DATA NEXUS

CONSULTING

DOCUMENTACION

INDICE

1. CONSIDERACIONES GENERALES

1.1. Introducción

Las opiniones de los usuarios son una fuente valiosa de información que crece constantemente gracias a plataformas de reseñas como *Yelp* y *Google Maps*. *Yelp* permite a los usuarios dejar reseñas de diversos negocios. Estas reseñas reflejan las experiencias de los usuarios y son cruciales para que las empresas comprendan cómo son percibidas, midan su desempeño, y descubran áreas de mejora en sus servicios.

De manera similar, *Google Maps* ofrece una plataforma de reseñas integrada en su servicio de localización, donde los usuarios pueden compartir sus experiencias sobre diversos negocios. Esta información es esencial para que las empresas comprendan la imagen que tienen los usuarios de sus locales, lo que les ayuda a ajustar sus estrategias y mejorar la calidad de sus servicios.

En conjunto, el análisis de estas reseñas puede ser determinante para la planificación de estrategias empresariales, permitiendo a las empresas tomar decisiones informadas sobre dónde abrir nuevos locales, qué aspectos de sus servicios deben mejorar y cómo satisfacer mejor las expectativas de sus clientes.

1.2. Acerca del proyecto

Como consultora de datos, Datanexus realiza un análisis profundo del mercado estadounidense para un inversor dentro del conglomerado de empresarios gastronómicos y afines. El enfoque principal es realizar un análisis exhaustivo de restaurantes, utilizando datos de las plataformas *Google Maps* y *Yelp* que proporciona información valiosa para los inversores, ayudándoles en la toma de decisiones estratégicas.

1.3. Roles

Data Engineer: Gustavo Pardo.
Data Scientist: Yair Juarez, Jaime Gold.
Data Analyst: Eliana Larregola, Rocio Alaniz.

1.4. Alcance

Categoría: El análisis se centra exclusivamente en restaurantes. Otras categorías no son consideradas en el alcance de este proyecto.

Área: Luego de una investigación preliminar sobre los estados en Estados Unidos más poblados y con mayor densidad poblacional, el análisis se limita a California, Texas, Florida, New York, Pennsylvania.

Fuentes de datos: Se utilizan los datos proporcionados por las compañías *Google Maps* y *Yelp.*

1.5. Metodología de trabajo

Para el desarrollo del trabajo se adopta la metodología *SCRUM* como metodología ágil. Esto permite iterar rápidamente y mejorar el *MVP* a partir de la retroalimentación del *Product Owner* y los resultados obtenidos.

1.6. Objetivos

1.6.1. Objetivo general

Realizar un análisis exhaustivo de las reseñas de restaurantes en *Google Maps* y *Yelp* para proporcionar una visión clara de las oportunidades de mercado, determinar ubicaciones óptimas para nuevos establecimientos y desarrollar un sistema de recomendación que ayude a los usuarios a encontrar restaurantes basados en su ciudad y preferencias.

1.6.2. Objetivos específicos

1. Analizar popularidad y éxito de los locales gastronómicos:
    a. Evaluar las reseñas positivas de los usuarios hacia los restaurantes, identificando tendencias positivas en las opiniones para predecir los restaurantes que tendrán mayor crecimiento.
2. Identificar Ubicaciones Óptimas:
    a. Analizar datos geográficos para determinar las mejores ubicaciones para abrir nuevos locales de restaurantes.
3. Desarrollar un Sistema de Recomendación:
    a. Crear un sistema de recomendación que sugiere restaurantes a los usuarios, basado en la ubicación (ciudad) y sus preferencias.
    b. Crear un sistema de recomendación que sugiere ciudades a los empresarios, basado en el estado y sus preferencias.

1.7. Productos entregables

- *Dashboard* interactivo
- Sistema de recomendación
- Documentación (repositorio de *Github* y documento actual)
- Diccionario de datos

2. DESARROLLO

El presente proyecto se desarrolla en 3 *sprints* principales. Cada *sprint*, corresponde a un bloque que tiene una duración de una semana en los cuales se establecen diversos hitos y objetivos y finaliza con una demo con el *Product Owner(PO)*. El PO es el encargado de supervisar el avance del proyecto, para presentar el trabajo de la semana, ajustar detalles, recibir sugerencias y así redireccionar esfuerzos para encaminar los productos a lo solicitado por el cliente.

