Texto

Descripción generada automáticamente

Universidad Internacional de La Rioja

Escuela Superior de Ingeniería y Tecnología

Máster Universitario en Análisis y Visualización de Datos Masivos/ Visual Analytics and Big Data

Crecimiento de la Riqueza en el mundo: Factores y Perspectivas Futuras

|  |  |
| --- | --- |
| Trabajo fin de estudio presentado por: | Machado Rincón, Juan Felipe |
| Tipo de trabajo: | Desarrollo Software |
| Director/a: | Fuentes Lorenzo, Damaris |
| Fecha: |  |

Resumen

La pobreza ha sido un tema abarcado en mayor o menor medida por la mayoría de los países como política económica con el fin de lograr una reducción de esta. Pero la generación de riqueza y los factores que pueden influir en esta no ha sido un tema tan estudiado.

En este documento se realiza una investigación para desarrollar dos modelos de inteligencia basados en redes neuronales y árboles de clasificación diseñados para predecir y clasificar la riqueza per cápita por país a partir de datos macroeconómicos de los países como por ejemplo el comercio exterior, la deuda, o el propio mercado laboral interno con el fin de explorar y comparar la efectividad de diferentes enfoques de modelado para ofrecer una comprensión más profunda de los datos y mejorar la toma de decisiones basada en ellos.

Adicionalmente, se desarrolló un *dashboard* interactivo utilizando el software PowerBI que permite la visualización de los datos asociados a este trabajo, con el fin que el usuario pueda obtener sus propias conclusiones y análisis a partir de los datos.

Todo el desarrollo del trabajo se realizó bajo la metodología CRISP-DM y al final se podrá encontrar las conclusiones ya que ambos modelos tienen sus ventajas particulares y pueden ser aplicados en diversas situaciones dependiendo de los requisitos específicos del análisis.

**Palabras clave:** Riqueza, Árbol de decisión, Regresión, Minería de Datos, Redes neuronales.

Abstract

Poverty has been a topic addressed to a greater or lesser extent by most countries as an economic policy to achieve its reduction. However, the generation of wealth and the factors that may influence it have not been studied as extensively.

In this document, research is conducted to develop two intelligence models based on neural networks and classification trees designed to predict and classify per capita wealth by country using macroeconomic data from the countries, such as foreign trade, debt, or the internal labor market, to explore and compare the effectiveness of different modeling approaches to provide a deeper understanding of the data and enhance decision-making based on it.

Additionally, an interactive dashboard was developed using Power BI software, allowing the visualization of the data associated with this work, so that the user can draw their own conclusions and analyses from the data.

The entire development of the work was carried out under the CRISP-DM methodology, and in the end, conclusions can be found since both models have their particular advantages and can be applied in various situations depending on the specific analysis requirements.

**Keywords**: Wealth, Decision Tree, Regression, Data Mining, Neural Netework.

Índice de contenidos

[1. Introducción 1](#_Toc171102269)

[1.1. Motivación 2](#_Toc171102270)

[1.2. Planteamiento del trabajo 3](#_Toc171102271)

[1.3. Estructura del trabajo 3](#_Toc171102272)

[2. Contexto y estado del arte 4](#_Toc171102273)

[2.1. Contexto del problema 4](#_Toc171102274)

[2.2. Reporte de Riqueza global – Banco de Crédito Suizo 6](#_Toc171102275)

[2.3 Herramientas y técnicas 10](#_Toc171102276)

[2.3.1. Deep Learning 11](#_Toc171102278)

[2.3.2. Redes Neuronales 11](#_Toc171102279)

[2.3.3. Regresión 12](#_Toc171102280)

[2.4. Visualizaciones Actuales 13](#_Toc171102281)

[2.4.1. DANE 13](#_Toc171102282)

[2.4.2. Banco Mundial 15](#_Toc171102283)

[2.4.3. Credit Suisse Bank 16](#_Toc171102284)

[2.5. Estudios Actuales Sobre la Pobreza en Colombia 17](#_Toc171102285)

[2.5.1. Pobreza Monetaria y Multidimensional en Colombia 17](#_Toc171102286)

[2.5.2. Medición de pobreza en Colombia: Cruzando las medidas unidimensionales 18](#_Toc171102287)

[2.5.3. Percepción de la pobreza en Colombia entre los años 2003 y 2016 18](#_Toc171102288)

[2.6. Conclusiones 20](#_Toc171102289)

[3. Objetivos concretos y metodología de trabajo 21](#_Toc171102290)

[3.1. Objetivo general 21](#_Toc171102291)

[3.2. Objetivos específicos 22](#_Toc171102292)

[3.3. Metodología del trabajo 22](#_Toc171102293)

[3.4. Protección de datos 24](#_Toc171102294)

[4. Desarrollo específico de la contribución 24](#_Toc171102295)

[4.1. Entendimiento del negocio 24](#_Toc171102296)

[4.2. Entendimiento de los datos 26](#_Toc171102297)

[4.3. Preparación de los datos 29](#_Toc171102298)

[4.4. Modelado, Evaluación y Despliegue 33](#_Toc171102299)

[4.4.1. Python 33](#_Toc171102300)

[4.4.2. PowerBI 34](#_Toc171102301)

[4.4.3. Anaconda 34](#_Toc171102302)

[4.4.4. MongoDB 35](#_Toc171102303)

[4.4.5. Selección e implementación de Modelos 35](#_Toc171102304)

[4.4.6. Modelo de red Neuronal con Función de perdida *Sparse Categorical Crossentropy* 36](#_Toc171102305)

[4.4.7. Modelo de clasificación *DesicionTree* 40](#_Toc171102306)

[4.4.8. Modelo de clasificación *Randomforest* 42](#_Toc171102307)

[4.4.9. Modelo de regresión basado en redes neuronales 44](#_Toc171102308)

[4.4.10. Modelo del Dashboard – Power BI. 48](#_Toc171102309)

[5. Conclusiones 58](#_Toc171102310)

[6. Limitaciones y prospectiva 59](#_Toc171102311)

[6.1. Limitaciones 59](#_Toc171102312)

[6.2. Trabajo Futuro 60](#_Toc171102313)

[Referencias bibliográficas 62](#_Toc171102314)

[Anexo A. Código Python 64](#_Toc171102315)

Índice de figuras

[Figura 1: Opciones de visualización DANE (Fuente: DANE, 2024). 14](#_Toc171102316)

[Figura 2: Reporte pobreza multidimensional (Fuente: DANE, 2024). 14](#_Toc171102317)

[Figura 3: Pobreza extrema en Colombia (Fuente: Banco Mundial, 2024). 16](#_Toc171102318)

[Figura 4: Reporte de riqueza (Fuente: Credit Suisse Bank, 2024). 17](#_Toc171102319)

[Figura 5: Ciclo de vida del data mining (Fuente: IBM, 2024). 23](#_Toc171102320)

[Figura 6: Ejemplo de la fuente de datos. (Fuente: Credit Suisse, 2022). 30](#_Toc171102321)

[Figura 7: Tipo inicial de los datos usados en los modelos. 37](#_Toc171102322)

[Figura 8: Transformación de los datos para el modelo. 37](#_Toc171102323)

[Figura 9: Tipo final de los datos usados en el modelo de red neuronal. 38](#_Toc171102324)

[Figura 10: Resumen estadístico de los datos usados en el modelo de red neuronal. 39](#_Toc171102325)

[Figura 11: Vistas que componen el dashboard propuesto. 49](#_Toc171102326)

[Figura 12: Vista 1 Portada Dashboard. 50](#_Toc171102327)

[Figura 13: Vista 1 Portada Dashboard con filtro. 50](#_Toc171102328)

[Figura 14: Vista 1 Portada Dashboard con tooltip. 51](#_Toc171102329)

[Figura 15: Vista 2 Vista\_País Dashboard. 51](#_Toc171102330)

[Figura 16: Vista 3 Resumen\_Países Dashboard sin filtrar. 53](#_Toc171102331)

[Figura 17: Vista 3 Resumen\_Países Dashboard con filtro. 54](#_Toc171102332)

[Figura 18: Vista 4 Comparativa\_Países Dashboard con filtro. 55](#_Toc171102333)

[Figura 19: Vista 5 Vista\_Regiones Dashboard. 56](#_Toc171102334)

[Figura 20: Vista 6 Relaciones Dashboard. 57](#_Toc171102335)

[Figura 21: Vista 7 Acerca de Dashboard. 58](#_Toc171102336)

Índice de tablas

[Tabla 1: Insumo HBS para el reporte global de riqueza (Fuente: Credit Suisse, 2022). 7](#_Toc171102337)

[Tabla 2: Insumo distribución de la riqueza para el reporte global de riqueza (Fuente: Credit Suisse, 2023). 8](#_Toc171102338)

[Tabla 3: Resultado de riqueza promedio per cápita por país (Fuente: Credit Suisse, 2022). 9](#_Toc171102339)

[Tabla 4: Evolución histórica de la riqueza per cápita (Fuente: Credit Suisse, 2022). 10](#_Toc171102340)

[Tabla 5: Distribución de la población según la riqueza (Fuente: Credit Suisse, 2022). 10](#_Toc171102341)

[Tabla 6: Extracción del reporte de riqueza del Credit Suisse. 26](#_Toc171102342)

[Tabla 7: Resumen de variables del data set. 29](#_Toc171102343)

[Tabla 8: Insumo inicial para el data set. 31](#_Toc171102344)

[Tabla 9: Muestra tomada del data set final. 33](#_Toc171102345)

[Tabla 10: Matriz de confusión resultante DecisionTree. 41](#_Toc171102346)

[Tabla 11: Reporte de clasificación DecisionTree. 41](#_Toc171102347)

[Tabla 12: Matriz de confusión resultante Randomforest. 42](#_Toc171102348)

[Tabla 13: Reporte de clasificación Randomforest. 43](#_Toc171102349)

[Tabla 14: Resumen estadístico de los datos usados en el modelo de regresión. 44](#_Toc171102350)

[Tabla 15: Estructura de la red neuronal para regresión. 46](#_Toc171102351)

# Introducción

A lo largo de los países, tanto los denominados sub-desarrollados, como los que se encuentran en vías de desarrollo e incluso en los mismos países desarrollados, la reducción de la pobreza ha pasado de ser una propuesta de campaña electoral a ser una política económica estructural, hasta el punto que en el año 2.000, los gobiernos firmaron una serie de ocho puntos, denominados “*Objetivos del milenio (ODM)”*, en donde en el primer apartado podemos encontrar el apartado sobre erradicar la pobreza extrema y el hambre (Organización Mundial de la Salud, 2018).

Posteriormente, en el año 2.015, las Naciones Unidas, transformaron los ODM, para convertirlos en los *“Objetivos del Desarrollo Sostenible (ODS)”*, cuyo primer punto resalta nuevamente el tema la pobreza indicando que se debe buscar su fin (Programa de las Naciones Unidas para el Desarrollo, 2024).

Desafortunadamente, aunque los gobiernos han hecho esfuerzos para cumplir estos objetivos y reducir sus niveles de pobreza, el Banco Mundial estima que casi el 9% de la población vivía bajo la línea de pobreza extrema (USD $2,15 por día) y casi el 50% no vive con más de USD$6,85 por día (Banco Mundial, 2013).

A partir de este panorama global, que se mueve entre lo esperanzador y lo insuficiente, evidentemente Colombia no es ajena al concepto de pobreza, ya que según el Departamento Nacional de Estadística en Colombia (DANE), para el año 2.022 el 36,6% de la población vive bajo la línea de pobreza.

Para dicho año, la línea de pobreza monetaria per cápita en Colombia se situó en COP$396.864 (USD$101 aprox.); es decir, que un hogar de 4 personas que reciba ingresos de igual o menos de COP$1’587.456 (USD$407 aprox.) se considera pobre. Si por otro lado evidenciamos que el salario mínimo para dicho año fue de COP$1’117.172 (USD$286 aprox.) y el 41% de los asalariados del país de un total de veintidós millones de personas reciben más de este salario mínimo por sus actividades laborales (DANE, 2024), podemos empezar a preguntarnos si esa línea de pobreza va de acuerdo con la realidad del país.

Dado que se puede evidenciar que gran parte de la población colombiana vive muy cerca de los límites de la pobreza, lo que se busca con este trabajo es identificar los factores que ayuden a la población colombiana a aumentar sus niveles de riqueza per cápita, para que ya no solamente se considere pobre, sino que logre mejorar su calidad de vida y fortalecer sus perspectivas futuras.

Dicha riqueza per cápita, según el Banco de Crédito Suizo (*Credit Suisse*), estimaba que para el año 2.021 en Colombia se encontraba en USD$17.079, lo que significa cerca de 34 veces menos que la riqueza del estadounidense promedio (*Credit Suisse*, 2022).

Según estos escenarios, los datos obtenidos por dicho banco son una fuente muy valiosa para evaluar las políticas económicas nacionales, dado que reflejan si están favoreciendo el mejoramiento de los niveles económicos de su población más allá de solo poder satisfacer unas necesidades básicas.

Por esto, finalmente se propone no solo un estudio que resalte aquellas políticas que favorezcan el crecimiento de la riqueza, sino desarrollar una herramienta visual e interactiva que permita poder ver estos datos para poder apalancar la toma de decisiones.

## Motivación

Los recursos de los países son un bien finito y entre ellos se encuentra el tiempo. Los esfuerzos que realizan los gobiernos, cuando tienen políticas económicas definidas o al menos enfocadas en el bienestar de la población no siempre dan resultados en el corto plazo. Es por esto por lo que los esfuerzos se deben enfocar en las políticas que aseguren resultados estructurales.

En los entornos económicos, la pobreza ha despertado tanto interés, que esto se evidencia en la cantidad de variaciones y metodologías que existen para medir su evolución. Dado que no existe un criterio unificado, hay discusiones entre académicos y gobiernos acerca de aquellas dimensiones en las que una persona debe estar privada para ser considerada pobre (Jhon Fredy Ariza et al., 2020).

A partir de estos escenarios, los gobiernos han llegado a fijar sus propios criterios, por lo que se genera una ambigüedad en las definiciones, con lo cual es posible decir que una persona pobre en Colombia no es igual a una persona pobre en México. Pero no existe otra región en el mundo comparada con Latinoamérica en que se aplique un método multidimensional integrado, ya que, en los Estados Unidos, Unión Europea e incluso para la Organización para la Cooperación y Desarrollo Económico (OCDE), sigue siendo solo privación de ingresos (Julio Boltvinic, 2014).

Desafortunadamente, aunque estas mediciones son valiosas, no aseguran que las personas que no caen dentro de estos indicadores pero que estén más cerca de la línea de pobreza, necesariamente tengan satisfechas sus necesidades básicas.

A partir de esto surge la propuesta plasmada en este estudio, no sobre cómo reducir el indicador de pobreza en Colombia a partir de una medición de ingresos, sino cómo impulsar la riqueza per cápita en este país.

## Planteamiento del trabajo

De la realidad encontrada en las mediciones de la pobreza en Colombia, en donde se identifica quién vive bajo las líneas de pobreza monetaria, se puede concluir que, aunque provee información valiosa sobre la efectividad de las políticas económicas implementadas en el país, no refleja en muchos casos la realidad de aquellas personas que no son consideradas pobres, pero se encuentran muy cerca de dicha línea y más cuando hay entornos de alta inflación y desaceleración económica.

El siguiente estudio, busca identificar los principales factores que han impulsado que la riqueza promedio mundial prácticamente se haya duplicado en los últimos 10 años con los que se cuenta información.

Adicionalmente, se propone un *dashboard* que permita la fácil manipulación de los datos relacionados con la evolución de los niveles de riqueza per cápita con el fin de hacerlos accesibles de manera interactiva más allá del documento del reporte anual del Banco de Crédito Suizo generado por el mismo banco.

## Estructura del trabajo

A continuación, se presentan las secciones en las cuales se divide el estudio realizado en este trabajo:

*Capítulo 1 - Introducción:* Es la primera sección del documento, en la cual se incluye un resumen de los temas que se van a tratar en el estudio, así como una breve descripción del problema actual y la importancia de una solución.

*Capítulo 2 - Contexto y estado del Arte:* En este capítulo, se incluye el contexto y estado del arte relacionado con las mediciones de pobreza existentes. Adicionalmente se describe brevemente los estudios que se han realizado, las herramientas que se van a utilizar y las visualizaciones disponibles hoy en día.

*Capítulo 3 – Objetivos y metodología del trabajo*: Esta sección describe cuál es el objetivo general que se debe resolver en el estudio, así como los objetivos específicos que ayudarán a responder las necesidades del objetivo general. Adicionalmente, se incluye una breve descripción de la metodología a seguir para el desarrollo del estudio propuesto.

