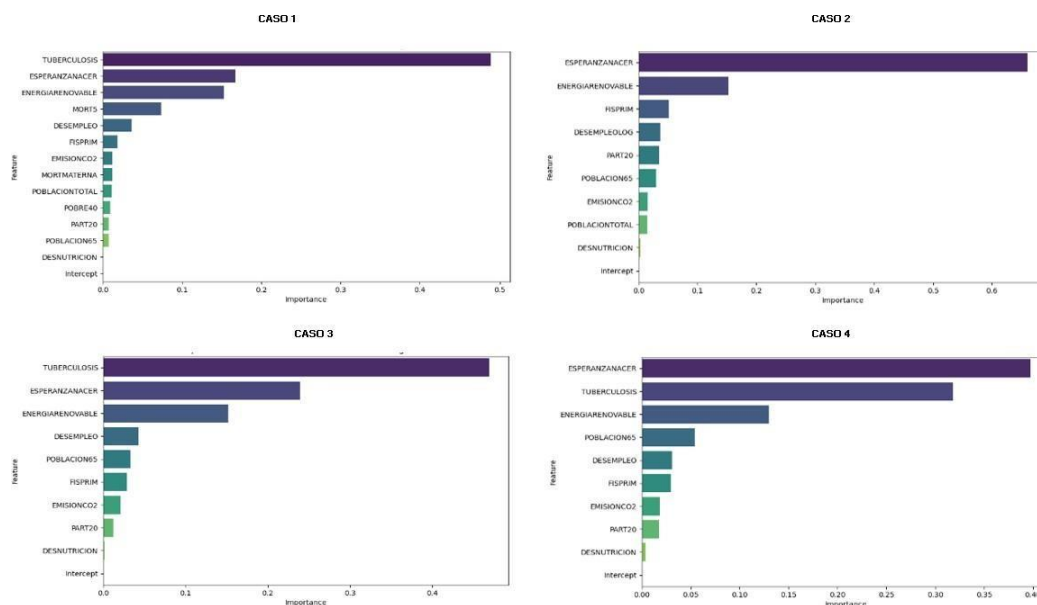


Gráfico 18: Compilado gráficas de importancia de variables en Random forest tree.
(Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Para entender mejor esta gráfica y consolidar los resultados se genera la siguiente Tabla:

Variables	Random forest tree				Frecuencia
	C1	C2	C3	C4	
Tasa de desempleo	✓	✓	✓	✓	4
Esperanza de vida al nacer	✓	✓	✓	✓	4
Consumo de energía renovable	✓	✓	✓	✓	4
Tasa de finalización nivel primario	✓	✓	✓	✓	4
Población mayor de 65 años		✓	✓	✓	3
Incidencia de tuberculosis	✓	✓		✓	3
Participación de ingreso del 20% peor remunerado		✓	✓		2
Emisiones de CO2		✓		✓	2
Tasa de mortalidad, menores de 5 años	✓				1
Población total					0
Tasa de mortalidad materna					0
Tasa de incidencia de pobreza por debajo del 40%					0
Prevalencia de desnutrición					0
Total	6	8	6	7	

Tabla 21: Variables importantes en algoritmo Random forest tree (Creación propia en base a datos obtenidos desde World Bank (2000-2020))

Se puede observar que las variables con mayor frecuencia de importancia en los casos estudiados concuerdan con las variables significativas, lo que nos da un indicio de cuáles son las variables que más pesan en el modelo. A continuación, se presenta la importancia de cada variable en términos de contribución porcentual al modelo:

Variables	Importancia				Promedio
	C1	C2	C3	C4	
Incidencia de tuberculosis	42.50%	48.11%	-	32.75%	41.12%
Esperanza de vida al nacer	21.99%	22.67%	66.02%	39.70%	37.59%
Consumo de energía renovable	14.97%	15.19%	14.98%	13.11%	14.56%
Tasa de mortalidad, menores de 5 años	8.37%	-	3.50%	-	5.94%
Tasa de desempleo	3.80%	4.10%	3.70%	3.26%	3.72%
Tasa de finalización nivel primario	2.40%	2.80%	5.72%	2.40%	3.33%
Población mayor de 65 años	0.83%	3.30%	2.97%	5.53%	3.16%
Participación de ingreso del 20% peor remunerado	0.72%	1.39%	3.50%	1.30%	1.73%
Emisiones de CO2	1.08%	2.03%	1.30%	1.60%	1.50%
Población total	1.02%	-	1.40%	-	1.21%
Tasa de mortalidad materna	1.07%	-	-	-	1.07%
Tasa de incidencia de pobreza por debajo del 40%	0.98%	-	-	-	0.98%
Prevalencia de desnutrición	0.13%	0.18%	0.26%	0.27%	0.21%

Tabla 21: Contribución de variables importantes al modelo de Árboles de regresión (Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Observamos que en términos porcentuales tenemos 3 variables que se destacan siendo “Incidencia de tuberculosis”, “Esperanza de vida al nacer” y “Consumo de energía renovable”, el caso 2 será tomado como referencia de mejor modelo, debido al comportamiento que ha experimentado frente a otros modelos.

Un punto importante que considerar es la consideración de la variable “Incidencia de tuberculosis” como variable importante y siendo la que mayor contribución tiene, esta variable no se mostraba significativa en modelos de regresión lineal, esto se puede explicar debido a que esta variable tiene un comportamiento logarítmico frente a PIB per cápita, nuestra variable dependiente. Pudo ocurrir que los modelos de regresión lineal no logaran capturar la no linealidad de esta variable no considerándola significativa, en el caso de *Random forest tree* este si logra capturar la tendencia no lineal de la “Incidencia de tuberculosis” y se posiciona como la variable más influyente en el modelo. También en este punto se puede entender la disminución en métricas de rendimiento en el caso 3, ya que luego a aplicar transformaciones logarítmicas esta variable tomo una alta correlación con “Esperanza de vida al nacer” es por esto que fue eliminada del modelo, la ausencia de esta variable que explica el comportamiento de la variable dependiente se puede traducir como una disminución de R-squared y aumento de MSE debido a que el modelo no se ajusta bien a los datos y tiene un rendimiento menos eficiente en la predicción de la variable dependiente.

6.5 Datos de panel

Como se detalló en la metodología se deben verificar los supuestos para definir si datos de panel es una metodología apropiada y cuál de sus variantes (Efectos fijos y efectos aleatorios) es la que posee mejores rendimientos

Homocedasticidad

Para que se elija Datos de panel como modelo se debe verificar la presencia de Heterocedasticidad, lo que nos haría rechazar la hipótesis nula de homocedasticidad. Esto se puede identificar de forma gráfica. Para visualizar la presencia de la heterocedasticidad se puede usar el gráfico de los residuales para detectar una varianza creciente.

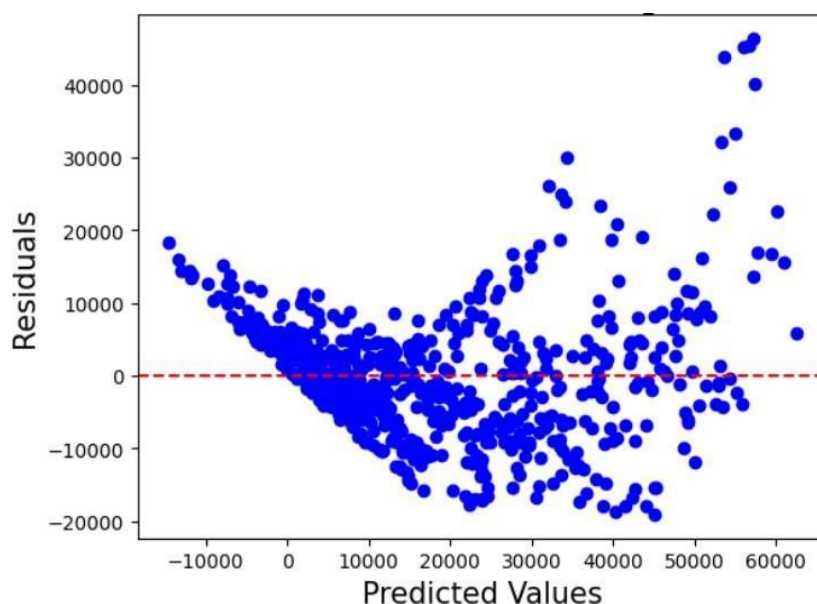


Gráfico 18: Gráfica de prueba de homocedasticidad de los residuales (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

La presencia de heterocedasticidad nos muestra una variabilidad no constante de los errores del modelo para los valores de las variables independiente. En un gráfico de residuales, lo ideal es que los puntos se distribuyan de manera aleatoria alrededor del valor 0 en el eje horizontal. En este caso se puede apreciar una tendencia en cierta parte de los datos entre los valores -10.000 y 20.000, en donde se observa una tendencia lineal, luego en los valores superiores a 20.000 se ve un aumento en los valores asemejándose a una forma de cono. Por lo tanto, estaríamos en presencia de heterocedasticidad. Al no cumplirse la condición de homocedasticidad nos da indicios de que la metodología de datos de panel puede ser apta para nuestros datos

Homocedasticidad y Autocorrelación.

A continuación, se detallan los valores de pruebas de homocedasticidad y no autocorrelación:

Breusch-Pagan test P-value			
C1	C2	C3	C4
1.33E-30	4.46E-25	1.58E-28	4.46E-25
Durbin Watson			
C1	C2	C3	C4
0.21	0.18	0.18	0.18

Tabla 22: Prueba de homocedasticidad y no autocorrelación (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

La prueba de Breusch Pagan nos indica la presencia de heterocedasticidad, el valor p obtenido es mucho menor al valor $p = 0.05$ el cual es el umbral comúnmente utilizado para considerar heteroscedasticidad, es decir el modelo posee heterocedasticidad como se pudo visualizar en el gráfico 18. Por otro lado, la prueba de Durbin-Watson nos arroja valores entre 0.18 y 0.21 lo que nos indica una autocorrelación negativa, no obstante, como se vio en el caso de regresión lineal múltiples esta tiene un valor cercano a 2 lo que nos indica no autocorrelación, estas discrepancias pueden explicarse por las diferentes metodologías o librerías utilizadas y puede generar afectos negativos en los resultados.

Teniendo en cuenta eso. Si bien se cumple la presencia de heterocedasticidad, no es del todo claro si existe autocorrelación, por lo tanto, esto nos podría indicar que datos de panel puede no ser un buen modelo para nuestros datos. De igual forma se procederá con la implementación del modelo en sus 2 variantes.

Efectos fijos y variables

Se realiza la prueba de Hausman la cual nos indica endogeneidad, entendiendo como endogeneidad como la característica de un modelo en donde algunas variables son determinadas de forma interna en lugar de ser influenciada por factores externos, la hipótesis nula es que la covarianza entre las variables predictoras y el término alfa es cero. Cumpliéndose la hipótesis nula se prefiere el modelo de efectos aleatorios a efectos fijos. En este caso un valor p pequeño indica que se puede rechazar la hipótesis nula y preferir el modelo de efectos fijos muestra la endogeneidad en el modelo.

En el caso del valor chi-squared si esta toma valores grandes existe una gran discrepancia entre las frecuencias observadas y esperadas. La hipótesis nula de chi-squared nos indica independencia de variables, si los resultados observados difieren significativamente de los resultados teóricos se puede rechazar la hipótesis nula, lo que indica que las variables están asociadas.

Hausman test - chi-squared			
C1	C2	C3	C4
49.2	61.61	57.02	61.6
Hausman test - P – value			
C1	C2	C3	C4
8.19E-06	1.79E-09	1.32E-08	1.79E-09

Tabla 23: Hausman test (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Los resultados obtenidos nos indican que los valores p son muy pequeños para todos los casos (<0.05) por lo que se rechaza la hipótesis nula, por lo tanto, el modelo de efectos fijos sería el más adecuado. Así mismo, el valor de chi-squared nos indica que las variables están asociadas. Por lo que se procede a comparar las variaciones de los modelos de dato de panel en términos de métricas de rendimiento.