2.1. *Sprint* 1: Puesta en marcha del proyecto y Trabajo con Datos

El *sprint* 1, es una etapa fundamental, ya que sirve para plantear las generalidades del proyecto, así como también dividir tareas, establecer cronogramas y crear flujos de trabajo entre otros.

En primer lugar, todo lo presentado en el capítulo 1 de generalidades, corresponden a lo realizado en el *sprint* 1.

Por otro lado, a continuación se anexa la información extra trabajada en este *sprint*.

2.1.1. Repositorio de GitHub

Se crea un repositorio público en *Github* donde se trabaja colaborativamente. El repositorio del proyecto contiene un *README* con consideraciones generales del proyecto, también se encuentran distintos archivos de código divididos en carpetas diferentes que se crearon durante el desarrollo del proyecto.

2.1.2. Análisis Exploratorio de Datos (EDA)

Al iniciar se realiza un análisis exploratorio de los datos para entender la naturaleza de los datos obtenidos de *Google maps* y *Yelp*. Este proceso permite identificar patrones, tendencias, *outliers* y relaciones de los datos. Informes de EDA.

2.1.3. Diagrama de Gantt y tareas en Trello

Cómo equipo, coordinar y aprovechar las fortalezas individuales para lograr sinergia y avanzar en el desarrollo del proyecto es indispensable. A lo largo de cada semana se definen y asignan distintas tareas para cada integrante del equipo, las cuales son determinantes para completar y presentar la propuesta de trabajo final al cliente.
En el diagrama de Gantt (Figs. 1, 2 y 3) se observa el cronograma con la distribución de las tareas de los diferentes *sprints* a través de las distintas etapas del proyecto. Tablero de Trello.

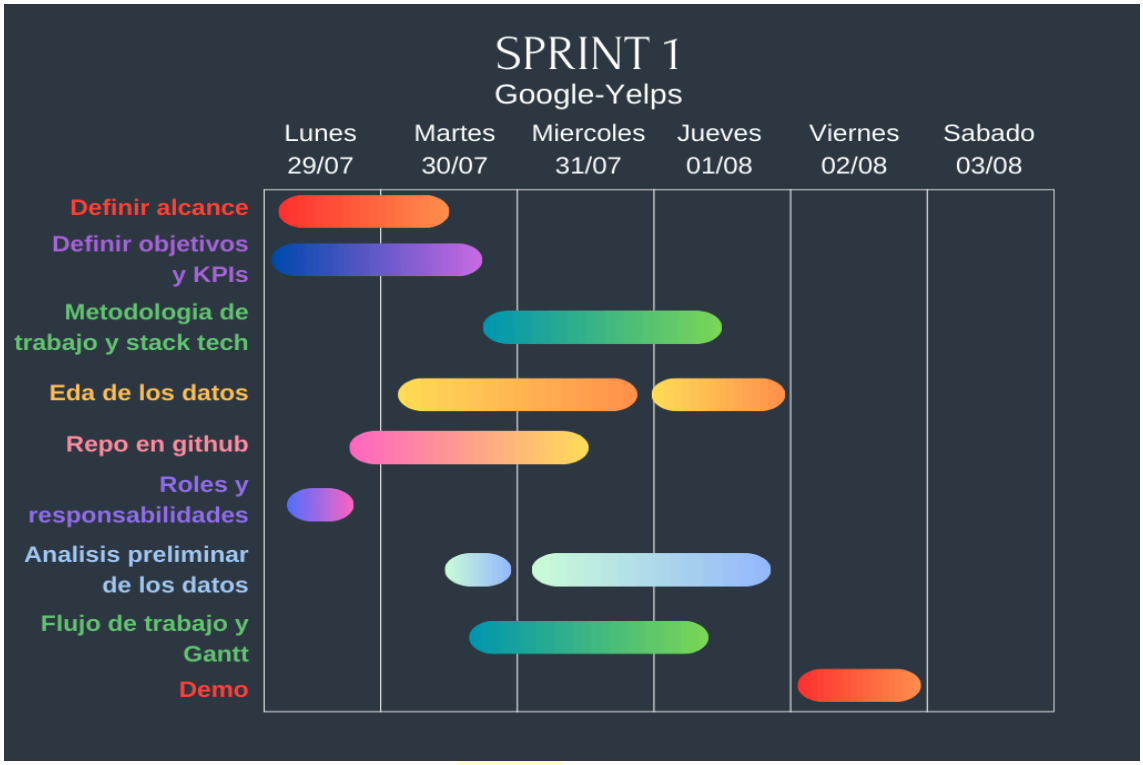

Fig. 1: Diagrama de Gantt del Sprint 1.