*Capítulo 4 - Desarrollo específico de la contribución:* Esta sección recopila todo el desarrollo de las fases definidas en la metodología de trabajo. Se puede encontrar detalladamente el despliegue de la solución propuesta con todas sus actividades desde la identificación de la necesidad, el procesamiento de los datos hasta la fase final de despliegue tanto de los modelos como de la solución visual propuesta.

*Capítulo 5 - Conclusiones:* En esta parte del trabajo, se puede encontrar las deducciones e inferencias que fueron posibles realizar a partir de la elaboración de este trabajo. De igual manera se pueden encontrar los resultados obtenidos en cada fase y la respuesta al cumplimiento o no de los objetivos propuestos.

*Capítulo 7 - Limitaciones y prospectiva: Esta sección corresponde a la parte final de este trabajo. Dicha sección incluye un resumen de aquellas limitantes que se encontraron a lo largo del desarrollo del trabajo y que pudieron influir de alguna manera en la solución propuesta.*

*Adicionalmente se describe cuáles podrían ser los estudios futuros que se podrían realizar y cuyo punto de partida es este trabajo con el fin de complementar los hallazgos iniciales.*

# Contexto y estado del arte

En el presente capítulo, se busca presentar las definiciones que servirán para llevar a cabo el análisis propuesto, junto con un contexto acerca de la realidad de la pobreza en Colombia para finalmente complementar con estudios actuales sobre esta realidad.

## Contexto del problema

Colombia es un país de cerca de cincuenta y un millones de habitantes, ubicado en la región caribe de Suramérica. Al igual que prácticamente el resto de los demás países de Latinoamérica, en las últimas dos décadas ha venido presentando un crecimiento económico sostenido que le ha permitido pasar de un PIB en el año 2.001 de USD$98.000 millones a USD$343.000 millones en el año 2.022 (Banco Mundial, 2024).

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfica 1: Evolución histórica del PIB de Colombia (Fuente: Banco Mundial, 2024).

Este crecimiento le ha permitido posicionarse como la cuarta economía de la región en términos nominales del PIB estando solamente por detrás de Brasil, México y Argentina.

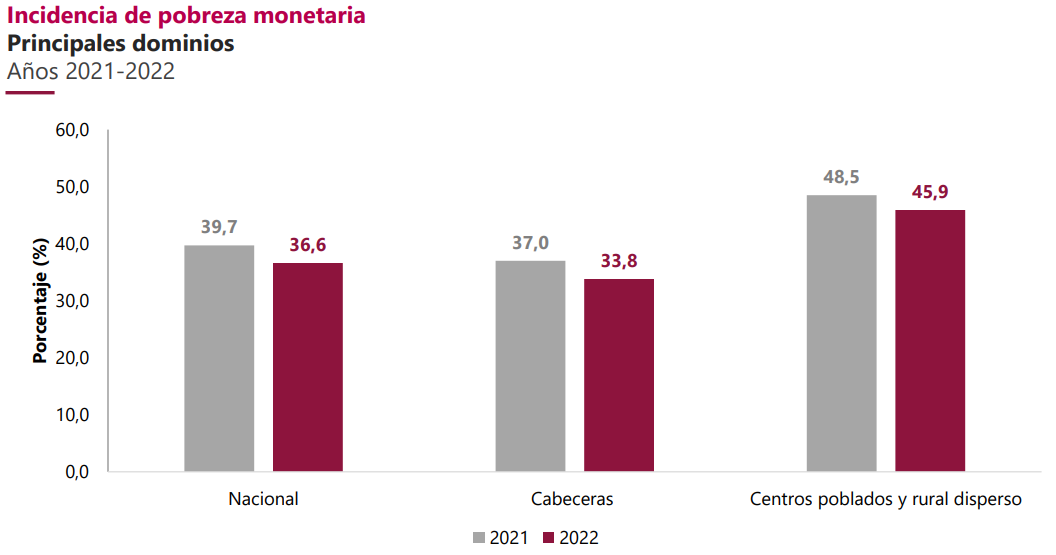
Adicionalmente, dicho crecimiento ha favorecido la reducción de ciertos indicadores como el de la pobreza que hace 20 años se encontraba en cerca del 53% y para el año 2.019 se encontraba en el 32% y que con la pandemia del Covid-19 tuvo un efecto rebote para ubicarse en el año 2.022 en el 36,6% (DANE, 2024).

Dicha pobreza se mide principalmente identificando al grupo poblacional que se encuentra bajo la línea de pobreza, que en términos de la Comisión Económica para América Latina (CEPAL), la línea de pobreza es un valor monetario que en resumen considera el costo de adquirir una canasta básica de alimentos de dos mil cien calorías por día y el costo de los demás bienes y servicios.

Según el Departamento Nacional de Estadística en Colombia (DANE), la actualización de la metodología del cálculo de la línea de pobreza siguió el método propuesto por Martin Ravallion (1998).

Luego de fijar dicho método, con algunas variaciones especiales sugeridas por la CEPAL, para el año 2.022, se fijó en Colombia que la línea de pobreza extrema que es cuando una persona tiene los medios para garantizar una canasta básica (ONG Acción contra el hambre, 2024) era de COP$198.698 per cápita y la línea de pobreza monetaria per cápita quedó fijada en COP$396.864.

A partir de estos valores, la población que cumplía las condiciones para considerarse pobre, a nivel nacional era del 36,6%, siendo la situación más crítica en las zonas rurales, ya que presenta una incidencia 12,1 puntos porcentuales más alta que en la ciudades o cabeceras municipales (DANE, 2023).



Gráfica 2: Variación de la pobreza monetaria en Colombia (Fuente: DANE, 2023).

## Reporte de Riqueza global – Banco de Crédito Suizo

En los últimos 13 años, el Banco de Crédito Suizo (*Credit Suisse*), ha sido una referencia global para la riqueza global de los hogares.

Sobre el reporte del año 2.022, que abarca doscientos diecisiete países y territorios autogobernados, se construyó siguiendo principalmente tres pasos:

1. Establecer el nivel de riqueza promedio para cada país. Para esto se tiene datos de cincuenta y un países que generan un reporte HBS (*Household Balance Sheet*) que es similar a un balance contable corporativo pero aplicado a un hogar o a una persona en un tiempo determinado.

Para ciento dieciocho países es necesario usar técnicas de econometría tales como una regresión. Finalmente, para los demás países, que solo representan el 2.5% del total de la población adulta mundial, la estimación se basa en que su ratio riqueza/PIB es el mismo al de la sub-región a la que pertenecen (*Credit Suisse*, 2022).



Tabla 1: Insumo HBS para el reporte global de riqueza (Fuente: Credit Suisse, 2022).

1. El segundo paso busca construir un patrón de distribución de la riqueza y la distribución de los ingresos. Hay información directa de cuarenta países, mientras que para ciento cuarenta países se extrapola su ratio riqueza/PIB, ya que se tiene información de la distribución del ingreso, pero no de la propiedad.

Finalmente, para los países restantes sobre los que no se tienen tantos datos, se utiliza la metodología de las sub-regiones (*Credit Suisse*, 2022).



Tabla 2: Insumo distribución de la riqueza para el reporte global de riqueza (Fuente: Credit Suisse, 2023).

1. El tercer paso busca corregir la ineficiencia de los datos sobre distribución de la riqueza. Para esto, se usa la lista de FORBES de billonarios, para ajustar los patrones de distribución en los rangos de riquezas altas (*Credit Suisse*, 2022).

Finalmente, a partir de reporte, se estima que, para este año, la riqueza mundial sea de cerca de 455 billones de dólares; para hacernos una idea, es cerca de 16 veces el PIB de Estados Unidos que es la mayor economía mundial (*Credit Suisse*, 2022).

Del informe del año 2.021, el país cuya riqueza per cápita es mayor es Suiza. En general su población posee una riqueza promedio de USD$696.604.

La mayor potencia mundial, Estados Unidos, se encuentra en tercer lugar con una riqueza promedio USD$579.051. Pero es importante aclarar que este país posee el 31,45% de toda la riqueza mundial.

La segunda potencia mundial, China, tiene una riqueza promedio de USD$76.639 y tiene el 18,36% de la riqueza mundial. Si se suma la riqueza de Estados Unidos y la de China, se llega a casi la mitad de la riqueza mundial.

Colombia, tiene una riqueza promedio de USD$17.079. Si ponemos en perspectiva, un suizo tiene cerca de 40 veces más riqueza que un colombiano.

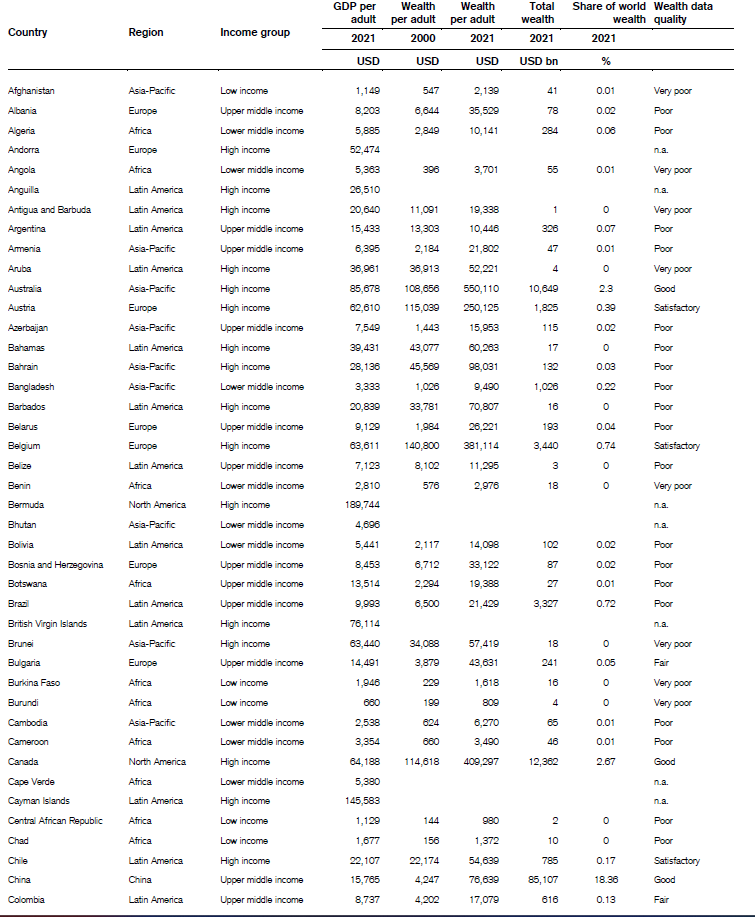


Tabla 3: Resultado de riqueza promedio per cápita por país (Fuente: Credit Suisse, 2022).

A pesar de esta diferencia, la riqueza promedio de un colombiano ha pasado de USD$4.202 en el año 2.000 a los USD$17.079 en el año 2.021.

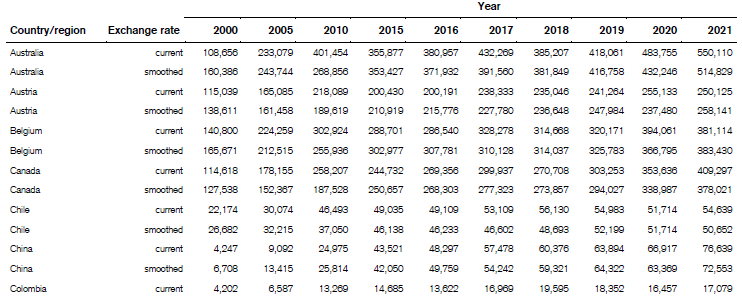


Tabla 4: Evolución histórica de la riqueza per cápita (Fuente: Credit Suisse, 2022).

Al desagregar la población en Colombia, de un total de treinta y seis millones de adultos, el 73% de la población, es decir 26,6 millones de adultos no tiene una riqueza mayor a los USD$10.000, mientras que hay treinta y seis mil personas con una riqueza de más USD$1’000.000.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Tabla 5: Distribución de la población según la riqueza (Fuente: Credit Suisse, 2022).

## Herramientas y técnicas

Para el desarrollo del trabajo, fue necesaria la utilización de varias herramientas y técnicas de inteligencia artificial, así como de programación y estadística las cuales son explicadas a continuación.



### Deep Learning

La técnica *Deep Learning* abarca una serie de algoritmos de aprendizaje automático que mediante el uso de múltiples capas extrae características de un nivel superior al de la entrada bruta.

Dichas capas pretenden emular el comportamiento del cerebro humano, sin llegar a igualarlo, pero sí con una evidente capacidad de aprender a partir de gran cantidad de datos.

A diferencia de la técnica *Machine Learning*, el *Deep Learning* es capaz de trabajar con datos no estructurados como lo son imágenes con el fin de extraer patrones. Esta característica permite eliminar la dependencia humana que se requiere en *Machine Learning* para clasificar y catalogar los datos con el fin de que sean utilizables para los modelos.

Los algoritmos de *Deep Learning*, trabajan sobre varias capas de nodos que están conectados sobre la capa anterior para refinar y corregir las predicciones. Dicho avance entre capas se denomina propagación hacia adelante, la cual difiere de la propagación inversa (*Backpropagation*), que consiste en algoritmos que se mueven hacia “atrás” en las capas con el fin de corregir errores y ajustar los sesgos. Estos dos comportamientos permiten que los modelos sean cada vez más precisos.

Hoy en día las aplicaciones de *Deep Learning* van desde la identificación de fraudes a partir del estudio de movimientos transaccionales para identificar comportamientos sospechosos, así como la evaluación de riesgos de mercado, hasta las aplicaciones relacionadas con servicio al cliente como los *chatbots,* que soncapaces de usar lenguaje natural para responder a las necesidades del usuario (IBM, 2024).

### Redes Neuronales

Las redes neuronales están basadas en el Perceptrón de Rosenblatt (1958), que está inspirado en las neuronas biológicas humanas y representan la unidad básica de las redes neuronales, al funcionar como una neurona artificial.

Se pueden clasificar como un subconjunto del *Machine Learning* y la base para el *Deep Learning,* ya que, a nivel práctico, una red neuronal de más de tres capas es un algoritmo de *Deep Learning*.

Los algoritmos basados en redes neuronales están inspirados en los medios de comunicación de las neuronas, que han demostrado dar buenos resultados cuando el conocimiento es impreciso o variable.

Adicionalmente, se pueden considerar algoritmos adaptativos, los cuales son aquellos capaces de modificar su conducta durante su ejecución, basados no solo en el entorno sino en la información disponible. Para el caso de las redes neuronales, estas son capaces de ajustar los pesos de los nodos en el método de propagación inversa o por ejemplo cuando el modelo está entrenado, ser capaz de recibir datos que no han sido ingresados nunca en el modelo y poder adaptarse a dichos datos.

Son capaces de realizar tareas de aprendizaje que incluso pueden resultar imposibles para el cerebro humano, como por ejemplo clasificar una gran cantidad de imágenes en un tiempo menor en el que lo haría una persona.

Finalmente, toda red neuronal consta de una capa de entrada, varias capas ocultas y una de salida. El flujo a través de la red se da a partir de los umbrales de cada capa. Es decir que si la salida de una capa es mayor al valor del umbral de dicho nodo este se activa y envía los datos a la siguiente capa.

### Regresión

Las regresiones son un modelo de predicción estadístico sobre los valores de cierta variable *Y* la cual actúa como variable dependiente o explicada a partir de los valores de una variable *X* que es la variable independiente o explicativa entre las que se asume una relación.

Para interpretar la fortaleza de la regresión se puede revisar el valor de r, que puede tomar un rango entre -1 y 1, en donde un valor 1 representa una correlación perfecta.

Las regresiones son una herramienta básica para hacer predicciones de variables sobre las cuales no conocemos del todo su comportamiento, pero de las que sí disponemos de cierto conocimiento previo.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfica 3: Comportamiento del valor r en una regresión (Fuente: Bioestadística: Métodos y aplicaciones, p. 82).

## Visualizaciones Actuales

Existen muchas fuentes de datos con información relacionada a la pobreza, pero visualizaciones interactivas y que adicionalmente estén disponibles para una diversa gama de usuarios no son tan diversas. A continuación se describe brevemente algunas de las fuentes halladas con información relacionada y que ofrecen alguna visualización.

### DANE

El Departamento Nacional de Estadística (DANE), es una fuente oficial de datos en Colombia, la cual posee una gran variedad de análisis estadísticos de distinta índole tales como económicos, demográficos, sociales entre otros.

Pero en cuanto a sus opciones de visualizar dichos estudios, lo que uno puede encontrar principalmente son reportes publicados en su página web y que se encuentran en archivos planos tales como documentos PDF o algunos datos agrupados en archivos XLS para que los usuarios hagan sus propios análisis.



Figura 1: Opciones de visualización DANE (Fuente: DANE, 2024).

En la Figura 2 mostrada a continuación, se puede observar en la parte izquierda de la imagen, el comportamiento del índice de pobreza multidimensional en los últimos cuatro años en las principales siete regiones de Colombia. Claramente se puede observar que desde el año 2.020 se mantiene una tendencia a la baja.

