AyC

EMPRESA CONTRATANTE:

Banco IPB

OBJETIVO DEL PROYECTO:

Asesoramiento Integral, estudio de mercado y desarrollo de modelo predictivo acerca del Riesgo Crediticio

Julio 2024

Índice

1. **INTRODUCCIÓN**
2. **OBJETIVOS**
   1. Objetivo General
   2. Objetivos Específicos
3. **FUNDAMENTACIÓN**
4. **METODOLOGIA DE TRABAJO**
   1. **Sprint 0:**
   2. **Sprint 1:**
   3. **Sprint 2:**
   4. **Sprint 3:**
5. **EQUIPO DE TRABAJO**
6. **STACK TECNOLÓGICO**
7. **DESARROLLO**
   1. Extracting, Transforming, Loading
   2. Exploratory Data Analisys
   3. Key Performance Indicators
   4. Machine Learning
   5. Deploy
8. **CONCLUSIONES**
9. **DICCIONARIO DE DATOS**
10. INTRODUCCIÓN
11. OBJETIVOS
    1. Objetivo General

* La construcción de un marco de análisis y predicción que permita la toma de decisiones en referencia al otorgamiento o no de créditos en base al perfil y estado de la empresa-potencial cliente y su cálculo de riesgo crediticio.
  1. Objetivos Específicos
* La construccion de un marco analitico donde se puedan realizar la toma de decision principal (el otorgamiento y autorizacion del credito) tomando en cuenta riesgo y garantia.
* El estudio exhaustivo de la situacion general y potencial de la empresa en cuestion, que abarque aspectos de capital, financieros, contables, de sector, de produccion, etc
* La construccion de un modelo o sistema que permita una mirada concisa, precisa y predictiva del comportamiento de la empresa para/con el cumplimiento de pago del prestamo.
* El asesoramiento de ideas y conclusiones que complementen el marco analitico con informacion util y argumentos concretos para la toma de decisiones.

1. FUNDAMENTACIÓN

Respecto a la decisión de los datos:

El contexto economico y politico en Argentina plantea varios aspectos puntuales a considerar de manera particular para la seleccion de nuestros datos: El primero es el marco juridico que viene a ser implementado por leyes y regulaciones, entre los que se encuentra el RIGI (Régimen de Incentivo para las Grandes Inversiones), el segundo a destacar es el descubrimiento y exploracion del yacimiento en Vaca Muerta para el sector de Hidrocarburos y su reelevancia, y el tercero es la necesidad de una gran inversion por parte de empresas de hidrocarburos para explotar el potencial de recursos en Argentina en general, y en Vaca Muerta en particular.  
  
Por estos motivos nos decantamos por utilizar a YPF como objeto de estudio principal, poniendo en contexto su analisis mediante la obtencion de datos de relevancia del mercado internacional, de diversos aspectos referidos a la industria de hidrocarburos en el ambito nacional y de aspectos financieros, contables y de capital en referencia a la empresa en cuestion (que pasa a ocupar en este caso, el rol de empresa que pide el credito a nuestro cliente)

Para el desarrollo de nuestro estudio, obtenemos datos de diversas fuentes nacionales (en referencia al estado nacional) y de acceso publico por parte del gobierno, de distintas bases de datos que contienen ficheros y archivos que hacen referencia a data del ambito internacional concerniente a hidrocarburos. Y por ultimo, ingresamos a una plataforma que nos brinda el acceso publico a datos contables, financieros y de capital de YPF en los ultimos años.

Mencionamos las fuentes de los datos:

* <https://kaggle.com/>
* <https://datos.gob.ar/>
* <https://www.argentina.gob.ar/>
* [https://nasdaq.com](https://nasdaq.com/)
* <https://inversores.ypf.com/informacion-financiera.html>
* <https://datos.enerdata.net/productos-petroliferos/estadisticas-consumo-mundial-petroleo-consumo-domestico.html>
* <https://datos.enerdata.net/petroleo-crudo/datos-produccion-energia-mundial.html>

Respecto a la elección del stack tecnológico:

El entorno de Google Colaboratory nos brinda los recursos necesarios para el análisis y el acceso a el trabajo colaborativo en conjunto con la facilidad de uso, la conexión nativa con Google Drive, y su compatibilidad con el lenguaje de programación Python (Adecuado para su uso en el análisis de datos)

Respecto al proyecto realizado:

Vertical Tecnológica: Data – BI

* Descripción

La vertical de Data-BI se enfoca en la recopilación, análisis y visualización de datos para tomar decisiones informadas y estratégicas en una organización. Esta vertical abarca desde la recopilación y limpieza de datos hasta el análisis avanzado y la presentación de insights a través de herramientas de Business Intelligence (BI) y Data Visualization.