Fig. 2: Diagrama de Gantt del Sprint 2.

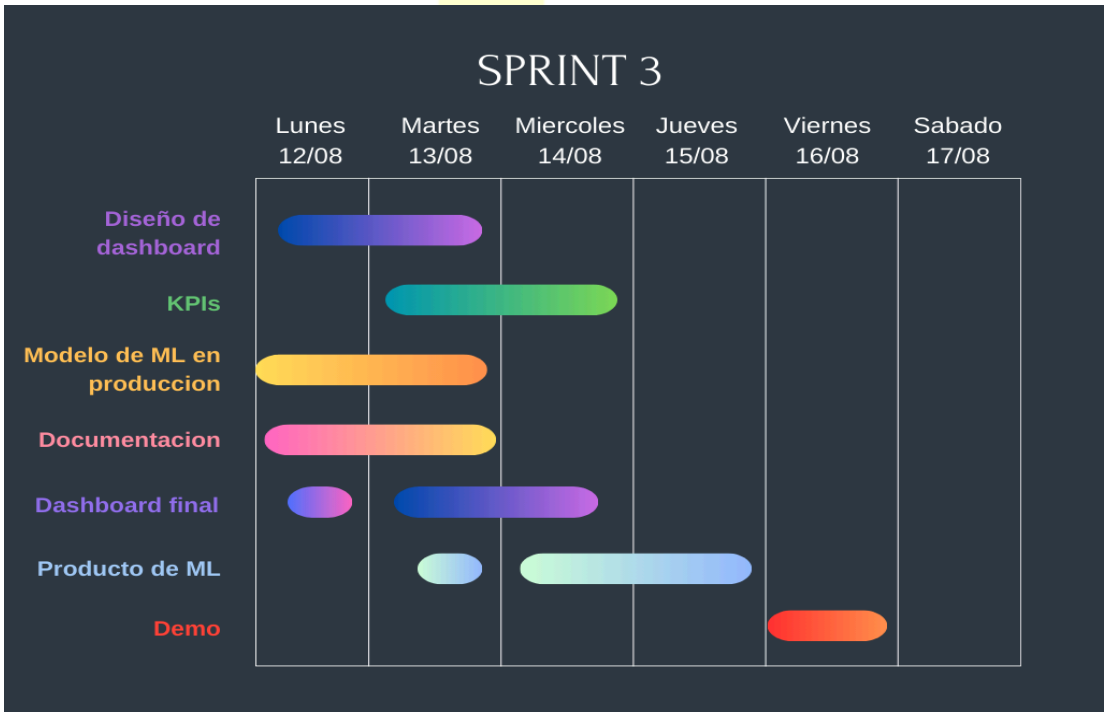

Fig. 3: Diagrama de Gantt del Sprint 3.

2.1.4. Stack Tecnológico

Una parte muy importante de la solución propuesta es definir qué herramientas se utilizan para la arquitectura del proyecto (Fig. 4). Seleccionando una pequeña porción de los datos disponibles y de un análisis preliminar, se resuelve proponer las siguientes tecnologías:

Fig. 4: Stack tecnológico utilizado.

2.1.5. Indicadores clave de rendimiento (KPI)

Del entendimiento de la problemática planteada en el proyecto, surgen tres propuestas de KPIs que se buscan resolver con el análisis de los datos y las herramientas utilizadas para evaluar su cumplimiento (Ver sección 2.3.2 para más detalle). Ellos son:

- KPI 1: Aumentar 1% la cantidad de restaurantes en las ciudades más propicias en un plazo de un año.
- KPI 2: Mejorar el rating promedio de restaurantes que tienen entre 3 y 4 estrellas en un 4% en el plazo de un año.
- KPI 3: Aumentar en un 4% las reviews positivas de restaurantes en el plazo de un año.

2.1.6. Flujo de trabajo

Definir los pasos y procesos de una idea junto con las tecnologías a utilizar, son una parte fundamental para completar un proyecto de manera eficiente. El siguiente flujo (Fig. 5) define cómo se mueven las tareas, los recursos y la información a través de cada etapa del mismo.