	R – squared				MSE			
Modelo	C1	C2	C3	C4	C1	C2	C3	C4
Efectos fijos	42.5%	41.8%	42.2%	41.8%	2.14E+07	2.10E+07	2.16E+07	2.17E+07
Efectos variables	40.6%	39.9%	40.5%	39.9%	2.34E+07	2.37E+07	2.36E+07	2.37E+07

Tabla 24: Métricas de rendimiento para datos de panel de efectos fijos y variables(Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000- 2020))

Como se mencionó el modelo más adecuado sería el de efectos fijos obteniendo mejores rendimientos en términos de R^2 y un menor valor de MSE. No obstante, como se mencionó anteriormente no queda del todo claro el valor de la prueba de Durbin-Watson la cual nos indica la autocorrelación de las variables. Si los resultados de datos de panel de esta prueba son incorrectos esto indicaría que no cumple con el supuesto de autocorrelación y no sería óptimo ocupar datos de panel, ya sea para efectos fijos o variables.

Según los resultados obtenidos en métricas de rendimiento nos da indicios que la prueba de Durbin-Watson esta errónea, lo que explica el mal rendimiento del modelo y los bajos coeficientes de determinación obtenidos en comparación con el modelo de regresión lineal. Por lo tanto, para efectos de este problema se tomará el valor obtenido de Durbin- Watson obtenido en modelos de regresión lineal el cual es cercano a 2 lo que nos indicaría que no hay correlación entre las variables.

7. DISCUSIÓN

En esta etapa de discusión, realizaremos un análisis profundo de los resultados obtenidos, explorando el impacto de diversas variables sociales claves en el comportamiento del Producto interno bruto per cápita en conjunto con el análisis de los diferentes modelos de regresión, su impacto en las métricas de rendimiento y en la elección de variables significativas e importantes. Variables que se destacan en la contribución en la mejora del modelo predictivo y tienen significancia en el modelo son “Esperanza de vida al nacer”, “Incidencia de tuberculosis”, “Consumo de energía renovable”, “Tasa de desempleo” y “Participación del 20% peor remunerado”.

La relación entre el PIB per cápita y el bienestar social es bidireccional. Por un lado, el crecimiento económico puede permitir la inversión en las políticas de salud, educación y bienestar social mejorando la calidad de vida de los ciudadanos.

Cuando se ve la dirección desde cómo el desarrollo humano influye en el PIB per cápita en la investigación de Ranis y Stewart (2002) encontramos respaldo empírico de que a medida que la población es más sana, mejor alimentada y más instruida, contribuyen más al crecimiento económico. Los resultados de investigaciones a nivel económico han demostrado que un mejoramiento en la salud y nutrición influye directamente en la productividad de la mano de obra, especialmente entre las personas más pobres.

En el ámbito de la salud, según Flores (2006) esta se puede ver en forma de capital humanos, por ende, como un insumo y como un producto del proceso de crecimiento; los países con una población sana y con mejor educación tienen mayores posibilidades de prosperar.

Nos indica también *“Un niño que nace hoy en el mundo en desarrollo puede esperar vivir 20 años más que aquellos que lo hicieron hace 40 años y, además, desde 1965, han sido más que duplicados los ingresos per cápita del mundo en desarrollo”*.

Las variables que tienen mayor contribución al modelo son “Esperanza de vida al nacer” “Incidencia de tuberculosis” teniendo un promedio de contribución de 38% y 41% respectivamente, estas variables están intrínsecamente ligadas a la calidad de vida de la población. “Esperanza de vida al nacer” es un indicador clave de salud general de la población, un aumento en esta sugiere mejoras en la atención médica, mejora en la educación, la nutrición, disponibilidad de servicios, condiciones de vida y factores ambientales. Una población más saludable contribuye a disminuir la carga de enfermedades y mejorar la calidad de vida. Barro (1997) mostró en su informe que la esperanza de vida está significativamente correlacionada con el crecimiento económico, después de la segunda guerra mundial estimó que un aumento de 10% en la esperanza de vida aumentó un 0.4% por año el crecimiento económico.

La Tuberculosis es una enfermedad infecciosa provocada por bacteria *Mycobacterium tuberculosis*, que afecta principalmente a los pulmones, aunque puede manifestarse en otros órganos. A lo largo de la historia, la tuberculosis ha sido una de las enfermedades más prevalentes y ha tenido un impacto significativo en la humanidad. A menudo es conocida como la “Plaga blanca” y ha dejado una huella duradera en comunidades de todo el mundo. Su impacto se extiende más allá de la esfera médica, afectando aspectos sociales y económico de la vida humana, este indicador refleja directamente la eficacia de los sistemas de salud en la prevención y tratamiento de enfermedades infecciosas.

De acuerdo con Morales y Beldarraín (2015) se aprecia que los diversos patrones epidemiológicos están muy íntimamente relacionados con los factores económicos y sociales. El hecho de que una enfermedad ancestral perdure y vuelva a surgir en las comunidades se relaciona con situaciones de inequidad social, falta de prioridad de los programas de control y medidas de contención. La Tuberculosis supone una carga social y financiera importante para las personas que padecen la enfermedad, así como para sus familias y cuidadores.

Los mayores casos de Tuberculosis recaen en adultos jóvenes productivos, que a menudo se vuelven incapaces de trabajar en el caso de ancianos y niños, por lo general es una persona encargada o cuidador quién también ve los efectos de la TB incurre en pérdidas. A nivel mundial, el gasto de los hogares en la enfermedad puede llegar a representar entre el 8 y 20% de los ingresos anuales del hogar.

En cuanto al medioambiente, un aumento en el consumo de energía renovable puede traer consecuencias positivas para la sostenibilidad ambiental, pero también influir en la inversión y la innovación del sector energético. Al impulsar el crecimiento en industrias de energía renovable puede esta contribuir al PIB per cápita y ampliar la cantidad de empleos relacionados a ese rubro. En palabras de Posso, Acevedo y Hernández (2014) la creciente participación de energías renovables en el mercado mundial está teniendo consecuencias importantes como beneficios ambientales y sociales, sino también por su efecto positivo en la economía de los países, este efecto se materializan en flujos significativos de capital hacia la generación de energía renovable, un aumento en el empleo local, ya que se requiere personal constantemente tanto en la fase de diseño como en la de operación y un ahorro de recursos económicos originalmente destinados a combustibles fósiles.

Un aumento en la tasa de desempleo puede indicarnos una disminución en el poder adquisitivo de la población, lo que afectaría negativamente al consumo y en consecuencia al PIB per cápita. Por otro lado, una tasa de desempleo más baja se asocia a un mayor crecimiento económico, ya que las industrias al crecer tienen más puestos de trabajo y por ende una mayor producción. De acuerdo con Vargaz, Zuñiga, Vega-Granda, y Sotomayor (2021) el desempleo es un fenómeno económico que trae en consecuencia muchas problemáticas a la sociedad, como es la pobreza, precariedad, delincuencia y problemas psicológicos, afectando de manera directa a la satisfacción de necesidades humanas.

Siguiendo con esta línea de pensamiento Silva (2017) nos indica que el

desempleo tiene una seria repercusión en el desarrollo social y económico del país, este se ve reflejado en parámetros como la productividad y la contratación.

Finalmente, la participación del 20% peor remunerado puede indicar una distribución más equitativa de ingresos. La reducción de desigualdad económica puede tener efectos positivos en el consumo al tener más paridad en el poder adquisitivo; aquellos con ingresos más bajos, al tener mayor participación del ingreso total, no solo tienen acceso a lo necesario para sobrevivir, sino también pueden considerar una gama más amplia de productos y servicios. Este fenómeno puede impulsar una dinámica positiva de consumo, donde el peor remunerado pueda participar más activamente en la economía, generando demanda en diversos sectores. El acceso a opciones más allá de las necesidades básicas estimula la producción y el crecimiento económico general, incluyendo una mejora en la calidad de vida de los ciudadanos. Según Stiglitz (2012), 'Se argumenta que podríamos tener más igualdad, pero solo a costa de pagar el elevado precio de un crecimiento más lento y un PIB menor. La realidad es exactamente, al contrario: tenemos un sistema que ha estado trabajando horas extra a fin de trasladar el dinero desde los niveles inferiores y medios hasta el nivel más alto, pero el sistema es tan ineficiente que lo que salen ganando los de arriba es mucho menos de lo que pierden los de en medio y los de abajo. En realidad, estamos pagando un elevado precio por nuestra creciente y desmesurada desigualdad: no solo un crecimiento más lento y un PIB menor, sino incluso más inestabilidad'

En resumen, nuestra discusión destaca la conexión e influencia de las variables sociales y el PIB per cápita. Hemos investigado cómo aumentar la

participación equitativa en los ingresos y reducir la desigualdad no solo puede mejorar el bienestar general de la sociedad, sino que también puede mejorar la dinámica económica. Al tomar en cuenta indicadores como la distribución de ingresos, la salud y la educación, no solo estamos a favor de un mayor equilibrio social, sino que también reconocemos su importancia para impulsar el crecimiento económico. La investigación sostiene que una población más sana, educada y equitativamente remunerada tiene un impacto significativo en el desarrollo económico y el fortalecimiento del PIB.

8. CONCLUSIÓN

En conclusión, nuestro análisis exhaustivo revela la interconexión crítica entre variables sociales clave y el Producto Interno Bruto per cápita (PIB). Exploramos detalladamente el impacto de factores como la salud, la educación, el medio ambiente y la distribución de ingresos en el comportamiento económico, utilizando modelos de regresión y evaluando su significancia.

En el ámbito de la salud, observamos que variables como la “Esperanza de vida al nacer” y la “Incidencia de tuberculosis” emergen como contribuyentes significativos al modelo predictivo. La calidad de vida de la población, intrínsecamente ligada a estos indicadores, se vincula directamente con el crecimiento económico, respaldando la noción de que una población más sana favorece la productividad y contribuye al desarrollo económico.

En cuanto al medio ambiente, destacamos la influencia positiva del aumento en el consumo de energía renovable. Este factor no solo tiene beneficios ambientales, sino que también puede contribuir al PIB per cápita mediante flujos significativos de capital y la generación de empleo en la industria de energías renovables.

La tasa de desempleo también se presenta como una variable crucial, evidenciando una relación directa con el poder adquisitivo de la población y, por ende, con el PIB per cápita. Una tasa de desempleo más baja se asocia con un mayor crecimiento económico, contribuyendo a la producción y el consumo.

Finalmente, la participación del 20% peor remunerado emerge como un indicador de una distribución más equitativa de ingresos. La reducción de la desigualdad económica no solo mejora el bienestar social, sino que también puede impulsar una dinámica positiva de consumo, generando demanda en diversos sectores y contribuyendo al crecimiento económico general.

Esta investigación encuentra resultados similares a las investigaciones respaldadas por autores como Ranis, Stewart, Flores, Barro y Stiglitz, concluimos que el bienestar social y económico están intrínsecamente entrelazados. Una población saludable, educada y equitativamente remunerada no solo mejora la calidad de vida, sino que también fomenta un entorno para el desarrollo económico sostenible. Este análisis respalda la importancia de abordar integralmente las variables sociales para comprender y potenciar el crecimiento económico de manera inclusiva y sostenible.

En torno a los modelos utilizados, el análisis detallado de diversos modelos para ajustar nuestros datos ha proporcionado una comprensión clara de cómo cada algoritmo se desempeña en nuestro contexto. Las regresiones lineales múltiples, a pesar de ofrecer métricas de rendimiento adecuadas para explicar parte del comportamiento de la variable dependiente, muestran limitaciones al no poder capturar las tendencias no lineales presentes en nuestro estudio. Se destaca Regresión lineal múltiple Ridge, que al abordar la multicolinealidad, mejora significativamente las métricas de rendimiento y reduce el número de condición.

En el caso de los datos de panel, se observan rendimientos deficientes, posiblemente debido a la no satisfacción de uno de los supuestos fundamentales para utilizar esta metodología. Hay discrepancia entre el estadístico de Durbin-Watson, que indica autocorrelación positiva según el código implementado, y los resultados de las regresiones lineales que sugieren lo contrario, por lo tanto, se descarta este modelo.

La metodología de árboles de decisión destaca al obtener métricas de rendimiento superiores, gracias a su capacidad para capturar relaciones no lineales, mediante decisiones de poda, demuestran eficacia al reducir el sobreajuste del modelo.

