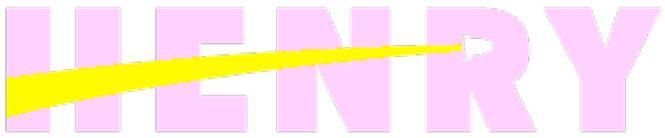

Reporte del Proyecto Final: Expectativa de vida

ML: Gonzalo Ducca <gonzaloducca@gmail.com>
DA: Juan P. Bertone <bertonejpb@gmail.com>
DE: Juan E. Flórez-Coronel <juan.florez@upr.edu>
DA: Valentino Caputa <caputavalentino@gmail.com>
ML: Carlos Madoery <ccmadoery@gmail.com>

Product Owner:
Melina Griffo <meligriffo@gmail.com>

Henry Mentor:
Pía Ruiz <mpiaruiz@gmail.com>

HENRY Data Science Part Time 3

Revision 2.0

15 de noviembre de 2023

Cuadro 1. Historia de Revisión del Documento

Revisión	Fecha	Razón para los Cambios
0.0	10-20-23	Creación del documento
1.0	11-01-23	Documentación Sprint 1
2.0	11-15-23	Documentación Sprint 2

Índice

Índice de cuadros

Índice de figuras

1. Planteamiento del problema

Consultamos bases de datos del Banco Mundial y decidimos utilizar como referencia temas clave que proporcionan indicadores que influyen en la esperanza de vida de los habitantes de un país. Estos temas seleccionados son la educación, la salud, la economía, el desarrollo de la ciencia y la tecnología y el ámbito social. Planeamos recopilar indicadores de cada tema, crear las bases de datos con las que queremos trabajar y luego vincular estas variables para establecer relaciones. Visualizamos este proyecto como un producto que se puede ofrecer a una empresa que busque invertir en ciencia y desarrollo. Una vez completado todo el trabajo de análisis de datos, crearemos modelos de aprendizaje automático especificando ciertos parámetros, como el porcentaje del PIB invertido en educación o ciencia. Estos modelos luego nos proporcionarán una esperanza de vida estimada para un país en particular.

1.1. Objetivos

Analizar la expectativa de vida en 30 paises en base a 5 factores:

- Educación
- Salud
- Economía
- Ciencia y Tecnología
- Social

1.2. Público objetivo

El público obejtivo de este proyecto incluye entidades públicas y privadas con intereses creados en el tema y usuarios cotidianos que quieran conocer la esperanza de vida esperada en un país determinado. Las instituciones públicas, como las agencias gubernamentales, pueden utilizar los conocimientos y modelos generados para informar decisiones políticas y asignar recursos de manera efectiva en áreas relacionadas con la educación, la atención médica y la investigación científica. Por otro lado, las empresas privadas, especialmente aquellas involucradas en industrias relacionadas con la ciencia y la tecnología, pueden beneficiarse de los hallazgos del proyecto al tomar decisiones de inversión e iniciativas de responsabilidad social corporativa. Este proyecto está dirigido a una amplia gama de organizaciones que buscan tomar decisiones informadas que impacten el bienestar y el desarrollo de los países.

1.3. Alcance del proyecto

El alcance del proyecto implica un análisis exhaustivo de la esperanza de vida en 30 países de todo el mundo durante los últimos 35 años. Este análisis abarcará cinco temas clave dentro

de cada uno de los cinco factores influyentes, a saber, educación, salud, economía, ciencia y tecnología y ámbito social. El objetivo es proporcionar una visión detallada y basada en datos de los factores que afectan la esperanza de vida en estos países durante el período de tiempo especificado. El producto final incluirá un servicio web donde se pueda acceder a la información brindada por el análisis y un dashboard interactivo con insights claves para la interpretación de los datos.

2. Obtención de los datos

La información fue obtenida utilizando la siguiente API World Bank API.