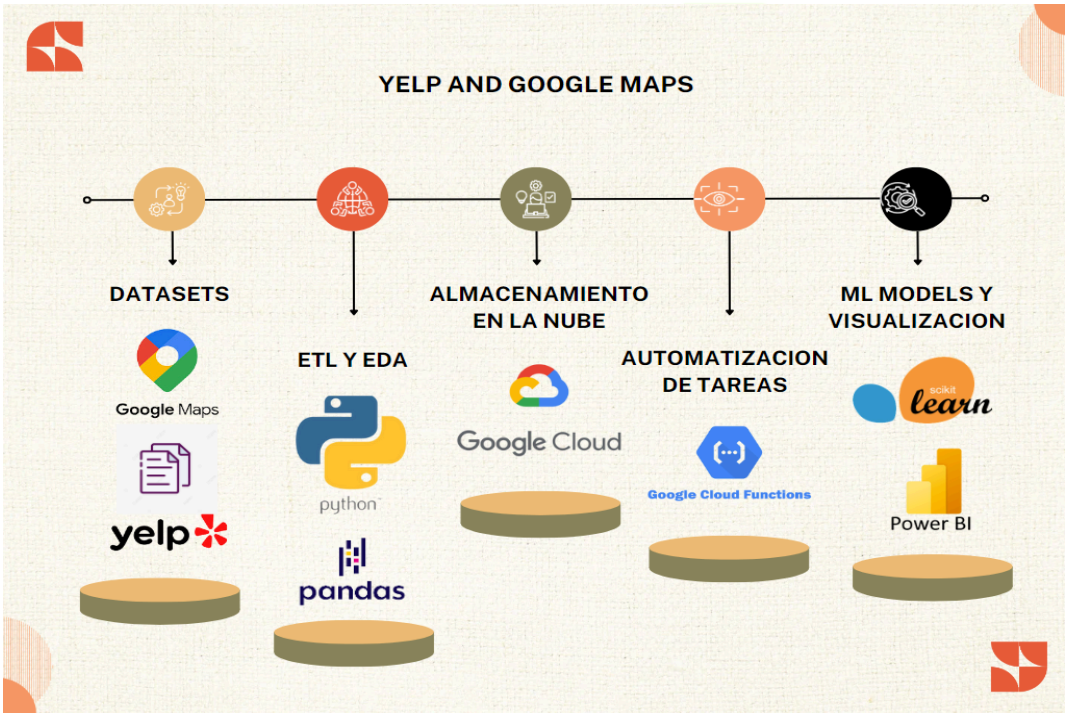

Fig. 5: Flujo de trabajo.

2.2. Sprint 2: Data Engineering

Durante el *sprint* 2 se trabaja principalmente montando la infraestructura del proyecto, con *pipelines* para realizar el proceso de ETL automatizado apuntando a una estructura de tipo *Data Warehouse*, contemplando la carga incremental de datos y la validación de datos a través de correo electrónico.

2.2.1. Diagrama E/R

Se crea Datanexus que es un *Data Warehouse* creado en *BigQuery* del servicio *Google Cloud Platform (GCP)* el cual presenta el siguiente diagrama E/R (Fig. 6).