* Objetivo de Proyecto

El objetivo del proyecto de Data-BI es utilizar datos para generar insights y conocimientos que impulsen la toma de decisiones informadas en una organización. Esto puede incluir la creación de informes ejecutivos, paneles interactivos, modelos predictivos y mucho más.

1. METODOLOGÍA DE TRABAJO

Sprint

El concepto de Sprint es el corazón de las metodologías ágiles que sirven para avanzar de forma ordenada y gradual hacia la consecución de los objetivos, mostrando el progreso de su proyecto en todos sus aspectos: funcional, estético, responsive, etc. La finalidad detrás es que el “cliente” pueda ver que el proyecto avanza y que el equipo no deje todo para el último día de la entrega final.

Cada sprint tendrá una demo con sus TLs y serán 4.

● Semana 1: Jueves.

● Semana 2: Jueves.

● Semana 3: Jueves.

● Semana 4: Último previo al Demo Day

Los principios de Lean Startup promueven la experimentación rápida y el aprendizaje validado. Esto implica estar preparados para enfrentar desafíos y conflictos, resaltando la importancia de la flexibilidad y la adaptabilidad como soft skills fundamentales para el éxito en proyectos tecnológicos.

No Country ejercerá el rol de “cliente”, esto tiene como intención simular el proceso de entorno laboral real donde el equipo debe realizar una demostración al cliente. En este contexto cada Team Leader y los diferentes canales de Slack de la comunidad representan los intereses del cliente.

MVP Mindset

Es decir, Producto Mínimo Viable (MVP, por sus siglas en inglés) como enfoque fundamental del proyecto. Esto implica la creación de soluciones que aborden las necesidades críticas de manera efectiva, sin complicaciones innecesarias. Durante la simulación, es crucial adoptar esta

mentalidad, enfocándonos en lo esencial para lograr resultados tangibles y valiosos.

Recuerda, el MVP es el primer paso hacia la innovación continua y la iteración basada

en el feedback del usuario.

Gantt y Kanban

La metodología Kanban se implementa por medio de tableros Kanban. Se trata de un método visual de gestión de proyectos que permite a los equipos visualizar sus flujos de trabajo y la carga de trabajo. En un tablero Kanban, el trabajo se muestra en un proyecto en forma de tablero organizado por columnas.

¿Qué es un diagrama de Gantt? Un diagrama de Gantt es una herramienta de gestión de proyectos que ilustra el trabajo realizado durante un período de tiempo en relación con el tiempo previsto para el trabajo.

Github nos provee en su entorno estas herramientas, que dan soporte a la organización de tareas y segmentación de trabajos a lo largo del periodo de desarrollo

1. EQUIPO DE TRABAJO

Nuestro equipo de trabajo se compone de los siguientes integrantes, cada uno abocado a una etapa y cumpliendo un rol esencial en el desarrollo del proyecto:



|  |  |  |
| --- | --- | --- |
| Marco Caro  Data Engineer/Data Scientist | Ángel Troncoso  Business Intelligence | Malena Jara  Developer |

|  |  |
| --- | --- |
| Ana Herrera  Data Analyst | Jatzua Chacon Rico Data Analyst |

**Marco Caro:** Como Data Engineer, tuvo como responsabilidad la búsqueda y extracción/descarga de los datos y material de trabajo para el Desarrollo del proyecto.

Tuvo a su cargo todo el proceso de ETL de la data, la estructuración del Directorio de ficheros (clasificando los mismos en raw y stage), la construcción de las funciones necesarias para el proceso, y la construcción y Desarrollo de los diferentes informes en las distintas instancias en que se encontraba el proceso de formateo/normalizacion/estructuración de la data (como Podemos ver en los diferentes informes preliminares a los hitos)

En su rol de Data Scientist,

A su vez, tuvo como responsabilidad la gestion de las diferentes Herramientas de organizacion y gestion del proyecto, en sus distintas etapas e hitos. Estos fueron:

* Construcción y gestion de pizarra de tareas (Kanbas)
* Construcción y gestion de Diagrama de Gantt
* Gestion de Repositorio en Github
* Gestion de repositorio temporal en Drive (para la construcción de entorno de trabajo del proceso de Data Engineering)