2.1. World Bank API

La información de la API de World Bank tiene los siguientes tópicos:

id	value
1	Agriculture & Rural Development
2	Aid Effectiveness
3	Economy & Growth
4	Education
5	Energy & Mining
6	Environment
7	Financial Sector
8	Health
9	Infrastructure
10	Social Protection & Labor
11	Poverty
12	Private Sector
13	Public Sector
14	Science & Technology
15	Social Development
16	Urban Development
17	Gender
18	Millenium development goals
19	Climate Change
20	External Debt
21	Trade

Figure 1. Tópicos World Bank

Y las siguientes economías:

	id	value	aggregate	longitude	latitude	region	adminregion	lendingType	incomeLevel	capitalCity
0	ABW	Aruba	False	-70.0167	12.51670	LCN		LNX	HIC	Oranjestad
1	AFE	Africa Eastern and Southern	True	NaN	NaN					
2	AFG	Afghanistan	False	69.1761	34.52280	SAS	SAS	IDX	LIC	Kabul
3	AFW	Africa Western and Central	True	NaN	NaN					
4	AGO	Angola	False	13.2420	-8.81155	SSF	SSA	IBD	LMC	Luanda
...	...	...	...	...	...	...	...	...	...	...
261	XKX	Kosovo	False	20.9260	42.56500	ECS	ECA	IDX	UMC	Pristina
262	YEM	Yemen, Rep.	False	44.2075	15.35200	MEA	MNA	IDX	LIC	Sana'a
263	ZAF	South Africa	False	28.1871	-25.74600	SSF	SSA	IBD	UMC	Pretoria
264	ZMB	Zambia	False	28.2937	-15.39820	SSF	SSA	IDX	LMC	Lusaka
265	ZWE	Zimbabwe	False	31.0672	-17.83120	SSF	SSA	IDB	LMC	Harare

Figure 2. Economías World Bank

Cada tópico tiene factores asociados:

	id	value
0	BM.GSR.ROYL.CD	Charges for the use of intellectual property, ...
1	BX.GSR.ROYL.CD	Charges for the use of intellectual property, ...
2	TX.VAL.TECH.MF.ZS	High-technology exports (% of manufactured exp...
3	TX.VAL.TECH.CD	High-technology exports (current US$)
4	IP.PAT.NRES	Patent applications, nonresidents
5	IP.PAT.RESD	Patent applications, residents
6	GB.XPD.RSDV.GD.ZS	Research and development expenditure (% of GDP)
7	SP.POP.SCIE.RD.P6	Researchers in R&D (per million people)
8	IP.JRN.ARTC.SC	Scientific and technical journal articles
9	SP.POP.TECH.RD.P6	Technicians in R&D (per million people)

Figure 3. Factores Ciencia y Tecnología World Bank

2.2. Tópicos Posibles

Contando valores faltantes:

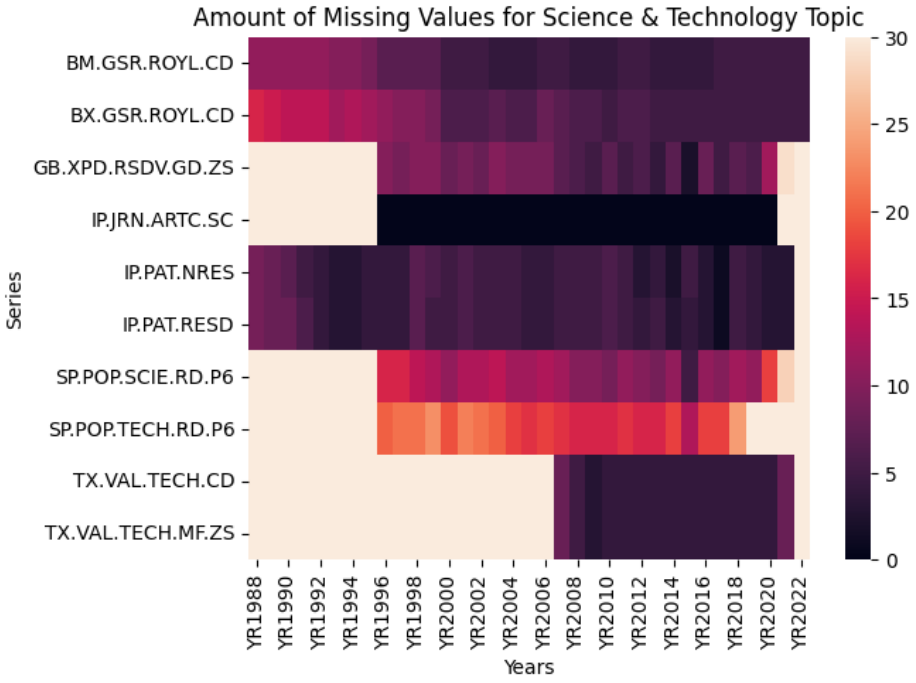

Figure 4. Análisis de valores Nan

3. Stack Tecnológico

Hemos diseñado cuidadosamente nuestro stack tecnológico en base a la experiencia previa en estas herramientas y, lo que es más importante, en el valor que aportarán a nuestro proyecto.

Comencemos con el corazón de cualquier proyecto de data science: el lenguaje de programación. En este caso, hemos elegido Python. ¿Por qué Python? Primero y ante todo, es ampliamente utilizado en data science debido a su rica colección de bibliotecas. Pero más allá de su popularidad, es el lenguaje con el que estamos más familiarizados. Durante nuestra carrera, hemos trabajado extensamente con Python, lo que nos brinda la confianza necesaria para aprovechar al máximo sus capacidades y lograr un rendimiento óptimo en nuestro proyecto.

Para trabajar en este entorno, hemos optado por dos herramientas que consideramos esenciales: Visual Studio Code y Jupyter Notebook. Estas no son elecciones aleatorias, sino que se basan en nuestra experiencia previa. Visual Studio Code es un entorno de desarrollo altamente personalizable y eficiente, ideal para la escritura de código y la gestión de proyectos. Por otro lado, Jupyter Notebook es perfecto para la exploración de datos y la creación de informes interactivos. Al usar estas herramientas, podemos maximizar nuestra productividad y colaboración en el proyecto.

En cuanto a librerías y frameworks, hemos seleccionado las que se han convertido en pilares de data science. NumPy y pandas son esenciales para la manipulación y el análisis de datos, mientras que scikit-learn es nuestro aliado para el machine learning y el modelado. Además, Matplotlib y Seaborn son nuestras herramientas de confianza para la visualización de datos, permitiéndonos comunicar de manera efectiva los resultados de nuestro análisis.

Pero, ¿cómo compartiremos estos resultados? Aquí es donde Power BI entra en juego. Esta potente herramienta nos permitirá crear visualizaciones avanzadas y paneles interactivos, brindando un valor adicional a nuestros informes y ayudando a las partes interesadas a comprender mejor los resultados.

La gestión de proyectos es una parte crucial de cualquier emprendimiento. Para este propósito, hemos optado por Trello, una herramienta de seguimiento de tareas ágil y sencilla. Y, por supuesto, GitHub será nuestro centro de control de versiones y colaboración en código. No solo nos brinda un control de versiones sólido, sino que también nos permite mantener una documentación completa del proyecto, lo que es esencial para la reproducibilidad y la colaboración efectiva. Por último, Monday. Esta plataforma nos permite llevar de forma organizada nuestro diagrama de Gantt.

En lo que respecta a la infraestructura y el alojamiento de recursos, hemos elegido Google Cloud Platform (GCP). GCP nos proporciona una infraestructura sólida y servicios en la nube que son esenciales para alojar nuestros datos y recursos, asegurando que estén disponibles cuando los necesitemos.

Finalmente, la comunicación y la colaboración son fundamentales para el éxito de cualquier proyecto. Por eso, utilizaremos Slack para la comunicación en equipo, Google Meet para

nuestras reuniones y Presentaciones de Google para colaborar en tiempo real en documentos.

