
INFORME DE AVANCE - SPRINT 1

Aylén Sol Guzman, Edith Priscila Muniz Cantu, Maria Marcela Diaz, Jorge Andres Jola Hernandez, Franco Dylan Damian Luna Pedroso.

Introducción

En la era digital, la toma de decisiones respecto a dónde comer, comprar o disfrutar de servicios se ha transformado significativamente. Plataformas como Google y Yelp se han convertido en referentes fundamentales para usuarios en búsqueda de experiencias auténticas y recomendaciones confiables.

Las opiniones se han transformado en monedas de gran valor. Los comentarios y calificaciones no solo reflejan la calidad de un negocio, sino que también influyen en la toma de decisiones de otros usuarios. La importancia de las reseñas radica en su capacidad para proporcionar una visión auténtica y personalizada, guiando a los consumidores hacia elecciones informadas y satisfactorias.

El análisis de sentimientos se convierte en la herramienta clave para desentrañar las emociones y percepciones expresadas en estas reseñas. A través de esta metodología, se busca comprender no solo la satisfacción o insatisfacción general de los usuarios, sino también los matices emocionales que pueden revelar aspectos más profundos de la experiencia.

Con el análisis de sentimientos como base, el proyecto busca construir un sistema de recomendación inteligente. Este sistema no solo se limitará a sugerir lugares similares a los que han recibido reseñas positivas, sino que también considerará las preferencias individuales de los usuarios, ofreciendo recomendaciones personalizadas que se alineen con sus gustos y expectativas únicas.

En resumen, este proyecto se sumerge en el universo de opiniones que estos dos gigantes ofrecen, buscando extraer conocimientos valiosos mediante análisis de sentimientos y la creación de un sistema de recomendación el cual pretende no solo entender la experiencia del usuario, sino también anticipar sus preferencias futuras.

Objetivos

1. Análisis Integral de Presencia y Reputación:

- Descripción: Evaluar la presencia y reputación de negocios de la categoría de restaurantes en Yelp y Google Maps en los estados seleccionados.
- Meta: Obtener una comprensión completa de la distribución geográfica y la relevancia de los restaurantes en las plataformas.

2. Desarrollo de un Sistema de Análisis Predictivo:

- Descripción: Implementar un sistema que utilice técnicas de análisis predictivo para anticipar tendencias y comportamientos del mercado gastronómico.
- Meta: Proporcionar a los stakeholders herramientas predictivas para la toma de decisiones estratégicas.

3. Creación de un Dashboard Interactivo y Personalizable:

- Descripción: Desarrollar un dashboard interactivo que permita a los usuarios explorar datos, filtrar información y obtener insights visuales.
- Meta: Facilitar el acceso a información clave de manera intuitiva y personalizada.

4. Identificación de Oportunidades de Expansión:

- Descripción: Identificar áreas geográficas y nichos de mercado específicos que representen oportunidades de expansión para nuevos restaurantes.
- Meta: Proporcionar datos estratégicos para la toma de decisiones sobre la apertura de nuevos locales.

Productos (Necesidad del cliente)

Modelo de recomendación: se propone un sistema de recomendación de restaurantes innovador que combina análisis de sentimientos y un modelo de filtrado colaborativo. Este producto busca mejorar la experiencia del usuario al ofrecer sugerencias personalizadas basadas en evaluaciones sentimentales y experiencias anteriores, permitiendo a los usuarios descubrir nuevos lugares culinarios dentro de su ciudad. Con un diseño modular y un pipeline automatizado, se busca proporcionar flexibilidad y escalabilidad en la introducción de nuevas categorías o estados para obtener recomendaciones específicas y actualizadas.

Alcance del proyecto

Antes de proceder, se realizó un recuento del número de negocios asociados a cada tipo específico de actividad comercial. Se identificó que las categorías de restaurantes y comida fueron los tipos de negocios más frecuentes y recurrentes en los datos recopilados de Yelp.

