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/ETL Developer/Power BI Developer/Developer | **Ángel Troncoso**  Business Intelligence | **Malena Jara**  Developer |

|  |  |
| --- | --- |
| **Ana Herrera**  Data Analyst/Business Intelligence | **Jatzua Chacon Rico** Data Analyst/Business Intelligence |

**Marco Caro:** Como Data Engineer y ETL Developer, 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/normalización/estructuración de la data (como podemos ver en los diferentes informes preliminares a los hitos).

Así mismo se encargó de la construcción del modelo ER (Modelo Entidad-Relación) y en construir las relaciones entre las distintas tablas y ficheros para su posterior carga y establecimiento en algún tipo de BBDD Relacional, como también para facilitar el manejo de dichos datos en el Desarrollo de una interfaz gráfica (en este caso, en Power BI).

En su rol de Data Scientist, se encargó de realizar la ingeniería preliminar necesaria para el modelado y normalización de la data a consumir por el modelo (feature engineer), realizó las pruebas de elección del modelo puntualizando en la naturaleza de los datos, el feature tomado como target, el tamaño y heterogeneidad de la muestra así como el rendimiento puntual de cada modelo a través de pruebas y calibraciones (Grid Search, Validación Cruzada) y la observación de los resultados de las distintas métricas y errores de los mismos.

Una vez elegido el modelo, se encargó del entrenamiento, el testeo y estudios generales de Rendimiento del modelo en producción.

A su vez, tuvo como responsabilidad la gestión de las diferentes herramientas de organización y gestión del proyecto, en sus distintas etapas e hitos, así como diversas partes puntuales del producto final. Estos fueron:

* Construcción y gestión de pizarra de tareas (Kanban).
* Construcción y gestión de Diagrama de Gantt.
* Gestión de repositorio temporal en Drive (para la construcción de entorno de trabajo del proceso de Data Engineering).
* Construcción y desarrollo de documentación del proyecto.
* Construcción, desarrollo y gestión del repositorio del proyecto.
* Creación de cliente, consultora y todo lo relacionado con los mismos (información, logos, banners, etc.).
* Creación y Desarrollo de Reporte Dashboard.
* Creación y Desarrollo de app para despliegue de Modelo Predictivo de Riesgo Crediticio
* Alta del servicio y despliegue de modelo predictive en Render.

**Ángel Troncoso:** Realizó el estudio contable y financiero de la empresa YPF y contribuyo a su informe de balance y estado de resultado. También realizó la construcción preliminar de los ratios de la empresa y contribuyó a la búsqueda de información financiera y contable de empresas del sector de hidrocarburos.

Dio apoyo a la construcción de la presentación del producto final realizando una presentación preliminar y aportando ideas para el speech de la misma.

**Malena Jara:** Se encargó del desarrollo de las presentaciones para las diferentes etapas del desarrollo del proyecto, así como de la presentación del producto final y su disponibilización en formato de documento.

Colaboró en la búsqueda de fuentes de datos para el desarrollo del proyecto y de estados contables y financieros de empresas del sector de hidrocarburos, para su análisis y posterior utilización en la etapa analítica y de construcción del modelo predictivo.

**Ana Herrera:** Estuvo a cargo de una parte del Análisis Exploratorio de los datos (EDA), donde desarrollo el estudio y análisis de los datos concernientes al ámbito local del sector de hidrocarburos (recordemos que el proyecto se desarrolla teniendo en miras la medición del riesgo crediticio de la empresa YPF, por lo que contexto local refiere a Argentina), y también de los datos concernientes a la empresa YPF.

A su vez, se ocupó de la tarea de búsqueda de datos referentes al estado financiero y contable de diferentes empresas del sector de hidrocarburos. Fue la encargada principal del análisis financiero y contable, de la construcción de los ratios de las empresas y del estudio de las diferentes condiciones que presentaban cada una, para establecer un marco de estudio y contexto en donde realizar la comparación del estado de situación contable/financiera de la empresa YPF.

Por último, realizo la presentación formal del producto terminado a los clientes, realizando una puesta en escena de una síntesis del recorrido atravesado por el equipo en las distintas etapas del desarrollo (acompaña dicha presentación con la muestra de diapositivas referentes a la presentación de negocio de nuestra consultora, así como del producto realizado)

**Jatzua Chacon Rico:** Estuvo a cargo del desarrollo del Diccionario de los datos trabajados en nuestro proyecto, y a su presentación en un formato inteligible y practico para cualquier usuario.