9. ANEXOS

Anexo 1:

Europa:

- Alemania (DEU)
- Armenia (ARM)
- Belorusia (BLR)
- Chipre (CYP)
- Dinamarca (DNK)
- Eslovenia (SVN)
- España (ESP)
- Estonia (EST)
- Finlandia (FIN)
- Georgia (GEO)
- Hungría (HUN)
- Italia (ITA)
- Kazajstán (KAZ)
- Kirguistán (KGZ)
- Letonia (LVA)
- Lituana (LTU)
- Noruega (NOR)
- Polonia (POL)
- República Checa (CZE)
- República Eslovaca (SVK)
- Suecia (SWE)
- Ucrania (UKR)

América:

- Argentina (ARG)
- Colombia (COL)
- Costa Rica (CRI)
- Ecuador (ECU)
- El Salvador (SLV)
- Panamá (PAN)
- Paraguay (PRY)
- Perú (PER)
- República Dominicana (DOM)
-

Asia:

- Indonesia (IDN)
- Israel (ISR)

Anexo 2

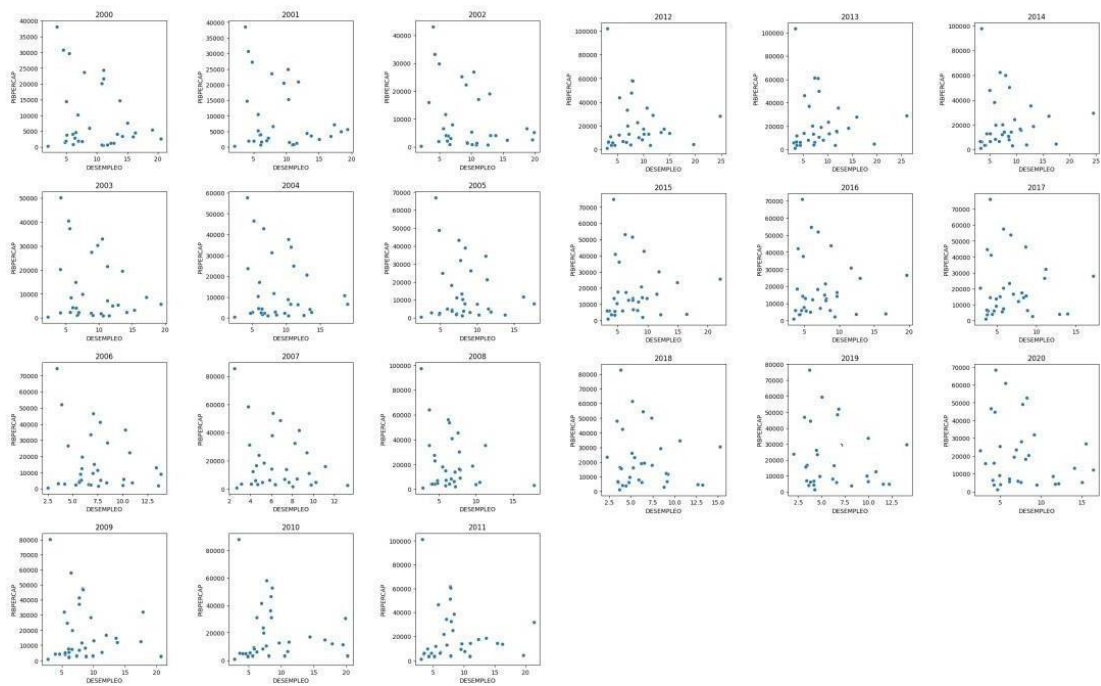


Gráfico 19: Gráficos de PIB per cápita en función de Tasa de desempleo para todos los años de estudio (Fuente: Elaboración propia en base a datos obtenidos desde World Bank(2000-2020))

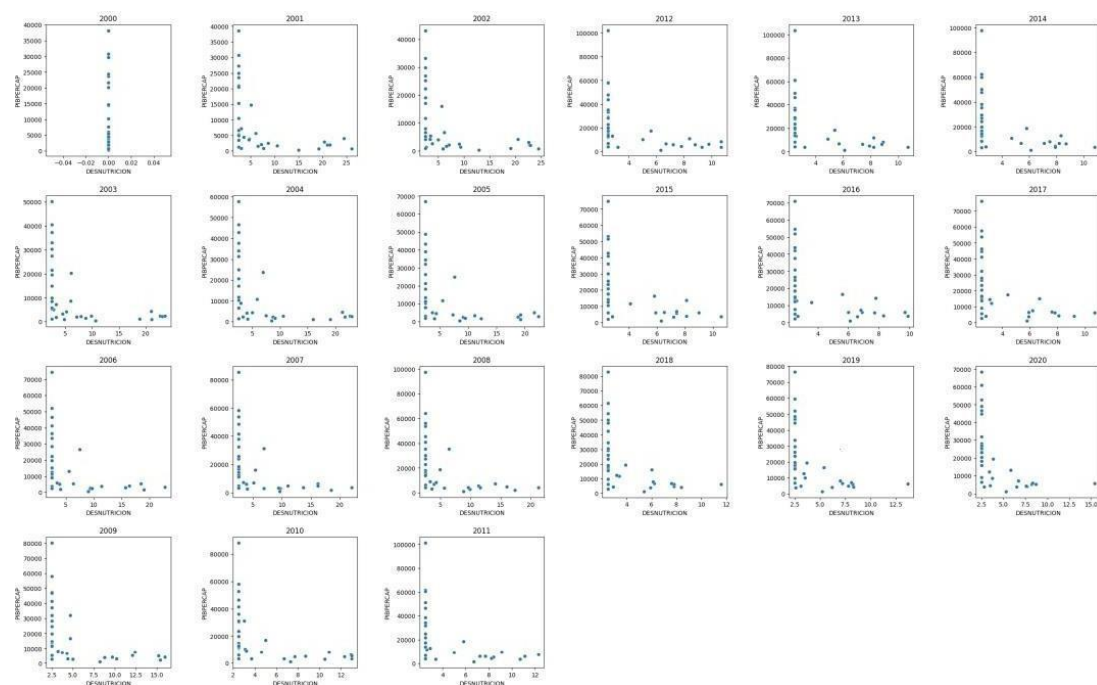


Gráfico 20: Gráficos de PIB per cápita en función de Prevalencia de desnutrición para todos los años de estudio (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

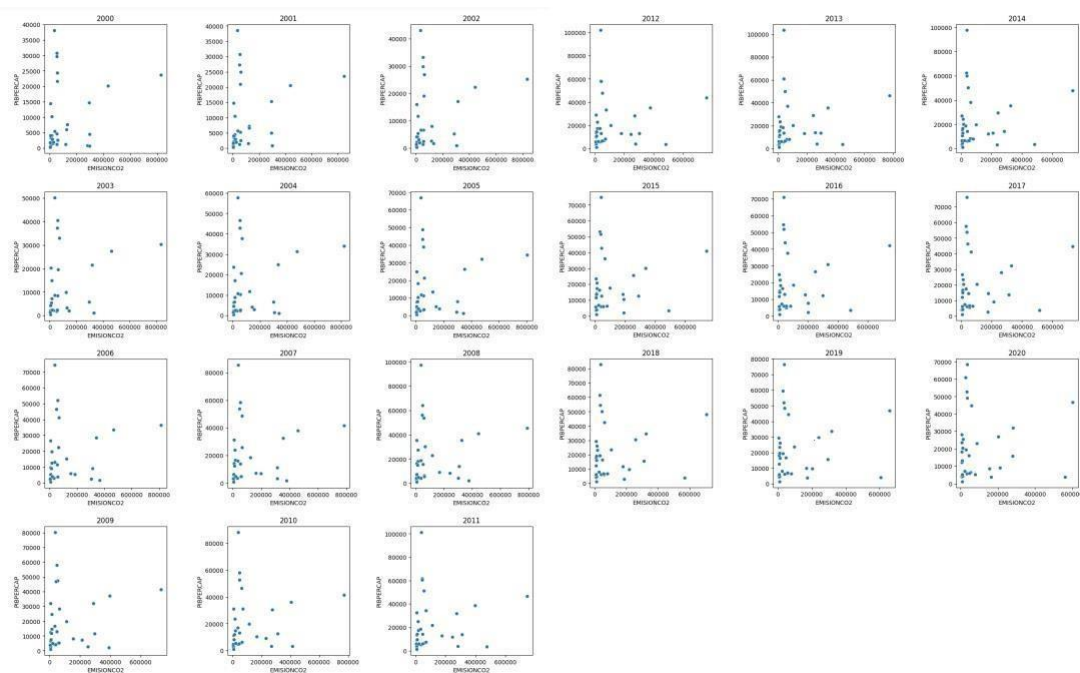


Gráfico 21: Gráficos de PIB per cápita en función de Emisiones de CO2 para todos los años de estudio (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

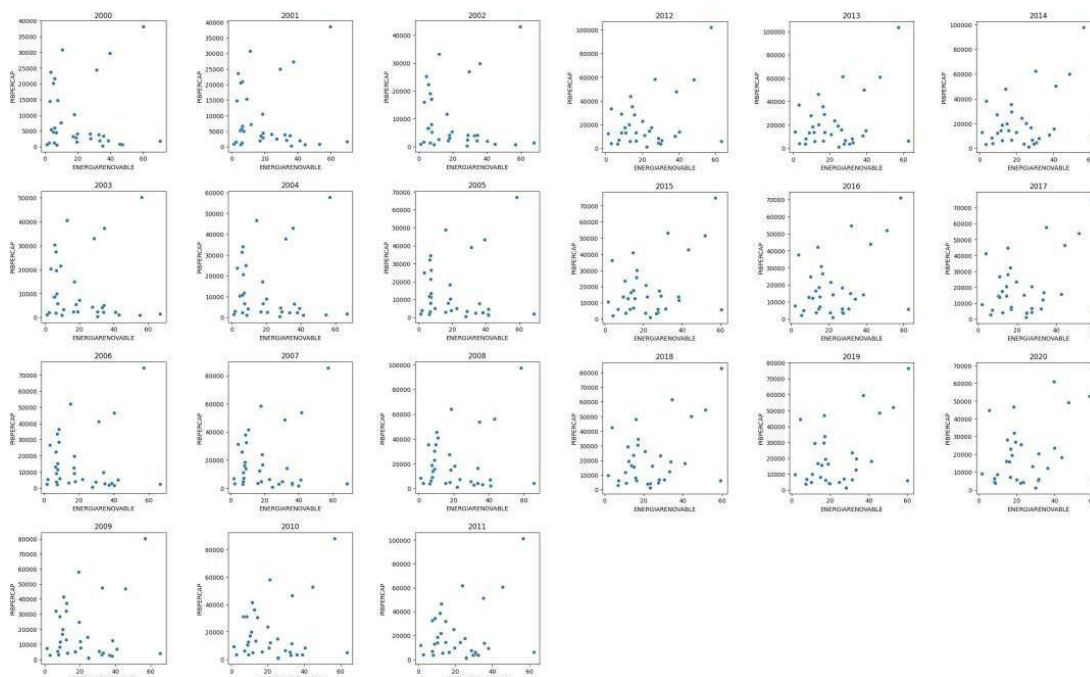


Gráfico 22: Gráficos de PIB per cápita en función de Consumo de energía renovable para todos los años de estudio (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

