

(27/08/2023 – 15/09/1023)

1.Integrantes:

Roman Aichino: Data Engineer

Max Jeffer: Data Engineer – Machine Learning

Diego Campos: Data Engineer

Eduardo Vivar: Data Analytics

2. Introducción:

Este es un proyecto que fue realizado desde la disciplina ‘ciencias de datos’, que busca recolectar datos para procesarlos, analizarlos y poder tomar decisiones en base a ellos. El proyecto cuenta con un EDA, una automatización de el ETL en la nube ‘Google Cloud Platform’, una carga incrementa a partir de los datos de una API, un dashboard con KPIs y por último, un modelo de Machine Learning que devuelve cual es la mejor opción para una decisión empresarial.

3. Quiénes Somos y Quien es nuestro cliente:

Nosotros somos Drem Data Insights, una empresa dedicada a la ciencia de los datos, y en este proyecto en particular nuestra tarea es crear un sistema que proporcione información sobre los restaurantes en Estados Unidos, en el estado de Florida, con el objetivo de que nuestro cliente, una empresa que quiere abrir un restaurante, pueda tomar buenas decisiones en cuanto a la ubicación y otros detalles del negocio.

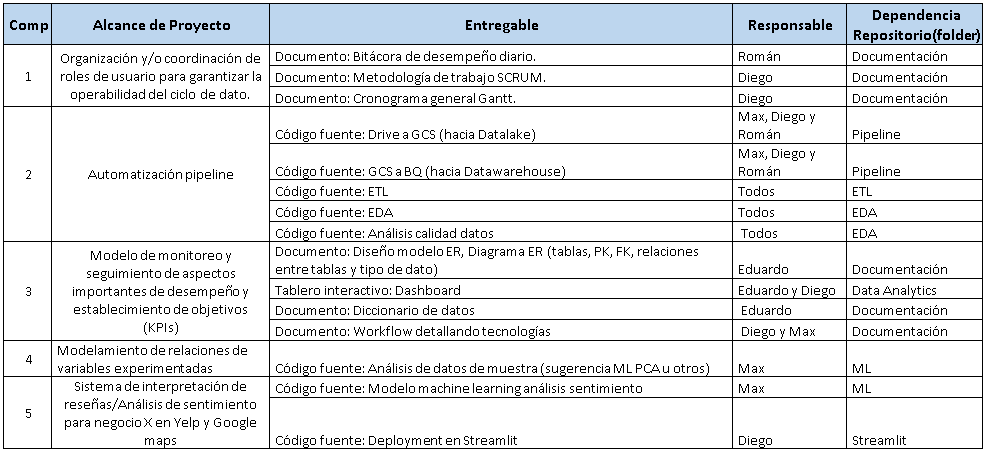
4. Objetivos:

Generales:

* Identificar ubicaciones y atributos estratégicos para abrir un restaurante mediante el análisis de datos a partir de la información recolectada, con el fin de maximizar la rentabilidad y la satisfacción del cliente.

Específicos:

* Definir metodología de gestión general (roles, tareas, diagrama de Gantt, etc).
* Definir el stack tecnológico.
* Automatización de pipelines (carga incremental, etl, envío de datasets al datawarehouse).
* Análisis exploratorio de los datos para la creación de KPIs.
* Diseño de un dashboard interactivo en Power Bi para realizar el seguimiento y monitoreo de los KPIs
* Análisis de sentimiento mediante un modelo de ML.

5. Alcance del Proyecto:

6. Planificación de Scrum:

Metodología Scrum: Roles y Planificación

**Roles en Scrum**

* Product Owner:

Responsable de definir y priorizar el Product Backlog.

Asegura que el equipo trabaje en las tareas más valiosas para el cliente.

Colabora con interesados y el equipo para mantener la visión clara del producto.

* Scrum Master:

Facilitadora del proceso Scrum.

Garantiza que el equipo siga las prácticas y principios de Scrum.

Elimina obstáculos y fomenta un entorno de trabajo colaborativo.

* Equipo de Desarrollo:

Campos Diego

Aichino Roman

Jeffer Max

Vivar Eduardo

Transforman elementos del Product Backlog en incrementos de producto entregables.

Autónomos y autoorganizados, toman decisiones sobre cómo abordar tareas.

Colaboran para lograr los objetivos del Sprint.

