22-3-2024

**Yelp & Google maps- Reviews and Recommendations**

Proyecto Final- Sprint 1

Kensit Cortes

Juan Carlos Sánchez

Facundo Blanco

José Efraín Pazos

Rafael Miranda

Data fiveBusiness Consultans

Contenido

[Introducción 2](#_Toc161948493)

[Objetivo General 2](#_Toc161948494)

[Objetivos específicos 2](#_Toc161948495)

[Alcance del proyecto 3](#_Toc161948496)

[Formulación de KPI´s 5](#_Toc161948497)

[Repositorio en GitHub 5](#_Toc161948498)

[Stack tecnológico 6](#_Toc161948499)

[Metodología de trabajo 6](#_Toc161948500)

[Diseño detallado 6](#_Toc161948501)

[Equipo de trabajo 7](#_Toc161948502)

[Cronograma general 8](#_Toc161948503)

[Análisis preliminar de los datos 9](#_Toc161948504)

# Introducción

En el mundo actual, la información generada por las opiniones de los usuarios en plataformas con Yelp y Google Maps se ha convertido en una fuente invaluable de datos pata comprender las preferencias y opiniones de los consumidores, siendo el análisis de estos un factor crucial para el éxito de los negocios. En este contexto, nosotros como consultora nos enfrentamos al desafío de realizar un análisis del mercado estadounidense, específicamente, los restaurantes de las 5 ciudades con mayor afluencia turística.

Este proyecto tiene como objetivo proporcionar a nuestro cliente, parte de un conglomerado de empresas de restaurantes, una comprensión del panorama actual y futuro del mercado, con el fin de optimizar sus estrategias comerciales y mejorar la experiencia del cliente

# Objetivo General

Desarrollar e implementar un sistema de gestión de datos integral y eficiente que permita la recopilación, depuración y disponibilidad de información proveniente de diversas fuentes, entre esas Yelp, con el fin de facilitar la creación de un DataWarehouse robusto y funcional. Este sistema también permitirá el análisis y la generación de reportes significativos sobre las opiniones de los usuarios en Estados Unidos en cuanto a restaurantes, así como el entrenamiento y despliegue de un modelo de machine learning enfocado en un sistema de recomendación.

# Objetivos específicos

* Recopilar, limpiar y estructurar los datos de diversas fuentes como Yelp y Google maps
* Implementar técnicas de depuración de los datos, incluyendo la identificación y tratamiento de valores atípicos, nulos y la estandarización de formatos con el fin de garantizar calidad y consistencia de los datos almacenados
* Diseñar e implementar un DataWarehouse que permita almacenar y organizar eficientemente los datos recopilados de diferentes fuentes, utilizando tecnologías de almacenamiento de datos adecuadas
* Integrar el modelo de machine learning en el sistema de gestión de datos para permitir la generación de recomendaciones en tiempo real basadas en las preferencias individuales de los usuarios.
* Realizar un dashboard intuitivo e interactivo, con el fin de proporcionar métricas, gráficos y brindar los resultados del análisis de datos.

# Alcance del proyecto

Alcance temporal

* Consta de 6 semanas, con duración de los sprint de 2 semanas.

**Sprint 1**

Desde: 4 de maro de 2024

Hasta: 22 de marzo de 2024

**Sprint 2**

Desde: 25 de marzo de 2024

Hasta: 5 de abril de 2024

**Sprint 3**

Desde: 8 de abril de 2024

Hasta: 20 de abril de 2024

Recursos

* Data scientist (una persona)15 horas de trabajo a la semana durante 6 semanas
* Data engineer (dos personas)10 horas de trabajo a la semana durante 6 semanas
* Data analyst (dos personas)15 horas de trabajo a la semana durante 6 semanas

Entregables

**Sprint 1**

* 3 KPI’s
* Documentación general del proyecto
* EDA de los datos
* Repositorio en GitHub
* Stack tecnológico seleccionado para trabajar
* Metodología de trabajo
* Diseño detallado
* Roles y responsabilidades del equipo de trabajo
* Cronograma general de trabajo, representado con un diagrama de Gantt
* Análisis preliminar de calidad de datos