También fue responsable de parte del Análisis Exploratorio de los datos (EDA), más precisamente al estudio y análisis de los datos referidos a un contexto general y global del sector de hidrocarburos y data relacionada.

Aportó fuentes de datos tanto en lo referido al sector de hidrocarburos, como a la situación contable y financiera de empresas del mismo sector.

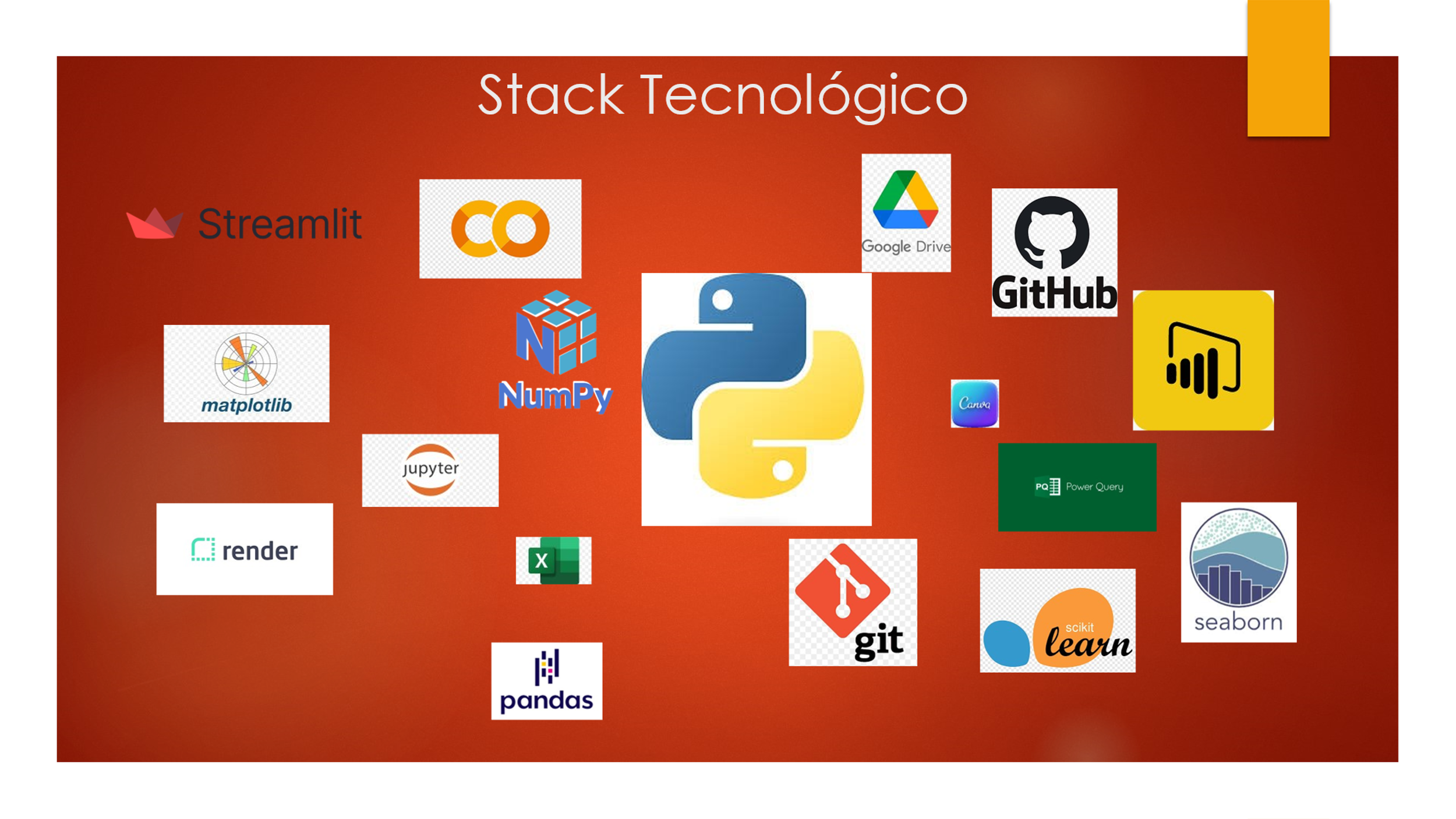
Por último, realizo la construcción del informe del análisis realizado, donde se presentan las conclusiones a las que llegó nuestro equipo a través del estudio de los datos y de los diferentes benchmarks del ámbito del petróleo y gas.