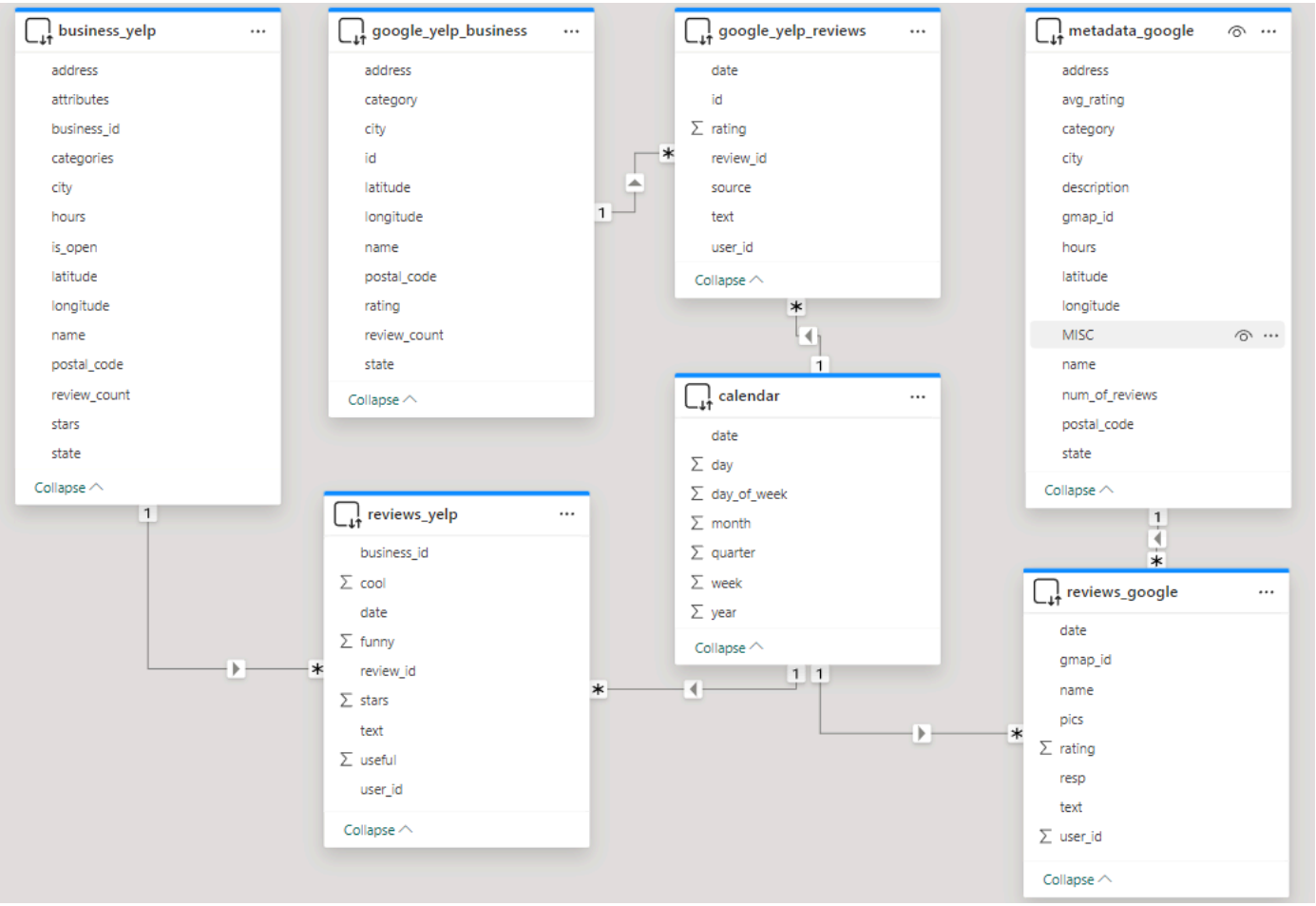

Fig. 6: Se muestra el diagrama E/R

2.2.2. Pipeline: ciclo de vida de lo datos

Este diagrama del ciclo de vida de los datos (Fig. 7), muestra el camino del proceso de los mismos desde el acceso a los datos de *Google maps* y *Yelps*, creación de un proyecto en la nube (GCP) y almacenamiento en *buckets*, revisar y dar permisos de control de procesos en la nube, proceso de ETL automatizado (*Cloud Functions*), creación y carga de datos en el *Data Warehouse (Bigquery)*, conexion con la base de datos en *Power BI*, conexión a la base de datos para el modelo de *machine learning* usando las credenciales que creamos GCP desde VSC, y luego *deployment* en *Streamlit*.