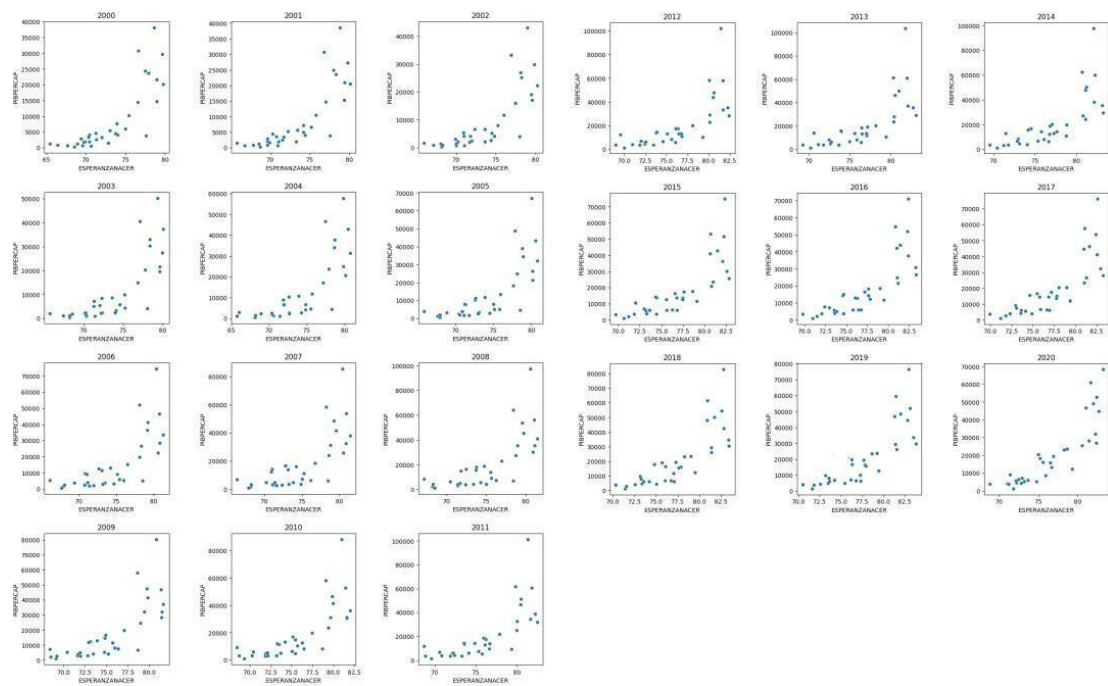


Gráfico 23: Gráficos de PIB per cápita en función de Esperanza de vida al nacer para todos los años de estudio (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

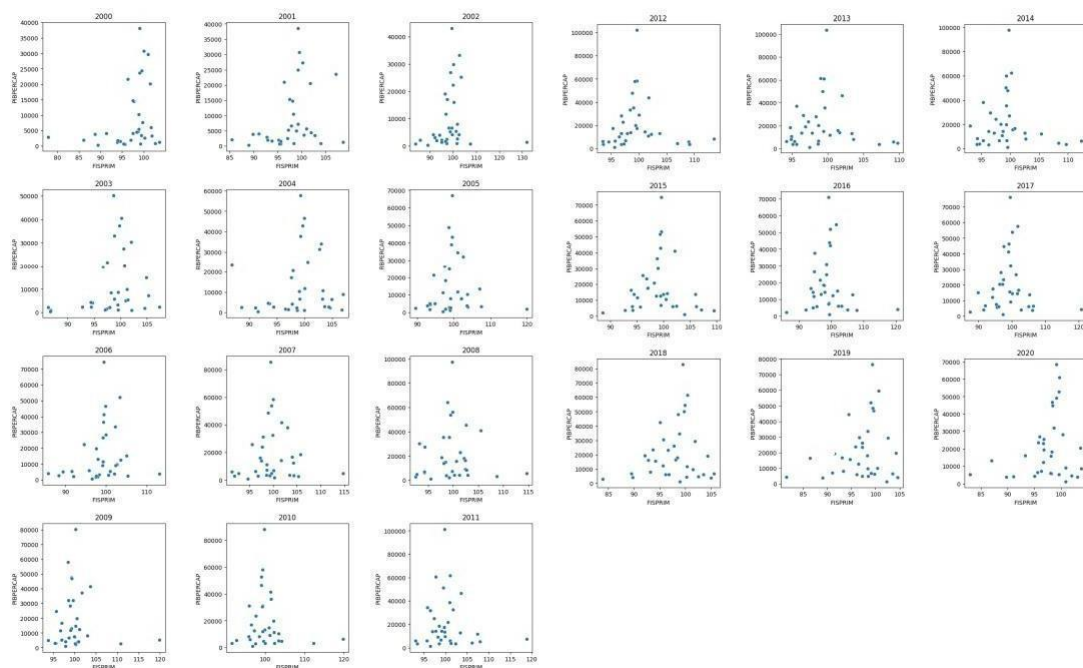


Gráfico 24: Gráficos de PIB per cápita en función de Tasa de finalización nivel primario para todos los años de estudio (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

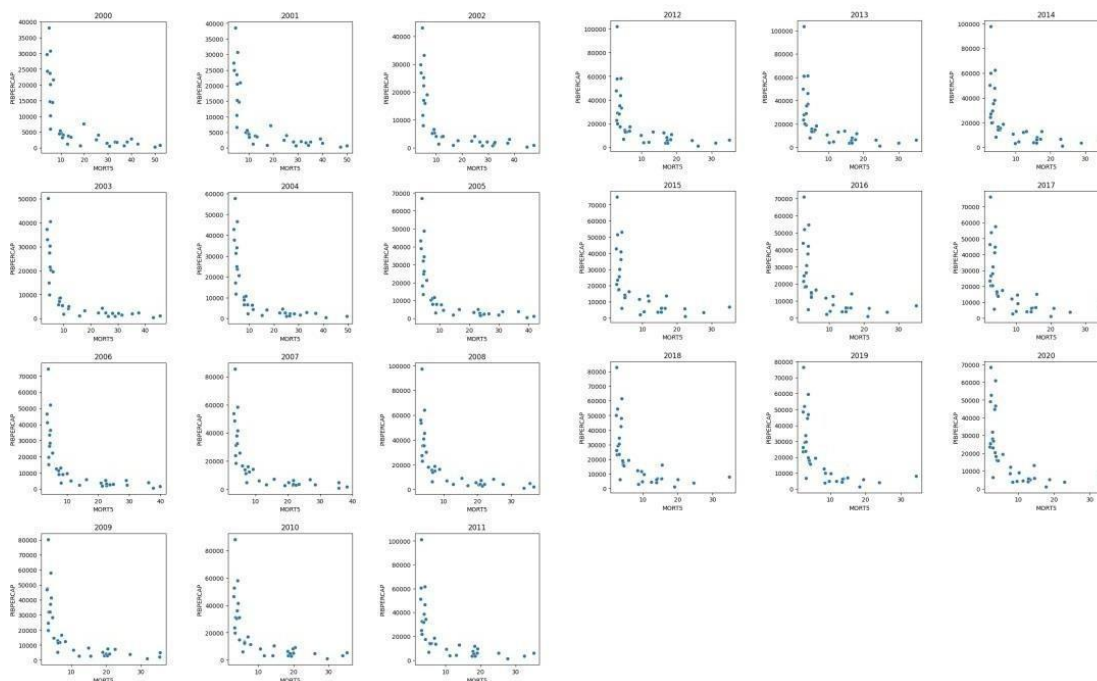


Gráfico 25: Gráficos de PIB per cápita en función de Tasa de mortalidad en menores de 5 años para todos los años de estudio (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

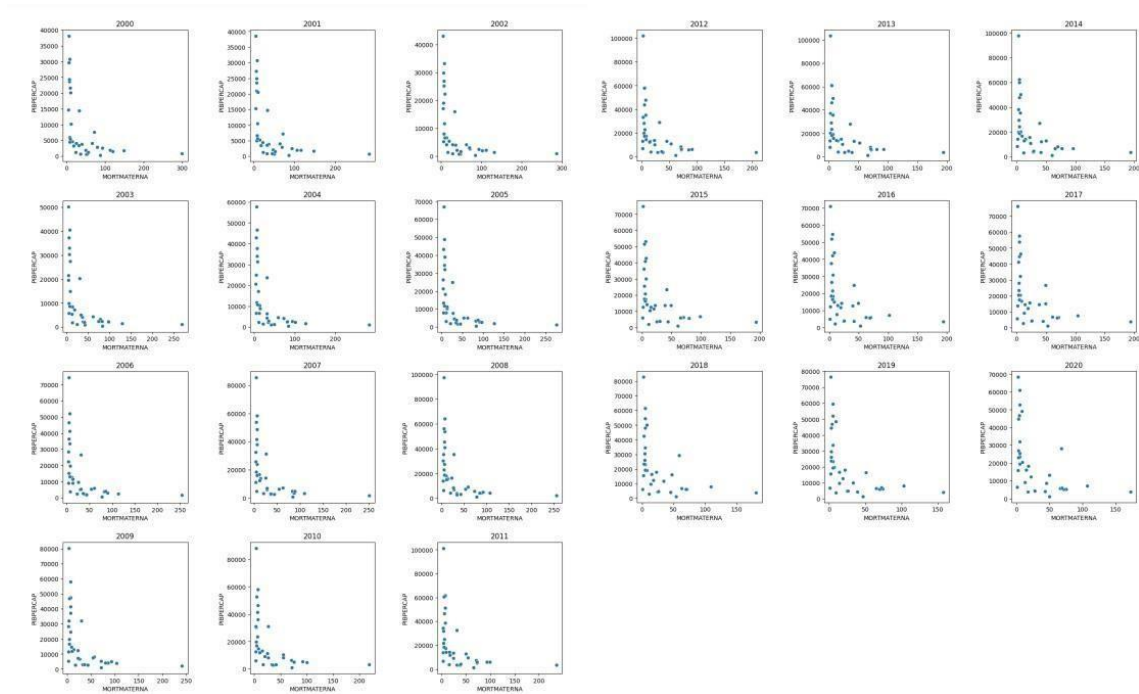


Gráfico 26: Gráficos de PIB per cápita en función de Tasa de mortalidad materna paratodos los años de estudio (Fuente: Elaboración propia en base a datos obtenidos desdeWorld Bank (2000-2020))

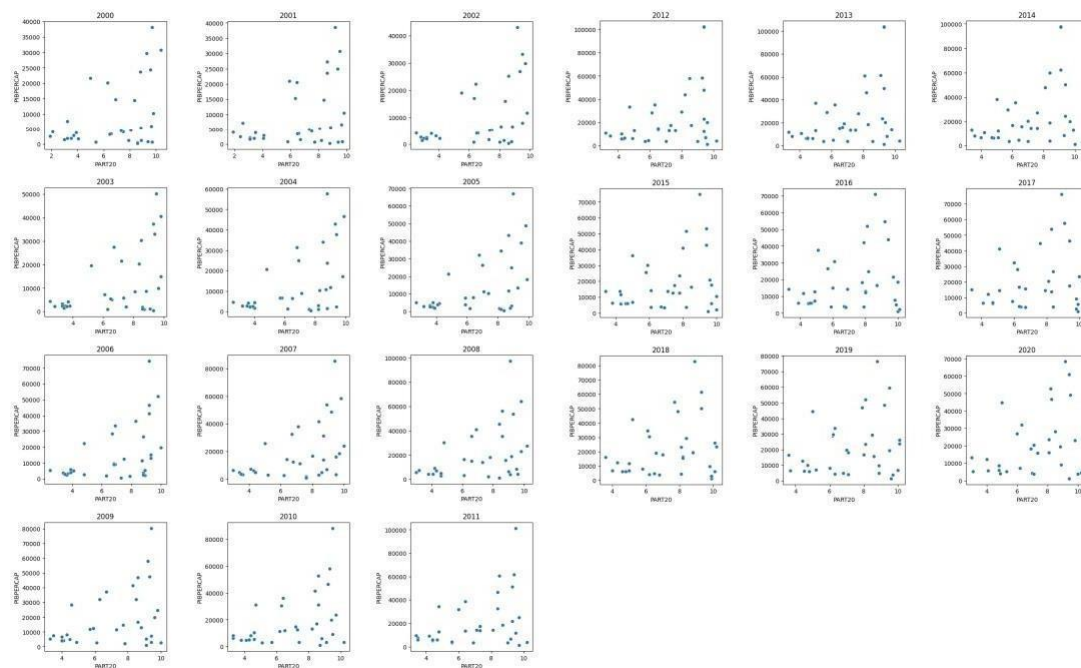


Gráfico 27: Gráficos de PIB per cápita en función de Participación del 20% peor remunerado para todos los años de estudio (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