A continuación, se procedió a filtrar el conjunto de datos utilizando la columna 'categories', seleccionando exclusivamente aquellos negocios clasificados como restaurantes. Posteriormente, se realizó un conteo utilizando la variable 'state' para identificar los estados con el mayor número de restaurantes. Es importante destacar que todas estas acciones se llevaron a cabo utilizando el conjunto de datos de negocios de Yelp.

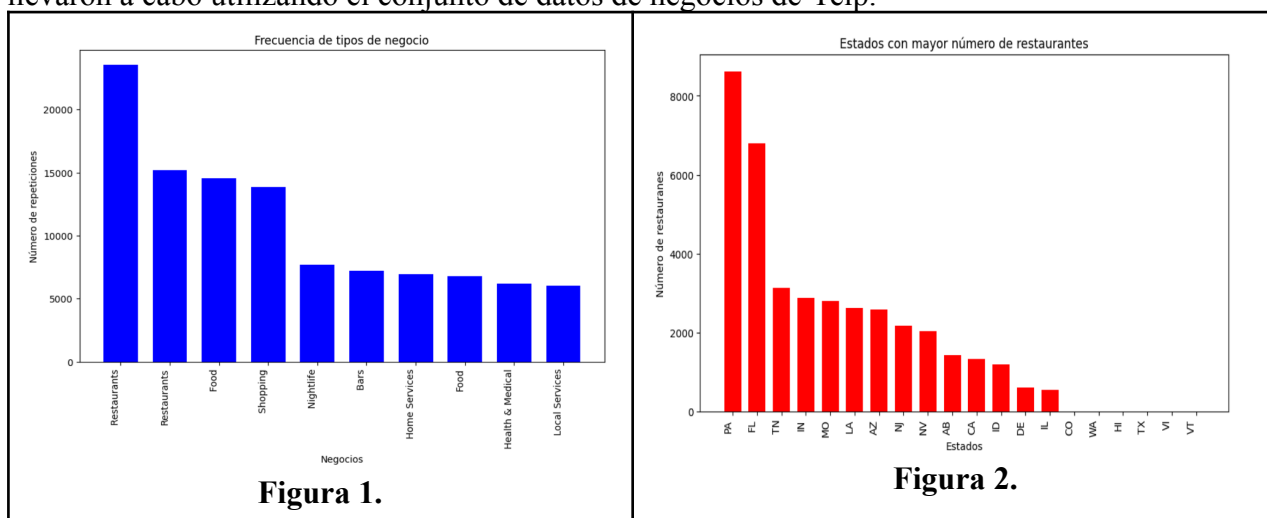


Figura 1. Conteo de número total de tipos de negocio
Figura 2. Conteo de restaurantes en cada uno de los estados

Equipo de trabajo y roles

Los roles experimentarán cambios a lo largo del proyecto, adaptándose a las necesidades específicas de cada sprint. No obstante, a continuación, se detallan de manera general los roles principales asignados a cada miembro del equipo en la siguiente tabla.

Aylén Sol Guzman	Data Science
Edith Priscila Muniz Cantu	Data Analytics
Maria Marcela Diaz	Data Analytics
Jorge Andres Jola Hernandez	Data Engineer
Franco Dylan Damian Luna Pedroso	Data Engineer