**Planificación en Scrum**

Product Backlog: Lista de tareas priorizadas

1. 4 KPI’s
2. Documentar alcance del proyecto
3. EDA de los datos
4. Repositorio en Github
5. Implementación stack tecnológico
6. Metodología de trabajo
7. Diseño detallado
8. Equipo de trabajo - Roles y responsabilidades
9. Cronograma general - Gantt
10. Análisis preliminar de calidad de datos
11. ETL completo
12. Estructura de datos implementada (DW, DL, etc). Pueden usar algún servicio
13. Pipeline ETL automatizado
14. Diseño del Modelo ER
15. Pipelines para alimentar el DW
16. Datawarehouse
17. Automatización
18. Validación de datos
19. Documentación

Diagrama ER detallado (tablas, PK, FK y tipo de dato)

Diccionario de datos

Workflow detallando tecnologías

1. Análisis de datos de muestra
2. MVP/Proof of Concept de producto de ML ó MVP/Proof of Concept de Dashboard
3. Diseño de Reportes/Dashboards
4. KPIs
5. Modelos de ML
6. Modelo de ML en producción
7. Documentación

Selección del modelo, feature engineering

Informe de análisis

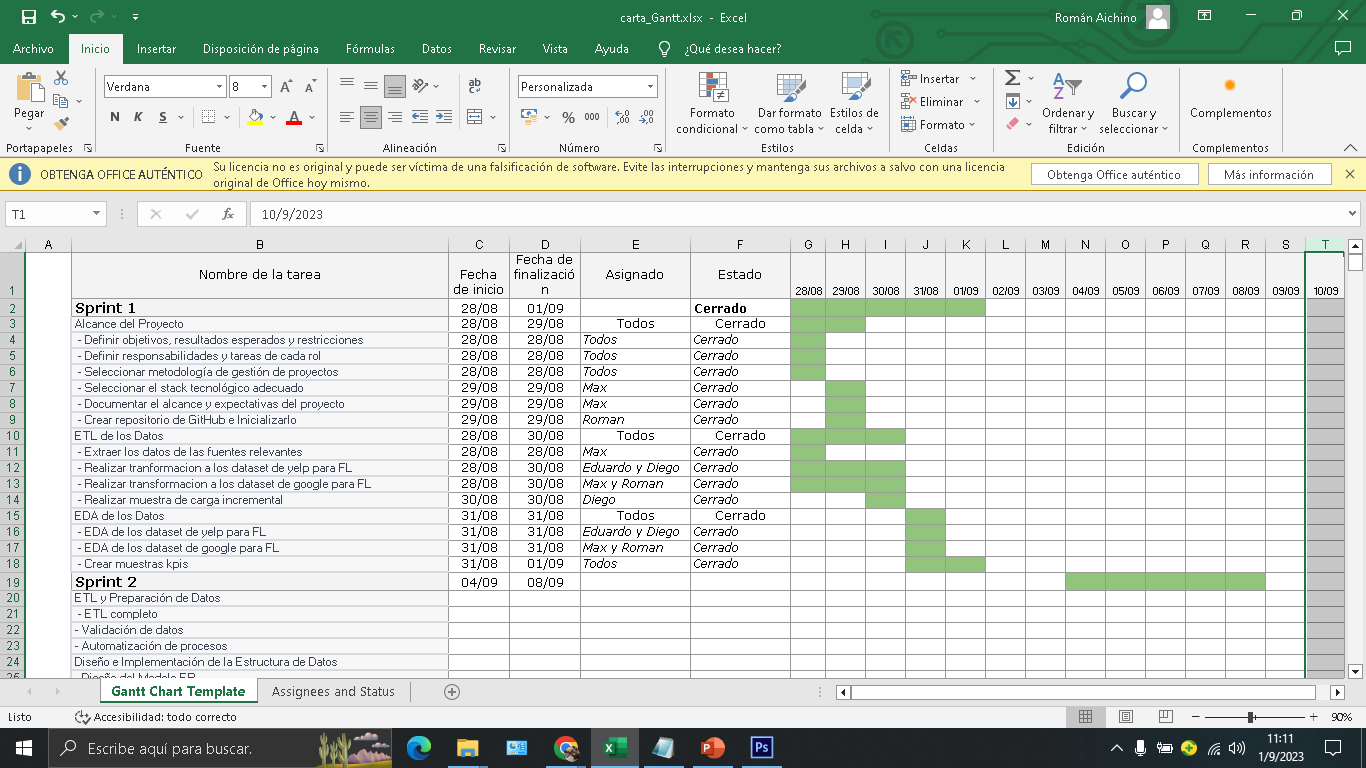
**Sprint Planning: Selección de tareas para el Sprint**

En la primera reunión, el equipo de desarrollo, junto con el Product Owner y el Scrum Master, seleccionó las siguientes tareas para el Sprint:

1. 4 KPI’s
2. Documentar alcance del proyecto
3. EDA de los datos
4. Repositorio en Github
5. Implementación stack tecnológico
6. Metodología de trabajo
7. Diseño detallado
8. Equipo de trabajo - Roles y responsabilidades
9. Cronograma general - Gantt
10. Análisis preliminar de calidad de datos

Estas tareas se han priorizado para este Sprint con el objetivo de lograr un incremento valioso del producto al final del período de desarrollo.

