

Proyecto de Inversión Gastronómica y Afines en Mercado De Estados Unidos



Integrantes

Víctor Vargas
Guillermo del Rio
Michael Martinez Chinchilla
Julian Scarpeccio
Benjamin Zelaya

Índice

Índice	1
1 Introducción	2
2 Objetivos	2
3 Arquitectura De Datos	3
3.1 LOCAL	3
3.2 AZURE:	4
4 Diagrama Entidad Relación	7
5 INDICADORES CLAVE DE DESEMPEÑO (KPIs)	7
6 DASHBOARD	9
7 MODELO DE MACHINE LEARNING DE PREDICCIÓN DE PROMEDIO RATING	12
8 API	13

1. Introducción

En este tercer sprint, avanzaremos en la fase crucial de nuestro proyecto, centrada en la creación de un panel de mando que incluye la generación de resultados y recomendaciones significativas, así como la adaptación necesaria de nuestro modelo de Machine Learning. Este avance representa un paso esencial para transformar nuestros datos en conocimiento valioso y en última instancia, en acciones estratégicas en beneficio de nuestro cliente.

Hasta este momento, hemos cimentado una base sólida mediante la creación y optimización de nuestro flujo de datos en el segundo sprint. Además, hemos puesto en marcha un almacén de datos automatizado y la carga gradual para garantizar la eficiencia y la constante actualización de nuestros datos.

Nuestro equipo ha subrayado la necesidad de realizar ajustes en la estructura global del proyecto y en la definición de los indicadores clave de rendimiento (KPIs) para asegurarnos de que nuestros resultados sean precisos y satisfagan plenamente las necesidades del cliente.

En este sprint, nos lanzamos con entusiasmo a la tarea de visualizar y comunicar nuestros descubrimientos de manera clara y efectiva. Estamos comprometidos en llevar a cabo un análisis exhaustivo y en ofrecer soluciones bien informadas para mejorar el rendimiento y respaldar la toma de decisiones de nuestro cliente.

2. Objetivos

Objetivos del proyecto

- Investigar y analizar los conjuntos de datos de Franquicias y sucursales compartidos en el mercado gastronómico de los Estados Unidos para identificar patrones y oportunidades de mercado.
- Entender la correlación entre los datos y las métricas con el fin de proporcionar una base sólida para las decisiones futuras relacionadas con la inversión en Franquicias del Mercado Gastronómico de los Estados Unidos.
- Proveer información relevante y confiable a nuestro cliente para respaldar su toma de decisiones sobre la inversión en franquicias.

Objetivos Específicos

- Recopilar y depurar datos de diferentes fuentes para crear una base de datos (DataWarehouse).
- Crear un dashboard interactivo y visualmente atractivo que integre los resultados del análisis exploratorio de datos
- Entrenar y poner en producción un modelo de machine learning para recomendar una inversión en el sector.

3. Arquitectura De Datos

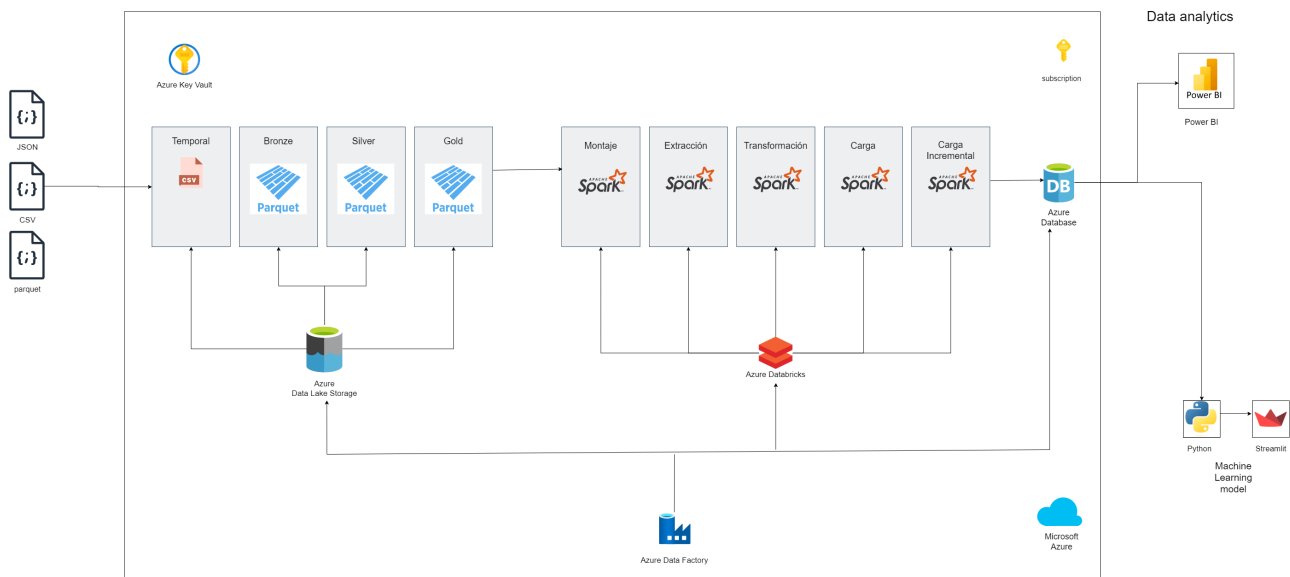


Figura 1:

3.1. LOCAL

ETL_LOCAL

Los datos entregados por la empresa y extraídos por nuestro equipo mediante api y web scraping se descargaron y son almacenados de manera temporal en el localhost de nuestra máquina, posteriormente se realizaron los siguientes pasos.