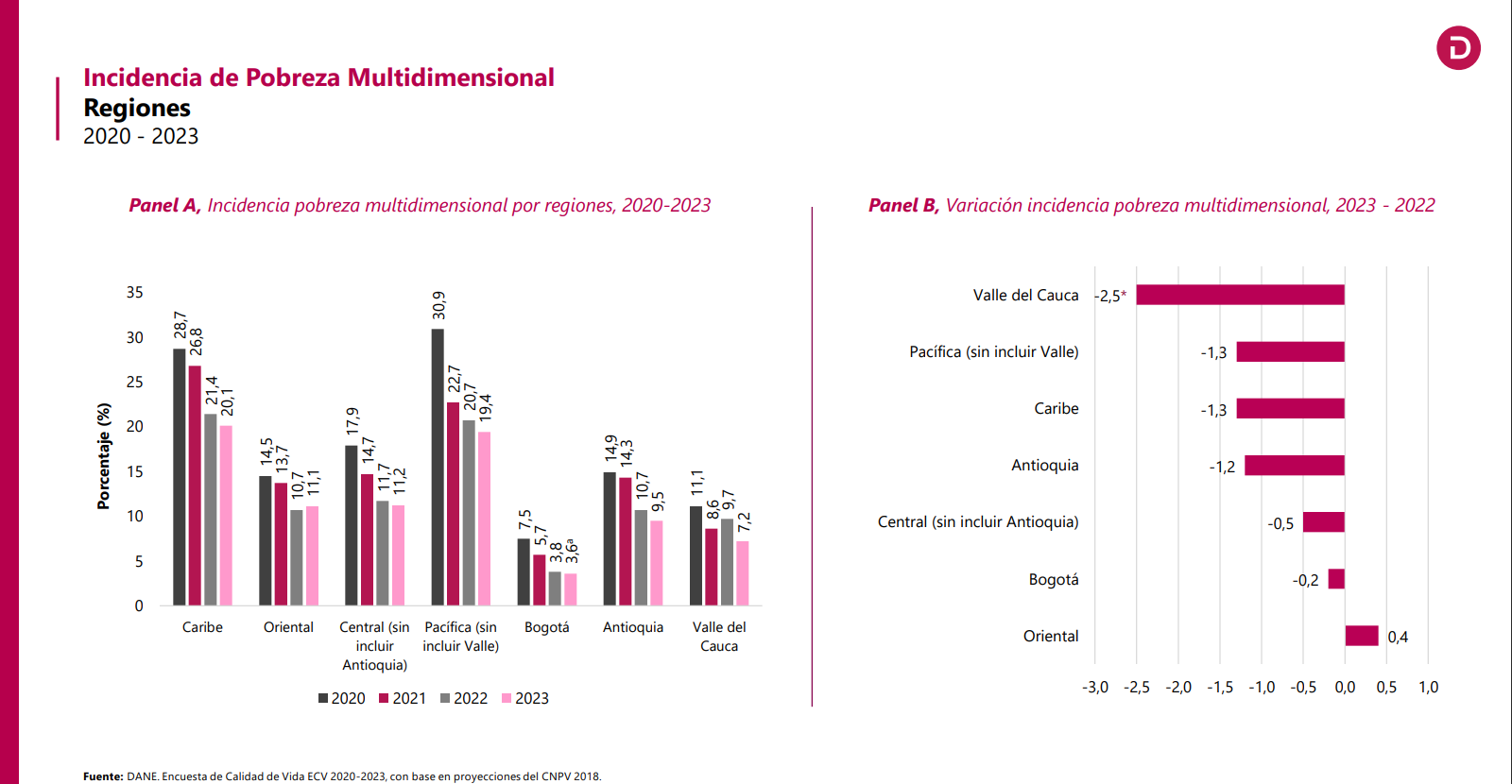


Figura 2: Reporte pobreza multidimensional (Fuente: DANE, 2024).

Por otro lado, en la parte derecha, se encuentra el porcentaje de variación de dicho índice entre los años 2.022 y 2.023 para cada una de las regiones, en donde se observa una disminución excepto en la región oriental que corresponde a la zona de los llanos hacía la frontera sur con la República Bolivariana de Venezuela.

Esta visualización, a pesar de que puede contener información relevante respecto a los logros obtenidos frente a la reducción de este indicador, no permite ninguna interacción del usuario más allá de lo que este puede encontrar en el archivo PDF y por esto la extracción de información y análisis por parte del lector se puede quedar en una capa muy superficial de interpretación.

Muchas opciones que ofrece el DANE, respecto a reportes de interés nacional, se pueden encontrar para descargar en opciones de archivos PDF como se puede ver en la Figura 1. Desafortunadamente mucha de la información aquí contenida no llega a ser leída dado que no es capaz de involucrar del todo al lector.

### Banco Mundial

El Banco mundial es otra gran fuente de datos para realizar análisis. Al igual que el DANE, posee una gran variedad de datos demográficos, sociales, etc. pero extrapolados a todos los países.

A diferencia del DANE, en la página web del banco mundial se pueden encontrar gráficas que permiten cierta interactividad, aunque limitada solo al aspecto de interés.

Se puede evidenciar directamente en la página que dichas gráficas están embebidas en lenguaje HTML.

Adicionalmente, el Banco Mundial también permite la descarga de los datos en formatos que permita a los usuarios hacer sus propios análisis.

En la Figura 3, se puede encontrar una extracción de la interfaz que ofrece la página web del banco mundial respecto al comportamiento de la pobreza en Colombia a lo largo de una línea histórica. A pesar de que ya nos encontramos en un plano digital más interactivo versus el explicado en el numeral anterior y que adicionalmente ya le permite al lector algún grado mayor de interacción, el usuario no puede relacionar la gráfica con alguna otra que complemente la información observada y que permita un mayor nivel de análisis con el fin de obtener mejores conclusiones.

El Banco Mundial ofrece muchas gráficas con información relevante acerca de los países, pero no permite al lector conectar los datos contenidos en cada gráfica ya que cada una funciona como una isla. Lo más que el usuario puede llegar a configurar es la comparación entre países de un mismo indicador.

Interfaz de usuario gráfica, Gráfico, Aplicación, Gráfico de líneas

Descripción generada automáticamente

Figura 3: Pobreza extrema en Colombia (Fuente: Banco Mundial, 2024).

### Credit Suisse Bank

Dicho banco, en su estudio anual de riqueza, provee en gran medida los datos que se utilizaron en este estudio. Pero a nivel de visualización, el banco solo genera un reporte anual en formato PDF y que se puede descargar directamente desde su página web.

Este reporte no permite ninguna interactividad de los usuarios y de hecho los datos utilizados tuvieron que ser extraídos del mismo archivo PDF.

A continuación, en la Figura 3, se encuentra una extracción del reporte generado por el *Credit Suisse*, que al igual que el mostrado en el numeral 2.4.1 se resume a un documento en PDF bastante completo en el análisis de la riqueza de la población en cada país, pero que no ofrece ningún grado de interacción, por lo que la capacidad de análisis e interpretación se puede ver reducida, ya que guía al usuario en todo momento a través de la lectura y esto impide que el usuario encuentre sus propias conclusiones adicionales a las que ofrece el documento.

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Figura 4: Reporte de riqueza (Fuente: Credit Suisse Bank, 2024).

## Estudios Actuales Sobre la Pobreza en Colombia

### Pobreza Monetaria y Multidimensional en Colombia

Este estudio publicado en el año 2.020 en el libro “Pobreza Monetaria y Multidimensional en Colombia” (Jhon Fredy Ariza et al., 2020) analiza variables que influyen en la pobreza colombiana desde una perspectiva de pobreza multidimensional.

Se analizaron variables a nivel departamental (división administrativa de la República de Colombia).

Las conclusiones más relevantes de este estudio se presentan a continuación, en donde el primer numeral indica un mayor impacto en la pobreza que el segundo y así sucesivamente.

* Un índice de Gini alto incrementa el índice de pobreza multidimensional.
* Pertenecer a una minoría como indígena, raizal, afro-descendiente incrementa la pobreza multidimensional.
* Ser mujer y jefe de hogar incrementa el índice de pobreza multidimensional
* Haber sido despojado de sus tierras incrementa la pobreza multidimensional.
* Haber tenido miembros de la familia desplazados por el conflicto armado incrementa la pobreza multidimensional.
* Vivir lejos de una capital departamental incrementa la pobreza multidimensional.

Como comentario final, se puede indicar que la pobreza se puede seguir reduciendo en la medida que el país mantenga una senda de crecimiento, aunque la desigualdad no disminuya.

### Medición de pobreza en Colombia: Cruzando las medidas unidimensionales

Nuevamente encontramos enfoques que indican que una medición multidimensional, genera una mejor perspectiva que una medición unidimensional, ya que permite evaluar factores más allá de una privación en los ingresos (Henry Laverde et al., 2016).

Para determinar la pobreza multidimensional, en este estudio se utilizan dos enfoques de Unión e Intersección (Tsui, 2002; Bourguignon et al., 2003), en donde si un individuo es privado de una de las dimensiones, se considera pobre, y entre mayor sea el número de dimensiones en que un individuo es privado, mayor es su probabilidad de ser pobre.

Los resultados principales de este estudio son:

* Existe una diferencia de cerca de 48 puntos porcentuales entre las mediciones de pobreza multidimensional en comparación con una unidimensional.
* La región Caribe colombiana presenta los peores resultados en todos los indicadores.
* Los principales determinantes de la pobreza se encuentran relacionados con el nivel de educación, acceso a las TIC, condiciones laborales y finalmente los ingresos. En donde estos últimos no explican completamente el fenómeno de la pobreza.

### Percepción de la pobreza en Colombia entre los años 2003 y 2016

Este estudio profundiza en la pobreza subjetiva. Es importante mencionar que la medición objetiva de los indicadores de pobreza presenta una correlación con las mediciones subjetivas de la pobreza (Diana Niño et al., 2018).

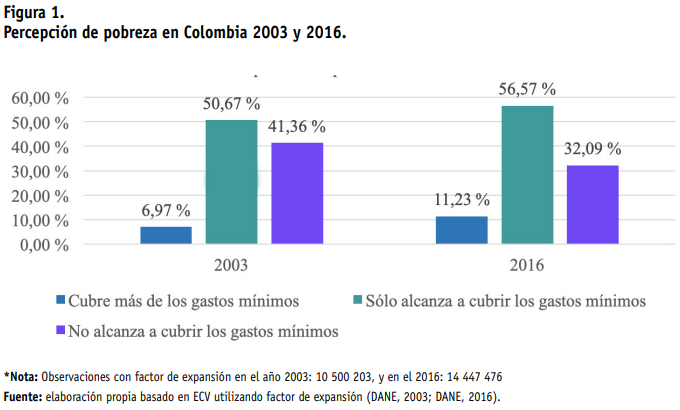
En la medición subjetiva, se incluyen mediciones de autopercepción acerca de la capacidad que tienen las personas para cubrir sus necesidades básicas. Aunque estas mediciones pueden generar incoherencias e incompatibilidades para agregar a las personas según (Martin Ravallion, 1998).

Esto se da debido a que cada persona puede asimilar su realidad socio-económica sin importar sus ingresos debido por ejemplo a un deseo de mejoramiento continuo.

Para realizar este estudio, se analizaron los datos obtenidos por las encuestas del Departamento Nacional de Estadística (DANE) para el año 2.003 y el año 2.016.

A nivel de pobreza subjetiva, se encontró que pasó de un 59,1% en el año 2.003 a un 39,6% en el 2.016.

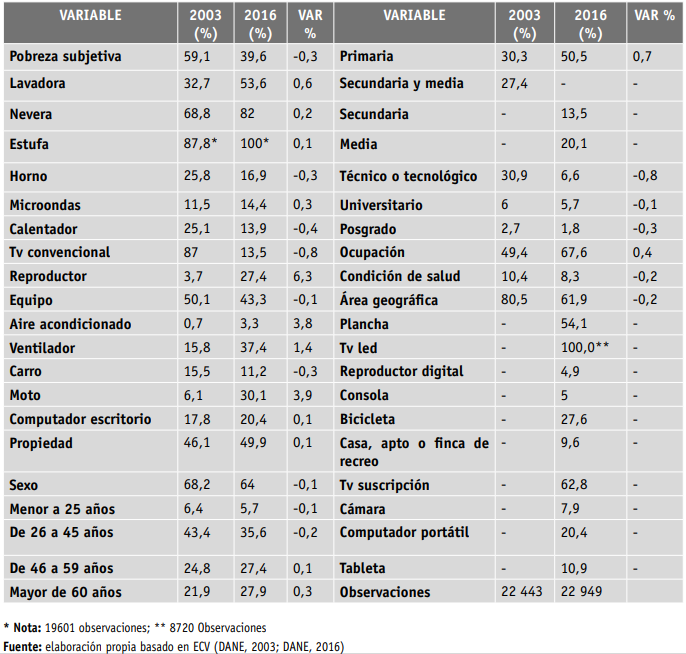
De igual manera, en el periodo del estudio, disminuyó la cantidad de jefes de hogar que consideraba que no le alcanzaban los ingresos para cubrir sus necesidades básicas.



Gráfica 4: Fuente: Niño, D. & Morera. N (2018, p.29).

Como conclusiones a este estudio, se encuentra que hay factores adicionales a sus ingresos, como la tenencia de ciertos electrodomésticos y bienes que mejoran la autopercepción de la pobreza.

Para finalizar, hay bienes intangibles que tienen peso en la percepción subjetiva de la pobreza. Por ejemplo, para el año 2.016, un factor relevante para considerarse pobre fue el solo tener educación primaria.



Gráfica 5: Fuente: Niño, D. & Morera. N (2018, p.30).

## Conclusiones

A partir de los escenarios identificados, es muy común encontrar estudios que analizan en profundidad las condiciones para determinar si una persona es pobre o no, sin importar los criterios.

Claramente se han identificado dos tipos de mediciones, unidimensional basada simplemente en una restricción en los ingresos para suplir necesidades básicas y un enfoque multidimensional que considera otras variables adicionales al ingreso para fijar un criterio de pobreza.

Adicionalmente, se concluye que el crecimiento económico es un factor determinante para reducir los indicadores tradicionales de pobreza.

No es común encontrar estudios que midan un crecimiento estructural en la riqueza de las personas, en comparación con los existentes respecto a la reducción de la pobreza.

De igual manera, aunque se puede encontrar información bastante diversa sobre los diferentes tipos de pobreza, así como los logros obtenidos al reducir los indicadores de pobreza en Colombia y en el mundo, dicha información normalmente se encuentra en archivos digitales que no ofrecen ningún grado de interacción, o se puede encontrar gráficas en sitios web que ofrecen más interacción que los archivos antes mencionados pero que en muchos casos no permiten al usuario relacionar los datos estudiados con otros que pueden complementar su ejercicio de análisis.

Debido a esta falta de visualizaciones, el usuario debe hacer un mayor esfuerzo de interpretación al tener que buscar diferentes fuentes que complementen los datos iniciales, con el fin de poder obtener conclusiones más robustas en su ejercicio de análisis.

A partir de los resultados del informe global de riqueza, se propone este nuevo enfoque que sugiere un cambio en los estudios, ya que, aunque las mediciones objetivas en la reducción de la pobreza siguen siendo una fuente valiosa para fijar políticas económicas, son mediciones que se van al límite inferior que necesita una persona para satisfacer unas necesidades básicas, por lo que en épocas de incertidumbre económica son propensos a dispararse, mientras que si se fijan políticas que garanticen un crecimiento sostenido en la riqueza de las personas, estas van a poder garantizar una estabilidad en su calidad de vida, reduciendo su grado de vulnerabilidad.

Estas políticas e indicadores deben estar enmarcados en soluciones ópticas soportadas por herramientas de visualización que sean capaces de permitir al usuario una mejor toma de decisiones y ofrezcan un panorama completo de los indicadores mostrados.

# Objetivos concretos y metodología de trabajo

En el apartado que se presenta a continuación, se describen el objetivo general y los objetivos específicos de este estudio.

## Objetivo general

El objetivo general de este trabajo es desarrollar una solución visual tipo *dashboard* que permita el análisis y el estudio de las variables que apalancan la creación de riqueza en el mundo con ayuda de modelos basados en inteligencia artificial.

## Objetivos específicos

Para la consecución del objetivo general previamente indicado, se trabajarán los siguientes objetivos específicos:

* Identificar los principales factores que han incidido en el aumento de la riqueza en la población a partir de un modelo basado en redes neuronales.
* Definir el algoritmo del modelo de regresión propuesto basado en redes neuronales.
* Construir el set de datos a partir del informe del *Credit Suisse Bank* y fuentes adicionales para alimentar el modelo.
* Evaluar el modelo para identificar las variables que mejor pronostican un aumento sostenido de la riqueza.
* Describir la situación actual de la riqueza a nivel mundial y cómo se puede comparar entre países.
* Evaluar la mejor alternativa de visualización para mostrar al lector un panorama completo entre la capa cruda de datos y los resultados del modelo y facilite el análisis de estos.