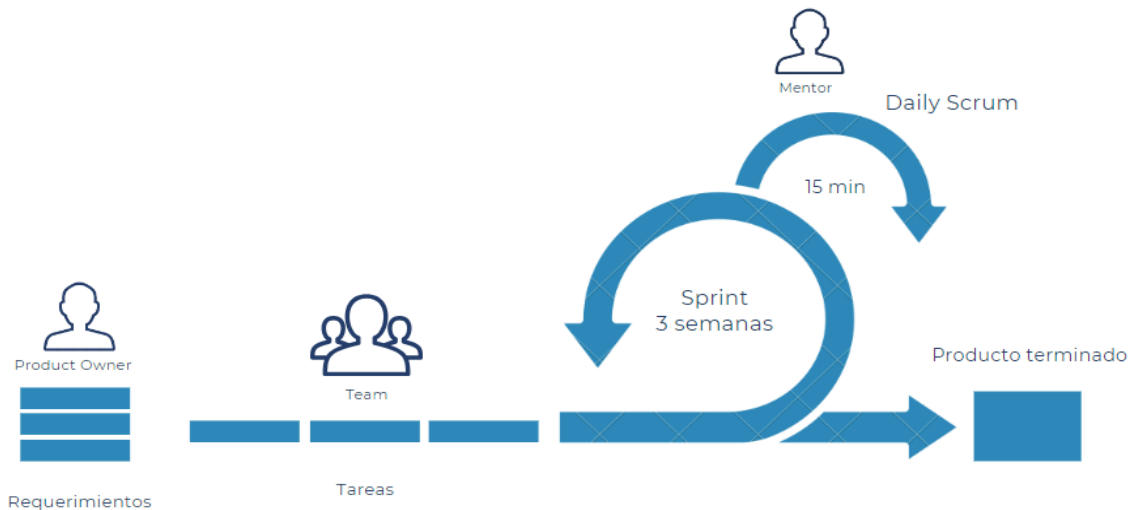
Tabla 1: Roles de cada integrante

Metodología de trabajo

Para el desarrollo de este proyecto, hemos optado por implementar la metodología SCRUM. Su objetivo principal es facilitar la entrega de productos de alta calidad de manera iterativa e incremental, basándose en principios fundamentales como la transparencia, inspección y adaptación. La estructura de SCRUM se organiza en torno a roles, eventos y artefactos clave.

El proyecto se ha dividido en tres sprints (etapas), cada uno con una duración de una semana. Cada miembro del equipo desempeña un rol específico en el proyecto, contribuyendo de manera especializada a los objetivos del equipo.

Nuestra metodología de trabajo incluye reuniones de planificación antes de comenzar el proyecto, con una duración de 30 minutos, donde se establecen las metas y asignaciones para el sprint. Asimismo, nos organizamos para realizar reuniones de revisión al final del día, con una duración de 15 minutos, para resumir y compartir los avances logrados durante la jornada. Este enfoque estructurado y colaborativo nos permite mantener una comunicación efectiva y garantizar un progreso constante hacia nuestros objetivos. **Figura 3. Tecnologías a utilizar en el presente proyecto**



Stack Tecnológico

Las principales tecnologías a utilizar se presentan a continuación:

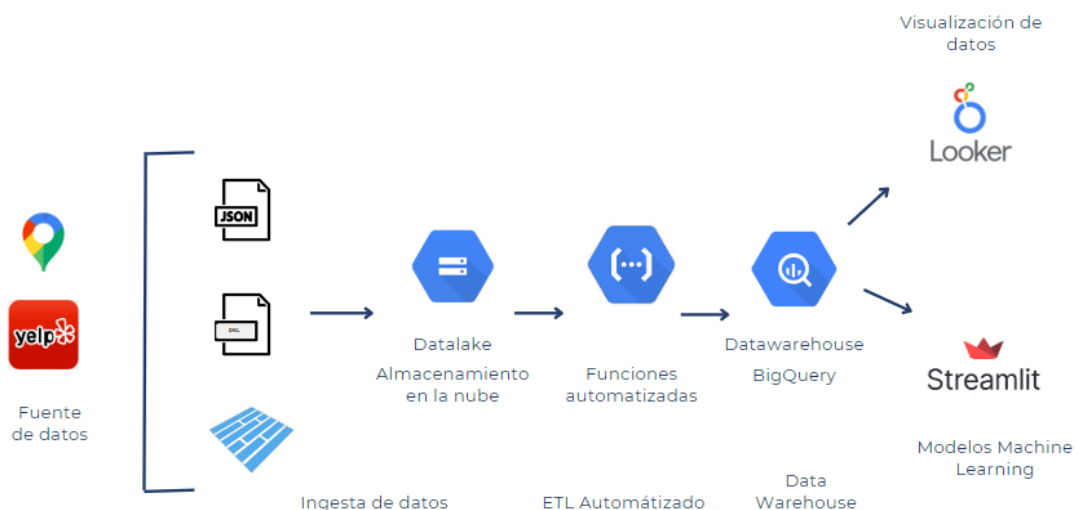


Figura 4. Tecnologías a utilizar en el presente proyecto

Cabe destacar que es probable que se incorporen nuevas tecnologías a medida que se avance en el desarrollo del proyecto.