En resumen, hemos construido nuestro stack tecnológico cuidadosamente, aprovechando nuestra experiencia y seleccionando las herramientas que consideramos más adecuadas para nuestro proyecto de data science. Al hacerlo, estamos seguros de que maximizaremos nuestra eficiencia y efectividad, permitiéndonos abordar los desafíos que se presenten en el camino.

1. Lenguaje de Programación
    - Python
2. Entorno de Desarrollo
    - Visual Studio Code
    - Jupyter Notebook
3. Librerías y Frameworks
    - NumPy
    - Pandas
    - scikit-learn
    - Matplotlib
    - Seaborn:
4. Herramientas de Visualización
    - Matplotlib
    - Seaborn
    - Power BI
5. Herramientas de Gestión de Proyectos
    - Trello
    - GitHub
    - Monday
6. Infraestructura y Cloud
    - Google Cloud Platform (GCP)
7. Comunicación y colaboración
    - Slack
    - Meet
    - Presentaciones de Google

4. Desarrollo del Proyecto

4.1. Roles del Equipo

Cuadro 2. Roles del Equipo

Role	Name
Machine Learning	Gonzalo Ducca
Machine Learning	Carlos Madoery
Data Analytics	Valentino Caputa
Data Analytics	Juan P. Bertone
Data Engineer	Juan E. Flórez-Coronel

4.2. Diagrama de Gaant

Esperanza de vida

Sprint 1

Nombre	Responsable	Fecha	Estado	Cronograma - Start	Cronograma - End
KPI	caputavalentino@gmail.com	2023-11-01	Listo	2023-10-23	2023-11-01
Stack tecnológico	Juan Pablo Bertone	2023-10-27	Listo	2023-10-23	2023-11-01
Documentación	Gonzalo Ducca, Carlos Madoery	2023-10-28	En curso	2023-10-23	2023-11-01
EDA preliminar	Juan E Florez-Coronel	2023-10-31	Listo	2023-10-23	2023-11-01
Github	Juan E Florez-Coronel	2023-10-27	Listo	2023-10-23	2023-11-01
		Desde el 2023-10-27 hasta el 2023-11-01		**2023-10-23**	**2023-11-01**

Sprint 2

Nombre	Responsable	Fecha	Estado	Cronograma - Start	Cronograma - End
DW automatizado	Juan E Florez-Coronel	2023-11-07	No iniciado	2023-11-01	2023-11-08
Tablas	Juan E Florez-Coronel, Gonzalo Ducca		No iniciado	2023-11-01	2023-11-08
Carga incremental	Carlos Madoery	2023-11-10	No iniciado	2023-11-08	2023-11-15
Big Data y servicios cloud.	Juan Pablo Bertone	2023-11-13	No iniciado	2023-11-08	2023-11-15
		Desde el 2023-11-07 hasta el 2023-11-13		**2023-11-01**	**2023-11-15**

Sprint 3

Nombre	Responsable	Fecha	Estado	Cronograma - Start	Cronograma - End
Storytellng	caputavalentino@gmail.com, Juan Pablo Bertone	2023-11-28	No iniciado	2023-11-15	2023-11-29
Dashbard y reportes	ıalentino@gmail.com, Juan E Florez-Coronel, Juan Pablo E	2023-11-25	No iniciado	2023-11-15	2023-11-29
Recomendaciones	Carlos Madoery	2023-11-29	No iniciado	2023-11-15	2023-11-29
KPIs y datos hallados	Carlos Madoery	2023-11-27	No iniciado	2023-11-15	2023-11-29
Retomar hitos	Gonzalo Ducca	2023-11-25	No iniciado	2023-11-15	2023-11-29
Modelos ML	Gonzalo Ducca, Carlos Madoery	2023-11-21	No iniciado	2023-11-15	2023-11-29
Reporte Vis. geografica	Juan Pablo Bertone	2023-11-23	No iniciado	2023-11-15	2023-11-29
		Desde el 2023-11-21 hasta el 2023-11-29		**2023-11-15**	**2023-11-29**