## Metodología del trabajo

La técnica de minería de datos (*data mining*) comprende el descubrimiento de conocimiento de datos o (KDD, *knowledge discovery in databases*), con el fin de buscar patrones o información valiosa a partir de grandes volúmenes de datos (IBM, 2024).

Hacia finales de la década de los 90, apareció la metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*), la cual fue concebida por tres grandes empresas (DaimlerChrysler, SPSS y NCR) en un mercado aún naciente relacionado con la minería de datos con el fin de estandarizar los procesos relacionados con este tema a lo largo de las industrias (*Data Science Process Aliance*, 2024).

Según IBM, como metodología posee descripciones de las fases básicas de un proyecto, tareas y relaciones y como proceso, ofrece un resumen del ciclo vital de la minería de datos.

Esta metodología comprende 6 fases, las cuales se pueden encontrar en la siguiente imagen.

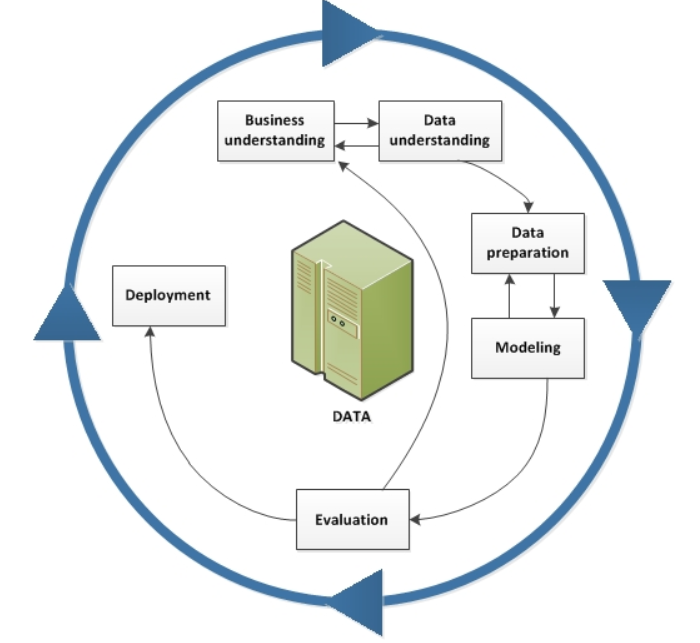


Figura 5: Ciclo de vida del data mining (Fuente: IBM, 2024).

* Fase 1 Entendimiento del negocio (*Business understanding*): Los proyectos empiezan con un entendimiento de las necesidades del cliente. Esta fase busca realizar una correcta definición de los requerimientos y objetivos del proyecto, así como los recursos disponibles.
* Fase 2 Entendimiento de los datos (*Data understanding*): Esta fase busca identificar, recolectar y analizar los conjuntos de datos necesarios para cumplir los objetivos del proyecto.

Esta fase abarca la recolección, descripción, exploración y revisión de calidad.

* Fase 3 Preparación de los datos (*Data preparation*): Se dice que el 80% de un proyecto es la preparación de los datos.

En esta fase se seleccionan los datos, se limpian, se integran y se formatean los datos.

* Fase 4 Modelado (*Modeling*): Normalmente es una fase corta. En esta sección se seleccionan las técnicas de modelado, se diseñan y testean los modelos y finalmente se comparan.
* Fase 5 Evaluación (*Evaluation*): En dicha fase, se busca identificar las mejores prácticas, así como los siguientes pasos.
* Fase 6 Despliegue (*Deployment*): Puede generarse un reporte o generar un nuevo proceso de minería de datos.

Incluye una definición de como desplegar el proyecto, una metodología de monitoreo y mantenimiento, un reporte final y finalmente la revisión del proyecto.

Una vez definida la metodología CRISP-DM, se determinó que ésta era la metodología más adecuada a seguir en la elaboración de estudio dada su sencillez y ya que fue concebida para facilitar la implementación y ejecución de proyectos relacionados con datos.

## Protección de datos

Este apartado se incluye como complemento a la metodología del trabajo, con el fin de seguir las normas generales del Reglamento General de Protección de Datos (RGPD), para no incurrir en infracciones sobre el manejo de datos personales.

En general los datos utilizados son aspectos macroeconómicos de las naciones, los cuales han sido publicados propiamente por cada uno de los entes gubernamentales o que por otro lado han sido determinados y publicados por organizaciones privadas para acceso al público.

A partir de lo anterior, en cumplimiento con el RGPD, dado que no se están usando datos personales que pueden ser considerados sensibles y se ha determinado que todos los datos utilizados en el estudio propuesto son de libre acceso y están disponibles para ser usados y consultados, se determina que no se requiere un tratamiento especial de uso de datos personales.

# Desarrollo específico de la contribución

Una vez definida la metodología de trabajo en el numeral 3.3, en esta sección se desarrolla el estudio siguiendo las seis fases de la metodología CRISP-DM comenzando con la fase del entendimiento del negocio y finalizando con el despliegue.

## Entendimiento del negocio

Como se explicó en capítulos anteriores, muchos gobiernos e instituciones realizan esfuerzos con el fin de reducir sus indicadores de pobreza, en donde Colombia no es la excepción.

Desafortunadamente, en un entorno social como los que se presentan en América Latina, en donde predominan los mercados informales, altos niveles de desempleo y deserción estudiantil sobre todo en las poblaciones menos favorecidas, difícil acceso a servicios básicos, entre muchas otras condiciones negativas pero que afortunadamente han mantenido tendencias a la baja en los últimos años, también se logra evidenciar que las reducciones en la pobreza, vienen acompañadas de lentos procesos en la mejora de la calidad de vida de las personas.

Los estudios actuales están enfocados en mediciones sobre reducción en la pobreza monetaria como en la pobreza multidimensional, que, pese a que son un buen indicador de la salud económica de los países, muchas veces camufla situaciones que no son tan evidentes y que ayudan a erosionar la calidad de vida las personas.

A partir de esto, el enfoque propuesto en este trabajo apunta a poder facilitar mediante una solución visual tipo *dashboard* la interpretación y el análisis de los resultados obtenidos en una serie de modelos que serán descritos más adelante, logrando ir más allá de reportes estáticos para poder aumentar el interés en los resultados y se alcance un mayor impacto.

Para poder alimentar el *dashboard*, se busca poder determinar mediante técnicas de *machine learning* así como el uso de redes neuronales aquellas características que influyen no en la reducción de la pobreza que se enfoca solamente en el ingreso, sino en el crecimiento de la riqueza ya que relaciona el patrimonio y la obtención de activos.

Para lograr dicha identificación, se parte de un grupo de datos que en este caso miden diversos aspectos del entorno macroeconómico de los países, así como aspectos social-demográficos y que toman el rol de variables explicativas para incluirlas en los modelos y revisar su importancia frente a la variable objetivo la cual en este caso es la riqueza per cápita obtenida del reporte anual del *Credit Suisse Bank,* y así lograr categorizar aquellos factores que son más relevantes para explicar dicha riqueza. El detalle de estas variables será explicado en la fase 2 correspondiente al entendimiento de los datos.

En general, lo que se quiere desarrollar es que, a partir de un primer modelo, sea posible identificar qué características son más determinantes en los países con niveles de riqueza más altos que por ejemplo Colombia, ya sean o no de la misma región, para poder contrastar mediante la ayuda de gráficas que ofrezcan al usuario una mayor interacción con los datos, si efectivamente es un aspecto en el que los países con menor riqueza con el fin de poder fijar mejores decisiones económicas basadas en datos y que dichas características sean impulsadas.

Posteriormente, a partir de un modelo, pero esta vez enfocado a una regresión basada también en redes neuronales, el determinar las variables que poseen una mayor correlación con el nivel de riqueza para poder llevar a cabo la construcción de un modelo que sea capaz de predecir el nivel de riqueza.

Dicho esto, a continuación, se explica la siguiente fase la cual corresponde a una revisión más a detalle de los datos que serán usados en el estudio.

## Entendimiento de los datos

Los datos se pueden considerar como el insumo principal de cualquier modelo, ya que su correcta integración, ayudará a obtener modelos más precisos y con mejores resultados.

Por esto, se procede a hacer una descripción de los datos que alimentarán los modelos propuestos y que, para practicidad del ejercicio, todos son datos de acceso público que se pueden obtener desde fuentes oficiales y que son reconocidas a nivel internacional.

En relación con esto, si nos referimos al origen de los datos utilizados en este trabajo, inicialmente hacemos énfasis en el grupo de datos que provienen del estudio de riqueza global generado por el *Credit Suisse Bank* para cada una de las naciones y regiones sujetas al estudio. De dicho reporte se toma el campo “*Wealth per adult 2021”*, que según este reporte anual es el dato más actualizado no solo para Colombia sino para la mayoría de las naciones y el cual representa a la variable objetivo para los modelos de redes neuronales que serán descritos en secciones posteriores en la fase de modelado según la metodología CRISP-DM.

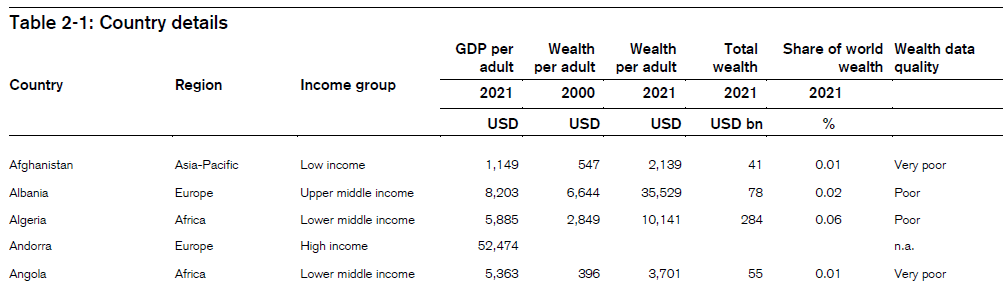


Tabla 6: Extracción del reporte de riqueza del Credit Suisse.

Adicionalmente, de esta misma fuente, se toma el campo de “*GDP per adult”* y “*Region”* para empezar a integrar aquellos datos que servirán como variables explicativas a la variable objetivo en el momento de definir los modelos.

Finalmente, de esta misma fuente de datos, se procedió a tomar de los campos “*Total wealth”* que representa la riqueza total en billones de dólares y “*Wealth per adult”* que representa la riqueza por adulto, solamente con el propósito de calcular el campo “*population”* para poder determinar el volumen de población adulta que compone cada país.

A continuación, siguiendo la línea de fijar las variables explicativas, como complemento al reporte del *Credit Suisse Bank*, se procedió a añadir nueve variables macroeconómicas de distinta índole y que provienen de diversas fuentes de acceso público, entre las que se encuentra el Banco Mundial, *trademap* yStatista.

Dichas variables se escogieron basadas no solo en su reconocimiento a nivel internacional, sino en la disponibilidad de fuentes existentes y que permiten garantizar la completitud de estas.

Por otro lado, dado que no se encontraron estudios previos similares, no hay una línea base para determinar qué variables pueden ofrecer una mejor explicación de la variable objetivo que otras. Fue necesario descartar algunos de los datos disponibles pero que puedan ser utilizados en estudios posteriores, con el fin de ampliar los estudios realizados.

Dicho esto, la definición de la estructura del conjunto de datos quedó fijada en un total de diecinueve variables que son explicadas a continuación en una matriz de resumen.

|  |  |  |
| --- | --- | --- |
| Variable | Definición | Formato |
| Country | Hace referencia al país al cual pertenecen las variables incluidas en este estudio y que fueron consideradas necesarias para la solución. | Object |
| Period | Esta variable indica el año en el cual fueron obtenidos los datos de las variables incluidas en este estudio | Object |
| Wealth\_per\_adult | Riqueza media por adulto medida en dólares americanos entre los años 2.012 y 2.021. Será la variable objetivo en los modelos propuestos. | Integer |
| GDP\_per\_adult | Es el producto interno bruto per capita por adulto. Se obtiene de dividir el Producto interno bruto (PIB) en la cantidad de población. | Integer |
| Income\_Group | Clasificación del grupo en el que se encuentran los ingresos en ese país. | Object |
| Region | Es una aproximación a la zona geográfica en donde se encuentra el país. En algunos casos puede coincidir con el continente. | Object |
| Population | Población adulta total en el país y sobre la cual fue obtenida la riqueza media. | Integer |
| Exports | Cantidad en dólares que el país recibió por conceptos de exportación en ese año. | Integer |
| Exports\_ratio | Porcentaje del Producto Interno Bruto que representan las exportaciones de ese país. | Integer |
| External\_Debt | Porcentaje de deuda per cápita que tiene la población de ese país. Se obtiene al dividir la deuda privada y la deuda pública en el total de población. | Integer |
| Main\_export | Principal producto de exportación que tiene ese país. | Integer |
| Class | Variable calculada a partir de Main\_export con el fin de agrupar en categorías los tipos de productos exportados. | Object |
| Gini | Índice Gini en ese país. Mide el grado de desigualdad en la distribución de la riqueza en un país. Un índice cercano a 1 equivale a una desigualdad total, mientras que un valor cercano a 0 corresponde a una repartición equitativa de la riqueza. | Double |
| Average\_wage | Salario medio en ese país medido en dólares americanos. | Integer |
| Income\_Group | Clasificación del nivel de los ingresos que tiene dicho país. | Object |
| Exp\_Education | Gasto en educación anual de cada país en relación con el Producto Interno Bruto. | Double |
| Democracy\_level | Índice de democracia en ese país. Un valor igual a 10 equivale a una democracia total. Valores cercanos a 0 indican dictaduras y/o regímenes totalitarios. | Double |
| Informal\_economy | Porcentaje de la economía que corresponde al sector informal, es decir no se rige bajo ningún marco legal y sobre el cual no se tiene un control bajo ninguna entidad regulatoria. | Double |
| Range | Variable calculada a partirde Wealth\_per\_adult, en donde se agrupa la riqueza por rangos. | Object |

Tabla 7: Resumen de variables del data set.

## Preparación de los datos

La fase que abarca la preparación de los datos es de las que más tiempo pueden requerir dada su importancia implícita. Esta fase busca tener los datos listos para su utilización y cualquier inconsistencia puede llevar a resultados erróneos, así como a llevar a conclusiones equivocadas de los datos.

Esta fase implica desde la tarea de obtención de los datos, su limpieza, organización y revisión de indicadores tales como completitud (busca que todos los atributos de un dato estén completos), unicidad (no debe haber registros duplicados) o precisión (los datos deben ajustarse a la realidad).

El data set requerido tanto para los modelos como para el *dashboard* no existe en la estructura propuesta en el apartado anterior. Es por esto por lo que uno de los objetivos específicos de este trabajo abarca la construcción de este.

Los datos requeridos para el ejercicio propuesto provienen de diversas fuentes y en diferentes formatos, lo cual implica un pre-procesamiento manual bastante amplio respecto a la obtención de estos, así como para la estandarización de los datos, ya que según la fuente puede por ejemplo variar el nombre del país lo cual dificulta la asociación.

En el data set final se abarca una profundidad histórica de diez años, que inicia en el año 2.012 y se extiende hasta el año 2.021 (inicialmente se cuenta con un reporte por año para un total de diez fuentes independientes) ya que este es el último reporte completo publicado por el *Credit Suisse Bank* acerca de los niveles de riqueza por adulto en cada uno de los países.

Dicho reporte es la principal fuente de información para este estudio, ya que proporciona siete de las variables requeridas, las cuales son: *Country, Period, Region, Income\_Group, Wealth\_per\_adult, Population* y *GDP\_per\_adult* (Para esta variable fue necesario complementar los datos faltantes con los que proporciona el Banco Mundial, ya que hay países en algunos de los reportes publicados sin este dato).

A partir de los diez reportes que en promedio tienen más de cien páginas, pero de las que solo se requieren aproximadamente cinco, el primer paso para la construcción del data set es la unificación de estos para luego proceder a completar las demás variables.

A continuación, se encuentra un ejemplo de como están los datos publicados. Se puede evidenciar que el manejo de estos no es sencillo a pesar de que contienen información valiosa sobre la cual poder realizar diversos tipos de análisis.