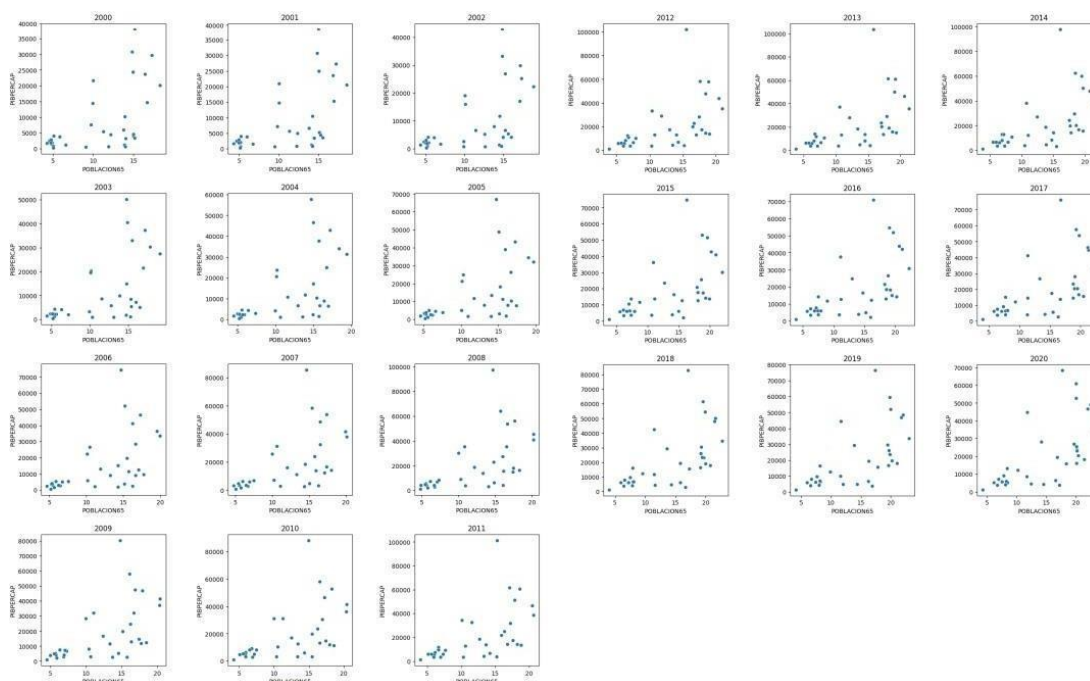


Gráfico 28: Gráficos de PIB per cápita en función de Población mayor a 65 años para todos los años de estudio (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

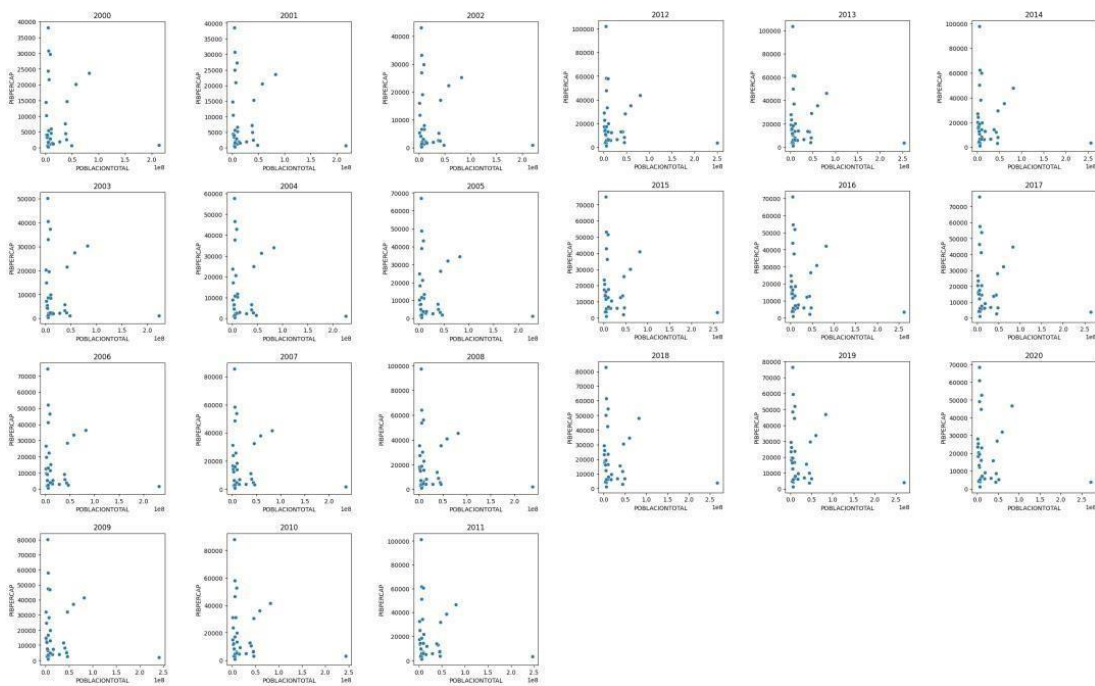


Gráfico 29: Gráficos de PIB per cápita en función de Población total para todos los años de estudio (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000- 2020))

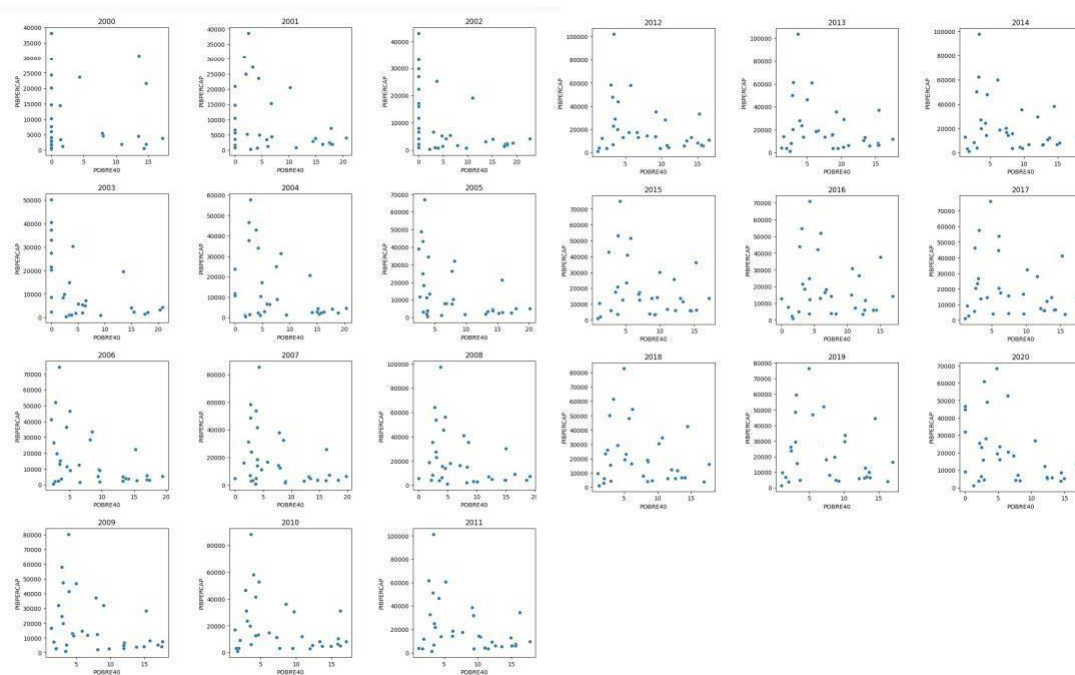


Gráfico 30: Gráficos de PIB per cápita en función de Tasa de incidencia de pobreza por debajo el 40% para todos los años de estudio (Fuente: Elaboración propia en base a datosobtenidos desde World Bank (2000-2020))

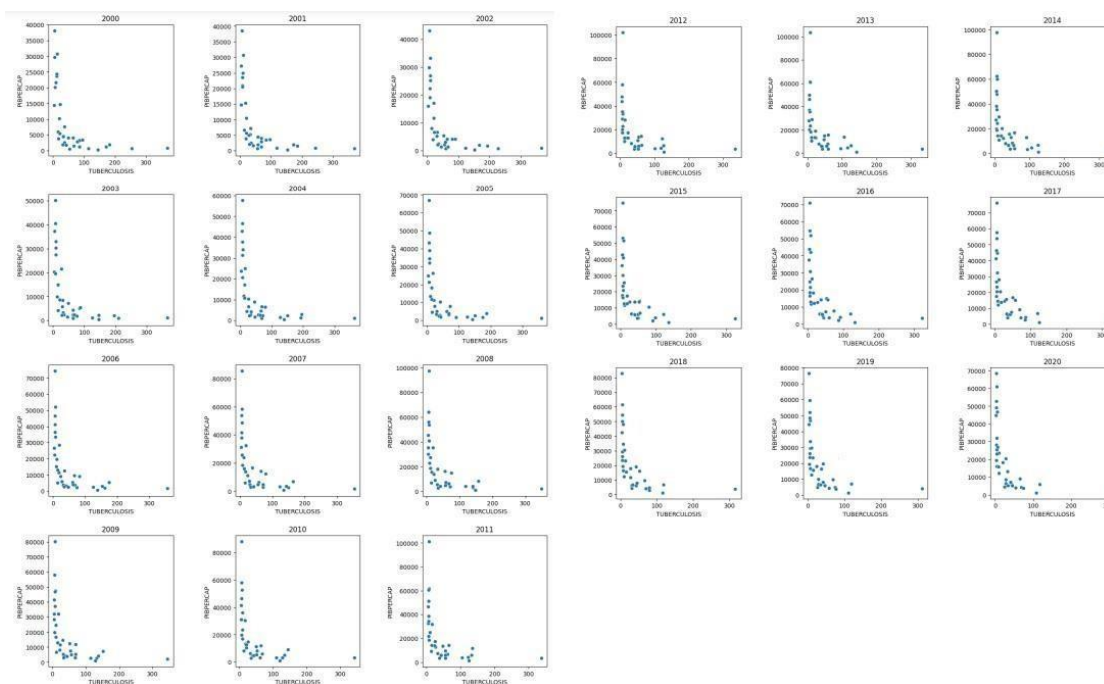


Gráfico 31: Gráficos de PIB per cápita en función de Incidencia de tuberculosis para todos los años de estudio (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Anexo 3

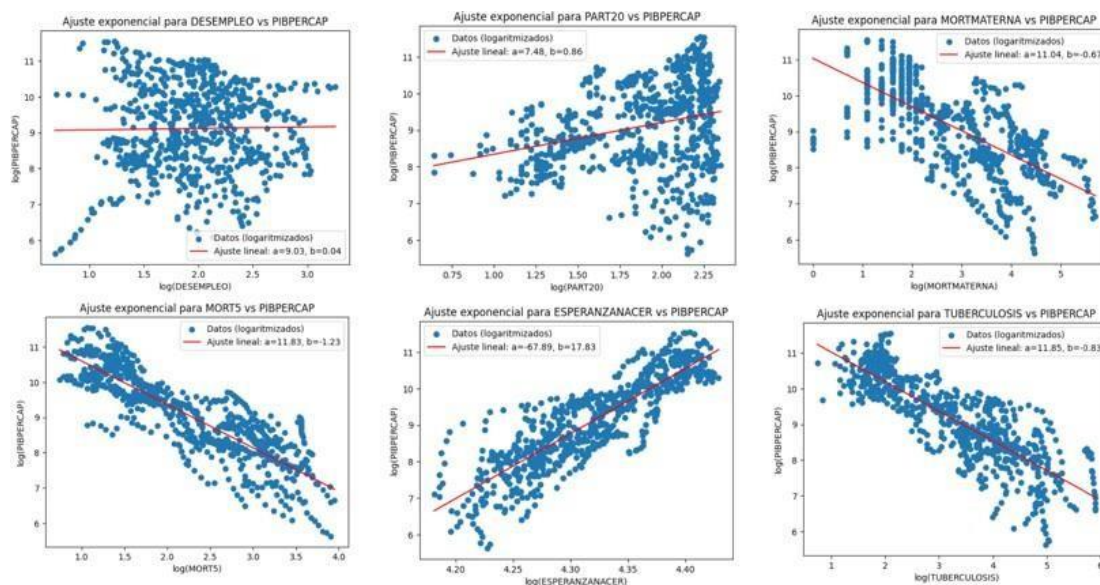


Gráfico 32: Ajuste en línea de tendencia de variables seleccionadas (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