Figure 5. Diagrama de Gantt

4.3. Demo 2

Hasta el sprint anterior trabajamos sobre los objetivos y alcance del proyecto, el stack tecnológico implementado, la elaboración de un EDA preliminar, determinar los KPIs y la construcción de un diagrama de Gantt que nos permita organizar el plan de trabajo. En esta nuevo demo se abordará: la realización de las bases de datos y el DER, el flujo de los datos, el servicio Cloud implementado junto a la carga incremental de los datos, un mockup de dashboard

y una idea general sobre el modelo de Machine Learning que ofreceremos cómo producto al cliente. El origen de las bases de datos es World Bank. Se eligieron 5 tópicos principales que consideramos afectan directamente a la esperanza de vida de un país: economía, educación, salud, desarrollo social y ciencia y tecnología. Vinculados a estos tópicos elaboramos los KPIs y se tomaron indicadores entre los que se encuentran: gasto público en educación, total (% del PIB), esperanza de vida al nacer (años), prevalencia de la desnutrición (% de la población), gasto en investigación y desarrollo (% del PIB), homicidios intencionados (por cada 100.000 personas), entre otros. Nos encontramos que muchos de los indicadores presentaban nulos en el primer intervalo de años o en la totalidad de los 30 años a estudiar, motivo por el cual decidimos descartar algunos. A la hora de elaborar el diagrama entidad-relación, se construyeron 4 tablas: 1 de hechos y 3 de dimensiones. Inicialmente se confeccionó una tabla para ser usada con el modelo ML que contenía los indicadores en la columnas y pensarla cómo tabla de hecho, pero se decidió modificarla con las columnas id, id_pais, id_indicador, año y valor. Continuando con el DER, se visualizan 3 tablas dimensionales: pais (id_pais, nombre, longitud, latitud, región, capital), indicador (id_indicador, id_topico y descripción) y tópico (id_topico, nombre).

4.4. Google Cloud Platform

4.4.1. Justificación

La decisión de implementar Google Cloud Platform (GCP) en nuestro entorno de trabajo responde a una estrategia destinada a optimizar el trabajo colaborativo. A continuación, se presentan algunas justificaciones clave para esta elección:

1. Escalabilidad y Flexibilidad Google Cloud Platform ofrece una arquitectura altamente escalable que nos permite ajustar nuestros recursos de manera dinámica según los objetivos y requisitos del trabajo.

2. Amplio Conjunto de Servicios GCP proporciona una amplia gama de servicios en la nube que abarcan desde cómputo, almacenamiento y bases de datos hasta inteligencia artificial, aprendizaje automático y análisis de datos.

3. Seguridad y Cumplimiento La seguridad es una prioridad fundamental, y GCP ofrece robustas medidas de seguridad para proteger nuestros datos y aplicaciones.

4. Colaboración y Conectividad GCP facilita la colaboración entre equipos al proporcionar herramientas y servicios diseñados para un desarrollo ágil y una integración continua.

5. Innovación Continua GCP se caracteriza por su compromiso con la innovación constante. Al adoptar esta plataforma, nos aseguramos de tener acceso a las últimas tecnologías y actualizaciones, lo que nos permite mantenernos a la vanguardia en un entorno tecnológico en constante evolución.

4.4.2. Implementación

En el contexto de nuestro trabajo práctico, hemos diseñado una estrategia integral que involucra tanto los servicios de Google Cloud Platform (GCP) como la integración de la API de World Bank. A continuación, se detalla la estructura de nuestra implementación:

1. API de World Bank En primer lugar, optaremos por la integración de la API de World Bank, la cual nos proporcionará un acceso directo y actualizado a los datos.