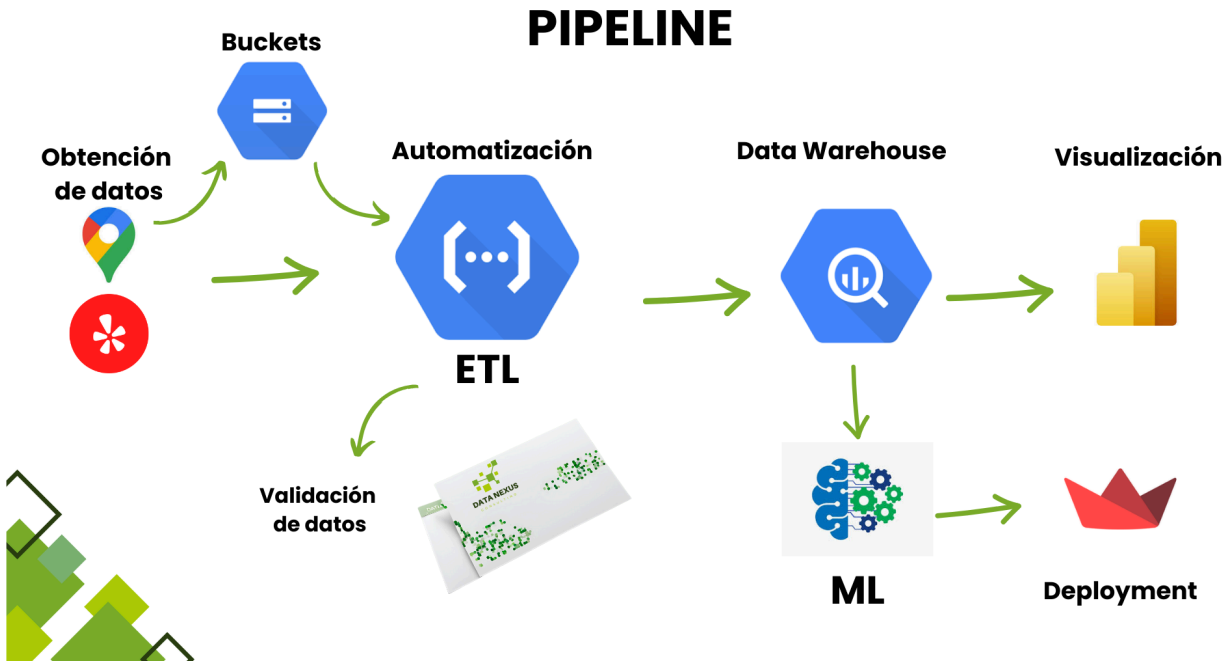

Fig. 7: Pipeline que representa el flujo de vida de los datos.

2.2.3. ETL

Para el proceso de ETL automatizado con los datos de *Google Maps* y *Yelp,* se crea una función para cada tipo de dato. Estas funciones comienzan a trabajar cuando un nuevo archivo se sube a los *buckets*, se cargan los datos y empieza el proceso de ETL devolviendo datos ya procesados y transformados. El proceso de ETL envuelve el filtrado de datos por restaurantes, los cuales corresponden al foco del proyecto.
El segundo paso del filtrado es para obtener los estados indicados anteriormente en el alcance del proyecto (California, Texas, New York, Pennsylvania y Florida). El proceso continúa con el acceso a los datos de columnas anidadas, eliminación de columnas irrelevantes, tratado de nulos, verificación de duplicados y cambio de tipos de datos. El siguiente paso es la carga de los datos limpios en su respectivas tablas y se almacenan en el *data warehouse* llamado Datanexus (*Bigquery*).

2.2.4. Validación de datos

Se implementa un sistema de validación que envía un reporte por email al administrador de la base de datos en caso de un error (Fig. 8) o que la carga haya sido exitosa (Fig. 9).

Fig. 8: Correo electrónico recibido en caso de error en la carga de datos.

Fig. 9: Correo electrónico en caso de que la carga de datos haya sido exitosa.

2.2.5. Diccionario de datos

Se crea un diccionario de datos, herramienta fundamental en la gestión de base de datos del proyecto, el cual proporciona una descripción detallada de la base de datos llamada Datanexus. El mismo permite comprender rápidamente qué datos están disponibles, cómo están organizados y relacionados, y cómo se pueden utilizar. Este diccionario es especialmente valioso para miembros del equipo o para aquellos que no están familiarizados con la base de datos.

2.2.6. MVP Dashboard

Se presenta una prueba de concepto del *dashboard* interactivo, donde el objetivo principal es lograr la conexión directa desde *Power bi* a la base de datos a través de *DirectQuery*, para que en un futuro al momento que ingresen nuevos datos, el reporte se actualice automáticamente.

Por otro lado, presentar un diseño preliminar es fundamental para obtener *feedback* y así ajustar detalles en el tercer *sprint* y presentar un producto final acorde a las necesidades del cliente.

2.2.7. MVP Modelo de machine learning

Se realiza la conexión a la base de datos para el modelo de *machine learning* usando las credenciales que creamos en GCP desde VSC, y luego se realiza el *deployment* en *Streamlit*.
El objetivo de este MVP es visualizar y probar el sistema de recomendación para los usuarios de *Google Maps* y *Yelp*, así como también simular dos casos de usos para presentar al PO.

2.3. Sprint 3: Data Analytics + ML

Las actividades del *sprint* 3 se basan principalmente en la mejora de los MVP tanto del *dashboard* interactivo como del sistema de recomendación. Se busca llegar al producto final, agregando las sugerencias propuestas por el *Product Owner* y realizando la puesta a punto de todos los entregables.

2.3.1. Diseño de reportes/dashboard final

El reporte final busca que el *dashboard* presente una interactividad que permita al usuario buscar la información que necesita de manera simple a partir de filtros que le permitan acceder a los estados, ciudades, valoraciones, locales y ubicaciones que necesite.

