Proyecto de Análisis de datos

Agustín Peri

16/04/2025

Introducción y contexto

"Empanadas Don Pedro" es una cadena familiar de empanadas fundada por Doña Carmen y Don Pedro hace 15 años. Lo que comenzó como un pequeño negocio local se ha expandido a 8 sucursales en diferentes barrios de la ciudad. Sus hijos Miguel (gerente general), Lucía (chef principal) y Daniel (responsable de marketing) se han unido al negocio familiar.

Sin embargo, en los últimos dos años, han experimentado algunos desafíos:

Fluctuaciones inesperadas en las ventas entre sucursales

Dificultad para predecir la demanda de diferentes tipos de empanadas

Dificultad para identificar la causa de las malas puntuaciones

Resultados mixtos de sus campañas promocionales

Incapacidad de tener un registro del estado de la compañía

La familia ha decidido implementar un enfoque basado en datos para optimizar sus operaciones y mejorar su rentabilidad.

Al querer implementar el enfoque mencionado los dirigentes de la empresa familiar identifican cierta dificultad por parte del personal en la carga de datos, además de esto la misma esta descentralizada ya que, mientras algunas sucursales cargan los datos desde una planilla de Excel las otras lo hacen en formato papel, lo que conduce inevitablemente a errores o falta de datos al momento de intentar unificarlos.

Problemática principal: Optimizar la gestión de la cadena de empanadas utilizando análisis de datos para mejorar la rentabilidad, predecir la demanda, optimizar el inventario y personalizar estrategias de marketing.

Preguntas específicas a resolver:

1. ¿Cuál es la sucursal con mayor recaudación?
2. ¿Cómo han evolucionado las ventas a lo largo del tiempo?
3. ¿Cuál es el método de pago más utilizado?
4. ¿Existen diferencias en las calificaciones entre sucursales?
5. ¿Cuál es el medio de pago más común?
6. ¿Qué tipo de empanada se vende en mayor cantidad?

Para abordar esta problemática se hará uso de librerías de visualización, análisis, procesamiento, limpieza y construcción de algoritmos de aprendizaje automático con Python.

Análisis y exploración del dataset

El dataset consta de 500 filas y 10 variables o columnas cuatro de ellas numéricas y seis tipo objeto.

Columna	Tipo de dato
fecha_venta	object
tipo_empanada	object
sucursal	object
medio_pago	object
metodo_compra	object
cliente	object
unidades	int64
precio_total	float64
calificación	float64
tiempo_entrega_min	float64

Columna	Descripción
fecha_venta	Fecha en que se realizó la venta. Es fundamental para analizar la evolución temporal de las ventas.
tipo_empanada	Sabor o variedad de la empanada vendida (por ejemplo: carne, jamón y queso, humita, etc.). Permite conocer las preferencias de los clientes.
sucursal	Nombre o ubicación del local donde se realizó la venta. Útil para analizar el rendimiento por tienda.
medio_pago	Forma de pago utilizada por el cliente (efectivo, tarjeta, etc.). Ayuda a entender hábitos de consumo.
metodo_compra	Canal a través del cual se realizó la compra (Presencial, Delivery, Web). Permite evaluar cuál canal tiene mayor impacto.
cliente	Tipo de cliente(Nuevo o Habitual) Se usa para análisis de comportamiento de clientes recurrentes.
unidades	Cantidad de empanadas vendidas en esa transacción. Es la variable objetivo secundaria para análisis de volumen.
precio_total	Importe total pagado por la compra. Variable clave para analizar ingresos y facturación.
calificación	Puntuación otorgada por el cliente (del 1 al 5). Indica el nivel de satisfacción con la experiencia de compra.
tiempo_entrega_min	Tiempo que demoró la entrega en minutos(delivery incluido si aplica). Útil para evaluar la eficiencia del servicio.

Tratamiento de variables

En primer lugar, realicé la **conversión de la columna fecha_venta al tipo datetime**, con el fin de facilitar análisis temporales como tendencias de ventas por mes. Esto me permitió luego agrupar las ventas en periodos mensuales y graficar su evolución a lo largo del tiempo.

Posteriormente, me enfoqué en el tratamiento de **valores inconsistentes y nulos**. Por ejemplo:

- En la columna sucursal, unifiqué mediante generalización a las etiquetas cambiando "Sucursal Palermo" por "Sucursal Norte", para evitar contar la misma sede como dos distintas.
- En la columna medio_pago, reemplacé "MercadoPago" por "Transferencia", estandarizando así las formas de pago.