**Ángel Troncoso:**

**Malena Jara:**

**Ana Herrera:**

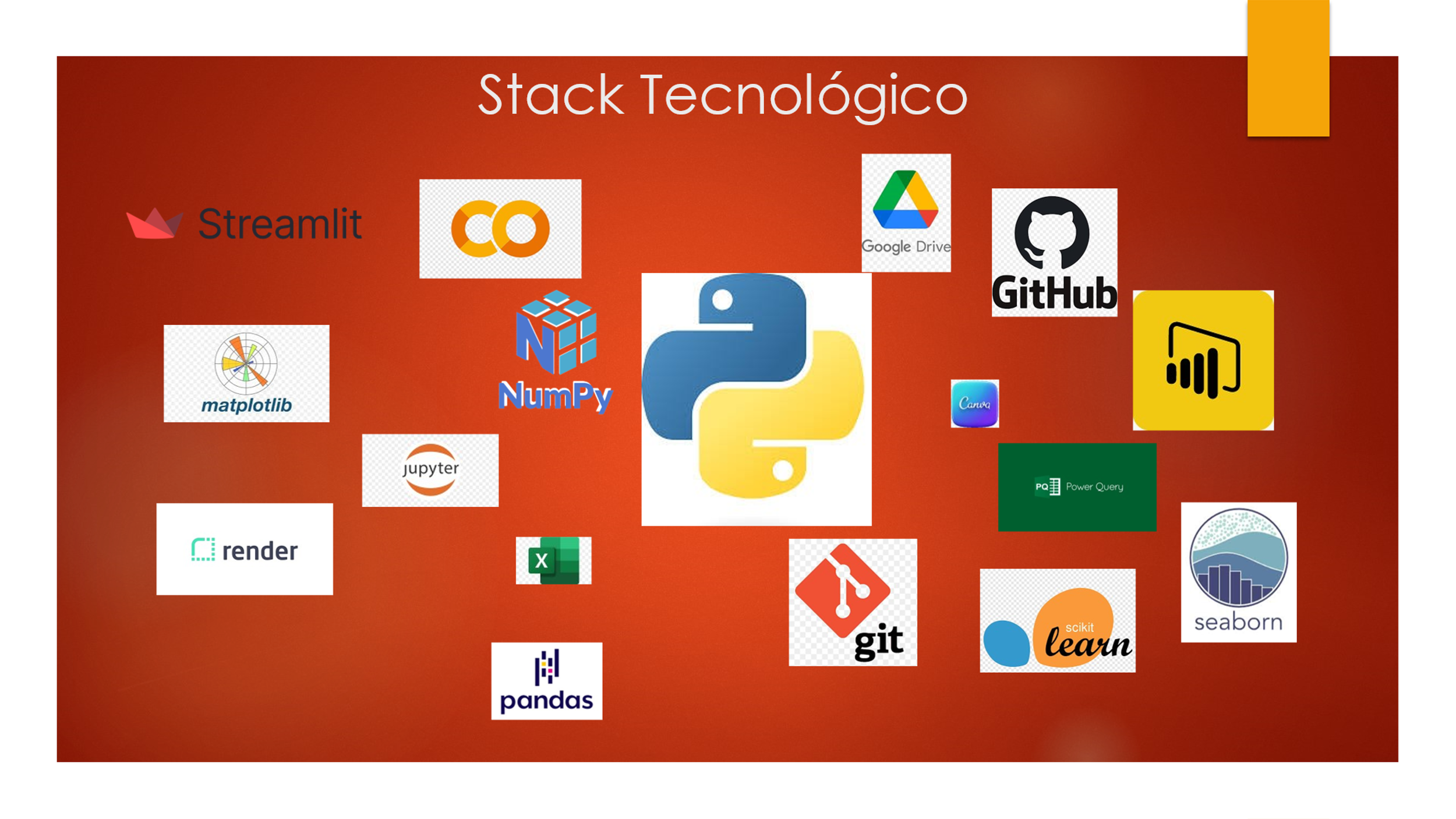
**Jatzua Chacon Rico:**

1. STACK TECNOLÓGICO

Citaremos algunas de las tecnologías, para entrar en el contexto de su funcionamiento.