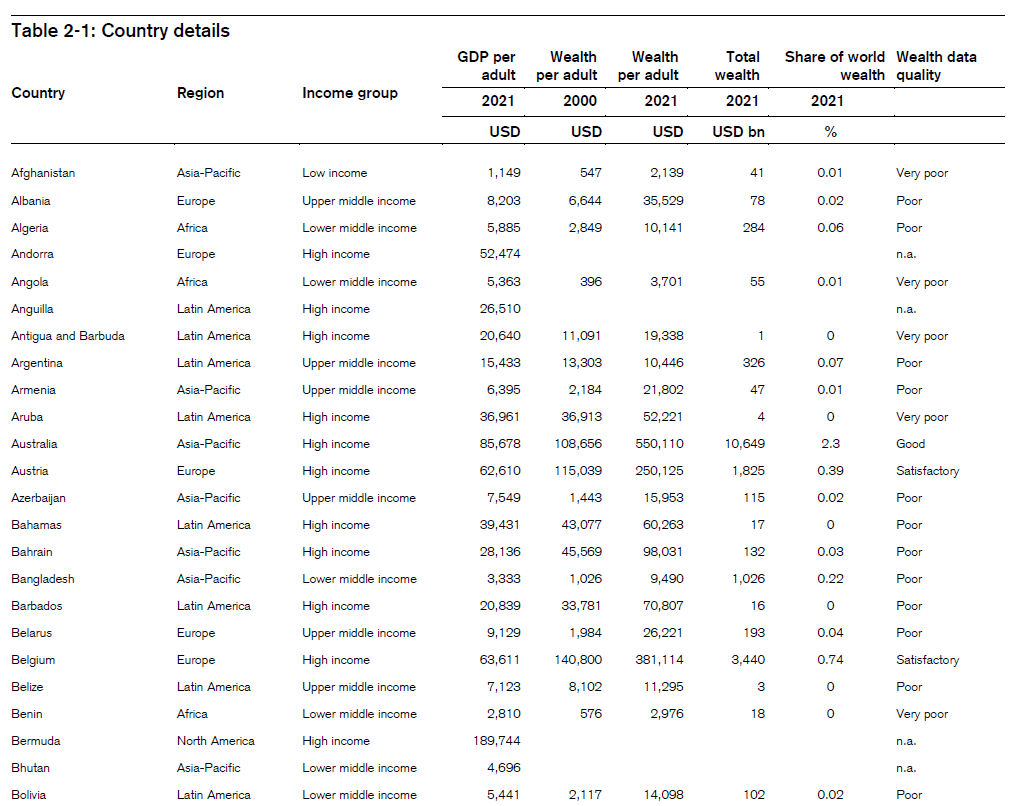


Figura 6: Ejemplo de la fuente de datos. (Fuente: Credit Suisse, 2022).

A partir de esto, luego de realizar la unificación inicial de insumos, se obtuvo una base inicial para la construcción del data set y la cual se muestra a continuación:

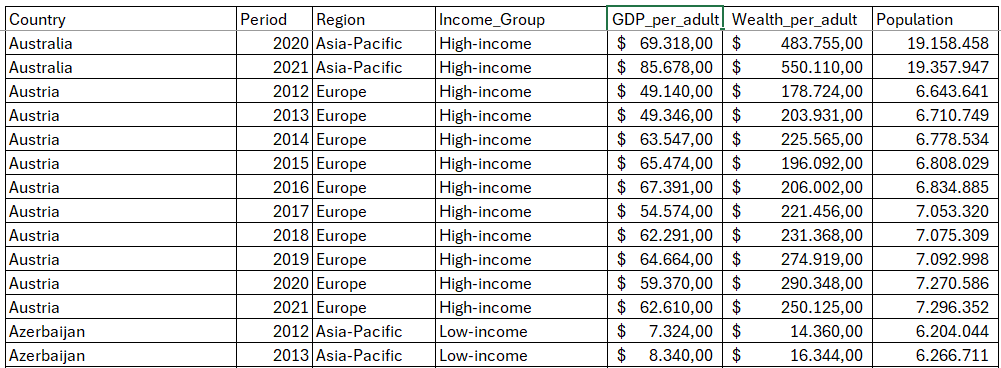


Tabla 8: Insumo inicial para el data set.

A partir de la información mostrada en la Tabla 7, se añadieron las demás variables indicadas en la Tabla 6. Para lograr esto, la variable *Country*, cumple la función de llave para relacionar las diversas fuentes.

Para tener una llave precisa, se debe realizar una limpieza con el fin de unificar parámetros, ya que según la fuente el nombre de los países puede variar en gran medida y la regla de asociación entre fuentes no ser la correcta. Esto se debe a que hay fuentes que se encuentran en idiomas diferentes y por ejemplo tener en una fuente un registro para “*United States*”, en otra fuente ser “*United States of America*” y por otro lado en una fuente adicional tener “*Estados Unidos*”.

Para la obtención de las demás variables para cada uno de los países en los diferentes periodos de tiempo, la fuente de las cuales se obtuvo la información requerida fue:

* *Exports*: Los datos relacionados con esta variable provienen principalmente de la plataforma WITS (*World Integrated Trade Solution),* la cual es una plataforma construida en colaboración entre el Banco Mundial, la OMC (Organización Mundial de Comercio), la UNSD (*United Nations Statistics Division*) y la UNCTAD (Conferencia de las Naciones Unidas sobre comercio y desarrollo (WITS, 2024).

Adicionalmente, con el fin de completar datos faltantes en las publicaciones de WITS, se recurrió a Datosmacro que es una plataforma de origen español, cuyo objetivo es consolidar las principales variables económicas y sociodemográficas de los países a partir de fuentes oficiales. (Datosmacro, 2024).

* *Exports\_ratio*: La fuente principal es el Banco Mundial, que es una organización conformada por 189 países, que es una de las mayores fuentes de financiamiento y conocimiento para los países en desarrollo (Banco Mundial, 2024).

Adicionalmente, para completar datos faltantes se recurrió a Datosmacro.

* *External\_Debt*: La fuente de esta variable, dada su completitud histórica fue Datosmacro.
* *Main\_export*: La fuente de esta variable fue el Banco Mundial.
* *GINI:* Principalmente los datos provienen de la plataforma Datosmacro y se complementó con datos procedentes de la plataforma WorldEconomics, cuya sede en Londres, tienen 30 años de experiencia en el desarrollo de series de datos económicos (WorldEconomics, 2024).

Adicionalmente, se complementó con datos provenientes de la plataforma Indexmundi, que es una plataforma especializada en recopilar datos estadísticos de diversas fuentes (Indexmundi, 2024).

* *Average*\_*Wage*: La fuente principal es WorldData dada su completitud que es una plataforma de origen alemán que compila datos provenientes de bases de acceso público (Worlddata, 2024).

Se complementó con datosmacro.

* *Exp\_Education*: La fuente principal es el Banco Mundial y se complementó con Datosmacro.
* *Democracy\_level*: La fuente principal es la organización intergubernamental IDEA (The International Institute for Democracy and Electoral Assistance) y que se encarga de promover la democracia en todo el mundo y que genera el reporte GSoD (Global State of Democracy) (IDEA, 2024)
* *Informal\_economy*: La fuente principal es el banco mundial y se complementó con la plataforma WorldEconomics.

Finalmente, una vez revisadas e integradas las fuentes, obtuvimos el data set que será la fuente para el modelo y el *dashboard* y cuya estructura final se puede evidenciar a continuación.

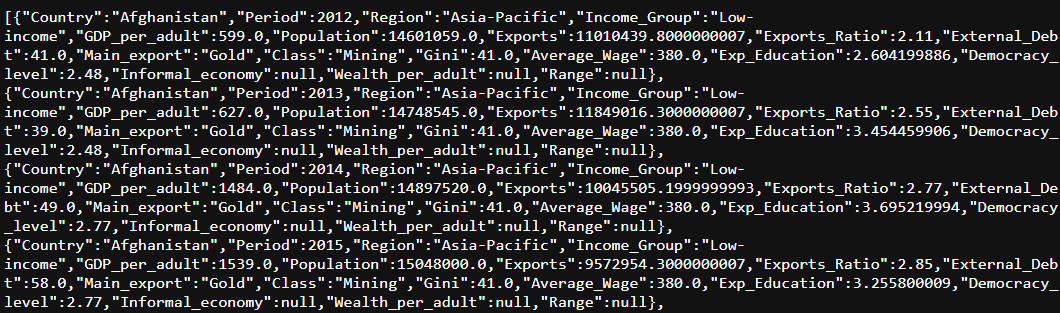


Tabla 9: Muestra tomada del data set final.

## Modelado, Evaluación y Despliegue

Esta es una de las fases que implica un desarrollo y un despliegue técnico. Es decir, entramos a consumir los datos que hemos recolectado a través de las herramientas que hemos considerado que son pertinentes para utilizar en nuestro estudio.

Por esto antes de entrar a explicar los modelos desarrollados, se realiza una introducción a las herramientas que serán usadas en el estudio e indicar la razón del por qué se decidió que eran las más pertinentes.

### Python

Es un lenguaje de programación con propósito general desarrollado por el neerlandés Guido Van Rossum a finales de la década de los años 80.

Dicho lenguaje es muy versátil para tareas relacionadas con ciencia de datos, no solo porque es un lenguaje práctico, sino que principalmente posee una serie de librerías especializadas para entrenar modelos de *machine learning* como lo es Tensorflow.

* Librería Tensorflow: Esta biblioteca de código abierto, fue desarrollada por Google, con el fin de resolver sus necesidades para entrenar redes neuronales y/o descifrar patrones (Tensorflow, 2024).

Es por esto que cumple las condiciones que se consideran idóneas para desarrollar el modelo que se quiere proponer, ya que adicionalmente posee librerías para el manejo de archivos como *Pandas* así como para la visualización gráfica de estadísticas como lo es *Seaborn.*

Esta va a ser la herramienta sobre la que se va a realizar el análisis exploratorio de los datos, así como los modelos mencionados en la fase 1 – Entendimiento del negocio.

### PowerBI

Esta es una herramienta registrada el entorno de Microsoft y que fue lanzada al público en el año 2.105. En resumen, es una colección de software y aplicaciones que funcionan en conjunto para convertir orígenes de datos sin aparente relación entre sí en información coherente, interactiva y visualmente atractiva que permite ser compartida a los usuarios que se desee (Microsoft, 2024).

Dicha herramienta permite expresar datos visualmente en una interfaz intuitiva mediante acciones de arrastrar y soltar, según las necesidades del desarrollador, así como sus capacidades en el desarrollo de soluciones visuales.

PowerBI es una herramienta que permite crear cuadros de mando (*dashboards)* a usuarios que no tienen conocimiento en el desarrollo gráfico de soluciones que implican saber programar como es el caso de la librería D3, la cual requiere saber técnicas de programación usando Java y HTML.

Adicionalmente ofrece una amplia gama de alternativas gráficas que permiten generar soluciones que se ajustan a una amplia gama de necesidades por lo que sumado a que es una herramienta que se podría catalogar como intuitiva, fue la escogida para plasmar la solución visual.

### Anaconda

Anaconda es una herramienta creada en el año 2.013 por Peter Wang y Travis Oliphant, con el fin de introducir el lenguaje Python en la analítica de datos de los negocios (Anaconda, 2024).

Anaconda maneja un entorno enfocado a la ciencia de datos, permitiendo soluciones escalables en una plataforma todo en uno y con un modelo de despliegue de un solo click.

Adicionalmente, Anaconda presenta la opción de desarrollar a través de sus *notebooks*, soluciones basadas en lenguajes de programación, como es Python en este caso, con el fin que sea la herramienta que se utilice para que a través de Python realicemos nuestro respectivo análisis exploratorio de los datos, así como el desarrollo y ejecución de nuestros modelos propuestos.

### MongoDB

Es un sistema de administración de bases de datos no relacional (DBMS *Database management system*) que utiliza documentos similares a un formato JSON lo cual permite una mayor flexibilidad frente a la estructura de tablas usadas en las bases de datos relacionales. (MongoDB, 2024).

Esto se debe a que Mongo provee un modelo de almacenamiento de datos que permite a los usuarios guardar y procesar datos multivariados con facilidad. Esto simplifica el almacenamiento y permite un entorno altamente escalable para aplicaciones y servicios.

Dado que los documentos son la unidad básica de datos en MongoDB, al estar formateados como objetos JSON (*Java Script Object Notation*), estos objetos pueden almacenar cualquier tipo de datos a lo largo de múltiples sistemas.

MongoDB cumple la función como al almacén de datos en este trabajo, pensando en que los estudios posteriores no tengan limitantes en el tipo de datos y adicional será la fuente que alimente el *dashboard* propuesto.

### Selección e implementación de Modelos

Desarrollando la fase de modelado, dado que no se evidenciaron modelos implementados previamente cuyo objetivo fuera clasificar la riqueza per cápita, pero basados en la naturaleza de los datos con los que disponemos, se decidió realizar un modelo de clasificación basado en redes neuronales y compararlo con un algoritmo *random forest*.

Ambos algoritmos antes mencionados en este caso son aplicados para clasificación y buscan determinar una variable objetivo, la cual es la variable de salida. Dicha clasificación se realiza a partir de una serie de variables, la cuales para este estudio son las explicadas en la fase de “Entendimiento de los Datos” y que obligatoriamente deben ser numéricas al momento de ingresar al modelo, por lo que si tenemos variables de texto dichas variables deben ser transformadas.

La diferencia radica en que el primer modelo se basa en redes neuronales y el segundo se basa en árboles de decisión.

Finalmente se diseñó un modelo de regresión basado en redes neuronales con una neurona de salida, con el fin de poder hacer predicciones, en este caso de la riqueza promedio por adulto.

Al finalizar, se evaluarán los resultados con el fin de poder determinar qué tan precisos pueden llegar a ser sin caer en el error del sobreajuste, el cual consiste en que los datos se adaptan perfectamente al modelo, pero no garantizan que se adapten a nuevas instancias.

### Modelo de red Neuronal con Función de perdida *Sparse Categorical Crossentropy*

Este es un algoritmo usado comúnmente en problemas de clasificación, como es el primer modelo propuesto en este estudio, se utiliza una función de perdida más eficiente que la función más básica *Categorical Crossentr*opy. Dicha función de perdida es una función matemática cuyo objetivo es cuantificar las diferencias entre los valores estimados y los valores reales.

Para su implementación, se utilizó la plataforma de desarrollo para *notebooks* ofrecida por Anaconda la cual se encarga de proveer el medio para estructurar el modelo usando el lenguaje de programación Python y que adicionalmente permite la ejecución paso a paso del modelo y poder revisar los resultados de cada uno de los pasos que componen el modelo.

Para esta solución, luego de haber leído desde el repositorio de origen los datos requeridos para el modelo y determinar la viabilidad de trabajar con estos, se hizo una exploración de todos los datos para entender su naturaleza, es decir el formato con el que llega cada variable, así como poder revisar cuantos registros presentan datos nulos o inexistentes y finalmente pero no menos importante saber con cuántas variables se puede trabajar.

Un resumen de dicha exploración de todos los datos realizada en Anaconda se puede ver a continuación:

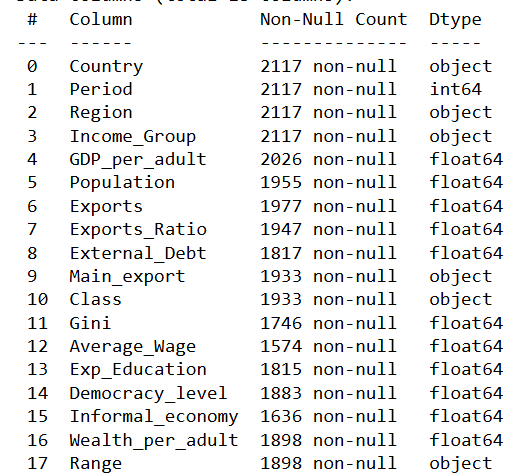


Figura 7: Tipo inicial de los datos usados en los modelos.

Como se evidencia en la Figura 7, hay datos de tipo *object* que en resumen son cadenas de texto, sobre los cuales se deberá hacer una transformación con el fin que queden definidos con formatos numéricos, y para esto fue utilizada la función *map*, provista por Python.

Esta transformación es fundamental dentro del procesamiento ya que estos modelos no reciben variables de texto.

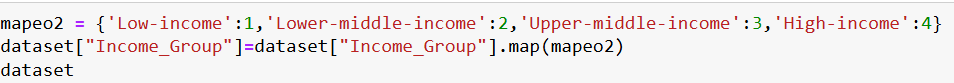


Figura 8: Transformación de los datos para el modelo.

Luego de haber hecho la respectiva transformación de datos, se procedió a dejar aquellas instancias que no tenían valores nulos en las variables que le interesan al modelo y adicionalmente se removieron aquellas que no iban a ser usadas en el modelo.

Dicha eliminación de valores nulos se realizó con el fin de no interferir con el resultado del modelo, ya que si se quiere obtener resultados más precisos lo mejor es trabajar con los datos reales .

Finalmente, dado que el objetivo del modelo es poder hacer una clasificación de la riqueza per cápita, dicha variable se puede clasificar como continua, por lo que es necesario hacer una transformación adicional para que ingrese al modelo como variable discreta.

