Sprint #1: Puesta en marcha del proyecto y Trabajo con Datos

1. **Entendimiento de la situación actual:**

   En el actual contexto empresarial, la retroalimentación de los usuarios se ha convertido en un activo invaluable, especialmente a través de plataformas de reseñas en constante expansión. En esta era digital, plataformas como Yelp y Google Maps desempeñan un papel crucial al permitir que los usuarios compartan sus experiencias y opiniones sobre diversos negocios, desde restaurantes hasta hoteles y servicios. Esta información ofrece una visión esencial para las empresas, ya que les brinda una comprensión directa de la percepción de los clientes y áreas de mejora.

   En nuestra capacidad como consultora de análisis de datos, hemos sido encomendados con un desafío significativo: realizar un análisis exhaustivo del mercado estadounidense, enfocado en el sector de restaurantes y actividades relacionadas. Nuestro cliente, un conglomerado de empresas, busca un entendimiento profundo de las opiniones de los usuarios en Yelp y Google Maps. Este análisis permitirá prever tendencias en el crecimiento o disminución de sectores empresariales y respaldará la toma de decisiones estratégicas.

   Adicionalmente, nuestro cliente tiene como objetivo determinar ubicaciones ideales para nuevos establecimientos. Además, está interesado en crear un restaurante con la temática de Messi, futbolista argentino de renombre. Esta decisión se basa en el furor que Messi ha generado desde su llegada al equipo de fútbol Inter de Miami. Este restaurante temático busca capitalizar la popularidad y admiración que Messi ha suscitado en la ciudad y más allá. Asimismo, nuestro cliente desea implementar un sistema de recomendación de lugares basado en las experiencias previas de los usuarios. Este sistema persigue enriquecer las experiencias de los usuarios, permitiéndoles descubrir nuevas opciones de acuerdo a sus preferencias.

2. **Objetivos:**

   a) **Recopilación y Depuración de Datos:** Inicialmente, nos enfocaremos en recopilar y depurar la información necesaria proveniente de múltiples fuentes, incluyendo datos provistos por nuestro cliente y otros adquiridos a través de fuentes externas. Crearemos un DataWarehouse que albergará estos datos, asegurando la calidad y consistencia de los mismos. Esto implicará la extracción de datos estáticos, consultas a APIs y técnicas de web scraping.

   b) **Análisis Significativos:** Nuestro siguiente paso es llevar a cabo análisis exhaustivos en las áreas de interés. Examinaremos las relaciones entre variables clave, identificando posibles patrones y correlaciones. Este análisis no solo nos permitirá descubrir relaciones relevantes, sino que también identificará los factores detrás de estas conexiones en el mundo real.

   c) **Machine Learning y Mejora de la Experiencia del Usuario:** Basándonos en la información recopilada y el análisis realizado, procederemos a entrenar y poner en producción un modelo de machine learning. Este modelo, ya sea de clasificación supervisada o no supervisada, abordará un problema específico que se vincule directamente con los objetivos del proyecto. Esto incluirá la creación de un sistema de recomendación para usuarios basado en sus preferencias y experiencias previas.

3. **Alcance:**

Dentro del marco temporal asignado para la ejecución del proyecto en curso, hemos tomado la decisión de precisar cuidadosamente los límites y alcances en relación a aspectos esenciales. Esta delimitación ha sido aplicada estratégicamente, considerando la industria objetivo, las fuentes de datos, la geografía involucrada y el periodo temporal abordado.

Con el objetivo de garantizar un enfoque preciso y la obtención de resultados concretos, hemos restringido el alcance del proyecto al ámbito de la industria gastronómica, centrándonos inicialmente en el estado de Florida. Esta elección se fundamenta en la amplitud, vitalidad y significativo impacto económico que la industria gastronómica ejerce en la economía de los Estados Unidos. Además, esta elección encuentra respaldo en la perspectiva de nuestro cliente de establecer un restaurante temático basado en la figura destacada del futbolista argentino Lionel Messi.

Cabe destacar que este enfoque se acota a los años entre 2015 y 2021, ya que estos años cuentan con datos particularmente significativos, considerando las fuentes seleccionadas, Google Maps y Yelp. La selección de la industria se basa en su variabilidad y relevancia en el contexto económico estadounidense.

4. **Objetivos y KPIs asociados:**

5. **Repositorio Github:** https://github.com/claudiacaceresv/pf_yelp_google.git

6. **Solución propuesta:**

Con el objetivo de cumplir de manera efectiva los objetivos de trabajo previamente establecidos, hemos desarrollado una metodología detallada que guiará nuestras acciones a lo largo del proyecto. Nuestra aproximación se basa en una combinación de enfoques, herramientas y roles asignados para lograr una implementación exitosa.

a) **Metodología de trabajo y Organización:**
Para lograr una ejecución coherente y productiva, hemos adoptado un enfoque iterativo y colaborativo, en línea con la metodología ágil Scrum. El equipo se organizará en tareas específicas, y cada miembro asumirá roles claramente definidos. Estos roles incluirán responsabilidades como la recopilación de datos, limpieza, análisis exploratorio, modelado de machine learning y presentación de resultados. Además, se establecerán reuniones regulares, como las reuniones diarias de Scrum, para monitorear el progreso y ajustar la estrategia según sea necesario.

- Claudia Caceres - *Data engineer / Data analyst*
- Virginia Tenorio - *Data engineer / Data analyst*
- Sebastián Bello - *Data engineer / Data analyst*
- Mariano Bernal - *Data analyst / Data scientist*
- Martín Varela - *Data analyst / Data scientist*

b) Productos y Entregables:

Nuestro trabajo resultará en una serie de productos claramente definidos, entregados en sprints de tiempo fijo.