1. **Extracción:** se descargaron los datos dados por la empresa en el cual son los de google maps y yelp, también se obtuvieron datos de fuentes externas que fueron los de estados que nos muestra la cantidad de población que hay en cada estado y el de inversión de franquicias que nos da a conocer la mínima y máxima inversión que se realiza a cada franquicia.
2. **Transformación:**

- **google maps:** se consolidó un dataset resultante que fue por la unión de dos dataset también consolidados que son estados y sitios unidos que posteriormente se filtró por la categoría de restaurantes que es el tipo de negocio que se estudiará. carpeta reviews-estados se consolidó un dataset resultante llamado estado en el cual se filtraron dos columnas a utilizar gmad id y estado. carpeta metadata-sitios se consolidó un dataset resultante llamado sitios_unidos en el cual se concatenaron los archivos .json en el cual nos muestra información importante de cada local comercial como las reviews, ubicación, estado, url, etc. Del dataset sitios unidos se filtraron cuatro columnas a utilizar category",avgrating", "gmap id", luego concatenarlo con el dataset estado por medio de su id y consolidar un dataset resultante llamado google y por último se filtró la categoría a utilizar que son los restaurantes.

yelp: se consolidó un dataset resultante llamado yelp en el cual se filtró por tres columnas a utilizar que son comentarios, latitud y longitud en el cual nos permite tener los comentarios de cada review dada y la ubicación de dicho local para posteriormente relacionarlo con el dataset de google.

- **población:** se consolidó el dataset población filtrando la columna estado y cantidad de población posteriormente realizando una columna nueva llamada categoría de densidad en el cual se divide por baja, media y alta la población de cada estado.
- **Inversión de franquicias:** se observó el dataset para un posterior análisis.

3.2. AZURE:

Grupo de recursos en azure a utilizar

Podemos observar los recursos creados en la cuenta de azure, en el cual tenemos.

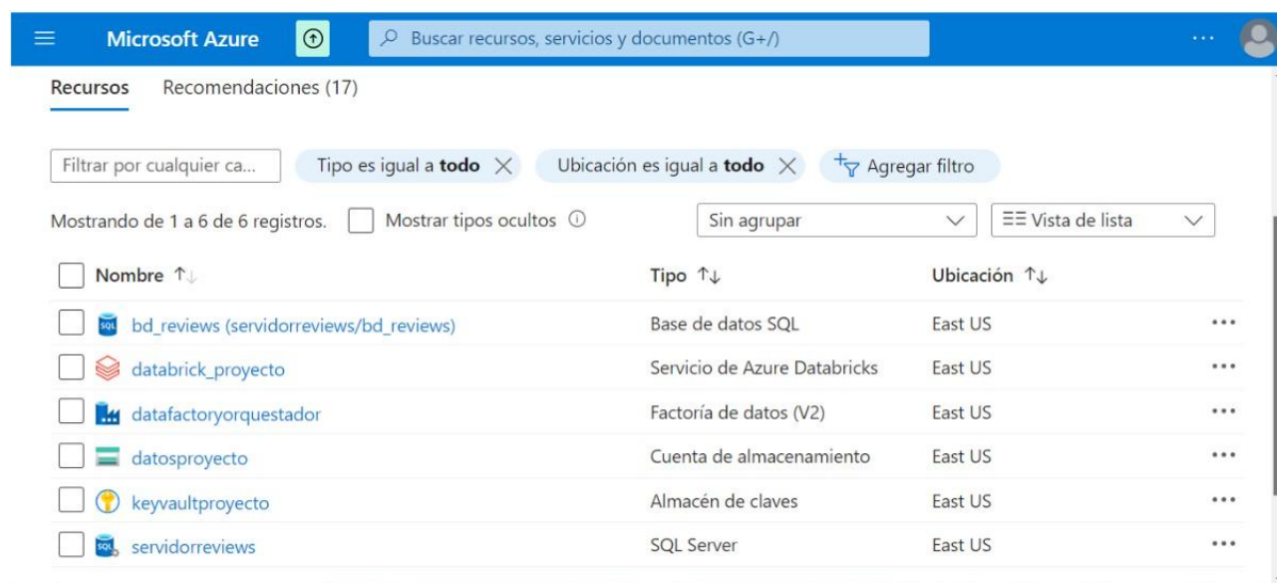


Figura 2:

- **datosproyecto:** cuenta de almacenamiento donde se crea el contenedor y se cargan los datos
- **keyvaultproyecto:** es nuestro almacenamiento de claves que contendrá todos nuestros secretos para que no sean visibles nuestras contraseñas en el documento.
- **databricks_proyecto:** nuestro lugar principal de trabajo en el cual realizaremos nuestras transformaciones y carga a la base de datos con pyspark.
- **servidor reviews:** es el servidor de la base de datos en el cual es de tipo sql server.
- **bd reviews:** es la base de datos de tipo sql en el cual contendrá nuestros datos limpios y organizados.
- **data factory orquestador:** es nuestro orquestador de datos para realizar la automatización y carga incremental.