Luego de estos pasos previos de limpieza y transformación de datos, ya se cuenta con el grupo de instancias que serán ingresadas al modelo. A continuación, se muestra el resultado de dicha transformación en donde cada una de las instancias solo tiene datos numéricos y no se presentan valores nulos.

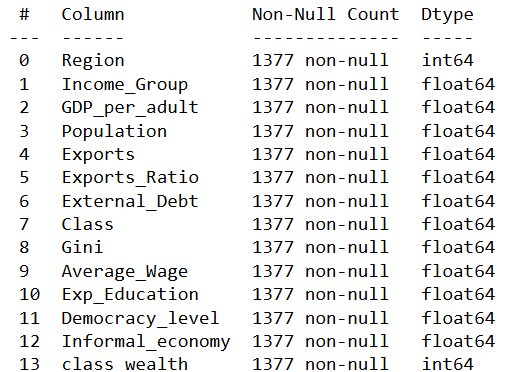


Figura 9: Tipo final de los datos usados en el modelo de red neuronal.

El siguiente paso requerido para la construcción del modelo es el de la normalización de los datos. Esto se debe ya que como se puede observar a continuación en la Figura 10, las variables usadas en el modelo se mueven entre rangos muy diferentes y la eficiencia del modelo aumenta cuando los datos se encuentran entre rangos similares. Por lo que en conclusión este paso tiene como objetivo optimizar el rendimiento del modelo con el fin de reducir los costos asociados al procesamiento. Una acción para normalizar los datos es tomar cada dato y proceder a restarle su media y dividirla por la desviación estándar, aunque esto no se realiza explícitamente sino a través de una función ofrecida por Python.

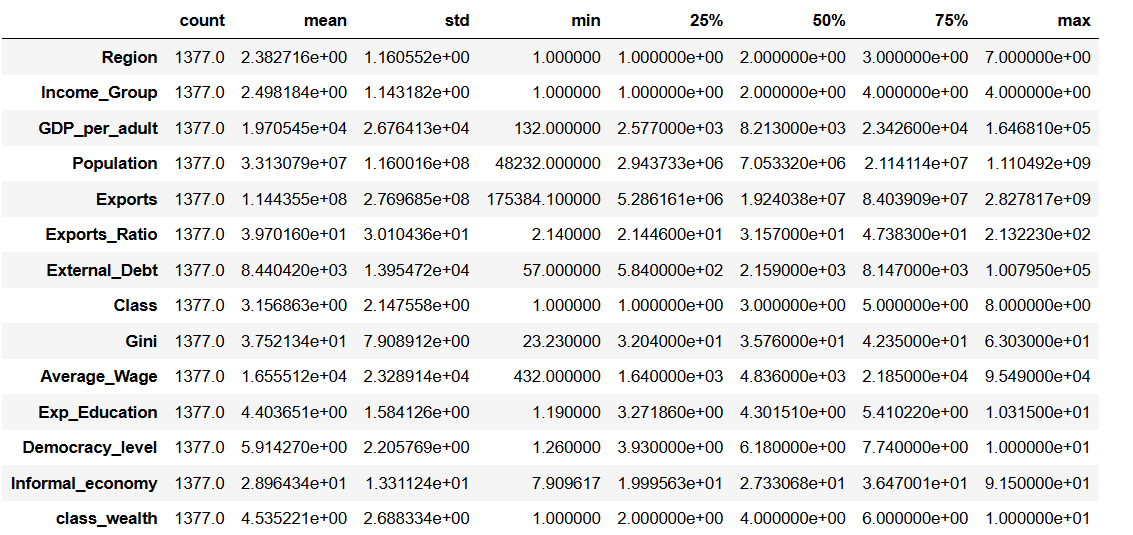
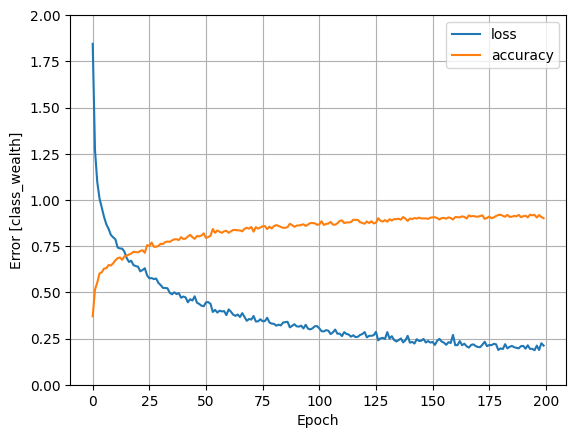


Figura 10: Resumen estadístico de los datos usados en el modelo de red neuronal.

Una vez teniendo el grupo de instancias, es necesario aclarar la distribución de los datos entre aquellos que serán usados para entrenar el modelo, como aquellos usados para probarlo. Esta distribución se determinó en que un 20% de las instancias serán usadas para entrenar el modelo y un 80% serán usadas para probarlo. A medida que aumenta las instancias usadas para entrenar el modelo mayor es el riesgo del sobreajuste dado que el modelo terminará ajustado solo para las instancias con las que se cuenta, más no con instancias futuras que ingresarían al modelo propuesto.

Continuando con la estructuración del modelo, se establecieron 4 niveles de neuronas en donde las 3 primeras tienen 128 neuronas cada una con función de activación “*relu*”, la cual solo requiere una comparación. Y finalmente la capa de salida con 11 neuronas según la cantidad de clases de salida establecidas con función de activación “*softmax*”, dada su eficiencia para clasificaciones multiclase como es nuestro caso que contamos con 10 posibles opciones.

Luego de entrenar el modelo con el 20% de las instancias en 200 iteraciones (*epoch*), los resultados graficados se muestran a continuación:



Gráfica 6: Resultado del entrenamiento del modelo neuronal propuesto.

Con el modelo iterado se está llegando a una precisión del 90% (Ver línea color naranja. El 90% de las predicciones serán correctas), lo cual se puede considerar aceptable y adicionalmente se llega a una perdida mínima cercana al 20%.

Es importante mencionar que el proceso se iteró cambiando los hiper-parámetros desde el número de capas de ocultas, la cantidad de neuronas en cada capa, el optimizados de aprendizaje y los resultados no fueron necesariamente mejores que el mostrado finalmente.

### Modelo de clasificación *DesicionTree*

Con el fin de mostrar cómo una red neuronal puede llegar a dar mejores resultados que un árbol de clasificación simple, se incluye este apartado con un algoritmo sencillo de clasificación basado en árboles.

Como vimos en el numeral anterior, se puede observar que luego de doscientas iteraciones, se logró una precisión del 90% en general para todo el modelo.

Para este modelo con algoritmo *Desiciontree*, al revisar la matriz de confusión, en la cual sobre la línea diagonal se observa las instancias del grupo de validación que fueron clasificadas correctamente, se concluye una precisión del 75%, lo cual representa aproximadamente 15 puntos porcentuales menos que en el modelo basado en redes neuronales.

Imagen que contiene Texto

Descripción generada automáticamente

Tabla 10: Matriz de confusión resultante DecisionTree.

Tabla

Descripción generada automáticamente

Tabla 11: Reporte de clasificación DecisionTree.

Respecto a la Tabla 11, se puede concluir que la precisión ponderada del modelo es del 75%, siendo menor en aquellas categorías en las que el número de instancias de entrenamiento fue menor, por lo cual se asume que se debería aumentar las instancias de entrenamiento para dichos grupos.

### Modelo de clasificación *Randomforest*

Este modelo de clasificación no basado en redes neuronales se incluye como apartado para comparar la efectividad versus el modelo descrito en el numeral anterior, ya que se parte en que es un algoritmo que optimiza los árboles de decisión.

Básicamente dicha optimización se logra partir de generar múltiples árboles de decisión para luego ponderarlos y así generar una única clasificación. Es decir que genera varias alternativas y las compara para luego entregar un árbol de clasificación mejorado.

Para este ejercicio, se definió como parámetro cien estimadores y se tomaron las mismas instancias para entrenamiento y validación que en el modelo basado en redes neuronales y que corresponden a una relación de 20% para entrenar el modelo y un 80% para probarlo.

El resultado del ejercicio genera una precisión del 77%, la cual es 2 puntos porcentuales superior que en el ejercicio anterior pero que sigue siendo casi 13 puntos porcentuales menor al logrado en el modelo basado en redes neuronales.

Respecto al resultado de las clasificaciones, a continuación, se muestra el resultado de la matriz de confusión para el modelo basado en el algoritmo *Randomforest,* en la cual sobre la línea diagonal vemos las instancias que fueron clasificadas correctamente.

Imagen que contiene Patrón de fondo

Descripción generada automáticamente

Tabla 12: Matriz de confusión resultante Randomforest.

Tabla

Descripción generada automáticamente

Tabla 13: Reporte de clasificación Randomforest.

A partir de la Tabla 13, vemos como la precisión cambia según la categoría de salida, llegando a valores superiores o cercanos al 90% para las categorías 1, 2 y 8 y a una precisión mucho más del 43% para la categoría 7 siendo la peor de todas.

Esta imprecisión puede estar determinada por la diferencia entre las muestras de entrenamiento para cada categoría, siendo similar el comportamiento que en el modelo basado en *DesicionTree*. Se concluye que este valor debería ser más alto al tener un número mayor de instancias de entrenamiento ya que por ejemplo para la categoría 7 solo entraron 5 instancias para entrenamiento.

Finalmente, es importante resaltar el resultado obtenido en el modelo de clasificación basado en redes neuronales en la precisión general del modelo versus los obtenidos en los árboles de clasificación (*DesicionTree y RandomForest)*, que aunque se pueden ajustar tanto en parámetros como en las instancias de entrenamiento para que obtengan mejores resultados, es evidente la ventaja del modelo inicial.

### Modelo de regresión basado en redes neuronales

A diferencia de los tres algoritmos previamente explicados que buscaban hacer una clasificación del rango de la riqueza per cápita, en este módulo se propone un modelo de regresión basado en redes neuronales con el fin de predecir la riqueza per cápita a partir de un grupo de variables explicativas.

Otro punto diferente respecto a los modelos anteriores es que la variable de salida no necesita ser estrictamente discreta, por lo que se mantiene la estructura original de variable continua.

Adicionalmente, se debe resaltar que al igual en los modelos de clasificación explicados previamente, se hizo la transformación de las variables que eran del tipo *Object* o tipo texto, con el fin que todas las variables explicativas fueran numéricas.

Finalmente, se escogieron las variables que se iban a ingresar al modelo, dado que no se ve la necesidad de usar todo el dataset. Es importante resaltar que no se debe obligatoriamente trabajar con todas las opciones sino con aquellas que pueden tener una correlación con la variable objetivo, que este caso es la riqueza per cápita (*Wealth\_per\_adult*).

A partir de lo anterior, la estructura del dataset usado para el modelo de regresión quedó definido bajo la estructura que se muestra a continuación.

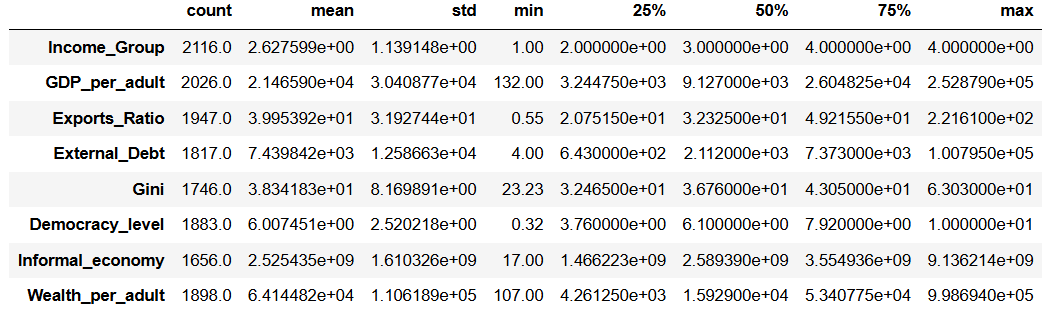
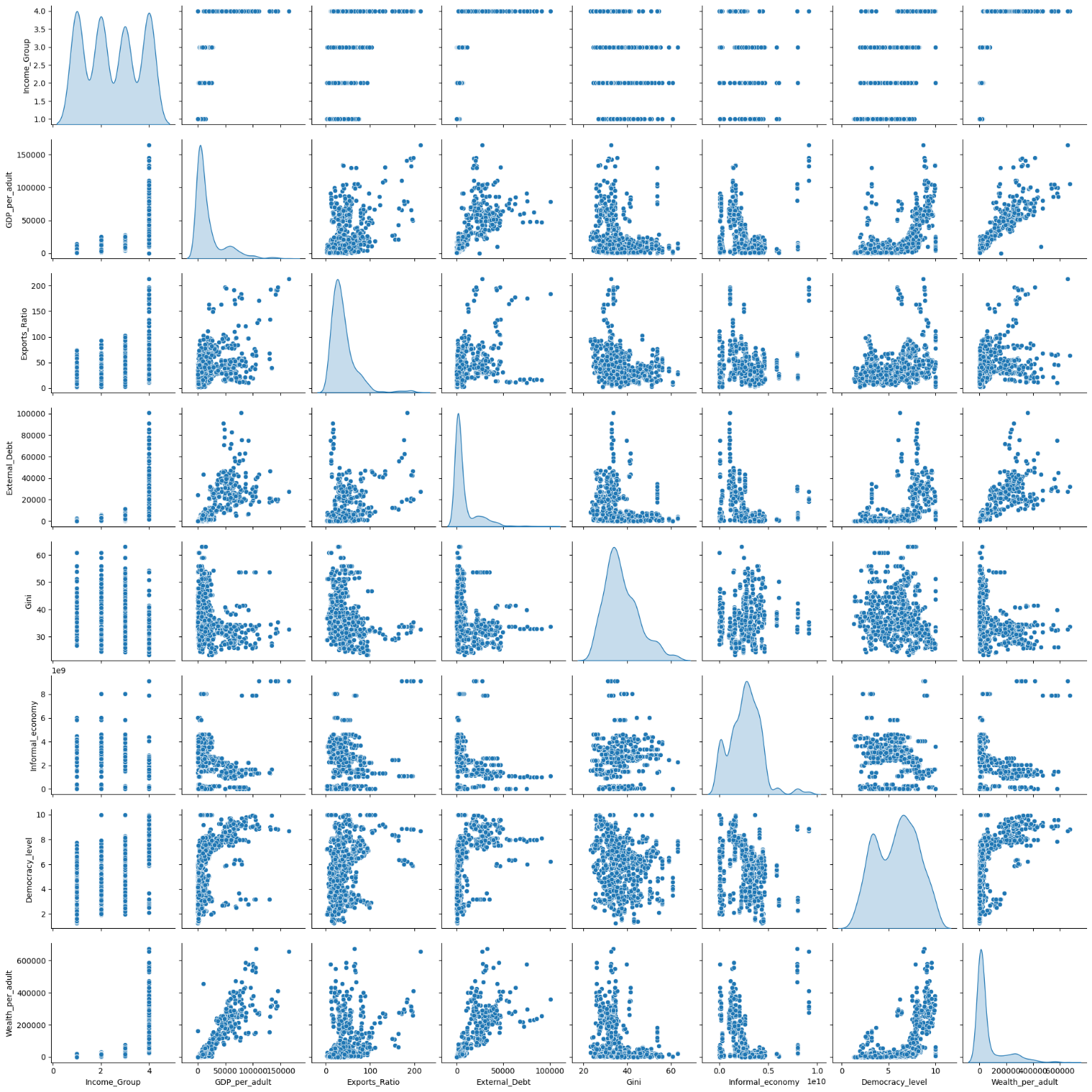


Tabla 14: Resumen estadístico de los datos usados en el modelo de regresión.

Con el fin de separar las instancias para el entrenamiento, se escogió nuevamente una proporción del 20% para datos de entrenamiento y un 80% para datos de validación. Esto se define, como se mencionó anteriormente para no caer en el sobreajuste del modelo.

A continuación, para poder explicar el por qué se escogió ciertas variables, se muestra una serie de gráficos de dispersión, para demostrar como la variable objetivo tiene correlaciones positivas y negativas más fuertes y evidentes con ciertas variables, mientras que con otras no se evidencia que tenga un correlación significativa por lo que fueron descartadas.



Gráfica 7: Gráficas de dispersión entre las variables escogidas para la regresión lineal.

A partir de la Gráfica 7, se evidencia en la sección resaltada en rojo, cómo la variable objetivo parece tener correlación positiva con las variables *GDP\_per\_adult, external\_debt, average\_wage*; adicionalmente, una correlación no lineal con la variable *Democracy\_index* y finalmente una correlación negativa con el índice de *Gini* e *Informal*\_*economy*.

Estas correlaciones cuando son positivas básicamente que mientras una crece, la otra también lo hará aunque no necesariamente en la misma proporción y por otro lado, en las correlaciones negativas se observa que mientras una de las variables crece, la otra decrecerá pero no necesariamente en la misma proporción.