Por otro lado, se presentan cada uno de los KPIs planteados con sus correspondientes medidores, de manera que se pueda acceder según el caso tanto a las métricas de años anteriores como a los objetivos planteados a futuro.

2.3.2. Indicadores clave de rendimiento (KPI)

- KPI 1: Aumentar 1% la cantidad de restaurantes en las ciudades más propicias en el plazo de un año.

$$KPI\ (1\%) = Cantidad\_actual\_restaurantes\_por\_ciudad * 1.01$$

En el caso del KPI 1, no se cuenta con la información de los años de apertura de cada uno de los locales gastronómicos. Esto lleva a que por el momento no se pueda acceder a las métricas de los años anteriores. Por esta razón, se decide proponer el objetivo al plazo de un año a partir de la cantidad de locales en el año más reciente de la información proporcionada por las empresas.

Por otro lado, el concepto de "ciudades más propicias" depende de la categoría del local y puede accederse a la propuesta a través del Sistemas de Recomendación de Ciudad, producto entregado en el presente proyecto.

- KPI 2: Mejorar el rating promedio de restaurantes que tienen entre 3 y 4 estrellas en un 4% en el plazo de un año.

$$KPI\ (4\%) = \frac{(Promedio\ rating\ del\ año\ siguiente - Promedio\ rating\ del\ año\ actual)}{(Promedio\ rating\ del\ año\ actual)} * 100$$

En el caso del KPI 2, se propone mejorar el rating promedio en restaurantes que cuentan con valoraciones entre 3 y 4 estrellas ya que se determinan como locales con mayor margen de crecimiento. En el *dashboard* interactivo se puede acceder a las métricas de años anteriores y se limitan los restaurantes entre 3 y 4 estrellas a través de un filtro de valoraciones. La decisión de realizarlo a través de un filtro es para que el cliente, si lo desea, pueda continuar con el seguimiento de ese local en específico aunque la valoración disminuya por debajo de 3 o aumente por encima de 4 estrellas.

- KPI 3 : Aumentar en un 4% las reviews positivas de restaurantes en el plazo de un año.

$$KPI\ (4\%) = \frac{(Total\ reviews\ positivas\ año\ siguiente - Total\ reviews\ positivas\ año\ actual)}{Total\ reviews\ positivas\ año\ actual)} * 100$$

En el caso del KPI 3, se propone mejorar la cantidad total de reseñas positivas de un año al otro. Para esto, se propone mejorar el servicio y la atención en los diversos locales debido a que en el Análisis Exploratorio de Datos (EDA) se determina como un *insight* la idea de que los clientes satisfechos suelen dejar más reseñas y a su vez reseñas positivas (Fig. 10).

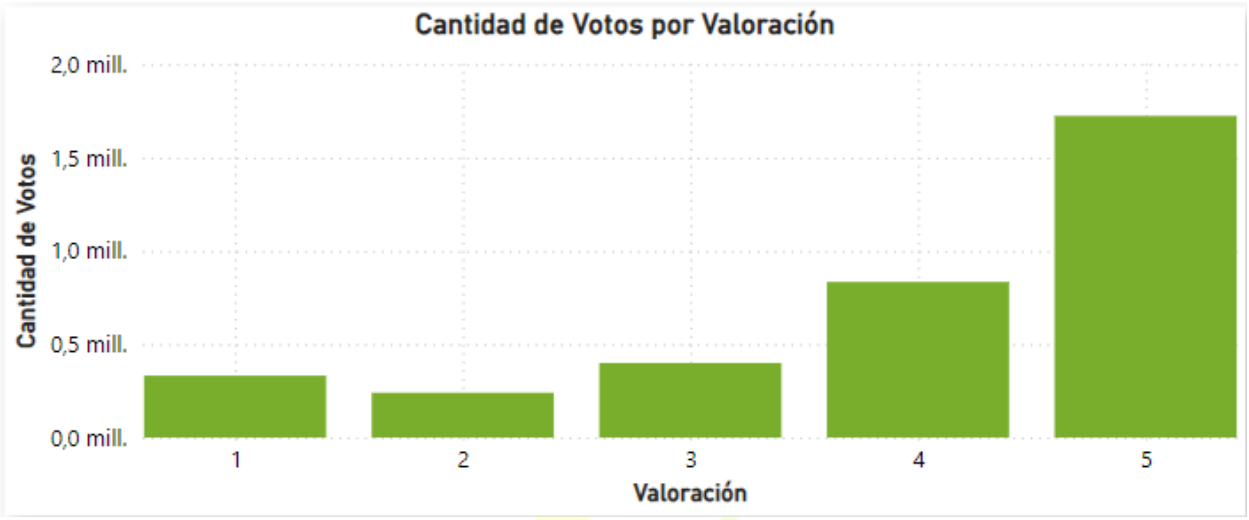

Fig. 10: Cantidad de votos por valoración, se observa claramente el insight anteriormente mencionado.