INGESTA DE DATOS AZURE

Dado que trabajaremos sobre el esquema de Microsoft Azure se creó un contenedor de tipo data lake donde se almacenan los datasets obtenidos del etl local en el cual se le seguirán haciendo transformaciones en la nube. Para esto, fue necesario crear una cuenta de trabajo en el portal de Azure. En dicha cuenta se crea un grupo de recursos donde incluimos una cuenta de almacenamiento con un contenedor. se cargaron de forma manual los cuatro dataset obtenidos del etl 01 local al contenedor de azure que está dividido en cuatro folders que son temporal,

bronze, silver y gold, en el que se cargaran estos dataset google, yelp, inversión franquicias y población en formato csv cada uno en el folder temporal.

Azure data lake storage: dentro del contenedor creamos cuatro folders que son.

- **Temporal:** se cargan los datos en crudo.
- **Bronze:** se hace una copia de los datos que están en temporal pero en formato delta para así mejorar el procesamiento de cómputo de databricks.
- **Silver:** tenemos los datos ya transformados y limpios.
- **Gold:** tenemos datos finales con agrupaciones que serán cargados a la base de datos

Conexión con Databricks

Una vez almacenados los datasets en el contenedor de Azure se procede a realizar la conexión con Databricks, nuestro lugar de trabajo principal. En el grupo de recursos previamente creado se añade un workspace de Databricks. Ahí se creará un cluster que permite computar nuestros datos (Single Node 10.4 LTS Apache Spark 14 GB Memory, 4 Cores), el criterio de selección es en base al alcance de nuestros recursos. Dentro de Databricks creamos un Notebook y lo conectamos con el cluster. En dicho Notebook establecemos las variables necesarias para la conexión con el contenedor.

ETL 02 azure:

se dividió los procesos en cuadernos de databricks:

1. **montaje:** se definen las variables para hacer la conexión hacia el azure data lake y su contenedor, se realiza el montaje y se listan las rutas de los folder que se trabajarán.
2. **extracción:** se definen las librerías y rutas a utilizar, se estructura cada dataset con sus nombres de columnas y tipo de datos para luego cargar y visualizar los datos de la capa temporal posteriormente se cargan estos datos a la capa bronze que son una copia de la capa temporal en formato delta para aprovechar los recursos de databricks y procese los datos con mayor rapidez.
3. **transformación:**
 - **población:** se carga el dataset de la capa bronze, se realiza una columna nueva llamada id estado que será nuestra llave primaria, se eliminan registros no necesarios y se carga el dataset limpio a la capa silver cambiando de nombre estado.
 - **franquicias:** se carga el dataset de la capa bronze, se eliminan duplicados, se realiza la columna id franquicia que será su llave primaria y guardó en el folder silver.
 - **google:** se carga el dataset de la capa bronze, se eliminan columnas innecesarias, se cambian registros con caracteres especiales, se crea columna id google que será su llave primaria y guardó el dataset ya limpio en el folder silver.
 - **yelp:** cargo del folder bronze, se eliminó columnas y duplicados que son innecesarios y guardó el dataset ya limpio en silver.
 - **Reviews:** se carga los dataset google y yelp de la capa silver y se realizan un join por medio de sus columnas latitud y longitud para formar un solo archivo llamado reviews que será nuestra tabla de hechos en el cual contiene toda nuestra información valiosa; posteriormente los relaciono con los dataset de estados y franquicias para añadir su llave foránea y poder relacionar los tres dataset finales, para finalizar también es guardado en silver. En conclusión se formaron tres dataset finales que son

llamados reviews, estados y franquicias, estos dataset son particionados en un 50 % cada uno y son creados 6 dataset para posteriormente cargados en el folder gold.

4. **carga:** se cargan los 3 primeros dataset finales de la capa gold en el cual se realiza una primera carga a la base de datos.
5. **carga incremental:** se cargan los 3 dataset finales restantes de la capa gold en el cual se realiza la segunda carga a la base de datos. previo a esto se realizó la configuración de conexión entre databricks y la base de datos sql para cargar estos dataframe por medio del protocolo jdbc.

Conexión con sql database (Almacenamiento):

Creada la SQL Database de Azure se realizará la conexión con Databricks por medio del protocolo jdbc. Los datos procesados se almacenan en una base de datos SQL, donde la estrategia de almacenamiento se divide en dos etapas. En la primera carga, se almacena el 50 % de los datos, y luego se procede con una segunda carga incremental para completar la base de datos. Esta metodología permite una gestión eficiente y escalable de la información procesada

Azure Data Factory (Proceso ETL Automatizado):