Continuado con la explicación de modelo, de igual manera que el modelo de clasificación basado en redes neuronales como en árboles de decisión, nuevamente fue necesario hacer una normalización de los datos, para que los rangos entre los que se mueven las variables sean más uniformes y el procesamiento llegue a ser óptimo.

Posteriormente, respecto al modelo de la red neuronal, se definieron tres capas intermedias de neuronas cada una con 64, 32 y 16 neuronas respectivamente, todas con función de activación ‘*relu’*, para luego tener una última capa de salida. Dicha función se determinó dado que hablando en términos computacionales permite un rápido procesamiento.

La última capa solamente posee una neurona, dado que como es un modelo de regresión lineal, lo que se espera al ingresar una serie de datos en una única salida con un valor predicho. Este aspecto difiere en los modelos de clasificación en los cuales se espera varias salidas.

Partiendo de lo anterior, la estructura de la red neuronal finalmente quedó definida la siguiente manera:

Tabla

Descripción generada automáticamente con confianza media

Tabla 15: Estructura de la red neuronal para regresión.

Finalmente, para entrenar el modelo, se definieron 3.000 iteraciones dada la complejidad de la regresión.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfica 8: Error medio absoluto generado en la regresión.

*Gráfico

Descripción generada automáticamente con confianza media*

Gráfica 9: Error cuadrado medio generado en la regresión.

Tanto en la gráfica 8 como en la 9, se puede ver el comportamiento de la regresión a lo largo de las 3.000 iteraciones. Claramente tenemos un error cuadrado medio cada vez acercándose más a 0, lo cual indica un aumento en la efectividad de la regresión. Dada la tendencia observada tanto en los valores del error, se podría pensar que al aumentar la cantidad de iteraciones se podría llegar a un menor error.

Por otro, se puede ver como el error absoluto parece llegar a un valor de 10.000 que, aunque parece alto, debemos tener presente que el rango observado en la variable objetivo se mueve entre los USD$100 y los USD$700.000.

Aunque es importante mencionar que pese a que es posible aumentar el número de iteraciones se llegue a un error menor, se presenta el riesgo de tener sobre un sobreajuste en el modelo y que este solo se adapte a las instancias con las que se cuenta pero no explique el comportamiento general.

### Modelo del Dashboard – Power BI.

El *dataset* final, así como el resultado de los modelos se presentan junto con una herramienta de visualización que permite al lector sacar sus propias conclusiones y adicionalmente interactuar con los datos crudos que sirvieron para alimentar el proceso. Dicha visualización se diseñó bajo un modelo de *dashboard* en el *software* PowerBI, desarrollado por *Microsoft.*

Teniendo la capa cruda de datos y revisando los potenciales estudios que se pueden realizar sobre estos al revisar perspectivas futuras, se decidió almacenar en la base de datos NoSQL MongoDB y a través de un *driver* de conexión enviar los datos crudos directamente a PowerBI. Esto dado que MongoDB permite una escalabilidad tanto en la cantidad de instancias con las que se puede contar, así como en una posible necesidad de agregar más variables sin afectar la estructura que se definió inicialmente

A partir de lo anterior, se incluye a continuación una descripción detallada del *dashboard* y cada una de las vistas que lo componen, así como una breve descripción de las gráficas que fueron escogidas para integrar dicho *dashboard*.

Para comenzar la presentación a nivel general el *dashboard* propuesto, lo primero que se debe mencionar es que estese compone de 7 vistas principales y entre las cuales el usuario podrá interactuar según su deseo. Dichas vistas se pueden encontrar en la Figura 11 mostrada a continuación desde la vista llamada “Portada” hasta la denominada “Acerca De”.

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente

Figura 11: Vistas que componen el dashboard propuesto.

La primera vista, denominada portada, además de contener el título del tablero, también presenta un mapa de coropletas en donde se puede evidenciar una clasificación por color de los países según la región a la que pertenece cada uno y que puede aproximarse o no a los continentes en los que se encuentran.

Adicionalmente, en la parte superior derecha se encuentran dos tarjetas de resumen con la sumatoria de la riqueza total, así como de la población adulta en la región o país seleccionado y que cambian la información mostrada según el usuario interactúe con la vista. Dichas tarjetas inicialmente muestran que la población mundial adulta total es 5.177 millones de personas las cuales poseen una riqueza acumulada de USD$458 billones.

Finalmente, en esta vista inicial se pueden encontrar dos filtros para que el usuario pueda interactuar con cada uno de los componentes de la vista. Dichos filtros son el periodo en el cual se relaciona los años en los cuales se tiene información y el usuario podrá escoger entre el año 2.012 y 2.021 y por otro la región que como se indicó previamente es la que determina el color para los países

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Figura 12: Vista 1 Portada Dashboard.

Por otro lado, en la misma vista el usuario, además de poder utilizar los filtros indicados anteriormente, podrá interactuar con el mapa directamente y escoger algún país en específico o un grupo de varios países para ver información de su interés.

Finalmente, el usuario también podrá si lo desea ver información de resumen de un país a través de un *tooltip* emergente, que aparecerá siempre al mover el cursor sobre este. Un ejemplo se muestra a continuación.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Figura 13: Vista 1 Portada Dashboard con filtro.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Figura 14: Vista 1 Portada Dashboard con tooltip.

Posteriormente, se encuentra la segunda vista denominada Vista\_País, la cual para empezar en la parte superior izquierda posee un filtro que permite escoger una sola opción de país, para que el usuario pueda escoger aquel sobre el que tenga interés para ver la información disponible de este.



Figura 15: Vista 2 Vista\_País Dashboard.

Una vez en esta vista, en la parte inferior izquierda, se encuentra un mapa en donde el usuario puede ver la ubicación del país que escogió previamente y el cual se encuentra resaltado con el color de la región al que pertenece.

Adicionalmente, en la parte superior derecha, se encuentran seis tarjetas en donde el usuario puede encontrar información macroeconómica adicional del país que escogió y que le permite hacerse una idea del entorno económico en el que se mueve el país de su elección. Dichas tarjetas contienen los siguientes datos:

* Promedio del total de las exportaciones en miles de millones de dólares de ese país en los últimos diez años.
* Promedio de la participación de las exportaciones de ese páis en el PIB en los últimos diez años y la cual se mide en porcentaje.
* Principal producto de exportación de ese país. En esta opción aparece por ejemplo materias primas como el petróleo, el oro o por ejemplo bienes como lo son carros u otros productos manufacturados.
* Promedio del gasto en educación de ese país en los últimos diez años en relación con su producto interno bruto y el cual se mide en porcentaje.
* Promedio del índice de GINI (Desigualdad) en los últimos diez años. Dicho valor va de 0 a 1, en donde un indicados cercano a 0 indica una distribución equitativa de la riqueza.
* Promedio del índice de democracia en los últimos diez años medido de 1 a 10, en donde un valor cercano a 10 indica una democracia perfecta y valores cercanos a 0 son para regímenes como dictaduras.

Finalmente, el usuario puede encontrar una gráfica histórica de líneas (la cual es adecuada para mostrar tendencias en el tiempo), la evolución del PIB per Cápita de ese país y la evolución de la Riqueza per Cápita para los años con los que se cuenta con información, para poder identificar posibles dependencias y comportamientos entre estas dos variables. Dicha gráfica permite al usuario interactuar con los periodos mostrados por si desea revisar un rango histórico puntual diferente al mostrado inicialmente.

En la tercera vista denominada Resumen\_Países, el usuario básicamente podrá explorar que países tienen las riquezas más altas o ver cómo se encuentra la riqueza de los países de una misma región.

En la parte superior el usuario puede encontrar los filtros de Periodo y Región, al igual que en la vista “Portada” y que fue explicada previamente y con los cuales se puede interactuar con los datos mostrados en esta vista.

En la parte inferior izquierda se encuentra un diagrama de barras horizontal, el cual se escogió ya que funciona para identificar jerarquías y en el cuál el usuario puede ver los países que lideran el top de riquezas per cápita a nivel mundial. Esto lo podrá hacer ya sea para todo el mundo y ver cómo se encuentran todos los países o para la región que le interese y ver solo el grupo de países que la componen.

En la parte inferior derecha, se encuentra una matriz con la información de la riqueza de todos los países entre los años 2.012 y 2.021 y que permite un primer acercamiento con los datos crudos que alimentan el *dashboard*.

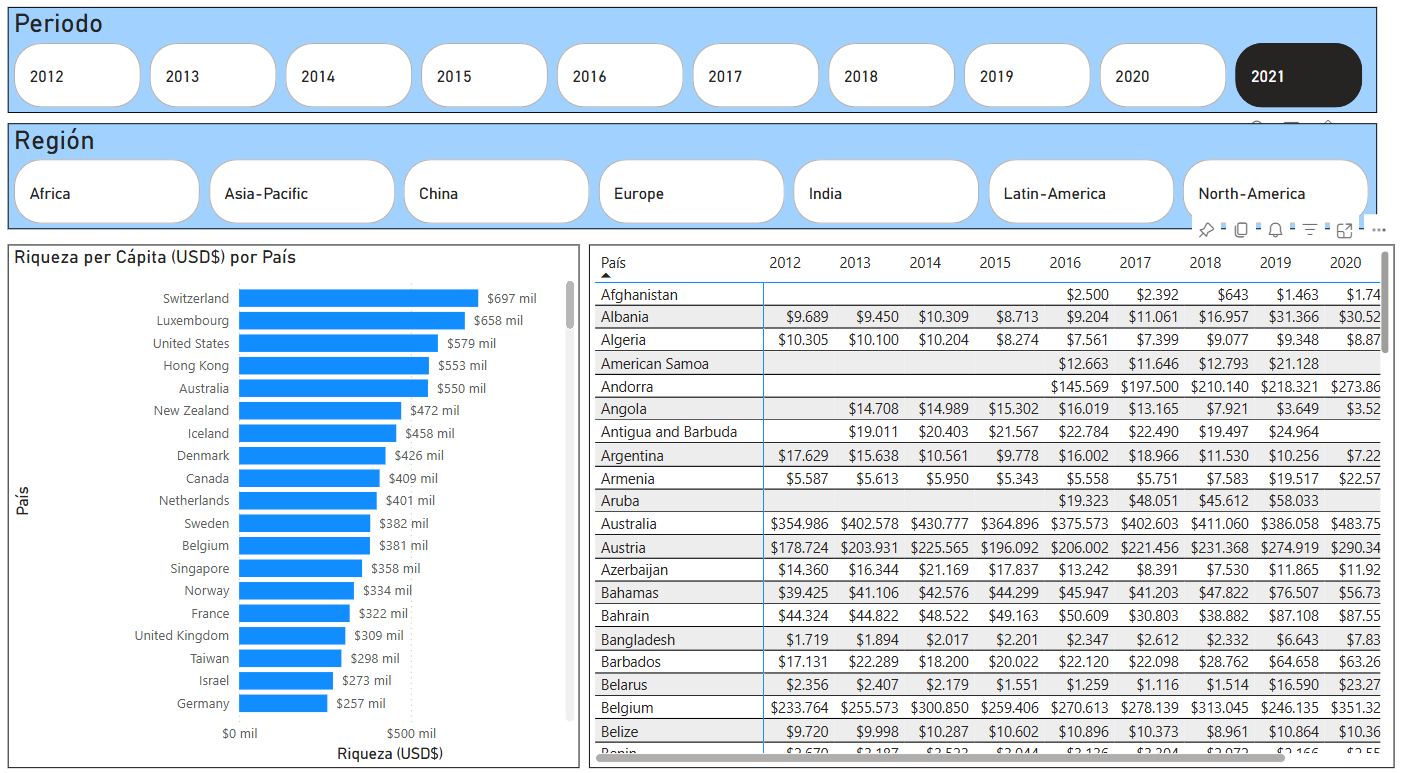


Figura 16: Vista 3 Resumen\_Países Dashboard sin filtrar.

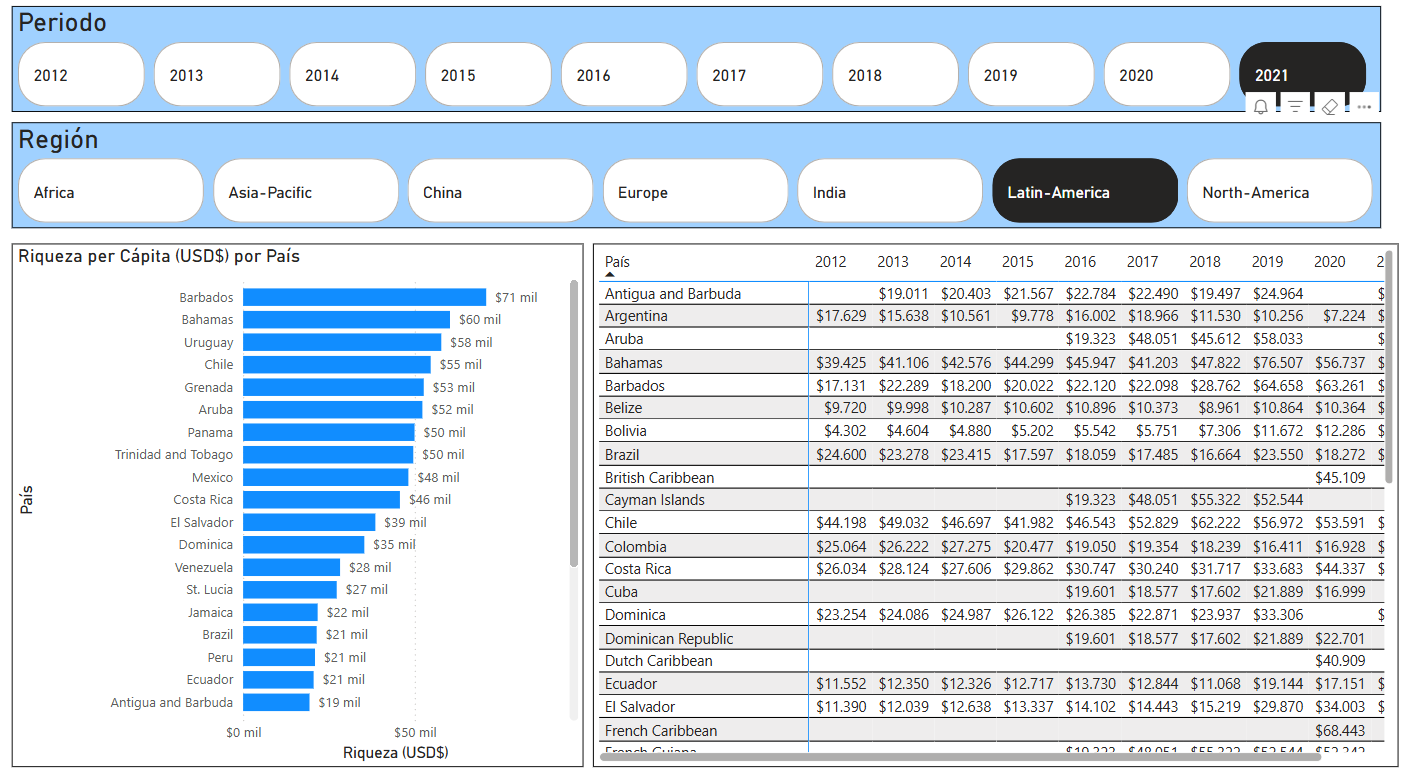


Figura 17: Vista 3 Resumen\_Países Dashboard con filtro.

Posteriormente, la cuarta vista denominada Comparativa\_Países, cumple con el objetivo de permitir al usuario el poder comparar la riqueza entre dos o más países que el usuario requiera.

En la parte superior izquierda, quien navegue por el *dashboard* podrá encontrar un panel de países con el cual podrá filtrar aquellos que desee.

En la parte inferior derecha, se encuentra un mapa en donde el usuario podrá ver la ubicación de los países que escogió previamente y en el cual nuevamente aparecen clasificados por color según la región a la que pertenezcan y que es consecuente con el color de la región en las vistas previamente explicadas.

Finalmente, en la parte izquierda de la vista, se puede observar una gráfica de líneas (que como se mencionó anteriormente son ideales para ver comportamientos en el tiempo) y en la cual se encuentra la información de la evolución de la riqueza per cápita de los países que el usuario ha escogido previamente.

Por ejemplo a continuación se muestra una comparativa entre Algeria en África y Armenia en Asia, en la cual se ve como Algeria tenía una riqueza superior hasta el año 2.018 cuando la población de Armenia disparó su riqueza hasta casi el doble que la de Algeria.

Interfaz de usuario gráfica, Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figura 18: Vista 4 Comparativa\_Países Dashboard con filtro.

A continuación, respecto a la siguiente vista denominada Vista\_Regiones, el usuario puede identificar esas zonas del mundo con mayor concentración de riqueza.

