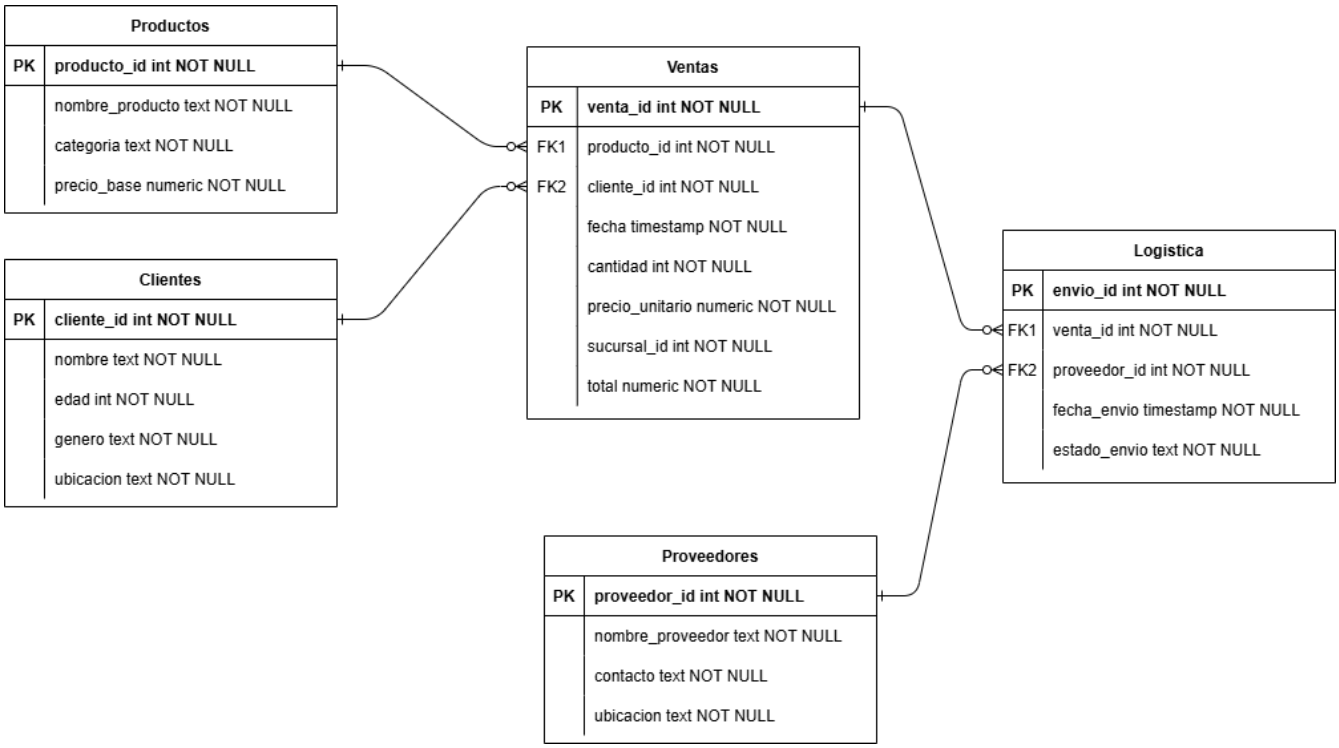
Caso Práctico Nexolutions 2025

Cristian Acosta

Diagrama de arquitectura

Modelo relacional de la base de datos SQL:



Decisiones de diseño y tecnológicas

Para el desarrollo de este proyecto se utilizó principalmente python y SQL con postgreSQL para manipulación y acceso a los datos.
Su uso python para la primera fase de ETL de los datos por su facilidad de uso y disponibilidad de bibliotecas como pandas.
Después se generó una base de datos relacional en postgres, pues se quería trabajar específicamente con SQL por ser el modelo de bases de datos más comúnmente usado por empresas y la facilidad de trabajar con postgres desde python.
Además, se usó postgres para una segunda fase de limpieza de los datos, ya que había inconsistencias a la hora de crear relaciones donde en una tabla se refería a un objeto externo con un identificador que no existía en su tabla correspondiente.