Análisis Exploratorio Preliminar - Calidad del Dato

En esta sección, llevamos a cabo un análisis exploratorio preliminar de los conjuntos de datos proporcionados para el proyecto, procedentes de Yelp y Google Maps. El propósito principal es obtener una comprensión preliminar de los datos disponibles, evaluando su calidad e identificando posibles procesos de transformación que la información deberá experimentar.

Datasets Yelp: Overview Business

Es un dataset con información geográfica, horarios de apertura y reviews de los negocios registrados en Yelp.

	Data_Types	%_Null	Qty_Null	Qty_No_Null	Total_Registros
business_id	object	0.00	0	31597	31597
name	object	0.00	0	31597	31597
address	object	0.87	276	31321	31597
city	object	0.00	0	31597	31597
state	object	0.00	0	31597	31597
postal_code	object	0.03	11	31586	31597
latitude	float64	0.00	0	31597	31597
longitude	float64	0.00	0	31597	31597
stars	float64	0.00	0	31597	31597
review_count	Int64	0.00	0	31597	31597
is_open	Int64	0.00	0	31597	31597
attributes	object	0.97	307	31290	31597
categories	object	0.00	0	31597	31597
hours	object	11.95	3777	27820	31597
business_id_1	object	100.00	31597	0	31597
name_1	object	100.00	31597	0	31597
address_1	object	100.00	31597	0	31597
city_1	object	100.00	31597	0	31597
state_1	object	100.00	31597	0	31597
postal_code_1	object	100.00	31597	0	31597
latitude_1	object	100.00	31597	0	31597
longitude_1	object	100.00	31597	0	31597
stars_1	object	100.00	31597	0	31597
review_count_1	object	100.00	31597	0	31597
is_open_1	object	100.00	31597	0	31597
attributes_1	object	100.00	31597	0	31597
categories_1	object	100.00	31597	0	31597
hours_1	object	100.00	31597	0	31597

Tabla 2. Overview Business Dataset

- Hay 31,597 registros
- Hay 14 columnas con el 100% de registros nulos
- Es necesario modificar el tipo de dato de las columnas 'object'

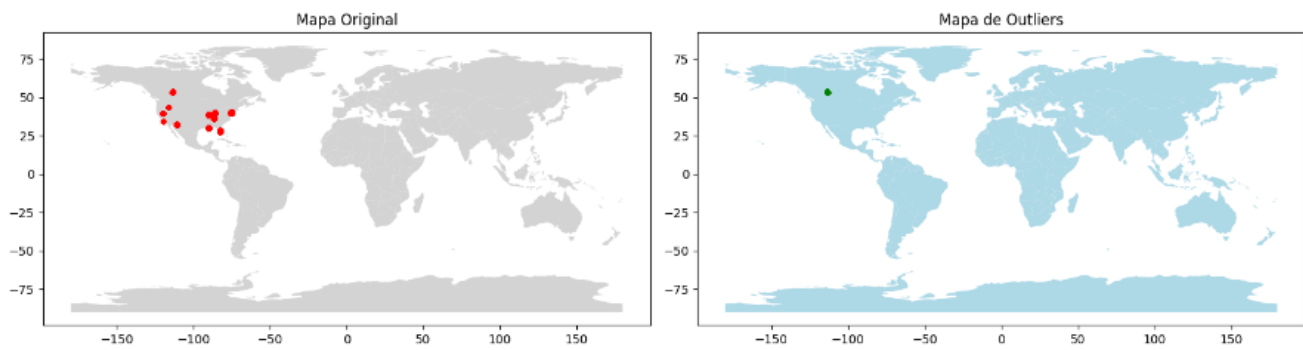


Figura 5. Mapa de Outliers - Yelp

Además al analizar las latitudes y longitudes, se observan outliers en Canadá, ya que el scope de este proyecto son ciudades de USA, por lo que será necesario posteriormente validar si realmente son outliers.