En la parte superior izquierda se puede encontrar un panel de Periodo con el cual se podrá filtrar el año del que se requiera información y que como se indicó previamente va desde el año 2.012 hasta el año 2.021.

Luego en la parte superior derecha se puede encontrar un diagrama de torta con la participación de cada región respecto al total de la riqueza en el mundo. Este diagrama funciona bien cuando se quiere ver una participación de una categoría frente al total y no son muchas las que están representadas (En este caso son siete regiones). Por ejemplo, para este caso se ve que entre Norteamérica y Europa suman más de la mitad de la riqueza mundial.

En la parte inferior izquierda, se encuentra un *treemap* con el volumen de riqueza que cada país posee frente a los otros países y en el mundo. Dicha representación funciona para resaltar categorías y claramente se ve como Estados Unidos y China marcan la diferencia respecto al resto de opciones.

Finalmente, en la parte inferior izquierda se encuentra un diagrama de doble eje en el cual se puede ver el volumen de riqueza de una región frente al total de población adulta que vive en dicha región. De aquí se pueden concluir aspectos como que África es un referente con mucha población, pero con muy poca riqueza.

Gráfico, Aplicación

Descripción generada automáticamente

Figura 19: Vista 5 Vista\_Regiones Dashboard.

Respecto a la vista Relaciones, en la parte superior el usuario podrá encontrar los paneles de filtrado por Periodo y Región y que como se puede concluir son las principales opciones que tiene el usuario para interactuar con el *Dashboard*.

Básicamente, en esta vista se puede encontrar seis diagramas de dispersión (Este tipo de gráficas funciona cuando se quiere evaluar correlaciones entre variables), entre la variable de riqueza per cápita y las siguientes variables:

* Deuda externa per Cápita.
* Índice de democracia.
* Índice de GINI.
* Gasto en educación en el PIB
* Porcentaje de economía informal
* PIB per Cápita.

Cabe resaltar que estas variables representadas en gráficos de dispersión fueron las mismas utilizadas en el modelo de regresión dado que indicaban cierta correlación ya fuera positiva o negativa con la riqueza per cápita.

En dichas gráficas se puede evidencias nuevamente la clasificación por color con la región, para así mantener la continuidad del *dashboard* y se conserve la estructura.

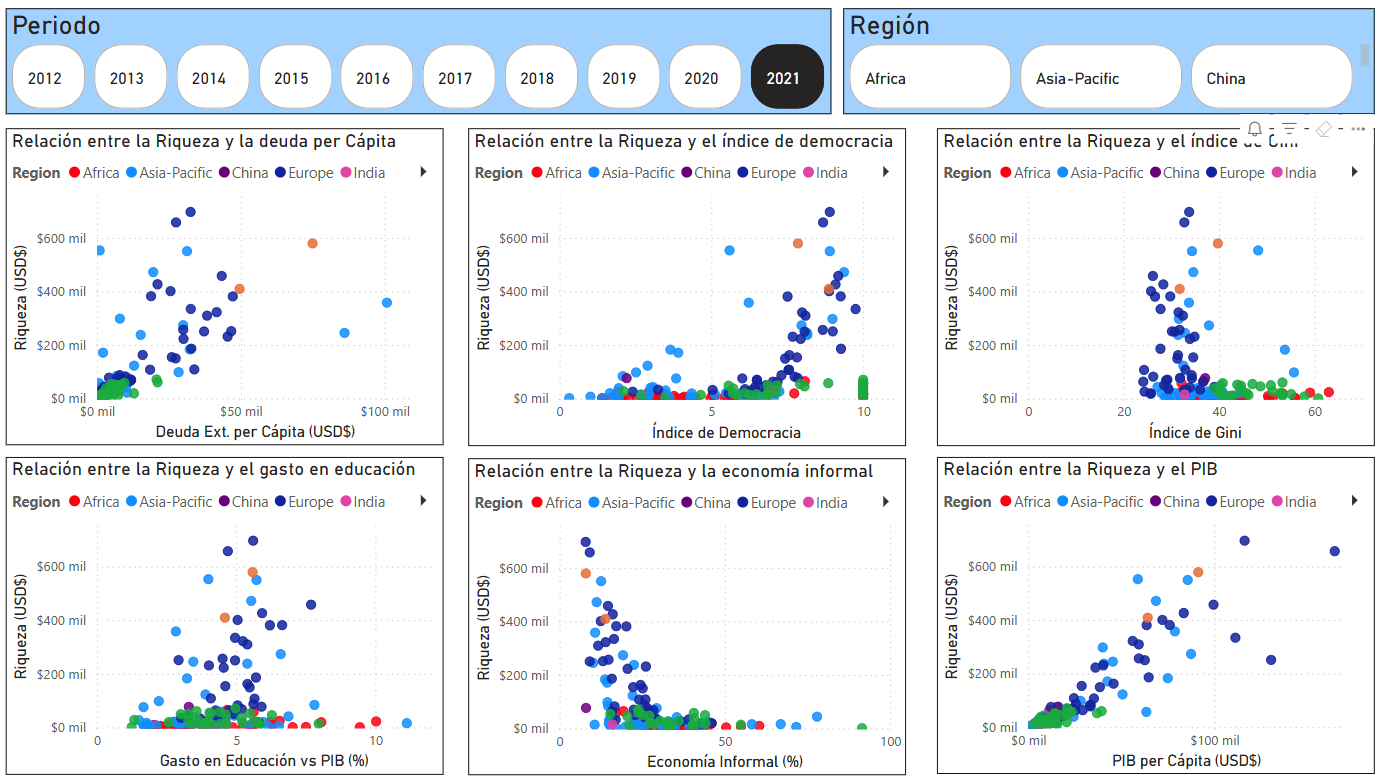


Figura 20: Vista 6 Relaciones Dashboard.

Para concluir, se añade una última vista completamente informativa llamada Acerca De, en la cual se encuentra un breve resumen del tablero y lo que se puede encontrar en cada una de sus vistas, así como la ruta del repositorio en *GitHub* en donde los usuarios pueden encontrar los datos crudos que alimentan el tablero en caso de que deseen descargarlos y hacer su propia analítica.

Dicha vista se incluye para dar un contexto de las fuentes, así como del objetivo de este para dar una idea al usuario de lo que puede encontrar a lo largo del *Dashboard*. Pese a que no presenta ningún grado de interactividad para quien revise esta solución visual si añade un grado de contexto para quien no esté familiarizado con la información encontrada.

Texto

Descripción generada automáticamente con confianza media

Figura 21: Vista 7 Acerca de Dashboard.

# Conclusiones

En este apartado se incluye las conclusiones a las que se llegaron en la realización de este trabajo. Hay conclusiones directamente sobre los objetivos planteados, así como conclusiones adicionales:

* Respecto al objetivo general planteado se logró el desarrollo de una solución visual a través del *dashboard* propuesto al permitir ver un panorama general de cómo el estado actual de los niveles de riqueza en el mundo. Dicha solución se complementó satisfactoriamente con los modelos propuestos al permitir a través de las variables propuestos explicar los niveles de riqueza con una precisión de hasta el 90%.
* A partir del modelo de regresión y como se vio en la visualización hay seis variables de las escogidas que se correlacionan positiva o negativamente con el nivel de riqueza en la población: deuda per cápita, índice de democracia, índice de Gini, gasto en educación, economía informal y el PIB per cápita.
* Respecto a la evaluación de los modelos, pese a que los árboles de decisión aportan soluciones valiosas, los mejores resultados se obtuvieron con los modelos basados en redes neuronales, con una precisión casi trece puntos porcentuales más alta.
* La situación mundial es muy desigual. Claramente los Estados Unidos poseen casi un tercio de la riqueza mundial, teniendo solo el 4% de la población adulta. Si se suma Estados Unidos, China y Europa se llega casi a un 75% de la riqueza mundial. África tiene los peores niveles de riqueza a pesar de tener gran participación en el total de la población adulta.
* Los árboles de decisión son más fáciles de interpretar y de implementar, aunque estos factores son castigados por una menor precisión, aunque sin decir que es un mal resultado para estas ya que demuestran que también son una solución efectiva.
* Las redes neuronales demostraron dar mejores resultados al evaluar relaciones complejas entre variables.
* Es importante resaltar como en los países con regímenes autoritarios o dictaduras la población tiende a tener menor nivel de riqueza que aquellas con democracias más establecidas.
* Se sugiere que futuras investigaciones consideren la incorporación de datos adicionales y más recientes para mejorar la precisión y relevancia de los modelos. Además, la exploración de otros algoritmos de inteligencia artificial y técnicas de modelado podría ofrecer perspectivas adicionales y mejorar aún más los resultados obtenidos.
* Los modelos desarrollados y el dashboard interactivo tienen el potencial de convertirse en herramientas valiosas para los responsables de la toma de decisiones en políticas económicas. Al proporcionar una comprensión más profunda de los factores que influyen en la riqueza per cápita, se pueden diseñar e implementar estrategias más efectivas para promover el crecimiento económico y reducir la pobreza.

# Limitaciones y prospectiva

## Limitaciones

En esta sección se encuentra un resumen detallado de las limitaciones y problemas hallados durante el desarrollo de este estudio y que pudieron influir en los resultados obtenidos.

* Como primer punto, dado que no se encontraron estudios similares que abordaran las variables que incidían en los niveles de riqueza a nivel de modelos de clasificación ni de regresión en los países, durante la fase de construcción del *dataset*, en la cual se tuvo que consultar diversas fuentes para lograr la estructuración de esto y de esta manera lograr un mayor grado de completitud no fue posible encontrar datos publicados para ciertas regiones. Esto se debe a que dado que principalmente el *dataset* se compone de variables macroeconómicas hay muchos gobiernos que se encargan de ocultarlos, así como hay otros que ni siquiera los calculan, es por esto por lo que se encontró una primera dificultad para la realización de este trabajo.
* Por otro lado, como un segundo punto, se presentó una limitante en el momento de escoger las variables que serían usadas en el estudio. Esto fue un problema ya que se desconocía que variables podrían servir para alimentar los modelos y así obtener unos resultados confiables. Las variables, aunque se escogieron principalmente por estar relacionadas con la generación de ingresos en los países, mercado laboral interno o en la categorización de la riqueza, se desconocía si eran las adecuadas o no.
* Como tercer punto, durante la fase de construcción del *dataset,* se tenía como fuente principal los reportes de riqueza generados por el *Credit Suisse Bank*. Cabe aclarar que estos reportes son publicados en formato PDF en la página web principal de la entidad, por lo que su tratamiento y proceso de extracción implicó una manualidad para poder obtener los datos que se requerían. Adicionalmente, respecto a las otras fuentes, aunque presentaban un almacenamiento más sencillo de manejar como lo es por ejemplo los archivos con extensiones .csv, poseían una dificultad para poder estructurar un proceso de ELT para dichas fuentes ya que para esto se hubiera requerido la construcción de procesos tipo *Web Scrapping* que son los encargados de navegar a través de páginas web y así poder extraer datos de estas, pero que por la dificultad para la construcción de tantos procesos de este tipo se salían del alcance del proyecto y de esta forma quedan para trabajos futuros.

## Trabajo Futuro

Este apartado comprende aquellos puntos que a lo largo del desarrollo del documento se consideraron relevantes pero que no alcanzaron a ser incluidos o desarrollados.

* Como primer aspecto, se considera importante el poder aumentar las instancias disponibles para alimentar los modelos, ya que como se explicó anteriormente, hubo categorías en las que se obtuvo peores resultados que otras, precisamente porque no se contaba con suficientes instancias de entrenamiento.
* Como segundo punto, se sugiere evaluar otros algoritmos de inteligencia artificial como aquellos relacionados con *clustering* para poder ampliar los estudios relacionados con generación de riqueza en los países.
* Como tercer aspecto relacionado con líneas futuras, se sugiere la construcción de los procesos de *Web Scrapping,* para poder tener procesos automatizados encargados desde la extracción de la capa cruda de datos hasta la ejecución de los modelos y la actualización del tablero.
* Un cuarto punto relevante para tener en cuenta es la ejecución de los modelos propuestos, pero incluyendo nuevas variables a las definidas inicialmente, con el fin de identificar si los resultados obtenidos son mejores o peores a los que se lograron en este estudio.
* Como quinto punto, sería un gran avance que este tipo de estudios sirvieran para que los gobiernos tomaran decisiones económicas, ya que los estudios más comunes se enfocan en reducir la pobreza monetaria y no en mejorar los aspectos que pueden ayudar a las personas no solo a salir de ella sino en sentar las bases para un crecimiento sostenido de la riqueza.
* En un sexto punto se sugiere la modificación del *dashboard* propuesto para una versión móvil para que este pueda ser consultado desde dispositivos móviles.

Referencias bibliográficas

ANACONDA. (2024). We’re not just a company; we’re a movement. Obtenido de https://www.anaconda.com/about-us

Ariza, J. & Retajac, A. (2020). *Pobreza monetaria y multidimensional en Colombia: medición, evolución y determinantes,* 14-23, 147.

BANCO MUNDIAL. (2023). *Pobreza Panorama General.* Obtenido de https://www.bancomundial.org/es/topic/poverty/overview

BANCO MUNDIAL. (2024). *Quienes Somos*. Obtenido de https://www.bancomundial.org/es/who-we-are

CREDIT SUISSE. (2022). *Download the Global Wealth Reports and Global Wealth Databook.* Obtenido de https://www.credit-suisse.com/about-us/en/reports-research/global-wealth-report.html

DANE. (2024). *Pobreza monetaria y pobreza monetaria extrema*. Obtenido de https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/pobreza-monetaria/pobreza-monetaria-historicos

Data Science Process Alliance. (2024). *What is CRISP DM?* Obtenido de https://www.datascience-pm.com/crisp-dm-2/

DATOS BANCO MUNDIAL. (2024). *PIB (US$ a precios actuales).* Obtenido de https://datos.bancomundial.org/indicador/NY.GDP.MKTP.CD?end=2022&locations=CO&stars=2001

Datosmacro. (2024). *¿Quiénes somos?.* Obtenido de https://datosmacro.expansion.com/legal/acerca-de

IBM. (2024). *¿Qué es data mining?* Obtenido de https://www.ibm.com/mx-es/topics/data-mining

IBM. (2024). *¿Qué es el Deep learning?* Obtenido de https://www.ibm.com/es-es/topics/deep-learning

IBM. (2024). *¿Qué son las redes neuronales?* https://www.ibm.com/es-es/topics/neural-networks

IBM. (2024). *Conceptos básicos de ayuda de CRISP-DM*. Obtenido de https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

Indexmundi. (2024). *About*. Obtenido de https://www.indexmundi.com/about.html

IDEA. (2024). *Democracy Assessment.* Obtenido de https://www.idea.int/theme/democracy-assessment

Laverde, H. & Corredor, L. (2016). *Medición de la pobreza en Colombia,* 52-54, 78.

Medium. (2024). *CRISP-DM framework: A foundational data mining process model.* Obtenido de https://medium.com/@avikumart\_/crisp-dm-framework-a-foundational-data-mining-process-model-86fe642da18c

MICROSOFT, (2024). *¿Qué es PowerBI?*. Obtenido de https://learn.microsoft.com/es-es/power-bi/fundamentals/power-bi-overview

MongoDB. (2024). *¿Qué es MongoDB?.* Obtenido de https://www.mongodb.com/es/company/what-is-mongodb

Niño, D. & Morera, N. (2018). *Percepción de la pobreza en Colombia en los años 2003 y 2016,* 25-38.

ONG Acción contra el hambre. (2023). *Pobreza extrema: definición, causas y cómo erradicarla.* Obtenido de https://www.accioncontraelhambre.org/es/pobreza-extrema-definicion

Organización mundial de la salud. (2024). *Objetivo de desarrollo del milenio (ODM).* Obtenido de https://www.who.int/es/news-room/fact-sheets/detail/millennium-development-goals-(mdgs)

PNUD. (2024). *Objetivos de desarrollo sostenible.* Obtenido de: https://www.undp.org/es/sustainable-development-goals

Ravallion, M. (1998). *Poverty Lines in Theory and Practice. Living Standards Measurement Study Working Paper 133*.

Repositorio Virtual de los Datos. (2024). GITHUB JUFEMACHADO – tfm. Obtenido de https://raw.githubusercontent.com/jufemachado/TFM/main/Dataset\_tfm.csv

WITS. (2024). Acerca de WITS. Obtenido de: https://wits.worldbank.org/es/about\_wits.html

Worlddata. (2024). About us. Obtenido de https://www.worlddata.info/about.php

WorldEconomics. (2024). *What we do*. Obtenido de https://www.worldeconomics.com/Information/What-We-Do/

1. Código Python

El presente anexo presenta el código de los modelos de inteligencia artificial desarrollado en Python y ejecutado en el *notebook* de Anaconda.

https://raw.githubusercontent.com/jufemachado/TFM/main/TFM.ipynb