2. Cloud Storage Como componente esencial, emplearemos Cloud Storage para almacenar de manera segura y eficiente los datos recopilados a través de la API de World Bank. Será nuestro Data Lake, donde la información se encontrará almacenada tal cual se recibe de la API. La API será invocada con una frecuencia de 4 semanas, con el objetivo de mantener la base actualizada.

3. Cloud Functions: La capa de automatización estará respaldada por Cloud Functions, permitiéndonos ejecutar código sin servidor de manera eficiente. Al aprovechar este servicio, garantizamos la automatización de procesos clave en respuesta a eventos específicos, asegurando una operación fluida y eficiente de nuestro sistema. Cada vez que los archivos se actualicen por la invocación a la API, se dispararán las funciones del cloud. Estas funciones tienen la tarea de realizar la limpieza de los datos y de trasnferirlos en el formato de tabla a BigQuery.

4. BigQuery Como motor SQL, BigQuery continuará desempeñando un papel fundamental en nuestro enfoque. La información almacenada en el motor mediante las funciones del cloud será el punto de partida para el análisis de los datos y la generación de modelos.

5. Looker Studio La fase final de nuestra implementación involucrará Looker Studio para la visualización y presentación de resultados. Looker Studio nos brindará la capacidad de crear paneles visuales intuitivos, facilitando la interpretación y comunicación efectiva de los hallazgos derivados de nuestros análisis.

4.5. Propuesta Machine Learning

Se propone la creación de un modelo de Machine Learning (ML) destinado a predecir la esperanza de vida en un país determinado. Utilizando los indicadores definidos como el conjunto de variables predictoras. El modelo buscará ofrecer una herramienta de predicción precisa y útil para diferentes audiencias, desde usuarios cotidianos hasta gobiernos interesados en comprender el impacto de cambios específicos en variables socioeconómicas.

La base técnica de nuestro modelo se fundamentará en el aprendizaje supervisado de regresión. Este tipo de modelo es idóneo para predecir valores continuos, como la esperanza de vida, a partir de variables predictoras.

La entrada al modelo será una matriz de características que represente las variables predictoras definidas. Cada columna de esta matriz corresponderá a un indicador específico.

El diseño del modelo contempla la posibilidad de incorporar variables categóricas. Las variables categóricas pueden proporcionar información adicional y enriquecer la capacidad predictiva del modelo. Por ejemplo, variables como la región geográfica, género o el sistema de gobierno podrían ser consideradas como variables categóricas que influyen en la esperanza de vida. Estas se codificarán adecuadamente para su inclusión en la matriz de características.

Productos Diferenciados: El modelo de predicción se diseñará para producir dos resultados diferenciados, adaptados a distintas necesidades y audiencias:

1- Producto para Usuarios Cotidianos: Este producto estará diseñado para satisfacer las necesidades de los usuarios cotidianos que desean conocer la esperanza de vida esperada en un país específico y en un año determinado. El input para este usuario será simple: el nombre del país y el año de interés. La interfaz será amigable y fácil de usar, proporcionando de manera clara y rápida la información buscada.

2- Producto para Gobiernos: El segundo producto se orientará a gobiernos y entidades interesadas en entender el impacto de cambios específicos en las variables predictoras sobre la esperanza de vida. Permitirá definir ajustes porcentuales en una o más variables predictoras y visualizar la esperanza de vida resultante. Este enfoque proporcionará a los gobiernos una herramienta valiosa para la toma de decisiones informadas y la planificación estratégica.

Beneficios y Aplicaciones:

- **Información Accesible:** Ofrecerá a los usuarios una visión rápida y accesible de la esperanza de vida, proporcionando información crucial para la planificación personal y la toma de decisiones informadas.
- **Planificación Estratégica:** Permitirá a gobiernos y entidades planificar estrategias basadas en la comprensión del impacto de cambios en variables socioeconómicas, facilitando la toma de decisiones y políticas más efectivas.
- **Adaptabilidad:** El modelo será diseñado para ser adaptable a diferentes contextos y regiones, brindando información valiosa a nivel global.