Para conservar estos datos se mandaron a tablas nuevas que funcionan como un archivero, con el fin de tener la información disponible en caso de que se puedan reconciliar estas inconsistencias en el futuro.

Por último se generó un dashboard interactivo con streamlit, un modelo de ML con scikit y un API con fastAPI. Se usaron estas tecnologías por su fácil conectividad con python, con el fin de mantener el proyecto lo más basado en python posible.

Insights clave del análisis exploratorio

Durante el análisis exploratorio se encontró que hay una oportunidad de crear una relación entre la ubicación de los clientes, proveedores y sucursales. En la tabla de ventas existe un campo “sucursal_id” el cual podría ser utilizado para crear una relación con una tabla “sucursales” la cual tenga la información pertinente de cada sucursal, como su ubicación.
De hacer esto se podría optimizar la cadena de suministro intentando agilizar ventas teniendo los productos en las sucursales más cercanas a los clientes que más frecuentan comprarlos y con los proveedores que más los envían para minimizar tiempo y costos de envío.

Otro punto relevante en cuanto a la cadena de suministro es que hay diferencias en cuanto a la ubicación de los clientes con la de los proveedores, como se explora en el dashboard.
En cuanto a clientes, la concentración por ciudad va en el orden:
Tijuana > Puebla > CDMX > Monterrey > Guadalajara
Mientras que para proveedores es:
Monterrey > Tijuana > Guadalajara > Puebla > CDMX

Esto presenta la oportunidad de optimizar envíos, buscando tener los proveedores de los productos que más compran los clientes de una ciudad en la misma.

Por último, un descubrimiento clave en cuanto a ventas es que hay una baja de alrededor de 60% en total generado en ventas en el mes de diciembre, a comparación del resto de los meses.
Esto quiere decir que sería útil reducir la cantidad de productos que se tienen en inventario durante el mes de diciembre y priorizar tener los productos que más se vendan en ese mes.

Evaluación de modelos predictivos

Para la creación del modelo predictivo se utilizaron las siguientes herramientas en el código:

pandas: Una biblioteca para la manipulación y el análisis de datos, a menudo utilizada para manejar conjuntos de datos como CSV.

train_test_split: Una función de Scikit-learn que divide el conjunto de datos en conjuntos de entrenamiento y prueba.

LinearRegression: Una clase de Scikit-learn para crear y entrenar un modelo de regresión lineal.

mean_absolute_error, mean_squared_error, r2_score: Métricas de Scikit-learn utilizadas para evaluar el rendimiento del modelo de regresión.

Usando como fuente la información de la tabla de ventas, utilizando el 20% de los datos para las pruebas y el 80% para el entrenamiento y como variable objetivo el campo “cantidad”, que representa la cantidad de un dado producto vendida en una transacción, se obtuvieron los siguientes resultados:

MAE: 4.736118419791314
MSE: 29.989713366808203
R-squared: 2.863727801227789e-06

Estos resultados significan que:

- En promedio, las predicciones del modelo se alejan unas 4,74 unidades de la demanda. Es decir, si el modelo predice una demanda de 100 unidades, es probable que la demanda real se sitúe entre 95 y 105 unidades.
- El MSE es elevado por que penaliza más fuerte los errores debido al proceso de cuadratura.
- Un valor R^2 de 2,8637e-06 es muy cercano a 0, lo que indica que el modelo no se adapta muy bien a la varianza de los datos.

Recomendaciones de negocio basadas en los hallazgos

Dados los insights encontrados se recomienda obtener información de las sucursales, específicamente su ubicación para optimizar el proceso de cadena de suministro.
Reducir o priorizar los productos más vendidos en diciembre, para reducir la baja en total generado en ventas en ese mes, a comparación del resto del año.