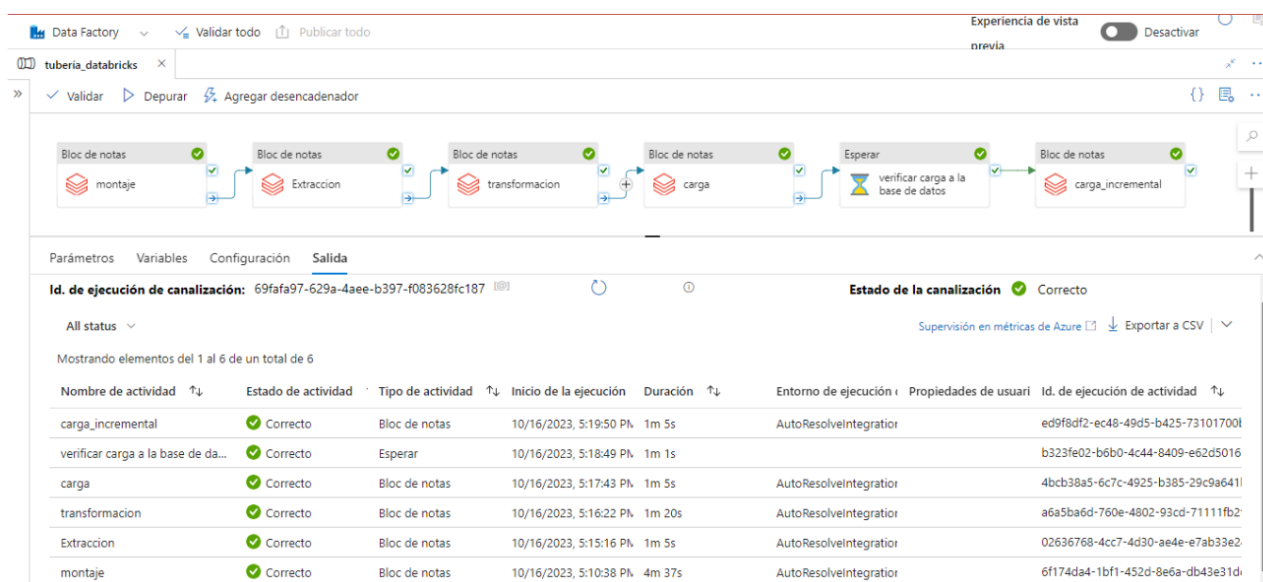


Figura 3:

Se utiliza Azure Data Factory como un orquestador de datos utilizando cada notebook de databrick en forma de tarea y secuencial. Esta combinación permite automatizar el flujo de datos y realizar una carga incremental de manera altamente eficiente.

Conexión con Power BI

La conexión se realiza mediante el conector de Azure SQL Database de PowerBI. Se ingresan las credenciales del servidor de base de datos y se cargan los datos ya sea por Direct Query o Import Data. Una vez que los datos se encuentran en la base de datos, se pueden analizar y visualizar utilizando Power BI. Esto permite identificar tendencias, patrones y obtener información valiosa para la toma de decisiones informadas.

Modelo de machine learning y streamlit

Se creó un modelo de Machine Learning en Python. Este modelo se implementa en una aplicación interactiva utilizando Streamlit, que permite a los usuarios utilizar el modelo y obtener

recomendaciones en tiempo real.

4. Diagrama Entidad Relación

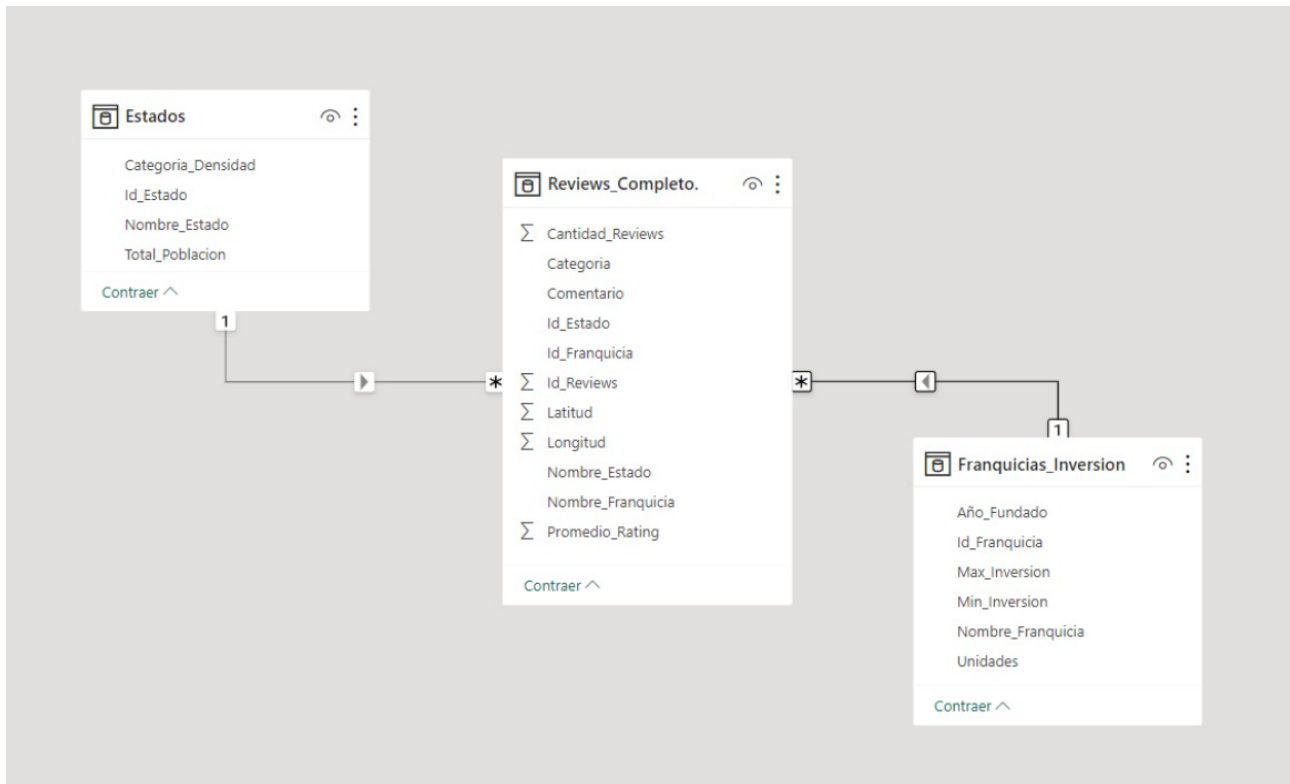


Figura 4:

Podemos examinar las relaciones entre nuestras tablas resultantes en nuestro modelo entidad-relación. En este contexto, la tabla Reviews_completo se convierte en nuestra tabla de hechos, que contiene información crítica sobre el negocio. Esta tabla de hechos tiene una clave primaria llamada id_reviews y utiliza claves foráneas, como id_franquicia e id_estado, para establecer conexiones con otras tablas de dimensiones, específicamente Estados y Franquicias_Inversion.