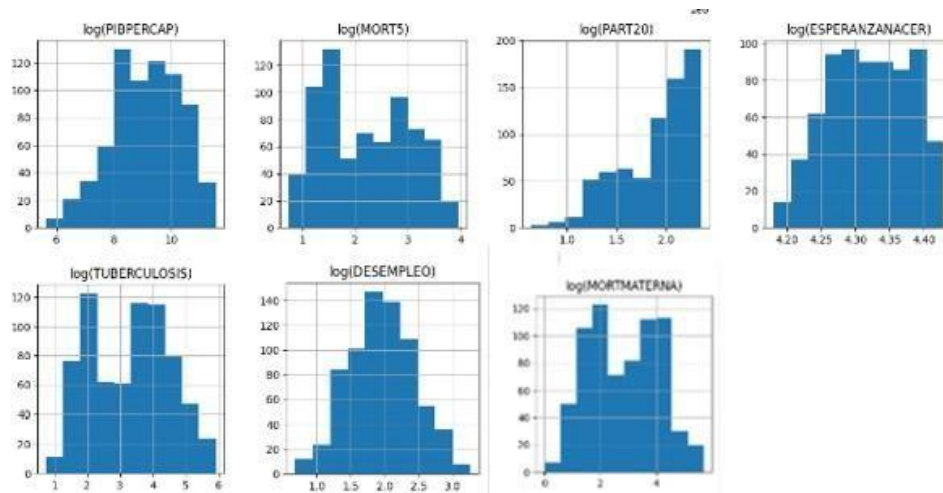


Gráfico 33: Histogramas de variables con ajuste en línea de tendencia (Fuente: Elaboración propia en base a datos obtenidos desde World Bank (2000-2020))

Anexo 4:

- Regresión lineal múltiple SKLearn



Gráfico 34: Gráfica de p-values para C1 Regresión SKLearn (Fuente: Creación en base adatos obtenidos desde World Bank (2000-2020))



Gráfico 35: Gráfica de p-values para C2 Regresión SKLearn (Fuente: Creación en base adatos obtenidos desde World Bank (2000-2020))

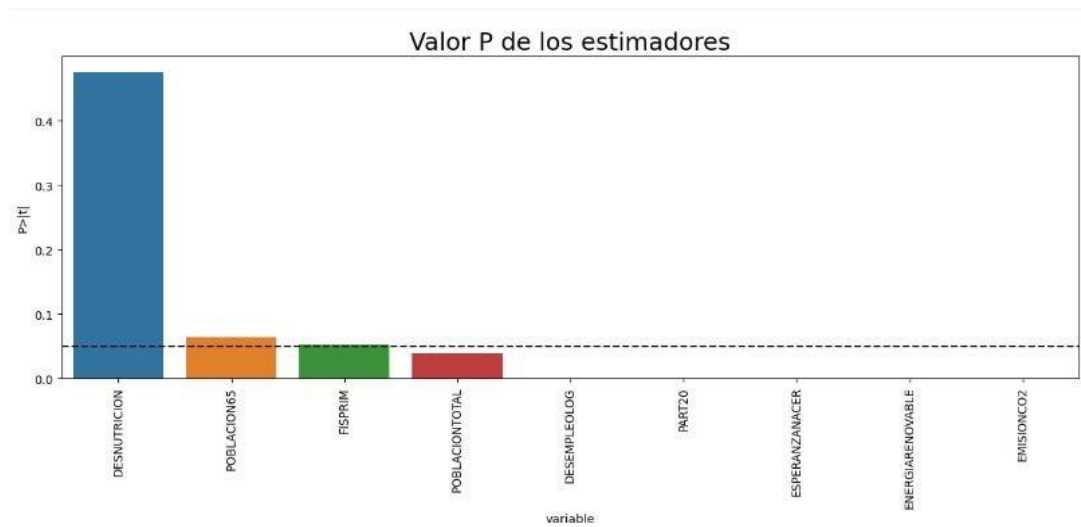


Gráfico 36: Gráfica de p-values para C3 Regresión SKLearn (Fuente: Creación en base adatos obtenidos desde World Bank (2000-2020))



Gráfico 37: Gráfica de p-values para C4 Regresión SKLearn (Fuente: Creación en base adatos obtenidos desde World Bank (2000-2020))

- Regresión lineal múltiple Stepwise

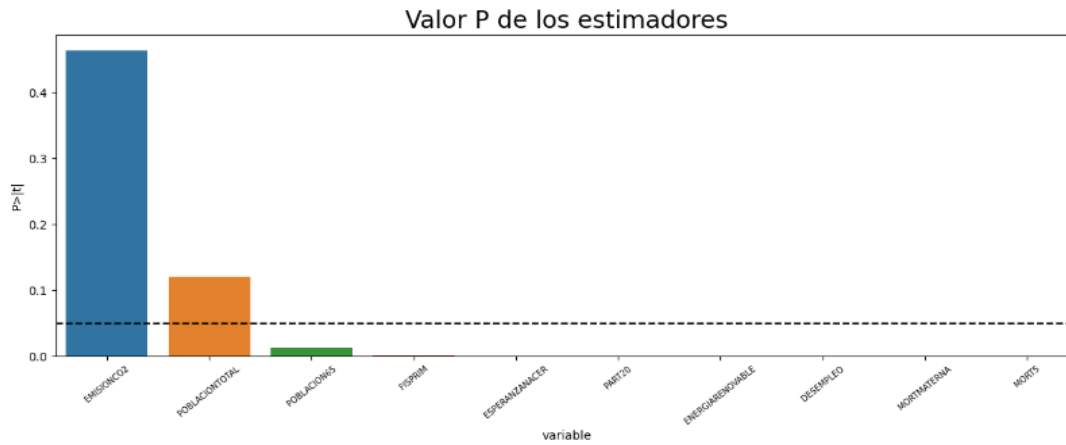


Gráfico 38: Gráfica de p-values para C1 Regresión Stepwise (Fuente: Creación en base adatos obtenidos desde World Bank (2000-2020))

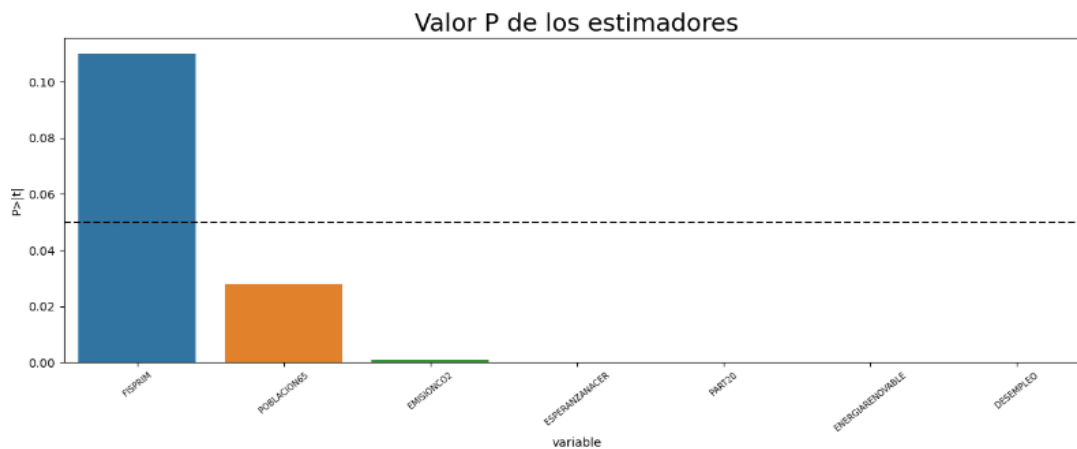


Gráfico 39: Gráfica de p-values para C2 Regresión Stepwise (Fuente: Creación en base adatos obtenidos desde World Bank (2000-2020))



Gráfico 40: Gráfica de p-values para C3 Regresión Stepwise (Fuente: Creación en base adatos obtenidos desde World Bank (2000-2020))

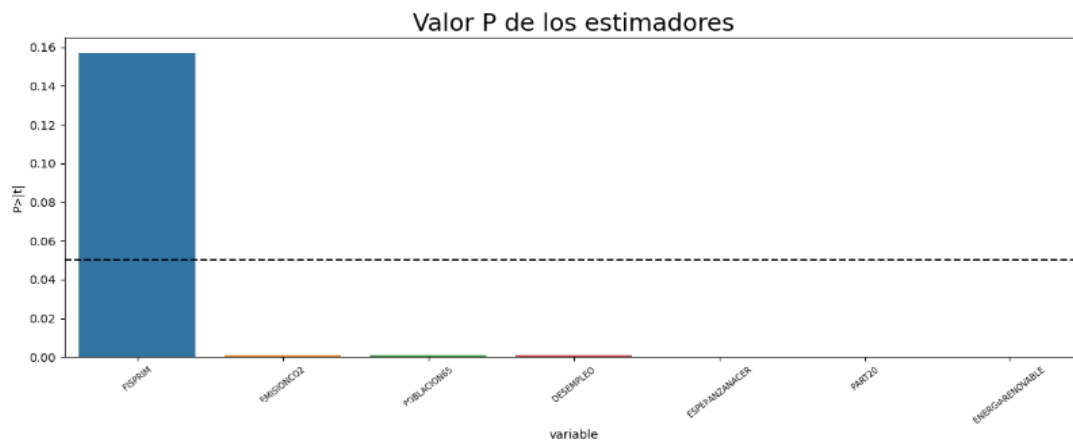


Gráfico 41: Gráfica de p-values para C4 Regresión Stepwise (Fuente: Creación en base adatos obtenidos desde World Bank (2000-2020))

- *Regresión lineal múltiple Ridge*

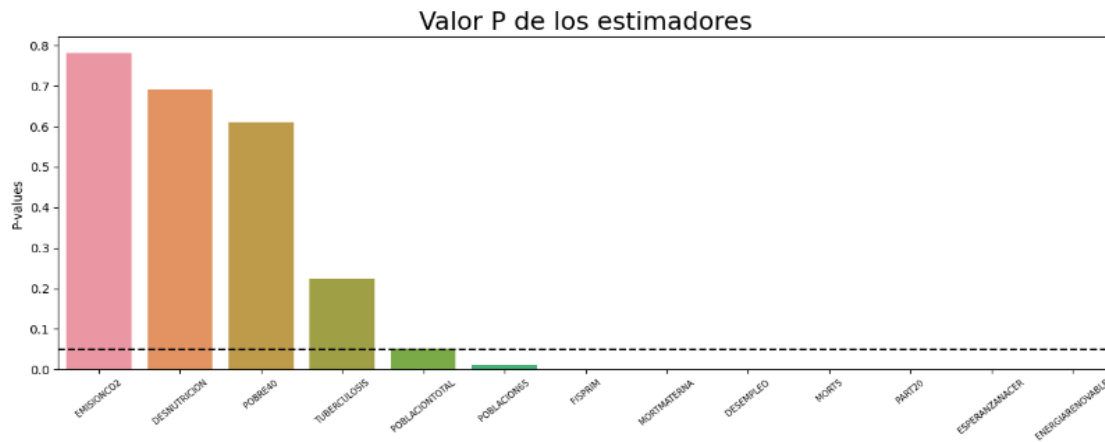


Gráfico 42: Gráfica de p-values para C1 Regresión Ridge (Fuente: Creación en base a datos obtenidos desde World Bank (2000-2020))

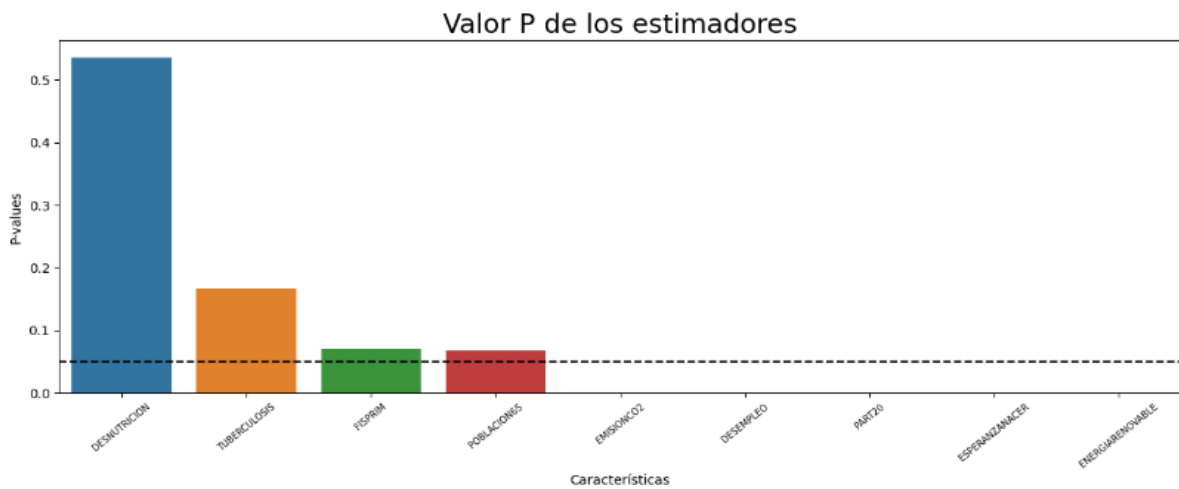


Gráfico 43: Gráfica de p-values para C2 Regresión Ridge (Fuente: Creación en base a datos obtenidos desde World Bank (2000-2020))