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



* 1. Extracting, Transforming, Loading (ETL)

En Principio, los datos con los que contamos y vamos a trabajar referidos a la tematica de proyecto, son de diferentes origenes, aunque principalmente se obtienen de la plataforma web de datos publicos de la nacion (tanto en Datos Argentina como en los diferentes ministerios o la plataforma de la nación).

Detallamos los origenes principales:

* <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>

Entre otras, vale la pena señalar que ademas se utilizó informes e informacion relacionada a la parte del proyecto que apunta mas precisamente a averiguar la capacidad de pago de crédito potencial (informes contables, estudios, balances y estados de resultado de empresas del sector, otros estudios del área, auditorias estatales, etc.) para hacernos una idea de la situación contable y financiera de la empresa.

Como se hace notar, nos centramos en la data de empresas petroleras en un contexto general medio. Apuntamos a el ambito dentro de argentina (el objeto de estudio se centra principalmente en la empresa YPF), aunque complementamos la misma con algunos datos referentes al área (de hidrocarburos) de otras empresas no radicadas en Argentina, y en el contexto global apuntamos a informacion general del sector de hidrocarburos en un contexto amplio para tomar como punto de referencia en diversos analisis.

La data apunta a el estado general de actividad productiva, el estado financiero y/o contable de la empresa y su capital-infraestructura, la proyeccion de produccion y crecimiento, las inversiones que realiza, la cotizacion en bolsa, etc., todo esto, pertinente para tener un panorama claro del estado general de la empresa a evaluar.

Los datos con los que se trabaja principalmente son:

* Inversiones de empresas petroleras en Argentina (desde el año 2020 hasta el año 2024)
* Informacion contable sobre YPF
* Cotizaciones de diversas empresas petroleras/ de hidrocarburos en bolsas de EEUU
* Produccion de gas y petroleo en Argentina (desde 2020 hasta 2024)
* Precio internacional del barril de crudo (desde el año 1970)
* Consumo internacional de hidrocarburos por pais
* Reservas de hidrocarburos en Argentina
* Pozos registrados en el país
* Rentas del petroleo en Argentina (en % de PBI)
* Informacion contable y financiera de diversas empresas de hidrocarburos

Mención aparte a los informes o analisis y benchmark de la temática que utilizamos y que no se registran como datos, ya que nos sirven para tener conocimiento y guia del tema en analisis y poner en contexto y terminologia, asi como mecanismos y utilizacion de herramientas y logica al material de datos con el que contamos.

Sumado a lo anterior, los informes y noticias también ponen en perspectiva la mirada a futuro en el desarrollo potencial de la entidad que pide el crédito, aunque esto se toma mas como un factor que agrega valor al análisis y a la construcción del contexto y marco en donde se produce el estudio, mas que como factor decisivo.

En algunos casos los ficheros y/o datos presentaban un formato no adecuado para su transformacion y manipulacion directa por lo que se debia realizar una instancia previa, la extraccion de los datos en formato de tabla, para su posterior manipulación y conversión.

Este proceso se llevo a cabo mediante la utilización de diferentes funciones desarrolladas a través del lenguaje Python, las cuales apuntaban a la extracción de manera integral de los datos pertinentes con formato de tabla y en su condición integra y detallada (señalamos esto a causa de que en muchas de las tablas dinámicas o de ficheros que contienen formatos extras a tablas, se establecían segmentaciones y filtros asi como diferentes tipos de macros e interfaces que complicaban su debida extracción)

El desarrollo de dichas funciones se puede ver mas detalladamente en el fichero [resources.py](https://github.com/No-Country-simulation/c19-108-m-data-bi/blob/main/notebooks/resources.py).

Hasta el momento se realizó una descripción de la fase de búsqueda y extracción de los datos, pasando a la fase de Transformación de Datos (ETL), se llevaron a cabo una serie de procedimientos para garantizar la preparación adecuada y la limpieza exhaustiva de los datos antes de su carga en el almacén de datos. Estas acciones incluyeron:

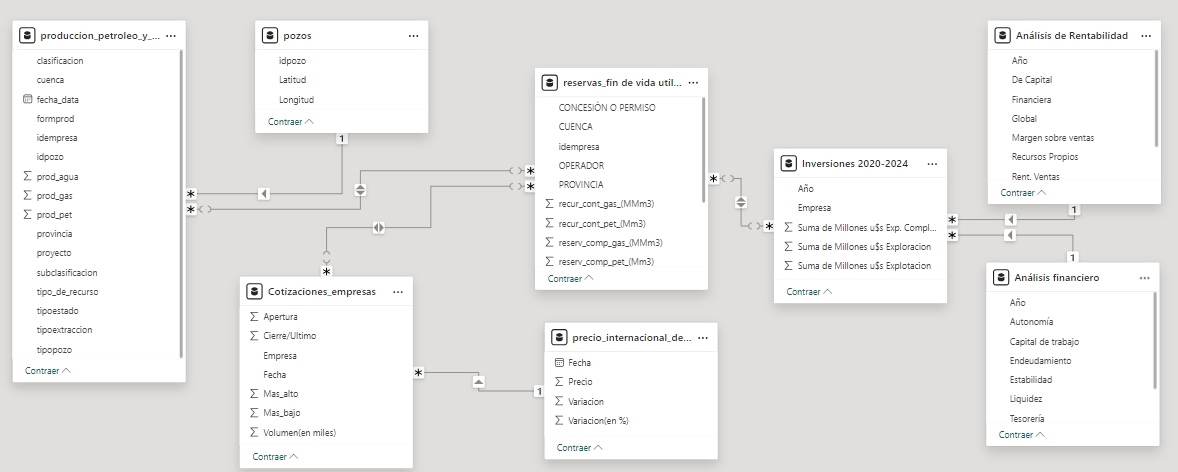
* Verificación del tipo de datos de cada columna: Se examinó minuciosamente el tipo de datos de cada columna para garantizar su coherencia y precisión en el análisis posterior.
* Análisis de la dimensionalidad de los datos: Se exploró la estructura del conjunto de datos para comprender su tamaño y complejidad, lo que permitió una mejor comprensión de la cantidad de registros y variables presentes.
* Manejo de valores nulos: Se identificaron y abordaron los valores nulos en el conjunto de datos mediante técnicas como la imputación de datos o la eliminación de registros incompletos, con el fin de evitar sesgos o distorsiones en el análisis posterior.
* Verificación visual de valores atípicos: Se realizó una exploración gráfica de los datos para detectar posibles valores atípicos o anomalías que podrían afectar la integridad y la precisión de los resultados.
* Indagación de consistencia de los datos: Se llevaron a cabo investigaciones exhaustivas sobre la consistencia de los datos, incluyendo la identificación de máximos, mínimos y rangos de valores para cada variable, lo que ayudó a garantizar la fiabilidad de los datos utilizados en el análisis.

Este proceso se llevo a cabo mediante la utilización de funciones creadas en lenguaje Python que pueden consultarse en [resource.py](https://github.com/No-Country-simulation/c19-108-m-data-bi/blob/main/notebooks/resources.py), el cual está dividido en 3 etapas/hitos los cuales 2 corresponden a informes de situación de los datos y uno al procedimiento en sí:

* El [Informe ETL\_preliminar](https://github.com/No-Country-simulation/c19-108-m-data-bi/blob/main/notebooks/Informe_ETL_Preliminar.ipynb), proporciona un análisis detallado de la calidad y las características de los datos originales, y ofrece funciones específicas para examinar y explorar los datos en profundidad. Estas funciones incluyen la visualización de valores nulos, análisis de frecuencia de palabras, boxplots numéricos y más, lo que facilita la identificación de patrones, tendencias y relaciones significativas en los datos.
* El [ETL\_ipynb](https://github.com/No-Country-simulation/c19-108-m-data-bi/blob/main/notebooks/ETL.ipynb), donde se realiza la normalización, formateo y estructuración de los datos que obtuvimos en referencia a la temática.
* El [Informe EDA preliminar](https://github.com/No-Country-simulation/c19-108-m-data-bi/blob/main/notebooks/EDA_preliminar.ipynb) se enfoca en una instancia intermedia entre la carga y disponibilización de los datos y su uso puntual en un informe EDA. En el utilizamos funciones que permiten ver gráficamente la distribución de valores de cada campo en cada fichero y su proporción en el conjunto total, complementando los mismos con estadísticas básicas de los mismos los que nos permiten tener un panorama general al momento de realizar un análisis más exhaustivo y complejo, es decir, nos permite aproximarnos a los datos de una manera general en un sentido mucho más estadístico/analítico previo a su estudio.

Para la carga de los datos, utilizamos como repositorio/base de datos de consulta el servicio de Google Drive, el cual también nos permite la creación de un “área de trabajo” colaborativa mediante el montaje de dicho repositorio al entorno de Google Colaboratory, donde se llevo a cabo el desarrollo de esta etapa.