Overview Review

Es un dataset que contiene información de los usuarios y los reviews que han dejado a los negocios dados de alta en la plataforma de Yelp.

	Data_Types	%_Null	Qty_Null	Qty_No_Null	Total_Registros
text	object	0.0	0	2989567	2989567
cool	Int64	0.0	0	2989567	2989567
stars	float64	0.0	0	2989567	2989567
date	datetime64[ns, UTC]	0.0	0	2989567	2989567
funny	Int64	0.0	0	2989567	2989567
review_id	object	0.0	0	2989567	2989567
useful	Int64	0.0	0	2989567	2989567
business_id	object	0.0	0	2989567	2989567
user_id	object	0.0	0	2989567	2989567

Tabla 3. Overview Review Dataset

- No tiene registros nulos
- Hay un total de 2,989,567 registros en la tabla
- Hay 31,597 negocios distintos
- Hay 1,098,921 usuarios únicos que comentaron dichos negocios.

Overview Users

Es un dataset que contiene un resumen de los usuarios y los reviews que ha dejado desde que se inscribió a la plataforma.

	Data_Types	%_Null	Qty_Null	Qty_No_Null	Total_Registros
user_id	object	0.00	0	289338	289338
name	object	0.00	0	289338	289338
review_count	Int64	0.00	0	289338	289338
yelping_since	datetime64[ns, UTC]	0.00	0	289338	289338
useful	Int64	0.00	0	289338	289338
funny	Int64	0.00	0	289338	289338
cool	Int64	0.00	0	289338	289338
elite	float64	83.62	241930	47408	289338
friends	object	0.00	0	289338	289338
fans	Int64	0.00	0	289338	289338
average_stars	float64	0.00	0	289338	289338
compliment_hot	Int64	0.00	0	289338	289338
compliment_more	Int64	0.00	0	289338	289338
compliment_profile	Int64	0.00	0	289338	289338
compliment_cute	Int64	0.00	0	289338	289338
compliment_list	Int64	0.00	0	289338	289338
compliment_note	Int64	0.00	0	289338	289338
compliment_plain	Int64	0.00	0	289338	289338
compliment_cool	Int64	0.00	0	289338	289338
compliment_funny	Int64	0.00	0	289338	289338
compliment_writer	Int64	0.00	0	289338	289338
compliment_photos	Int64	0.00	0	289338	289338

Tabla 4. Overview Users Dataset

- Se observa que la columna 'elite' tiene casi un 84% de valores nulos, esto podría ser esperado o explicado ya que el término 'Elite' hace referencia a un programa de Yelp el cual es un reconocimiento especial otorgado a ciertos usuarios de Yelp que han demostrado un compromiso activo y contribuciones significativas a la comunidad de Yelp, por ende se sobreentiende que no todos los usuarios van a pertenecer a este grupo.

GoogleMaps:

Overview Sitios

Este dataset contiene información de los negocios registrados en Google Maps, con su información geográfica, reviews, horarios de atención, precio, entre otros datos relativos al negocio.