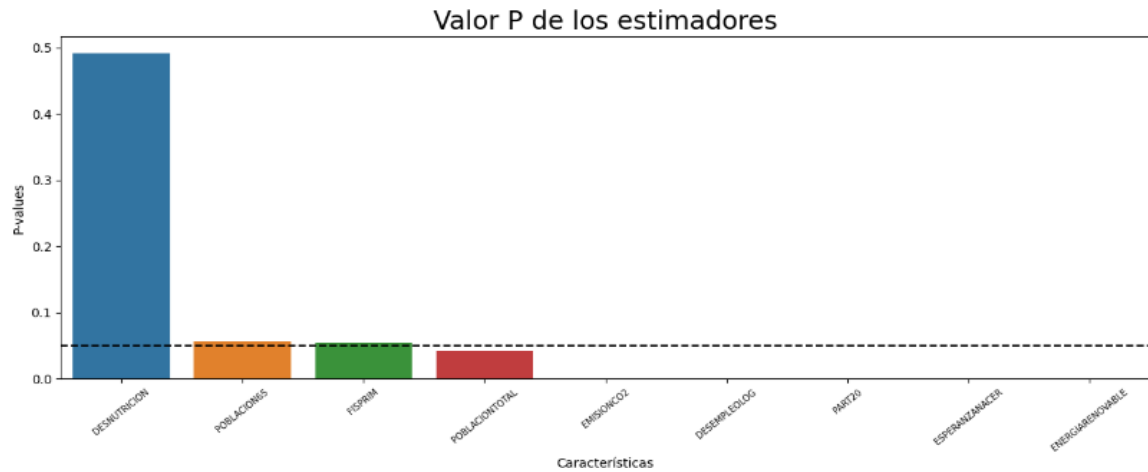


Gráfico 44: Gráfica de p-values para C3 Regresión Ridge (Fuente: Creación en base a datos obtenidos desde World Bank (2000-2020))



Gráfico 45: Gráfica de p-values para C4 Regresión Ridge (Fuente: Creación en base a datos obtenidos desde World Bank (2000-2020))

Anexo 5:

	Linear SKLearn			
Variables	C1	C2	C3	C4
Tasa de finalización nivel primario	313	151	161	127
Tasa de desempleo	-478	-444	-3,920	-390
Tasa de mortalidad, menores de 5 años	588	-	-	-
Tasa de incidencia de pobreza por debajo del 40%	-68	-	-	-
Participación de ingreso del 20% peor remunerado	2,585	2,624	2,564	2,227
Esperanza de vida al nacer	3,135	2,718	2,788	2,607
Tasa de mortalidad materna	-115	-	-	-
Incidencia de tuberculosis	-20	-14	-	5
Consumo de energía renovable	315	319	318	255
Prevalencia de desnutrición	48	70	78	-28
Emisiones de CO2	7.E-04	1.E-02	2.E-02	1.E-02
Población total	7.E-05	-	-3.E-05	4.E-05
Población mayor de 65 años	377	235	246	410

Tabla 25: Coeficientes de regresión lineal múltiple SKLearn (Fuente: Creación propia enbase a datos obtenidos desde World Bank (2000-2020))

	Linear Stepwise			
Variables	C1	C2	C3	C4
Tasa de finalización nivel primario	293	132	158	133
Tasa de desempleo	-479	-460	-4,014	-385
Tasa de mortalidad, menores de 5 años	543	-	-	-

Tasa de incidencia de pobreza por debajo del 40%		-	-	-
Participación de ingreso del 20% peor remunerado	2,620	2,501	2,516	2,262

Esperanza de vida al nacer	3,239	2,836	2,778	2,567
Tasa de mortalidad materna	-105	-	-	-
Incidencia d tuberculosis	-	-	-	-
Consumo de energía renovable	307	304	319	259
Prevalencia de desnutrición	-	-	-	-
Emisiones de CO2	4.E-03	9.E-03	2.E-02	1.E-02
Población total	-	-	-3.E-05	-
Población mayor de 65 años	372	266	223	401

Tabla 26: Coeficientes de regresión lineal múltiple Stepwise(Fuente: Creación propia enbase a datos obtenidos desde World Bank (2000-2020))

Anexo 6

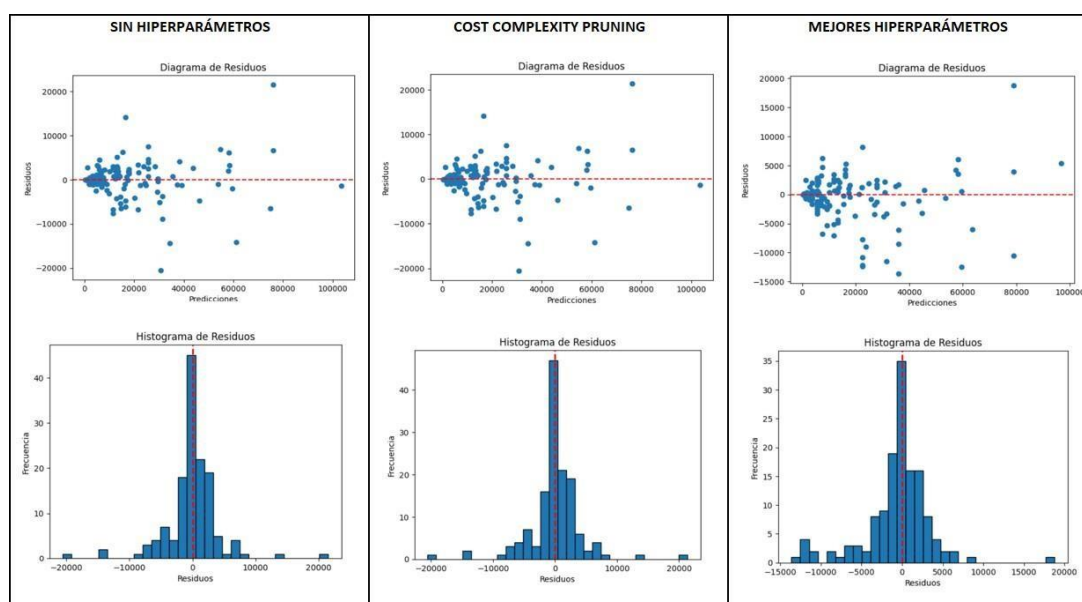


Figura 14: Diagrama e Histograma de Residuos – Caso 1 (Creación propia en base a datos obtenidos desde World Bank (2000-2020))

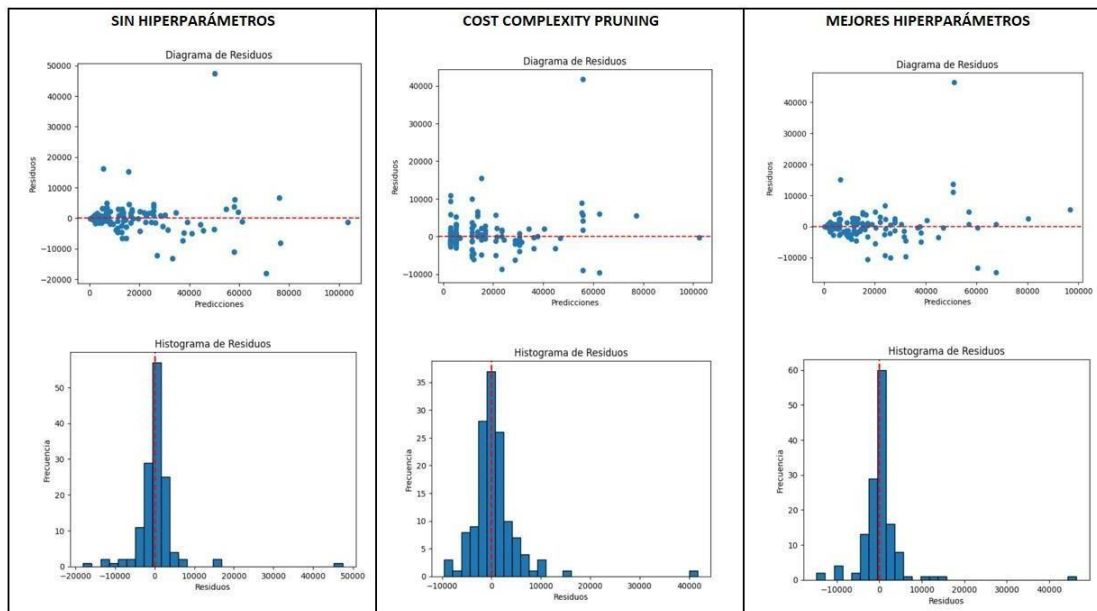


Figura 15: Diagrama e Histograma de Residuos – Caso 2 (Creación propia en base a datos obtenidos desde World Bank (2000-2020))

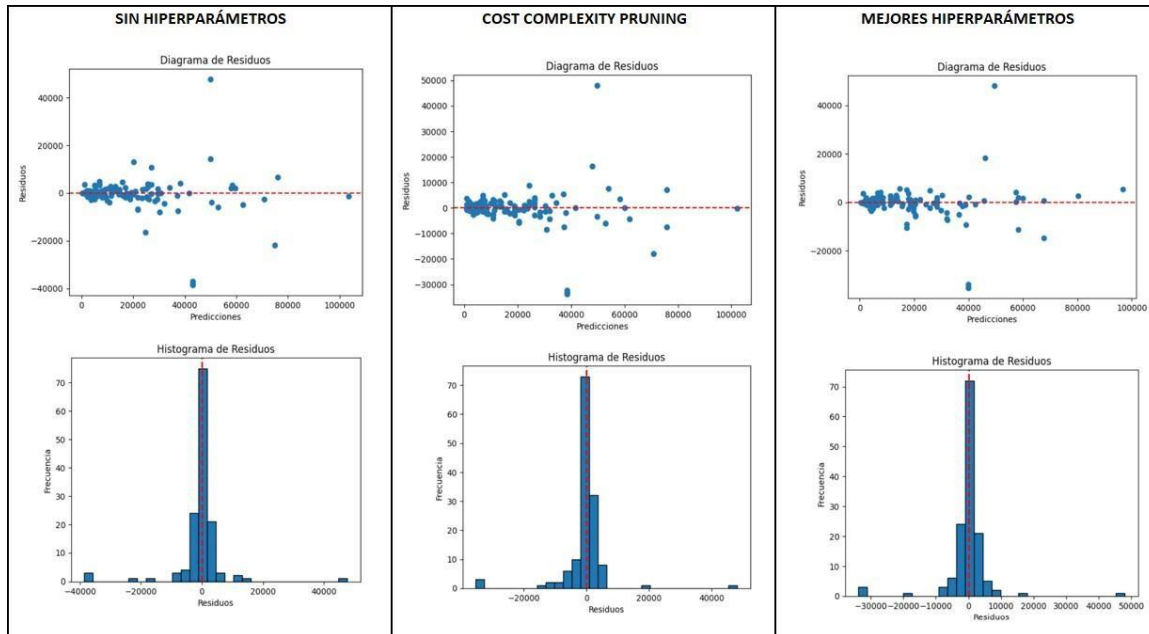


Figura 16: Diagrama e Histograma de Residuos – Caso 3 (Creación propia en base a datos obtenidos desde World Bank (2000-2020))