Cabe señalar que, en el proceso de ETL, se trabajó los datos con miras a establecer una estructura relacional, para una futura construcción de una Base de datos en un sistema gestor. En un principio, se lleva a cabo este procedimiento de manera conjunta y directa con el ETL, ya que proporciona una homogeneización de etiquetas de campos, tipos de datos, formato, etc lo que permite asignar y/o crear campos que actúen como PK y FK (Primary Key, Foreign Key).



Como vemos, el modelo ER muestra la conexión entre las diversas tablas de la base de datos, mediante sus KEY correspondientes. Detallamos brevemente la asignación en cada tabla de su Primary Key correspondiente y los campos que actúen como Foreign Key:

Tablas:

* Pozos: PK 🡪 idpozo
* Cotizaciones\_empresas: PK 🡪 Fecha
* Produccion\_petroleo\_y\_gas: PK 🡪 fecha\_data, FK 🡪 idpozo
* Reservas\_fin\_de\_vida\_util: PK 🡪 fecha\_data, FK 🡪 idempresa
* Precio\_internacional\_del\_crudo: PK 🡪 Fecha
* Inversiones 2020-2024: PK 🡪 Empresa
* Análisis de rentabilidad: PK 🡪 Año
* Análisis financiero: PK 🡪 Año
  1. Exploratory Data Analisys (EDA)

Con nuestra data ya estructurada y en condiciones óptimas, procedemos a la etapa de análisis de los mismos, lo que se conoce como EDA (Análisis Exploratorio de los datos).

En un principio dividimos esta etapa en dos grandes áreas:

* Análisis del contexto general-global.
* Análisis del mercado doméstico o local.

El análisis de estas dos grandes áreas se llevo a cabo por separado por dos integrantes de nuestro equipo, los cuales desarrollaron el estudio de los datos en los ficheros [Analisis\_EDA\_1.ipynb](https://github.com/No-Country-simulation/c19-108-m-data-bi/blob/main/notebooks/Analisis_EDA_1.ipynb) y [Analisis\_EDA\_2.ipynb](https://github.com/No-Country-simulation/c19-108-m-data-bi/blob/main/notebooks/Analisis_EDA_2.ipynb) respectivamente, utilizando para ello lenguaje Python (librerías Pandas, NumPy, Matplotlib, Seaborn, etc.).

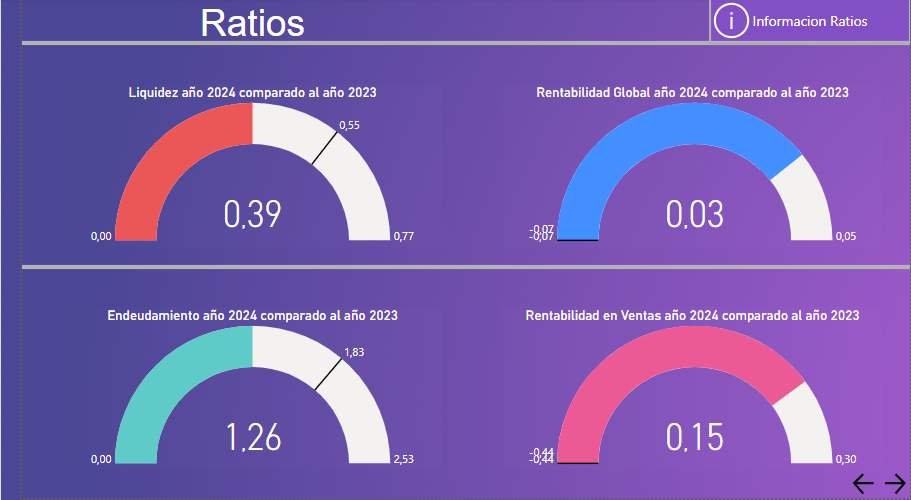
En resumen, en primera instancia se llevó a cabo un análisis de dispersión de frecuencias en los datos (frecuencias, proporciones, etc.), asi como se sacaron métricas básicas de los campos que integraban cada fichero.

Seguido a esto, se procedió a un análisis conjunto de 2 o mas variables cada vez en el cual se observaba su relación y se media el grado de relación entre ellas a trabes de diversos coeficientes y graficos y/o ploteos.

De este estudio EDA se obtuvieron diversas conclusiones respecto a los datos, los cuales podemos revisar en el [Informe Completo Evaluación del Riesgo Crediticio de YPF.pdf](https://github.com/No-Country-simulation/c19-108-m-data-bi/blob/main/docs/Informe%20Completo%20Evaluaci%C3%B3n%20del%20Riesgo%20Crediticio%20de%20YPF.pdf).