**Sprint 2**

* ETL completo
* Estructura del DataWarehouse
* Pipeline ETL automatizado
* Diseño del Modelo Entidad Relación detallado, incluye tablas, tipos de datos, llaves primarias y llave foránea
* Pipelines para alimentar el DataWarehouse
* Data Warehouse
* Automatización y validación de los datos
* Documentación
* Diccionario de datos
* Producto Mínimo Viable

**Sprint 3**

* Diseño de Reportes
* KPIs implementados en el dashboard
* Modelos de Machine Learning
* Modelo de ML en producción
* Documentación
* Selección del modelo, feature engineering
* Informe de análisis
* Video del proyecto realizado

Alcance geográfico

Del país de Estados unidos, se tomarán las 5 ciudades con mayor numero de turismo que son: New York, Miami, Orlando, Los Ángeles y San Francisco. Se tomarán empresas de restaurantes

# Formulación de KPI´s

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **KPI** | **Descripción** | **Formula** | **Periodicidad** | **Objetivo** |
| Índice de Satisfacción del cliente | Porcentaje de clientes satisfechos con su experiencia |  | Mensual | Aumentar el índice de satisfacción un 10% en 6 meses |
| Tasa de retención de clientes | Porcentaje de clientes que regresan a un restaurante |  | Trimestral | Aumentar la tasa de retención un 5% en 1 año |
| Valoración promedio de las reseñas | Calificación promedio que recibe un negocio en las reseñas |  | Trimestral | Aumentar la valoración promedio 5 puntos en 1 año |

# Repositorio en GitHub

https://github.com/FacuSB/PF\_Grupal

# Stack tecnológico

|  |  |  |
| --- | --- | --- |
|  | Herramienta | Utilidad |
| Lenguaje de programación | Python | Se utilizará para la creación del ETL y el EDA, y demás manipulaciones requeridas en la base de datos |
| Librerías | Numpy | Se utilizará para cálculos numéricos |
| Pandas | Se utilizará para la manipulación y análisis de los datos |
| Scipy | Se utilizará para estadísticas y modelado en Python |
| Matlotlib | Se utilizará para la generación de gráficos |
| Seaborn | Se utilizará para visualización estadística en Python |
| Scikit learn | Se utilizará para técnicas de machine learning |
| Entornos de desarrollo | Jupyter Notebook | Se utilizará en la exploración interactiva de los datos y creación de informes |
| Bases de Datos | SQL | Para crear y gestionar bases de datos relacionales |
| Herramienta de inteligencia de negocio | Power BI | Se utilizará para la creación de tableros |
| Máquina virtual | Google Cloud Platform | Se utilizará la plataforma en la nube de Google para procesamiento de datos |
| Automatización y control de versiones | Git | Control de versiones |

# Metodología de trabajo

Marco de trabajo ágil Scrum

# Diseño detallado

Sprint Planning: Sesión de planificación del Sprint

Se realiza una reunión previa al comienzo del Sprint, donde se definen el enfoque y alcance del proyecto, objetivos, entregables, actividades a realizar y roles de los integrantes del equipo

Comienzo del Sprint de Scrum:

Cada Sprint cuenta con una duración de dos semanas, donde cada miembro del equipo trabajara en las tareas pendientes que se establecieron en la sesión de planificación del sprint

Daily Stand Up:

Se organizan reuniones de 15 minutos los días lunes, miércoles y viernes con el mentor para informar con respecto al trabajo que se este realizando e identificar cualquier obstáculo e inquietudes que hayan surgido

Sprint Review:

Se presenta el trabajo una vez terminado el Sprint, es decir el viernes de la segunda semana. Tendrá una duración de 45 minutos, donde el equipo se reunirá con el cliente para hacer una revisión del Sprint y verificar que el producto entregable coincida con los objetivos trazados al comienzo de cada sprint