- **Entrega 1 - Stack Elegido y Fundamentación, y Flujo de Trabajo (Sprint 1):**
  Documentación detallada sobre las herramientas tecnológicas seleccionadas, su justificación y el flujo de trabajo propuesto.

  - ✓ 4 KPI's
  - ✓ Documentar alcance del proyecto
  - ✓ EDA de los datos
  - ✓ Repositorio en Github
  - ✓ Implementación stack tecnológico
  - ✓ Metodología de trabajo
  - ✓ Diseño detallado
  - ✓ Equipo de trabajo - Roles y responsabilidades
  - ✓ Cronograma general - Gantt
  - ✓ Análisis preliminar de calidad de datos

  Fecha de entrega: 01/09/2023.

- **Entrega 2 - Documentación y reporte realizado (Sprint 2):**
  Documentación exhaustiva sobre la implementación del Datalake, procesos de Extracción, Transformación y Carga (ETL) y la estructura del Data Warehouse. Este reporte ofrecerá una visión completa de cómo los datos son capturados, procesados y almacenados.

  - ✓ ETL completo
  - ✓ Estructura de datos implementada (DW, DL, etc). Pueden usar algún servicio
  - ✓ Pipeline ETL automatizado
  - ✓ Diseño del Modelo ER
  - ✓ Pipelines para alimentar el DW
  - ✓ Data Warehouse
  - ✓ Automatización
  - ✓ Validación de datos
  - ✓ Documentación:
    - Diagrama ER detallado (tablas, PK, FK y tipo de dato)
    - Diccionario de datos
    - Workflow detallando tecnologías
  - ✓ Análisis de datos de muestra
  - ✓ MVP/ Proof of Concept de producto de ML ó MVP/ Proof of Concept de Dashboard

  Fecha de entrega: 08/09/2023.

- **Entrega 3 - Dashboard y Modelo de Machine Learning (Sprint 3):**
  Desarrollo de un dashboard interactivo que muestre análisis exploratorios y resultados clave. Además, presentación del modelo de machine learning diseñado para proporcionar recomendaciones personalizadas a los usuarios basadas en sus preferencias.

  - ✓ Diseño de Reportes/Dashboards
  - ✓ KPIs
  - ✓ Modelos de ML
  - ✓ Modelo de ML en producción
  - ✓ Documentación:
    - Selección del modelo, feature engineering
    - Informe de análisis
  - ✓ Video del proyecto realizado, para ser votado y, en caso de ganar, ser presentado en la graduación final.

  Fecha de entrega: 15/09/2023.

c) Estimación de tiempo y Diagrama de Gantt:
Hemos trazado un cronograma detallado que abarca las tareas específicas y los hitos previstos para cada sprint. Esto nos permitirá monitorear y asegurar un progreso constante y entregas en intervalos regulares. Utilizaremos un diagrama de Gantt, adaptado al enfoque ágil de Scrum, para visualizar y administrar de manera efectiva los tiempos de ejecución de cada sprint.

d) Herramientas y Stack Tecnológico:
Nuestra arquitectura de proyecto se apoyará en una selección cuidadosa de herramientas tecnológicas:

- **Lenguaje de Programación:**
  Python
  SQL

- **Entorno de Desarrollo:**
  Visual Studio Code
  Google Colab
  Jupyter Notebook

- **Bibliotecas Esenciales:**
  Pandas
  NumPy
  Scikit-learn

- **Visualización de Datos:**
  Matplotlib
  Seaborn
  Power BI (para los dashboards)

- **Plataforma de Nube:**
  BigQuery (Google Cloud)
  Google Cloud Functions

- **Framework Web:**
  FastAPI (para el sistema de recomendación)

- **Control de Versiones:**
  GitHub

Este conjunto robusto de herramientas nos proporciona una base sólida para abordar cada aspecto del proyecto, desde la manipulación de datos hasta el desarrollo de modelos de machine learning y la implementación de sistemas de recomendación.

e) Análisis de Datos de Calidad:

En esta fase del proyecto, abordaremos el análisis exhaustivo de los datos para garantizar su calidad y confiabilidad. Hemos realizado un detallado análisis de los metadatos asociados con las fuentes de datos que utilizaremos en el proyecto. A continuación, proporcionamos información específica sobre las fuentes, columnas, tipos de datos, métodos de adquisición y fechas de adquisición y actualización:

- Google:

  - ✓ Dataset: metadata.sitios y reviews-estados
  - ✓ Última fecha de actualización: **[Fecha]**
  - ✓ Método de adquisición: Los conjuntos de datos fueron proporcionados directamente por nuestro cliente.

- Yelp:

  - ✓ Datasets: business, checkin, review, tip, user
  - ✓ Última fecha de actualización: **[Fecha]**
  - ✓ Método de adquisición: Los conjuntos de datos fueron proporcionados directamente por nuestro cliente.

Hemos establecido un sistema de documentación detallada para los conjuntos de datos que utilizamos en este proyecto. Se ha creado un diccionario exhaustivo que abarca cada uno de los datasets empleados, sus respectivas columnas, tipos de datos asociados y cualquier modificación realizada. Este diccionario proporciona una visión completa y organizada de la estructura y contenido de los datos utilizados en el análisis, garantizando la transparencia y trazabilidad en cada etapa del proceso.