* 1. Key Performance Indicators (KPI)

Ratios de la empresa YPF del año 2024 contrastados con el año 2023 (este último representado por la línea negra en cada barra)



En el transcurso del desarrollo del análisis, observamos que los distintos balances contables y estados de resultado de las empresas de hidrocarburos presentaban información que nos brindaba la posibilidad de configurar y establecer distintos ratios para ellas.

En estadística y análisis de balances y estados de resultados, los ratios financieros pueden considerarse similares a los KPIs (Key Performance Indicators). Ambos son métricas utilizadas para evaluar el rendimiento y la salud financiera de una empresa. Aquí hay una explicación detallada:

**Ratios Financieros**

Los ratios financieros son métricas que se calculan a partir de los estados financieros de una empresa y se utilizan para evaluar diversos aspectos de su desempeño financiero. Algunos ejemplos comunes incluyen:

* **Ratio de Liquidez**: Mide la capacidad de una empresa para cumplir con sus obligaciones a corto plazo. Ejemplos incluyen el ratio corriente (current ratio) y el ratio rápido (quick ratio).
* **Ratios de Rentabilidad**: Evalúan la capacidad de una empresa para generar ganancias. Ejemplos incluyen el margen de beneficio neto, el retorno sobre activos (ROA) y el retorno sobre el capital (ROE).
* **Ratios de Endeudamiento**: Miden el nivel de deuda de una empresa en relación con su patrimonio. Ejemplos incluyen el ratio de endeudamiento y el ratio de cobertura de intereses.
* **Ratios de Eficiencia**: Evalúan cómo una empresa utiliza sus activos y recursos. Ejemplos incluyen la rotación de inventarios y la rotación de activos.

**KPIs (Key Performance Indicators)**

Los KPIs son métricas clave que se utilizan para medir el rendimiento de una organización en relación con sus objetivos estratégicos y operativos. Los KPIs pueden abarcar una variedad de áreas, incluyendo finanzas, operaciones, marketing, ventas, etc. Algunos ejemplos de KPIs financieros incluyen:

* **Margen de Ganancia Bruta**: La diferencia entre los ingresos netos y el costo de los bienes vendidos.
* **Earnings Before Interest and Taxes (EBIT)**: Ganancias antes de intereses e impuestos.
* **Cash Flow Operativo**: Flujo de efectivo generado por las operaciones de la empresa.

**Similitudes entre Ratios Financieros y KPIs**

1. **Medición del Desempeño**: Ambos se utilizan para evaluar y monitorear el desempeño de una empresa.
2. **Toma de Decisiones**: Proporcionan información valiosa para la toma de decisiones por parte de la administración y los inversores.
3. **Comparación y Análisis**: Permiten comparar el desempeño actual con objetivos, estándares del sector o datos históricos.
4. **Monitoreo de la Salud Financiera**: Ayudan a evaluar la salud financiera y operativa de una empresa.

**Diferencias**

1. **Ámbito de Aplicación**: Los ratios financieros están más centrados en la evaluación financiera específica a partir de los estados financieros, mientras que los KPIs pueden abarcar una gama más amplia de métricas de desempeño en diferentes áreas de la empresa.
2. **Objetivos Estratégicos**: Los KPIs están más alineados con los objetivos estratégicos de una empresa, mientras que los ratios financieros suelen centrarse en la salud financiera y el rendimiento.

Por lo tanto en conclusión, los ratios financieros pueden considerarse como un subconjunto de los KPIs cuando se trata de análisis financiero. Ambos son esenciales para la evaluación y monitoreo del rendimiento de una empresa, aunque los KPIs abarcan un ámbito más amplio que incluye áreas no financieras.

En nuestro caso, hacemos hincapié en los ratios financieros ya que consideramos que para la problemática actual que analizamos, son la faceta principal a tener en cuenta en cuanto a su peso especifico en la construcción del panorama analítico sobre riesgo crediticio de una empresa.

Para ello elegimos:

* **RDT (Rentabilidad del Total de Activos)**:

Mide la eficiencia de una empresa en el uso de sus activos para generar ganancias. Se calcula como el resultado neto dividido por el total de activos.

* **RDC (Rentabilidad del Capital)**:

Indica la rentabilidad obtenida sobre el capital invertido en la empresa. Se calcula dividiendo el resultado neto entre el capital total.