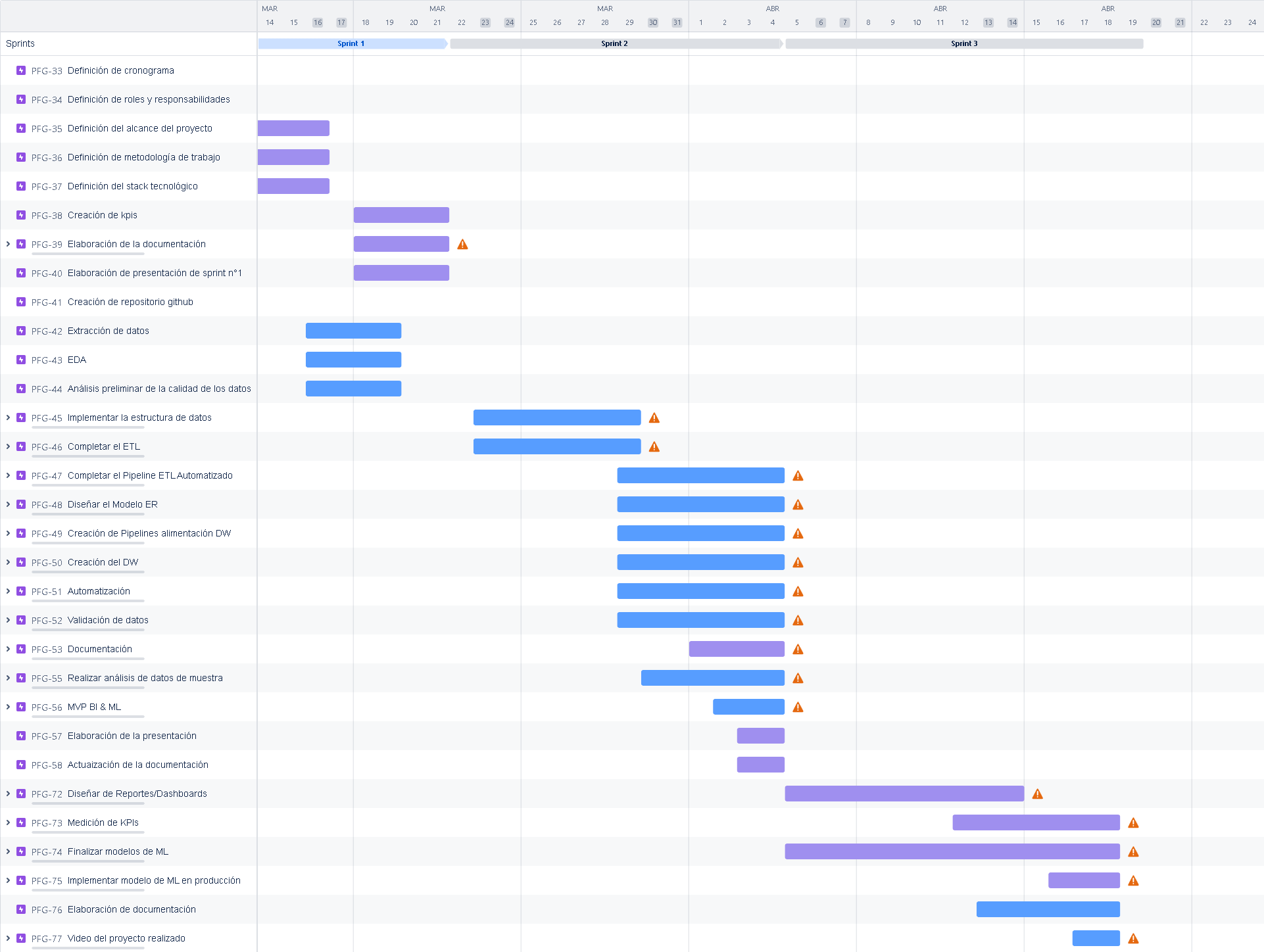
# Equipo de trabajo

|  |  |
| --- | --- |
| **Nombre y rol del integrante** | **Responsabilidades** |
| Facundo Blanco  **Data Engineer** | * Diseñar y construir la arquitectura del DataWarehouse * Diseñar e implementar un modelo de machine learning que utilice técnicas de recomendación para predecir las preferencias de los usuarios y ofrecer recomendaciones personalizadas * Supervisar el procesamiento de los datos |
| Juan Carlos Sánchez  **Data Engineer** | * Crear un modelo predictivo utilizando técnicas de machine Learning para realizar pronósticos sobre el crecimiento futuro de los negocios * Desarrollar el pipeline automatizado |
| Rafael Miranda  **Data Analyst** | * Realizar análisis exploratorio de los datos * Identificar patrones, tendencias y relaciones utilizando técnicas estadísticas * Generar informes y visualizaciones para comunicar patrones y tendencias |
| José Efraín Pazos  **Data Scientist** | * Explorar y limpiar los datos, identificar y tratar valores atípicos, manejar datos faltantes y transformar variables según sea necesario para mejorar la calidad de los datos * Realizar análisis descriptivos y la implementación del modelo predictivo * Seleccionar las tecnologías adecuadas para garantizar un procesamiento rápido y confiable de grandes volúmenes de datos |
| Kensit Cortes  **Data Analyst** | * Obtener datos del DataWarehouse creado * Realizar análisis descriptivos con los datos * Crear Data Storytelling |

# Cronograma general

https://proyectofinalhenryg5pt6.atlassian.net/jira/software/projects/PFG/boards/1/timeline?timeline=WEEKS

* Daily Stand up



# Análisis preliminar de los datos

1. **Validación de Datos**

Aunque los datos parecen precisos en una primera instancia, no están completamente actualizados, ya que solo contienen información hasta el año 2022. En el caso del DataFrame de Yelp, los años varían, pero para mantener la integridad de los archivos, considero apropiado tomar en cuenta los años que comprenden desde 2009 hasta 2018. En cuanto a los datos de Google, no disponen de una columna de fecha, lo que impide determinar la recencia de los datos con exactitud. Sin embargo, la integridad de los archivos parece bastante completa. Como prueba, busque los lugares que aparecen en el DataFrame combinado en Google Maps, donde sí aparecen las fechas de las reseñas. Por lo tanto, imagino que consultando los IDs en la API de Google Maps más adelante, podremos completar estos datos