Estas tablas de dimensiones proporcionan información detallada sobre cada estado y franquicia, lo que enriquece nuestros datos y permite construir un sólido modelo entidad-relación que refleja de manera efectiva la estructura y relaciones en nuestros datos.

5. INDICADORES CLAVE DE DESEMPEÑO (KPIs)

Los siguientes KPIs se utilizaron para evaluar el éxito de nuestras soluciones y la mejor opción de inversión de franquicia para nuestro cliente:

- **KPI 1: Rating Promedio y Cantidad de Reviews:**

Descripción: Mide el Rating Promedio de clientes en función del rating de opiniones recopiladas en plataformas como Yelp y Google Maps en el tiempo, pudiendo visualizar la oscilación del Rating a medida que pasa el tiempo.

Importancia: Evaluar la satisfacción del cliente es fundamental, ya que clientes satisfechos son más propensos a ser leales y atraer a otros clientes.

- **KPI 2: Porcentaje de Restaurantes con Alta Calificación:**

Descripción: Este KPI muestra el porcentaje de restaurantes que tienen una calificación por encima de un umbral específico (por ejemplo, 4 estrellas).

Importancia: Identificar restaurantes de alta calidad es esencial para mantener una buena reputación y atraer a clientes exigentes.

Fórmula: $\text{Porcentaje de Restaurantes con Alta Calificación} = (\text{Número de Restaurantes con Rating} = \text{Umbral}) / \text{Total de Restaurantes} \times 100$

- **KPI 3: Cantidad de Sucursales por Conglomerado de Estados:**

Descripción: Permite identificar cuantas sucursales tendremos por conglomerado de estados.

Importancia: Ayuda a los inversores a comprender la expansión geográfica de su inversión y la distribución de sucursales.

- **KPI 4: Cantidad de Usuarios por Franquicia:**

Descripción: Identifica la cantidad de Usuarios que tendremos por Franquicia pudiendo filtrar por Estado.

Importancia: Ayuda a los inversores a comprender el potencial de usuarios a los que puede llegar a impactar su inversión.

- **KPI 5: Porcentaje de Restaurantes con Alta Calificación:**

Descripción: Muestra el porcentaje de restaurantes que tienen una calificación por encima de un umbral específico (por ejemplo, 4 estrellas).

Importancia: Identificar restaurantes de alta calidad es esencial para mantener una buena reputación y atraer a clientes exigentes.

- **KPI 6: Top 5 Franquicias por Conglomerado de Estados:**

Descripción: identificar las mejores franquicias por conglomerado de estados.

Importancia: Ayuda a los inversores a centrarse en las franquicias líderes en regiones específicas, lo que puede impulsar el éxito de la inversión.

6. DASHBOARD

Visualizar los mismos en el siguiente Dashboard de Power Bi:

[Power Bi Icon Data Consulting](#)

Slide 1. Portada, donde se cuenta con botones para navegar por los distintos Indicadores presentados



Figura 5:

Slide 2. Población EEUU. Contamos con un mapa de burbujas dividido, donde podemos filtrar por Categoría de Densidad (Alta, Mediana y Baja), también podemos seleccionar el Estado que nos interese visualizar y además una tarjeta que muestra la población seleccionada.



Figura 6:

Slide 3. Top 5 Franquicias por Conglomerado de Estados. Aquí mostramos a través de un gráfico de barras horizontal el top 5 de Franquicias con mejor Rating promedio con la posibilidad de Filtrar por Categoría de Densidad y Año. Y también agregamos un desplegable con el Nombre de las Franquicias para tener acceso a buscar alguna Franquicia en particular que no se encuentre entre el top 5. Esto nos sirve para conocer las franquicias con mejor Rating y tener una perspectiva general para profundizar con los posteriores KPIs.



Figura 7:

Slide 4. Rating Promedio y Cantidad de Reviews. Se observa un Gráfico de Líneas con el promedio de Rating y la cantidad de Reviews, donde podemos filtrar por Franquicia, Estado y el periodo de tiempo seleccionado, también contamos con una tarjeta donde figura el Promedio Rating y la categoría de Densidad del Estado seleccionado. Este gráfico nos ayuda a entender cómo oscila en el tiempo el Rating Promedio de la Franquicia Seleccionada, pudiendo ser de ayuda para el Inversor para interpretar si la satisfacción fue en aumento o en decrecimiento y poder encontrar una tendencia o estacionalidad.