* **RDLP (Rentabilidad del Patrimonio)**:

Conocida también como ROE (Return on Equity), mide la rentabilidad obtenida sobre el patrimonio neto de los accionistas. Se calcula dividiendo el resultado neto entre el patrimonio neto.

* **Rotación de Inventarios**:

Mide cuántas veces se vende y se reemplaza el inventario de una empresa durante un período específico. Se calcula dividiendo el costo de las ventas por el inventario promedio.

* **Rotación de Activos Totales**:

Mide la eficiencia con la que una empresa utiliza sus activos para generar ventas. Se calcula dividiendo las ventas netas entre los activos totales promedio.

* **Rotación de Cuentas por Cobrar**:

Indica cuántas veces una empresa puede convertir sus cuentas por cobrar en efectivo durante un período específico. Se calcula dividiendo las ventas a crédito entre las cuentas por cobrar promedio.

* **Margen de Utilidad Neta**:

Mide el porcentaje de ingresos que se convierte en ganancias netas después de todos los gastos. Se calcula dividiendo el ingreso neto entre las ventas netas.

* **Rentabilidad Sobre los Activos (ROA)**:

Indica la rentabilidad que una empresa obtiene sobre sus activos totales. Se calcula dividiendo el ingreso neto entre los activos totales.

* **Rentabilidad Sobre el Patrimonio Neto (ROE)**:

Similar a la RDLP, mide la rentabilidad obtenida sobre el patrimonio neto de los accionistas. Se calcula dividiendo el ingreso neto entre el patrimonio neto.

* **Liquidez Corriente**:

Capacidad de la empresa para pagar sus deudas a corto plazo con sus activos a corto plazo.

* **Prueba Ácida**:

Capacidad de la empresa para pagar sus deudas a corto plazo con sus activos más líquidos (excluyendo inventarios).

* **Liquidez Inmediata**:

Capacidad de la empresa para pagar sus deudas a corto plazo con sus activos más líquidos disponibles inmediatamente (efectivo y equivalentes de efectivo).

Dicho esto, construimos nuestros gráficos de medición de ratios, segmentando sus valores por años y contrastando el presente año (2024) con el 2023 como meta.

* 1. Machine Learning

Para la construccion de nuestro modelo trabajamos con datos referentes a diferentes ratios de empresas de hidrocarburos, los cuales nos serviran para el proceso de entrenamiento del mismo.

Como primer paso, se realizó un feature engineer de la data, donde se transformo a un formato adecuado y se realizó la creacion de un campo que nos sirva de variable objetivo (Target) para nuestra predicción.

Este proceso esta trabajado en [feature\_engineer.ipynb](https://github.com/No-Country-simulation/c19-108-m-data-bi/blob/main/notebooks/feature_engineer.ipynb), un notebook segmentado paso a paso con comentarios de cada tarea y proceso.

Una vez tenemos la data en condiciones, y tenemos la elección de nuestro Target a predecir, realizamos un proceso de Grid Search, donde probamos el rendimiento de varios modelos de clasificación (nuestro target es una etiqueta), los que evaluamos con multiples hiperparametros y en el cual tambien se realiza un proceso de validación cruzada.

Esto nos llevó a identificar que el modelo de Logistic Regression era el modelo que mejor rendimiento mostraba, por lo cual procedimos a su construcción, entrenamiento, testeo y puesta a punto.

Como paso final, se creó una app utilizando la libreria streamlit la que nos proporciona una interfaz interactiva para nuestro modelo, como paso previo al despliegue del mismo.

* 1. Deploy

Para el despliegue de nuestro modelo, como primer paso debemos realizar la construcción de un app para embeberlo y poder interactuar con el a través de una interfaz.

La construcción de la app se llevó a cabo mediante la utilización de la librería streamlit, que es parte del lenguaje de programación Python.

Se construyó una interfaz apropiada en donde se habilita la carga de valores pertenecientes a los ratios correspondientes.



Como próximo paso se sube la aplicación a un repositorio (en este caso a GitHub), para su posterior despliegue en el servicio de Render.

Elegimos este servicio que brinda la posibilidad de desplegar nuestra aplicación en la nube a través de la conexión al repositorio donde se encuentra subida nuestra app.

1. CONCLUSIONES

El análisis realizado revela que el mercado petrolero es altamente volátil y está sujeto a la influencia de una multitud de factores, tanto internos como externos. YPF, como empresa líder en Argentina, se encuentra expuesta a estos riesgos. Si bien la compañía ha demostrado una cierta resiliencia en el pasado, su futuro dependerá en gran medida de su capacidad para adaptarse a las fluctuaciones del mercado, mejorar su eficiencia operativa y diversificar sus fuentes de ingresos. En este contexto, se recomienda a los acreedores evaluar cuidadosamente los riesgos asociados a la concesión de crédito a YPF y exigir garantías adicionales para mitigar dichos riesgos.