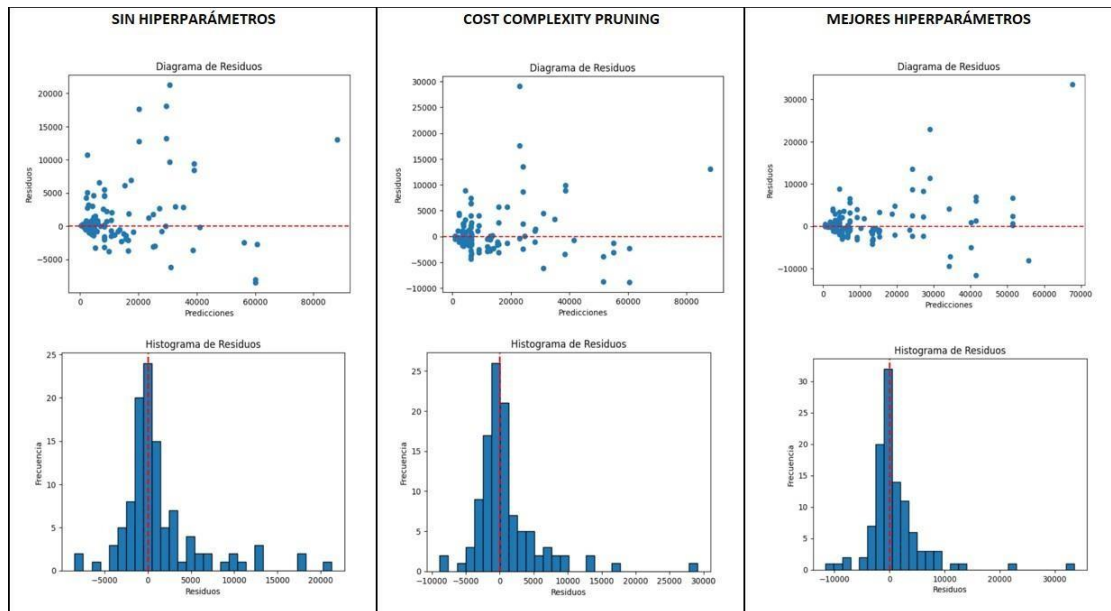


Figura 17: Diagrama e Histograma de Residuos – Caso 4 (Creación propia en base a datos obtenidos desde World Bank (2000-2020))

Bibliografía

- Our World in Data. (2023). *Poverty*. Obtenido de Our World in Data: <https://ourworldindata.org/poverty>
- Addlink. (19 de Junio de 2015). *Addlink Software científico* . Obtenido de Cómo interpretar los resultados del análisis de regresión: p-valores y coeficientes: <https://www.addlink.es/noticias/minitab/2540-como-interpretar-los-resultados-del-analisis-de-regresion-p-valores-y-coeficientes>
- Arévalo, E., Zurita, E., & González, M. (2020). *Crecimiento económico y bienestar social y ambiental. Construcción del índice de bienestar económico sostenible para América Latina periodo 2000-2018*.
- ASEPELT. (2000). Reunión Anual. Actas de la XIV Reunión: .
- Barro, R. (1997). *Determinants of Economic Growth* .
- Barro, R., & Sala, M. (2004). *Economic Growth*,. Seconf Edition MIT Press.
- Biblioteca Nacional de Chile (Servicio Nacional del Patrimonio). (2022). Obtenido de <https://www.memoriachilena.gob.cl/602/w3-article-3411.html#presentacion>
- Biblioteca Nacional de Chile. (2023). *El Estado de Bienestar social (1924-1973)*. Obtenido de Memoria Chilena: <https://www.memoriachilena.gob.cl/602/w3-article-3411.html>
- Brugger, B. (6 de Junio de 2021). *Una guía para regresión de datos de panel: teoría e implementación con python*. Obtenido de Towards Data Science : <https://towardsdatascience.com/a-guide-to-panel-data-regression-theoretics-and-implementation-with-python-4c84c5055cf8>
- Castillo Gatica, L., Ruiz Betancurt , C., & Valdebe. (2017). *Variables macroeconómicas y sociodemográficas que afectan el Bienestar Subjetivo de los países Latinoamericanos*.
- Chacín, R. B. (2019). Well-Being and development: Evolution of two concepts associated with the good living.
- Correa, F. (2004). *Crecimiento económico y medambiente: Una revisión analítica de las hipótesis de la curva ambiental de Kuznets*.
- Deaton, A. (2009). Ingresos, salud y bienestar en todo el mundo: evidencia de la encuesta mundial de Gallup.
- Diario Financiero . (2022). Concentración sector Retail Chile.
- Díaz, R., Masedaa, M., & Neira, I. (2011). Bienestar y Felicidad: Relación con la renta y el capital social en países europeos. *Revista Galega de Economía*.
- Dumitrescu Bogdan , A., Dedu , V., & Enciu , A. (2012). *THE CORRELATION BETWEEN UNEMPLOYMENT AND REAL GDP GROWTH*.
- Duran, J., & Peña, L. (2020). *Incidencia de la educación en el desarrollo económico de los municipios de Colombia*.
- Ejea, D. (2017). *Árboles de regresión: Algunos algoritmos y extensiones a métodos de consenso*.
- Escalona, E. (2006). Relación salud-trabajo y desarrollo social: visión particular en los trabajadores de la educación. *Revista Cubana de Salud Pública*.
- Esperanza, T. (2020). *Aprendizaje automático: Métodos y aplicaciones a los transtornos cerebrales* .
- Fernández, S. (2022). *Métodos de regresión y clasificación basados en árboles* .
- Finanzas & Desarrollo . (2017). Replantearse el PIB .
- Flores, G. (2006). *HEALTH AS A FACTOR OF ECONOMIC GROWTH*.
- Fundación IO. (2023). *Tuberculosis en Indonesia en 2022*. Obtenido de Fundación IO : <https://fundacionio.com/tuberculosis-en-indonesia/#:~:text=Indonesia%20ocupa%20el%20tercer%20lugar,de%2045%20a%2054%20a%C3%B1os>.
- García, J., & Céspedes, N. (2011). *Pobreza y crecimiento económico: tendencias durante la década del 2000*.
- García, M. (2018). *Análisis de sensibilidad mediante Random Forest*.
- Gascón, N. (2017). *Estudio del comportamiento de dos contrastes de normalidad a través de experimentos de Monte Carlo*.
- Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts,Tools and*

- Techniques to Build Intelligent Systems.*
- Guareño, J. (2016). *Support vector regression: Propiedades y aplicaciones* .
- Hackeling, G. (2017). *Mastering Machine Learning with scikit-learn*.
- Hernández, F. (2023). *Modelos predictivos*.
- Jiménez, L. A. (2018). *Criterio de Akaike para la selección de modelos con transformaciones* .
- Kohrsen, W. (10 de Enero de 2018). *Ajuste de hiperparámetros del bosque aleatorio en Python*. Obtenido de Towards data science: <https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74>
- Layard, R. (2005). *Happiness: Lessons from a new science*. Penguin Books/Penguin Group (USA).
- Madison, A. (1989). *The world economy in the 20th century*.
- Maridueña, A. (2018). Evolución e impacto del PIB y el IDH en un mundo desigual. *Revista Vinculando*.
- Mendez, J. (2019). *Rpubs*. Obtenido de https://rpubs.com/jorge_mendez/609253#:~:text=La%20regresi%C3%B3n%20paso%20a%20pas,o,modelos%20con%20cientos%20de%20variables.
- Ministerio de Desarrollo Social y Familia. (2017). *Evolución de la Pobreza 1990-2017 ¿Cómo ha cambiado Chile?*
- Morales, N., & Beldarraín , E. (19 de Abril de 2015). *Aspectos económicos y sociales relacionados con el comportamiento de la Tuberculosis en Cuba. Década de los 90*. Obtenido de Medisur: : <http://medisur.sld.cu/index.php/medisur/article/view/2864>
- Mwiti, D. (Septiembre de 2023). *Random Forest Regression*. Obtenido de neptune.ai.
- Pedregosa, F. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2825-2830.
- Pena Trapero, B. (2009). *La medición del Bienestar Social: una revisión crítica Estudios de Economía Aplicada*.
- PennState Eberly College of Science . (2023). *Minimal Cost-Complexity Prunning*. Obtenido de online.stat.psu.edu.
- Posso, F., Acevedo, J., & Hernández, J. (2014). *El impacto económico de las energías renovables*.
- Rajan, D. (2022). Un algoritmo eficiente de regresión de ridge con estimación de parámetros para el análisis de datos en aprendizaje automático . *SN Computer Science* .
- Ranis, G., & Stewart, F. (Diciembre de 2002). Crecimiento económico y desarrollo humano en América Latina. *Revista de la cepal* 78.
- Salmerón, R. (2020). *El problema de la multicolinealidad grave en modelos econométricos. Aplicación a las finanzas* . Universidad de Granada. Programa Oficial de Doctorado en Ciencias Económicas y Empresariales.
- Sandoval, F. (2011). *Durbin-Watson*.
- Silva, C. (2017). *Análisis de desempleo en Ecuador. Causas, consecuencias y posibles soluciones*.
- Sitiobigdata. (27 de Agosto de 2018). *Aprendizaje Automático ml: Métricas de regresión*,. Obtenido de Sitiobigdata: <https://sitiobigdata.com/2018/08/27/machine-learning-metricas-regresion-mse/>
- Stiglitz, J. (2002). Empleo, Justicial social y BInestar de la sociedad. *Revista Internacional del Trabajo*.
- Stiglitz, J. (2012). *El precio de la desigualdad: El 1% de la población tiene lo que el 99% necesita* .
- Talari, S. (1 de Noviembre de 2022). *Random Forest vs Decision Tree: Key Differences*. Obtenido de KDnuggets: <https://www.kdnuggets.com/2022/02/random-forest-decision-tree-key-differences.html>
- Vargaz, K., Zuñiga, F., Vega-Granda, A., & Sotomayor, J. (2021). *Ecuador: Determinantes de la tasa de desempleo periodo 1990-2020*.
- Veenhoven, R. (2000). *The Four Qualities of Life” adentro Journal Of Happiness Studies*.
- Veenhoven., R. d. (1991). *¿La felicidad es relativa? Investigación de Indicadores Sociales*.
- Wolrd Bank. (2023). *Tasa finalización nivel primario - Banco mundial* . Obtenido de Banco mundial: <https://datos.bancomundial.org/indicador/SE.ADT.LITR.ZS?view=chart>
- World Bank . (2023). *PIB per cápita - Banco mundial*. Obtenido de Banco mundial: <https://datos.bancomundial.org/indicador/NY.GDP.PCAP.CD?view=chart>

World Bank. (2023). *Consumo de energía renovable - Banco mundial*. Obtenido de Banco mundial:
<https://datos.bancomundial.org/indicador/EG.FEC.RNEW.ZS?view=chart>

World Bank. (2023). *Emisiones de CO2 - Banco mundial*. Obtenido de Banco mundial :
<https://datos.bancomundial.org/indicador/EN.ATM.CO2E.KT?view=chart>

World Bank. (2023). *Esperanza de vida al nacer - Banco mundial*. Obtenido de Banco mundial:
<https://datos.bancomundial.org/indicador/SP.DYN.LE00.IN?view=chart>

World Bank. (2023). *Incidencia de tuberculosis - Banco mundial*. Obtenido de Banco mundial:
<https://datos.bancomundial.org/indicador/SH.TBS.INCD?view=chart>

World Bank. (2023). *Participación de ingreso del 20% peor remunerado - Banco mundial*. Obtenido de Banco mundial : <https://datos.bancomundial.org/indicador/SI.DST.FRST.20?view=chart>

World Bank. (2023). *Población mayor a 65 años - Banco mundial*. Obtenido de Banco mundial:
<https://datos.bancomundial.org/indicador/SP.POP.65UP.TO.ZS?view=chart>

World Bank. (2023). *Población total - Banco mundial*. Obtenido de Banco mundial :
<https://datos.bancomundial.org/indicador/SP.POP.TOTL?view=chart>

World Bank. (2023). *Prevalencia de desnutrición - Banco mundial*. Obtenido de Banco mundial:
<https://datos.bancomundial.org/indicador/SN.ITK.DEFC.ZS?view=chart>

World Bank. (2023). *Tasa de desempleo - Banco mundial*. Obtenido de Banco mundial:
<https://datos.bancomundial.org/indicador/SL.UEM.TOTL.ZS?view=chart>

World Bank. (2023). *Tasa de mortalidad en menores de 5 años*. Obtenido de Banco mundial:
<https://datos.bancomundial.org/indicador/SH.DYN.MORT?view=chart>

World Bank. (2023). *Tasa de mortalidad materna - Banco mundial*. Obtenido de Banco mundial :
<https://datos.bancomundial.org/indicador/SH.STA.MMRT?view=chart>