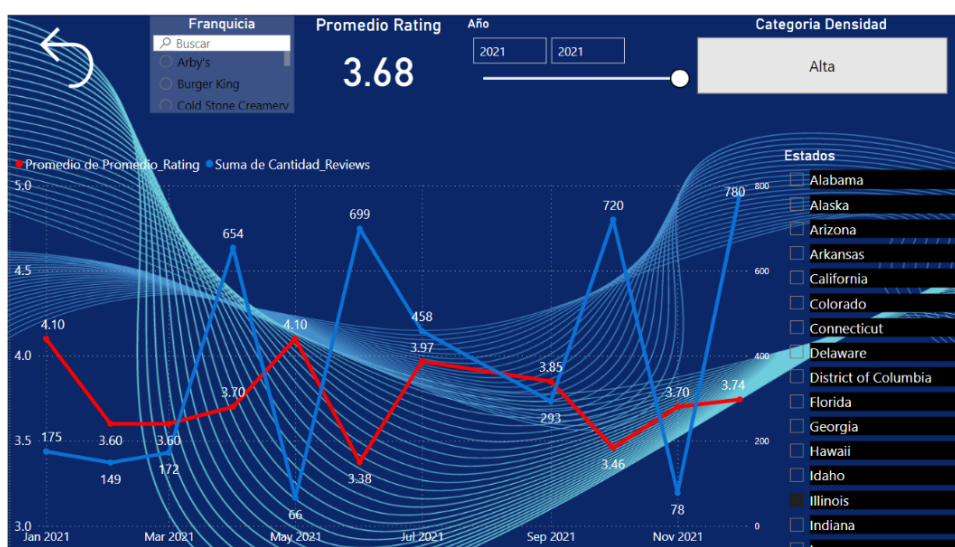


Figura 8:

Slide 5. porcentaje de Restaurantes con alta calificación. Aquí mostramos el porcentaje de restaurantes que tienen una calificación por encima de un umbral específico (tomamos como alta calificación a los restaurantes con 4 estrellas o más de Rating). Contamos con un Gráfico de Barras donde observamos el porcentaje de Restaurantes que tienen alta calificación distribuido por Estados, teniendo la posibilidad de filtrar la Categoría de Densidad y el Estado que seleccionemos. Esto nos sirve para saber en qué Estado y Categoría de Densidad se encuentran más porcentaje de Restaurantes con Alta calificación.

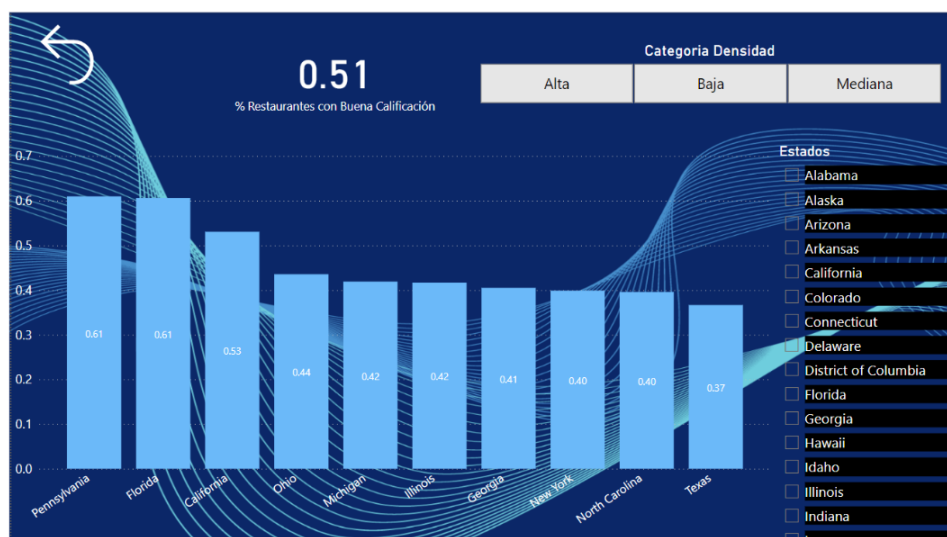


Figura 9:

Slide 6. Cantidad de Sucursales por Conglomerado. Aquí mostramos la cantidad de Sucursales que se encuentran pudiendo filtrar por Categoría de Densidad, y en el margen izquierdo encontramos un gráfico de Barra horizontal ordenado de Mayor a Menor los Estados y su cantidad de Franquicias. Esto nos sirve para saber en qué Estado y Categoría de Densidad se encuentran mayor cantidad de Franquicias.

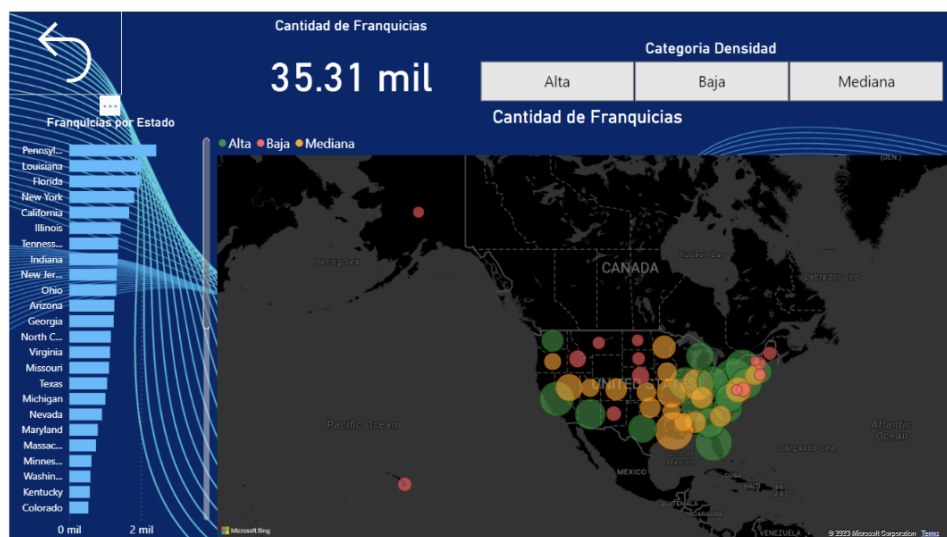


Figura 10:

Slide 7. Cantidad de Usuarios por Franquicia. Aquí mostramos la cantidad de usuarios por Franquicia que pudiendo filtrar si quisiéramos por Categoría de Densidad. Los visualizamos mediante dos gráficos de torta, el primero muestra el top 5 superior, y el segundo, el top 5 inferior. Este KPI nos sirve para conocer los Estados en donde hay una mayor/menor densidad de Usuarios por Franquicia dándonos una perspectiva de la cantidad de personas a las que podemos apuntar como Inversores gastronómicos.

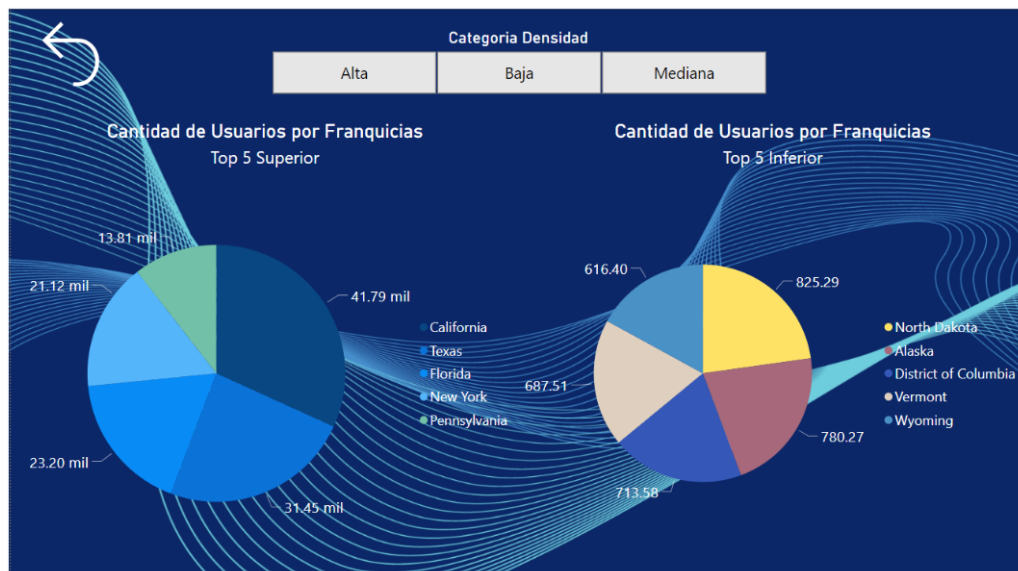


Figura 11:

7. MODELO DE MACHINE LEARNING DE PREDICCIÓN DE PROMEDIO RATING

El código es una guía para crear un modelo de aprendizaje automático que predice el promedio de rating de franquicias en función de ciertas características utilizando el algoritmo Random Forest Regressor”.

Preprocesamiento de Datos:

- Se eliminó la columna 'Categoría_Densidad' del conjunto de datos, ya que se considera que no es relevante para la predicción.
- Se definieron las características ('Latitud', 'Longitud', 'Min_Inversion', 'Max_Inversion' y 'Total_Poblacion') y la etiqueta ('Promedio_Rating') que se utilizarán para el modelo de predicción.

División de Datos:

- Los datos se dividen en dos conjuntos: uno para entrenar el modelo (conjunto de entrenamiento) y otro para probar el rendimiento del modelo (conjunto de prueba). Esto se hace para evaluar qué tan bien el modelo se desempeña en datos no vistos.

Entrenamiento del Modelo:

- Se creó un modelo de regresión de bosque aleatorio (Random Forest Regressor”) y se entrenó utilizando el conjunto de entrenamiento. Este modelo aprenderá de los datos para hacer predicciones.

Predicción Personalizada:

- Se definió una función llamada `custom_franchise_prediction` que toma como entrada características como la inversión mínima, inversión máxima, población, cantidad de revisiones, ID del estado, ID de franquicia y unidades. Esta función utiliza el modelo entrenado para hacer una predicción basada en esas características.