En conjunto con el analisis del estado financiero actual, observamos una leve mejora en la capacidad de pago inmediato y de mediano plazo con respecto a un posible credito, lo que ayuda a mejorar su capacidad de pago de un credito a futuro. Asimismo, a traves de nuestro modelo predictivo, vemos que el estado de la empresa en cuanto a pago crediticio se clasifica en su proyeccion como Riesgoso.

Para cerrar con la toma de decision con respecto al otorgamiento de credito, se toma en cuenta el contexto general y perspectiva general y potencial de la empresa. Esto viene siendo tanto como consecuencia de nuestro analisis y como investigación y noticias de mercado. destacamos entre ellas:

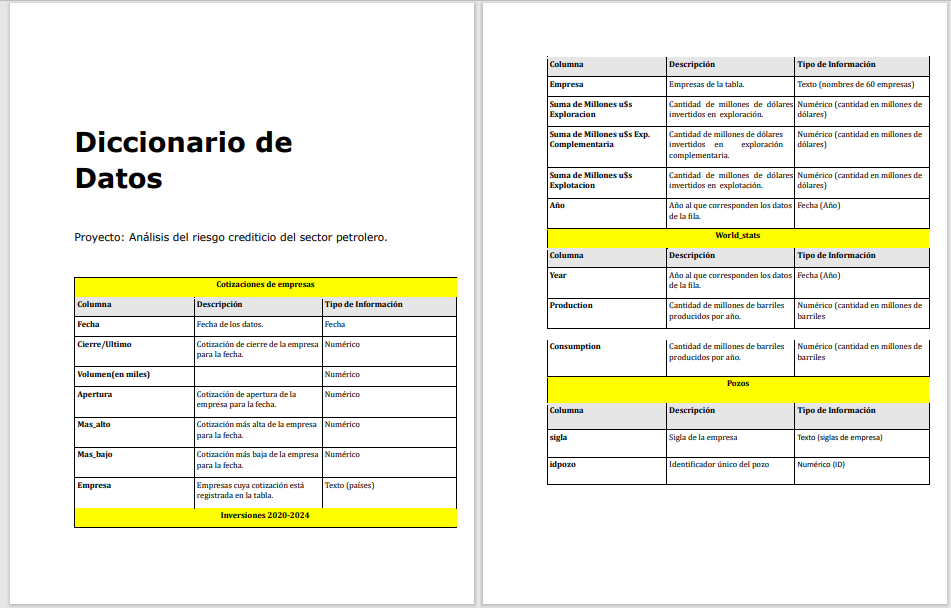
* Potencial ENORME de explotacion y produccion de recursos en reservas encontrados en Vaca Muerta, donde se estiman que esa región contiene alrededor de 3 veces las reservas totales actuales del país
* YPF cuenta actualmente con la concesion de explotacion de casi el 50% de la explotacion en la cuenca
* La planificacion de remodelacion del gasoducto y oleoducto Nestor Kirchner, que permitira aumentar el caudal de transito de la produccion para el pais
* la GRAN INVERSIÓN ya planificada, y aprobada de la construccion de tratamiento de GNL por parte de YPF y la construccion de puerto autónomo para su exportación.

Esto apoya nuestra conclusión final de que a pesar que el estado financiero y general de la empresa YPF para la concesión de un crédito la clasifican como Riesgosa, no podemos dejar de tomar en cuenta las proyecciones a 3-4-5 años hacia el futuro, que muestran un aumento considerable en la productividad y rentabilidad de la empresa a causa de su aumento en la explotacion, procesado y distribución-exportacion de recursos de Vaca Muerta lo que multiplicara su capacidad crediticia, volviendolo No riesgoso.

1. DICCIONARIO DE DATOS

Como fue mencionado anteriormente, el diccionario de los datos es un documento que se construyó específicamente para entender los datos, términos, campos y valores de este proyecto.

Se realiza una breve descripción de su naturaleza y se hacer referencia a su significado, y se especifica que tipo de dato es y su escala (en caso de se requiera)



El diccionario de los datos se encuentra disponible en el siguiente [link](https://github.com/No-Country-simulation/c19-108-m-data-bi/blob/main/docs/Diccionario%20de%20Datos.pdf).