	Data_Types	%_Null	Qty_Null	Qty_No_Null	Total_Registros
name	object	0.00	0	212014	212014
address	object	0.50	1058	210956	212014
gmap_id	object	0.00	0	212014	212014
description	object	65.86	139622	72392	212014
latitude	float64	0.00	0	212014	212014
longitude	float64	0.00	0	212014	212014
category	object	0.00	0	212014	212014
avg_rating	float64	0.00	0	212014	212014
num_of_reviews	Int64	0.00	0	212014	212014
price	object	52.83	112014	100000	212014
hours	object	10.42	22101	189913	212014
MISC	object	0.67	1415	210599	212014
state	object	10.03	21263	190751	212014
relative_results	object	19.37	41057	170957	212014
url	object	0.00	0	212014	212014

Tabla 5. Overview Sitios Dataset

- Se observan 212,014 registros
- Hay 2 columnas con más del 50% de valores nulos
- Hay 5 columnas con menos del 20% de valores nulos

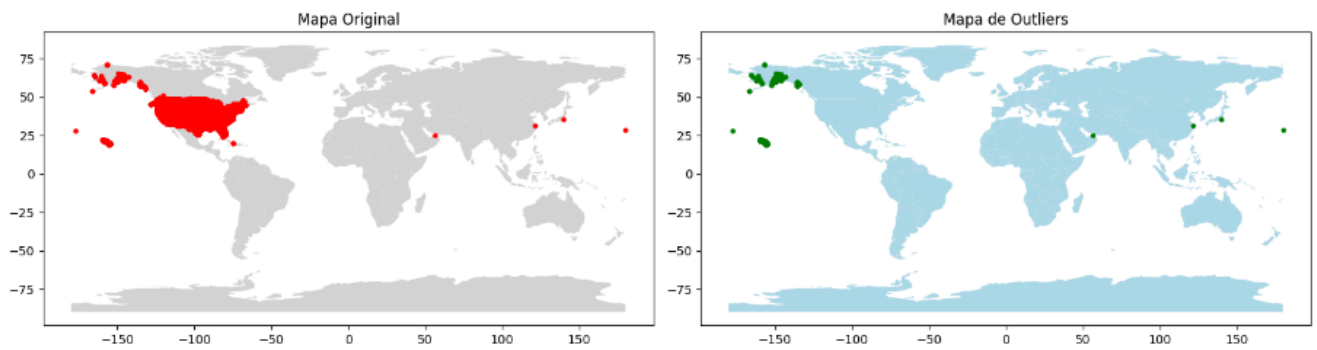


Figura 6. Mapa de Outliers - Google Maps

Además al analizar las latitudes y longitudes, se observan outliers en Canadá, ya que el scope de este proyecto son ciudades de USA, por lo que será necesario posteriormente validar si realmente son outliers.

Overview Estados

Se analizó un archivo de cada estado de los elegidos, y todos tienen la misma estructura y cantidad de registros:

	Data_Types	%_Null	Qty_Null	Qty_No_Null	Total_Registros
gmap_id	object	0.00	0	150000	150000
rating	Int64	0.00	0	150000	150000
text	object	40.23	60347	89653	150000
resp	object	82.91	124366	25634	150000
time	Int64	0.00	0	150000	150000
pics	object	0.00	0	150000	150000
name	object	0.00	0	150000	150000
user_id	float64	0.00	0	150000	150000

Tabla 6. Overview Estados Dataset

Estos datasets contienen información de cada estado, el id del negocio, su rating, el review, lo que el negocio contestó, algunas fotografías, entre otros datos.

En estos datos, las columnas con valores negativos son 'text' y 'resp' ya que no necesariamente los usuarios dejan un review escrito (sólo el rating de estrellas) y tampoco necesariamente les contestan los negocios a los reviews.

Calidad de los Datos

Para efectuar el análisis de calidad de los datos se utilizó Dataplex de Google Cloud Platform, en el que en base al EDA preliminar de cada uno de los datasets, se efectuó una pruebas específicas, por ejemplo para los datos de estados de Google Maps:

Dimensiones correctas
 Completeness ✔ 0 Errores
 Validity ✔ 0 Errores

Rules

Filtro Filtrar elementos										
Nombre de la columna ↑	Nombre de la regla	Tipo de regla	Estado	Evaluación	Dimensión	Parámetros	Filas con errores	Umbral	Consulta para obtener re	
gmap_id	-	Null Check	✔ Aprobado	Por fila	Integridad		0%	100 %	SELECT * FROM 'my-pr	▼
gmap_id	-	Row Condition Check	✔ Aprobado	Por fila	Validez	(LENGTH('gmap_id') >= 34 A...	0%	100 %	SELECT * FROM 'my-pr	▼
name	-	Null Check	✔ Aprobado	Por fila	Integridad		0%	100 %	SELECT * FROM 'my-pr	▼
name	-	Row Condition Check	✔ Aprobado	Por fila	Validez	(LENGTH('name') >= 1 AND L...	0%	100 %	SELECT * FROM 'my-pr	▼
rating	-	Null Check	✔ Aprobado	Por fila	Integridad		0%	100 %	SELECT * FROM 'my-pr	▼
rating	-	Value Set Check	✔ Aprobado	Por fila	Validez	set of: 5,4,1,3,2	0%	100 %	SELECT * FROM 'my-pr	▼
text	-	Row Condition Check	✔ Aprobado	Por fila	Validez	(LENGTH('text') >= 1 AND LE...	0%	100 %	SELECT * FROM 'my-pr	▼
time	-	Null Check	✔ Aprobado	Por fila	Integridad		0%	100 %	SELECT * FROM 'my-pr	▼
user_id	-	Null Check	✔ Aprobado	Por fila	Integridad		0%	100 %	SELECT * FROM 'my-pr	▼
user_id	-	Statistics Range Check	✔ Aprobado	Agregación	Validez	statistic: mean, range: [min: 1...	N/A	100 %	- This query executes t	▼
user_id	-	Range Check	✔ Aprobado	Por fila	Validez	min: 1.0000033022865685E2...	0%	100 %	SELECT * FROM 'my-pr	▼

Figura 7. Ejemplo pruebas calidad de datos para un estado

En este caso se realizaron varios tipos de reglas dependiendo el valor de la columna y el dataset, y se observa que todas las pruebas fueron aprobadas por lo que se determina que el dato es de calidad.

Este mismo proceso se realizó para todos los datasets provistos.

KPIs

En base a los datos previamente analizados, se determinaron los siguientes KPIs para este proyecto.

Índice de Satisfacción de Clientes (ISC):

$$ISC = \frac{\text{Cantidad de reseñas positivas de clientes}}{\text{Total de reseñas recibidas}} \times 100$$

- Descripción: Evaluar la satisfacción general de los clientes y fijar la meta de lograr un aumento del 7% en el ISC en los próximos 6 meses.

Tasa de Reseñas de 0 Estrellas en los Primeros 3 Meses (TR0E):

$$TR0E = \frac{\text{Número de reseñas de 0 estrellas}}{\text{Total de reseñas recibidas}} \times 100$$

- Descripción: Este KPI mide la proporción de reseñas que reciben 0 estrellas en comparación con el total de reseñas recibidas. La meta es mantener esta tasa en 0 durante los primeros 3 meses, lo que indica que no se han recibido reseñas de 0 estrellas en ese período.

Índice de Experiencia del Cliente:

$$IEC = \frac{\text{Puntuación de la Experiencia del Cliente}}{\text{Total de reseñas}} \times 100$$

- Descripción: Medir la percepción de los clientes que vienen en grupo, buscando mantener un promedio mensual de sentimiento positivo, representado por un valor superior a 0.25.

Incremento Mensual de Reseñas

$$\text{Incremento Mensual Reseñas} = \frac{\text{Reseñas Mes Actual} - \text{Reseñas Mes Anterior}}{\text{Reseñas Mes Anterior}} \times 100$$

- Descripción: Este KPI mide el crecimiento porcentual mensual en la cantidad de reseñas. Proporciona una evaluación cuantitativa del aumento en la participación y feedback de los usuarios en comparación con el mes anterior. El objetivo es alcanzar un aumento constante del 15% mensual en el número de reseñas, indicando un aumento positivo en la interacción y la retroalimentación de los clientes.

Planificación de Trabajo - División de esfuerzos

La planificación del desarrollo del proyecto se llevará a cabo mediante el uso del Diagrama de Gantt.