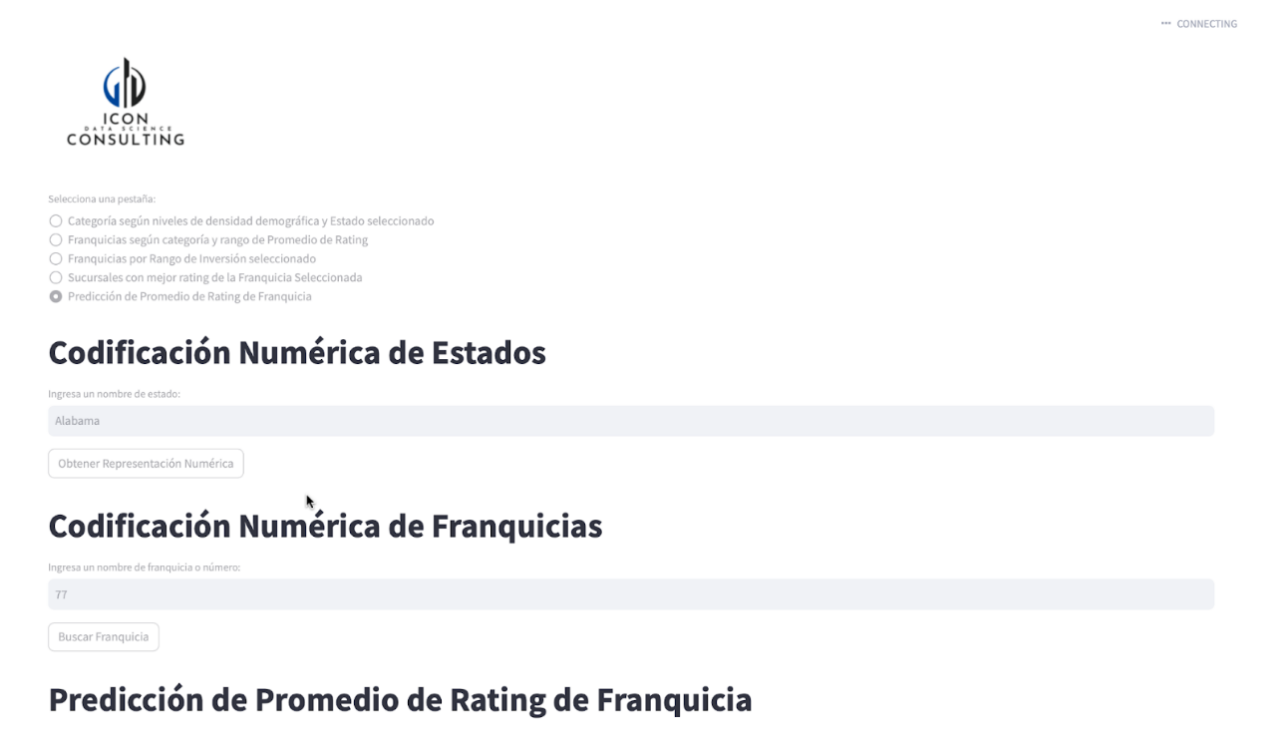
Interfaz de Usuario:

- Se creó una interfaz de usuario donde los usuarios pueden ingresar los valores de las características (como inversión, población, etc.) que desean utilizar para hacer una predicción. Esto se realiza a través de controles deslizantes y campos de entrada numérica.

Botón de Predicción:

- Se agregó un botón que, cuando se presiona, llama a la función `custom_franchise_prediction` con las características ingresadas por el usuario y muestra la predicción resultante en la interfaz.

Además, se incluyó una sección adicional que muestra cómo se puede utilizar un `LabelEncoder` para convertir nombres de estados y números de franquicias en representaciones numéricas, lo que puede ser útil para el procesamiento de datos. Este código en su conjunto crea un flujo de trabajo completo para entrenar un modelo de regresión y permitir a los usuarios realizar predicciones personalizadas basadas en las características que ingresan.



--- CONNECTING

ICON CONSULTING

Selecciona una pestaña:

- ☐ Categoría según niveles de densidad demográfica y Estado seleccionado
- ☐ Franquicias según categoría y rango de Promedio de Rating
- ☐ Franquicias por Rango de Inversión seleccionado
- ☐ Sucursales con mejor rating de la Franquicia Seleccionada
- ☒ Predicción de Promedio de Rating de Franquicia

Codificación Numérica de Estados

Ingresar un nombre de estado:

Alabama

Obtener Representación Numérica

Codificación Numérica de Franquicias

Ingresar un nombre de franquicia o número:

77

Buscar Franquicia

Predicción de Promedio de Rating de Franquicia

Figura 12:

8. API

Se ha desarrollado una API que permita a los usuarios obtener información sobre la Cantidad de Sucursales que tiene una determinada Categoría Gastronómica en el Mercado de Los Estados en función de la ubicación demográfica geográfica que elijan.

La API se ha desarrollado utilizando un conjunto de datos del mercado gastronómico de los Estados Unidos recopilados de diversas fuentes. Los datos se han procesado y almacenado en un formato que facilita su consulta a través de la API. Se ha desplegado en Streamlit junto con el modelo de Machine Learning, lo que facilita su uso por parte de los usuarios.

La API permite a los usuarios obtener información sobre los siguientes parámetros:

- Categorías más demandadas de un Estado determinado.
- Franquicias con mejor promedio de Rating de usuarios por Categoría
- Franquicias a invertir conforme un rango de inversión especificado por usuario, incluyendo porcentajes de ratio de inversión.
- Sucursales con mejor Rating de una Franquicia determinada.
- Franquicias existentes y categorías en un radio de 10km conforme ubicación proporcionada.

La API proporciona una herramienta valiosa para la toma de decisiones en materia de inversiones de franquicias en el Mercado Gastronómico de los Estados Unidos. Los usuarios pueden utilizar la API para obtener información sobre donde existen nichos de inversión sobre gustos diferentes de acuerdo a cada uno de los Estados que forman para del país, y saber cómo fueron puntuados por los usuarios, para posteriormente conforme su propia cartera de inversión tomar las decisiones correspondientes.

Para mejorar la utilidad de la API, se recomienda seguir recopilando datos sobre la puntuación de los usuarios de las diferentes sucursales de las franquicias que forman parte del mercado gastronómico, y la actualización de montos de inversión. Esto permitirá a la API proporcionar información más completa y precisa.