2.3.3. Selección del modelo de ML

Se desarrollan dos sistemas de recomendaciones:

- Recomendación de Restaurantes: Donde el usuario puede elegir entre la categoría y la ciudad en la que se encuentra y se le recomiendan los mejores 5 restaurantes según su selección.
- Recomendación de Ciudad: Donde el usuario puede elegir la categoría y el estado donde quiere abrir su negocio y le recomienda la mejor ciudad para abrir su restaurante según el éxito y la competencia en las ciudades del estado elegido

2.3.4. Modelo de machine learning en produccion

Se realiza el *deployment* del [sistema de recomendaciones] en el servicio en *cloud* de *Streamlit* para que el usuario pueda tener una interfaz gráfica interactiva (Fig. 11) donde pueda elegir todos los parámetros en los que esté interesado.

Fig. 11: Interfaz gráfica interactiva de los sistemas de recomendaciones.

3. CONCLUSIONES

Foco en Restaurantes con Calificaciones entre 3 y 4 Estrellas:

Los restaurantes con calificaciones entre 3 y 4 estrellas representan una oportunidad significativa para mejorar.
La mayoría de estos locales pueden estar a un paso de mejorar su calificación general con pequeños ajustes en su servicio o productos. Al enfocarse en este grupo, se podría lograr una mejora sustancial en la percepción del público, impulsando la reputación y atrayendo más clientes.

Importancia del Cliente Satisfecho:

El análisis de las reseñas muestra que los clientes que tienen una experiencia muy positiva son los más propensos a dejar una reseña. Esto indica que los restaurantes deben enfocarse en mantener altos niveles de satisfacción del cliente, ya que estos clientes no solo son recurrentes, sino que también ayudan a mejorar la visibilidad y reputación del local a través de sus reseñas. Un cliente satisfecho es un embajador de la marca que puede atraer a otros clientes.

Ubicaciones Óptimas para Nuevos Establecimientos:

El análisis de las ubicaciones geográficas muestra que algunos estados y ciudades son más propicios para la apertura de nuevos locales. El mapeo de los locales existentes en relación con las reseñas positivas y la densidad de población puede ofrecer una visión clara de dónde es más probable que un nuevo restaurante tenga éxito según su especialidad. Esto es de fácil acceso a través del Sistema de Recomendación de Ciudades.

Propuestas de mejora:

- Mejorar la Experiencia del Cliente:

Pequeños ajustes en la atención al cliente, la calidad de los ingredientes o la presentación de los platos podrían fomentar estas reseñas y obtener calificaciones más altas. Recomendar a los gerentes que implementen programas de capacitación continua para el personal puede ser clave para lograr este objetivo.

- Implementar Programas de Solicitud de Reseñas:

Es importante que los restaurantes establezcan estrategias activas para solicitar reseñas a sus clientes más satisfechos. Esto puede incluir el envío de recordatorios a través de correo electrónico, incentivos para aquellos que dejan una reseña, o incluso el uso de códigos QR en las mesas que lleven directamente a la página de reseñas del restaurante.

- Fomentar la Fidelización de Clientes:

Los programas de fidelización que recompensan a los clientes habituales pueden aumentar la cantidad de reseñas positivas y la lealtad a largo plazo. Incentivar a los clientes a través de descuentos o recompensas exclusivas por reseñar puede generar un mayor engagement y promoción del restaurante.

- Expansión Estratégica: Ej: Austin, Texas

Esta ciudad ha experimentado un crecimiento poblacional y económico significativo en los últimos años. La ciudad se ha consolidado como un centro tecnológico en expansión, atrayendo a una población joven y profesional con ingresos disponibles para gastar en ocio, incluidos restaurantes. A pesar de su crecimiento, Austin no presenta la misma densidad competitiva que otras grandes ciudades como Nueva York o Los Ángeles. Esto sugiere que aún hay espacio para que nuevos restaurantes se establezcan y prosperen.