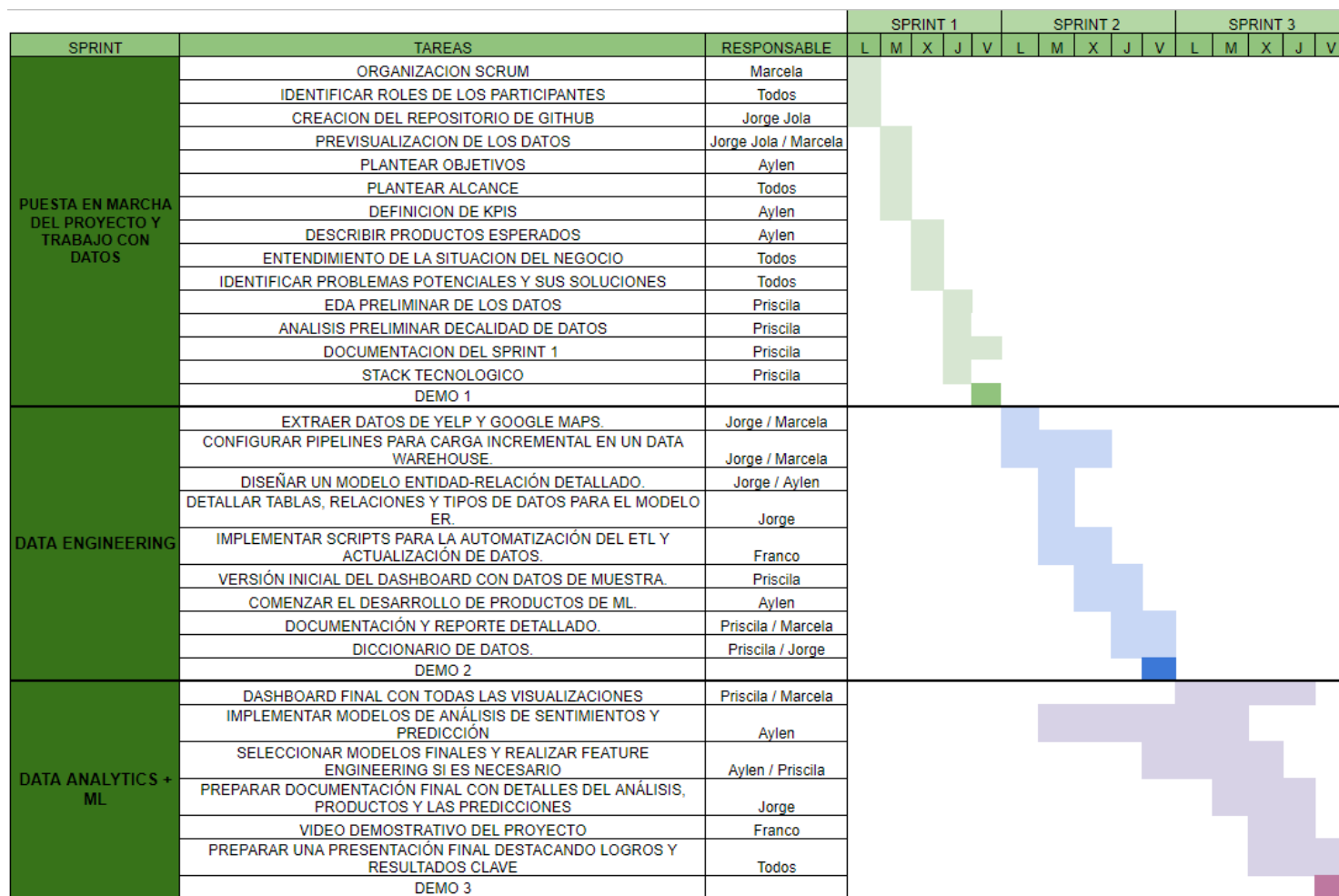


Figura 8 . Diagrama de Gantt para el proyecto