1. **Completitud**

En los datos de Yelp, se puede observar que, a partir de la mitad, las columnas en los datos de ‘business’ se repiten, pero con todos los datos nulos, lo que requiere una limpieza. Además, en el dataset de ‘users’, las columnas de ‘compliment’ en su gran mayoría también tienen datos vacíos. Considerando esto, estos datos no son relevantes para un análisis posterior. En los datos de Google Metadata, la gran mayoría de descripciones y precios también se encuentran vacíos, y en el dataset de Google Reviews, las columnas ‘pics’ y ‘resp’ también están vacías. Todas estas columnas podrán descartarse ya que están muy incompletas.

1. **Consistencia**

Es un poco difícil encontrar la paridad entre ambos datasets, pero existen coincidencias. Los problemas que se generan para establecer una relación entre los datos se deben a que solo estamos trabajando con una pequeña parte de estos. Por ejemplo, solo para hacer las pruebas, estamos trabajando con solo un metadata de Google y la totalidad de Yelp Business. A pesar de esto, se pueden encontrar coincidencias y se pueden emparejar las tablas a través de las columnas ‘name’, que contienen el nombre del local, o las coordenadas de las mismas. En cuanto a la relación entre cada uno de los datasets por separado, Yelp tiene sus datos bien emparejados a través de los distintos IDs como el ‘business\_id’ o el ‘user\_id’ y los datos de Google están emparejados a través de sus columnas ‘gmap\_id’.

1. **Unicidad**

Existen bastantes registros duplicados. Esto se tendrá que limpiar con una función para eliminar los registros duplicados de todos los datasets que se van a utilizar más adelante.

1. **Relevancia**

Actualmente, estamos en el proceso de evaluación para determinar qué datos serán prescindibles para nuestro producto final. Sin embargo, realizar un análisis en esta etapa inicial podría resultar en la eliminación de datos que podrían ser relevantes en etapas posteriores. Por lo tanto, la decisión sobre qué datos son prescindibles se tomará más adelante, basándose en su relevancia y disponibilidad para nuestro producto. Esta estrategia nos permitirá mantener la flexibilidad y asegurar que no descartamos información valiosa prematuramente.