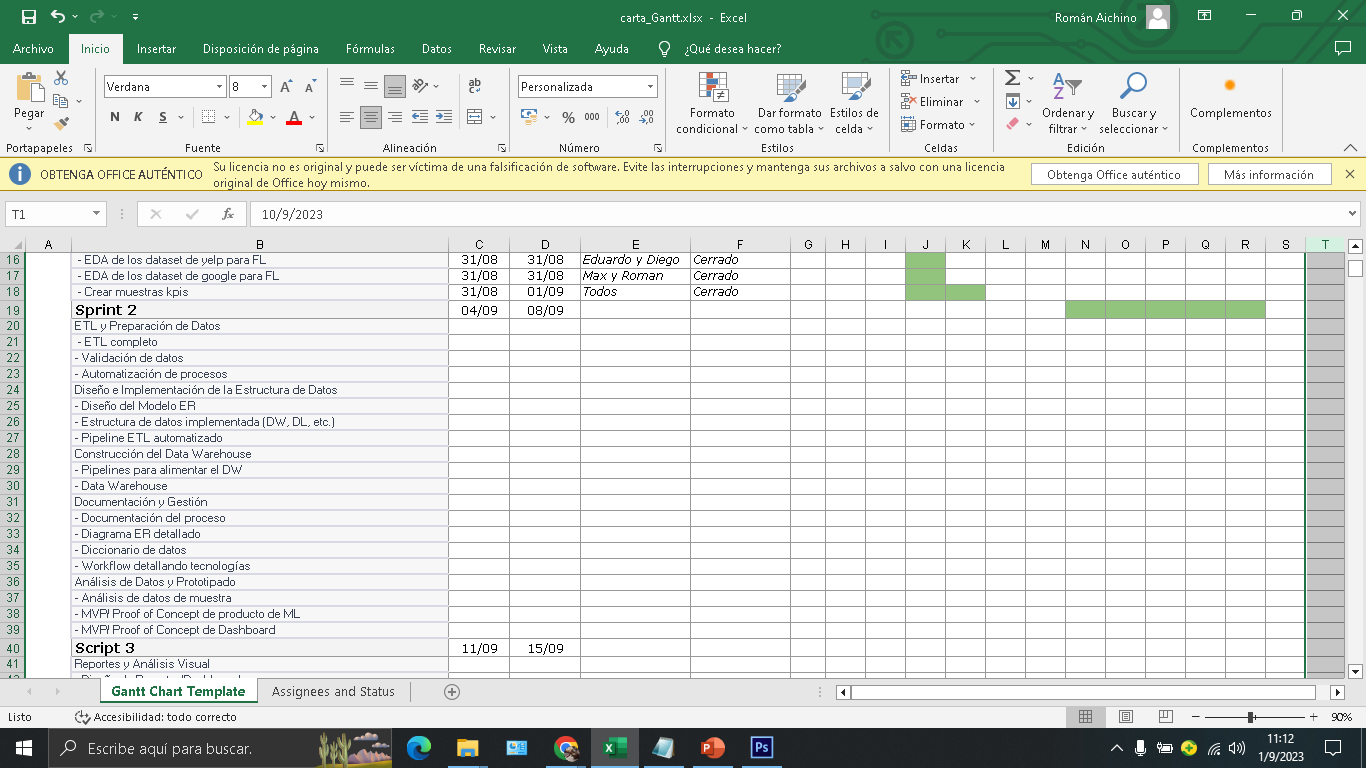
7. Diagrama de Gantt:



8. Stack Tecnológico:

**Parte técnica:**

Nube: Google Cloud Platform

Google Cloud Storage: para nuestro Datalake.

BigQuery: para nuestro Datawarehouse.

Google Colab: para la creación de archivos jupyter que automáticamente hagan la carga incremental, el ETL y la subida al Datawarehouse de todos nuestros datasets. Además, también lo utilizamos para los EDAs que quedarán subidos en el repositorio.

Cloud Functions: para la automatización de los archivos jupyter creados en Google Colab.

Pub/Sub y Cloud Scheduler: para el trigger de las funciones creadas en Cloud Functions.

Lenguaje de programación: Python, ya que es el que nuestra empresa utiliza habitualmente para procesar y analizar datos.

Librerías:

Pandas: la mejor para procesar datasets en forma de dataframes.