* **Python:** Es un lenguaje de programación de alto nivel, interpretado y multipropósito. Python será utilizado en todas las etapas del proyecto debido a su simplicidad, versatilidad y a las poderosas bibliotecas que ofrece. Dentro de ella trabajaremos con NumPy, Pandas, Matplotlib, Seaborn, Scikit-Learn, entre otras.
* **Numpy:** NumPy es una biblioteca para el lenguaje de programación Python que da soporte para crear vectores y matrices grandes multidimensionales, junto con una gran colección de funciones matemáticas de alto nivel para operar con ellas
* **Pandas:** Pandas es una librería de Python especializada en la manipulación y el análisis de datos. Ofrece estructuras de datos y operaciones para manipular tablas numéricas y series temporales, es como el Excel de Python
* **Matplotlib:** Matplotlib es una [biblioteca](https://es.wikipedia.org/wiki/Biblioteca_(programaci%C3%B3n)) para la generación de gráficos en dos dimensiones, a partir de datos contenidos en listas o arrays en el lenguaje de programación [Python](https://es.wikipedia.org/wiki/Python)
* **Seaborn:** Seaborn es una biblioteca para crear gráficos estadísticos en Python . Está basado en Matplotlib y se integra con las estructuras de Pandas. Es tan poderosa como Matplotlib pero aporta simplicidad y nuevas funciones. Nos permite explorar y comprender datos rápidamente. Se pueden capturar marcos de datos completos y las funciones internas para mapeo semántico y agregación estadística permiten convertir los datos en visualizaciones gráficas.
* **Scikit-Learn:** Scikit-learn (anteriormente scikits.learn y también conocido como sklearn) es una biblioteca de aprendizaje automático gratuita y de código abierto para el lenguaje de programación Python. Cuenta con varios algoritmos de clasificación, regresión y agrupamiento, incluidas máquinas de vectores de soporte, bosques aleatorios, refuerzo de gradiente, k-means y DBSCAN, y está diseñado para interoperar con las bibliotecas numéricas y científicas de Python NumPy y SciPy.
* **Google Colaboratory:** Colab es un servicio alojado de Jupyter Notebook que no requiere configuración para su uso y proporciona acceso gratuito a recursos informáticos, incluidas GPU y TPU. Colab es especialmente adecuado para el aprendizaje automático, la ciencia de datos y la educación.
* **Jupyter Notebook:** Es un entorno informático [interactivo basado en la web](https://es.wikipedia.org/wiki/Rich_Internet_application) para crear documentos de Jupyter notebook. El término "notebook" puede hacer referencia coloquialmente a muchas entidades diferentes, principalmente la aplicación web Jupyter, el servidor web Jupyter Python o el formato de documento Jupyter según el contexto. Un documento de Jupyter Notebook es un documento [JSON](https://es.wikipedia.org/wiki/JSON), que sigue un esquema versionado y que contiene una lista ordenada de celdas de entrada/salida que pueden contener código, texto (usando [Markdown](https://es.wikipedia.org/wiki/Markdown)), matemáticas, gráficos y texto enriquecidos, generalmente terminado con la extensión ".ipynb".
* **Google Drive:** Es un servicio de alojamiento y sincronización de archivos desarrollado por Google. Lanzado el 24 de abril del 2012, el servicio permite a sus usuarios almacenar archivos en la nube, sincronizar archivos entre dispositivos y compartir archivos.

* **Power BI:** Es una plataforma de análisis que permite visualizar y compartir datos de manera efectiva para tomar decisiones informadas. Ofrece una amplia gama de herramientas para la preparación de datos, visualización de datos, análisis y colaboración en un solo lugar.
* **Streamlit**: Es una biblioteca de python que permite crear aplicaciones web interactivas para el análisis de datos y la visualización de manera rápida y sencilla. Ofrece una amplia gama de widgets para la entrada de datos, gráficos interactivos y capacidades de visualización, lo que la convierte en una herramienta poderosa para la creación rápida de prototipos y la implementación de aplicaciones de análisis y ciencia de datos.
* **Github:** GitHub es una forja (o repositorio) para alojar proyectos utilizando el sistema de control de versiones Git. Se utiliza principalmente para la creación de código fuente de programas de ordenador.
* **Microsoft Excel:** Excel es una potente aplicación de hoja de cálculo desarrollada por Microsoft que permite a los usuarios crear, organizar y analizar datos en formato de tabla, lo que facilita tareas como cálculos matemáticos, compresión de estadísticas, creación de gráficos y gestión de información financiera.
* **Render:** Render es una nube unificada para crear y ejecutar todas sus aplicaciones y sitios web con certificados TLS gratuitos, CDN global, redes privadas e implementaciones automáticas desde Git.
* **Canva:** Canva es una plataforma de diseño gráfico que proporciona herramientas para crear gráficos para redes sociales, presentaciones, productos promocionales y sitios web. Lanzado en Australia en 2013, el servicio ofrece herramientas de diseño para particulares y empresas.



1. DESARROLLO
2. CONCLUSIONES
3. DICCIONARIO DE DATOS