Mathplotlib: para la creación de gráficos que nos permitan mirar los datos desde otra perspectiva más analizable.

Numpy: para ecuaciones matemáticas.

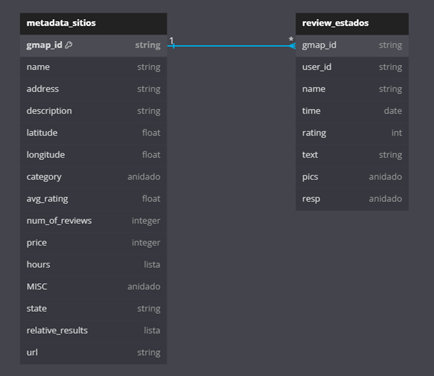
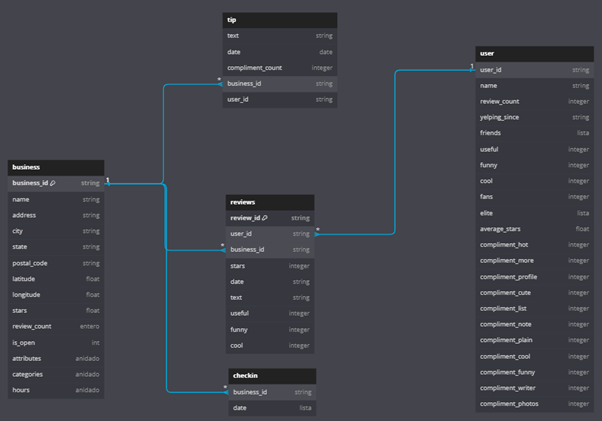
Sk-learn: para nuestro modelo predictivo.

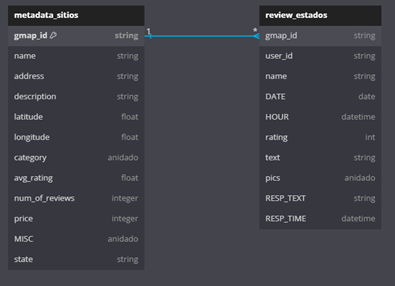
**Parte analítica:**

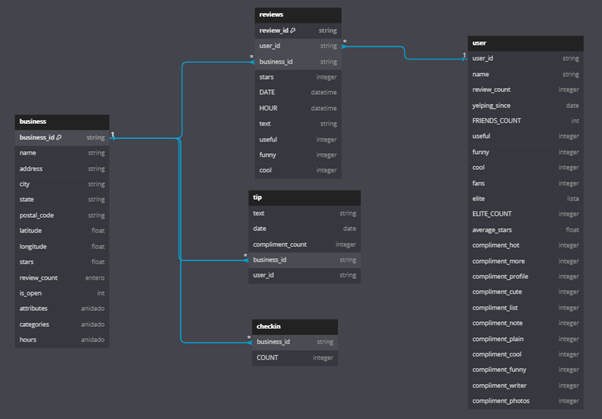
Power BI: para la creación del Dashboard interactivo.

Google Colab: para el análisis de los datos.

9. Diagrama Entidad-Relación (ER):







10. Análisis de Datos:

Descripción de las fuentes de datos utilizadas.

Proceso de limpieza, transformación y exploración de datos.

Resumen de los hallazgos clave del análisis de datos.

11. Ingeniería de Datos:

Diseño y arquitectura de la infraestructura de datos.

Proceso de extracción, transformación y carga (ETL) de datos.

Integración con sistemas existentes, si corresponde.

12. Modelo de Machine Learning:

Descripción del enfoque de aprendizaje automático utilizado.

Arquitectura del modelo, algoritmos y parámetros.

Métricas de rendimiento y validación del modelo.

13. KPIs (Indicadores Clave de Desempeño):

Definición de los KPIs utilizados para medir el éxito del proyecto.

Fórmulas y métodos de cálculo de KPIs.

14. Dashboard:

Descripción del panel de control o tablero utilizado para visualizar los datos y KPIs.

Capturas de pantalla o ejemplos del dashboard.

15. Conclusiones:

Resumen de los logros, desafíos y lecciones aprendidas en el proyecto.

Reflexiones sobre el éxito en relación con los objetivos.

16. Referencias:

(MAX AQUÍ VA EL LINK DE LA API)

17. Anexos:

Cualquier documentación técnica adicional, código fuente, capturas de pantalla, etc.

18. Agradecimientos:

Reconocimientos a los miembros del equipo, colaboradores, clientes, etc.

Recuerda que la documentación debe ser clara y concisa, y debe adaptarse a las necesidades de tu proyecto y audiencia. También es importante mantenerla actualizada a medida que el proyecto avanza y se desarrolla.