Respecto a los **valores faltantes**, implementé técnicas de imputación:

- En la columna calificación, completé los valores nulos con el **promedio de calificaciones según el método de compra**, asumiendo que la experiencia puede variar entre canales como delivery o presencial.
- Para tiempo_entrega_min, imputé los valores ausentes con el **promedio por sucursal**, dado que cada sede puede tener tiempos distintos de operación y logística.

Gracias a estos pasos de limpieza y estandarización, se logró preparar un conjunto de datos más coherente, útil tanto para el análisis como para el posterior modelado.

Análisis exploratorio

Durante el análisis exploratorio, comencé observando la distribución de variables clave como las calificaciones, los métodos de compra, medios de pago, y tipos de empanadas. Detectamos que la mayoría de las calificaciones se concentraban entre 3 y 5 puntos, lo que sugiere un nivel general de satisfacción aceptable.

Uno de los análisis más importantes fue la evolución de las ventas en el tiempo. Para esto, decidí agrupar las ventas por mes, transformando la variable fecha_venta a tipo datetime y graficando su comportamiento. Se observó una **tendencia general estable**, lo que podría reflejar un funcionamiento promedio.

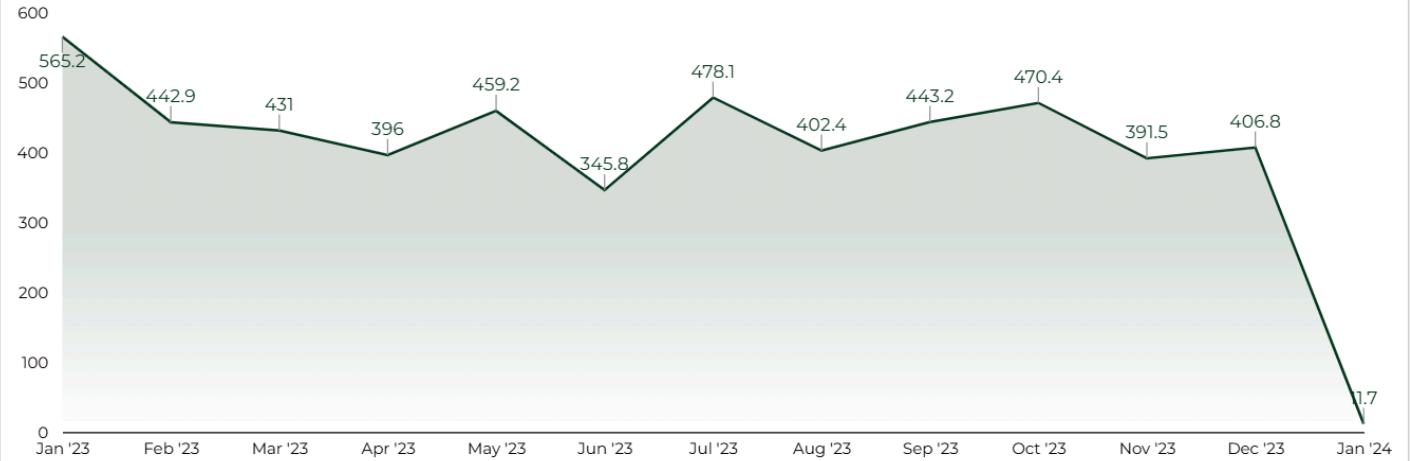
También se analizo la recaudación total por sucursal, identificando qué la sede que generó mayores ingresos fue claramente la sucursal norte con amplia diferencia frente a las demás.

Por último, revisamos la popularidad de los distintos tipos de empanadas. Las variedades clásicas como carne y jamón y queso fueron las más vendidas, mientras que las gourmet o vegetarianas tuvieron menor participación.

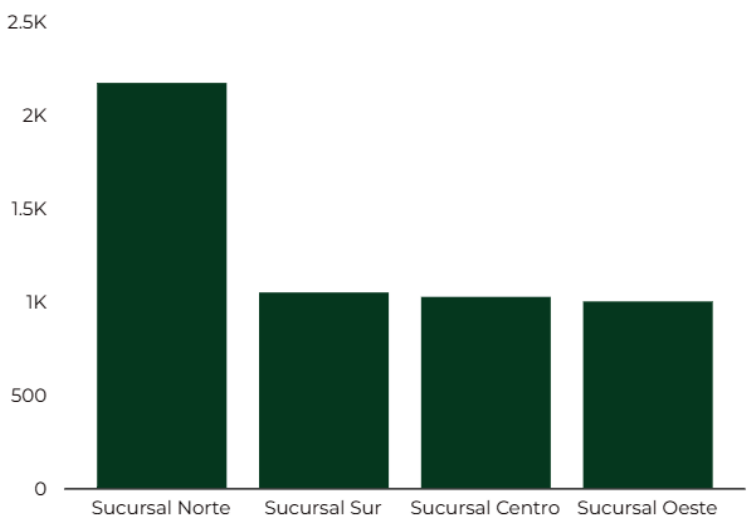
Preguntas respondidas:

1. **¿Qué tipos de empanadas se venden más?**
   Las empanadas de Humita y carne picante fueron las más populares.
2. **¿Cuál es la sucursal más rentable?**
   La sucursal norte fue la que más recaudó claramente destacada en el gráfico de recaudación total.
3. **¿Cómo evolucionan las ventas en el tiempo?**
   Se evidenció una tendencia estable con altos y bajos en las ventas bimensuales.
4. **¿Existe una relación entre el tiempo de espera y la calificación final?**
   No. No existe relación alguna entre el tiempo de espera por parte de los clientes y la calificación final.
5. **¿Cual fue la media de puntuación recibida por cada sucursal? ¿máximo? ¿mínimo?**
   Los mínimos y máximos entre ambas sucursales son exactamente los mismos con 1.0 como la peor calificación para cada una y 5.0 la mas alta. En cuanto al promedio se presenta un valor similar para todas las sucursales el cual varia entre 3.1 (sucursal centro) y 3.7 (sucursal oeste) puntos.
6. **¿Cual es el método de compra mas elegido?**
   Aunque no por amplio margen de diferencia, el canal de compra mas elegido es de forma presencial en la sucursal.
7. **¿Cual es el método de pago mas elegido?**
   El método de pago mas elegido es la tarjeta de debito mientras que el menos utilizado es el efectivo

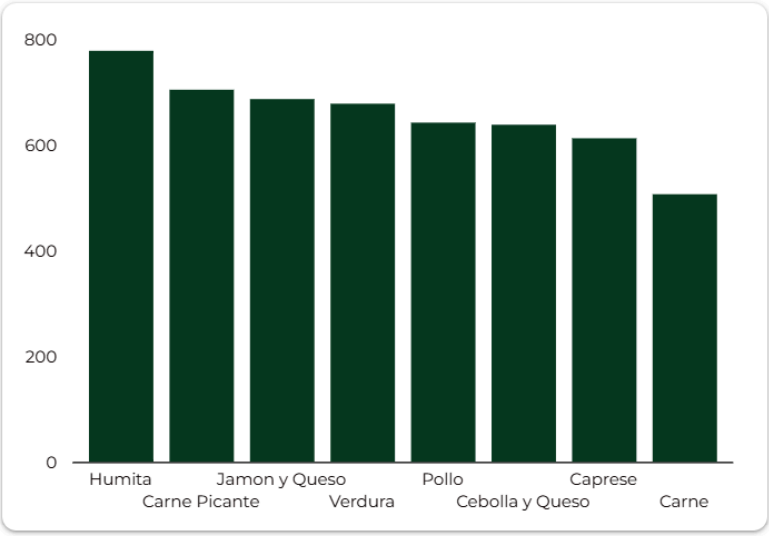
Evolucion de recaudacion



Recaudación por sucursal



Recaudación por producto



Preprocesamiento y desarrollo de modelo

En esta etapa, el objetivo fue corroborar si era posible **predecir la cantidad de unidades vendidas** utilizando como variables explicativas características del pedido como tipo de empanada, sucursal, medio de pago, cliente y método de compra.

Para ello, se eliminó la columna fecha_venta por no aportar directamente a la predicción debido a su tipo de dato, y se dividió el dataset en variables predictoras (X) y variable objetivo (unidades).

Se optó por un modelo de **Random Forest Regressor**, dada su capacidad para capturar relaciones no lineales y manejar variables categóricas codificadas. Luego de ajustar el modelo sobre un conjunto de entrenamiento (80% de los datos), se evaluó con el conjunto de prueba restante.

Los resultados mostraron un **RMSE (Root Mean Squared Error) ≈ 1.58**
Esto significa que, en promedio, el modelo se equivoca por aproximadamente 1.6 unidades de empanadas vendidas. Teniendo en cuenta que la mayoría de los pedidos estén en un rango de pocas unidades (por ejemplo, entre 1 y 10), es un error aceptable.

R^2 ≈ 0.777
Este valor indica que el **77.7% de la variabilidad en las unidades vendidas puede ser explicada por las variables predictoras**. Es un nivel bastante decente y sugiere que el modelo capta bien las relaciones en los datos por ende es posible aplicar un modelo que prediga la cantidad de ventas de empanadas por día, la precisión del mismo podría mejorar con una mayor cantidad de datos, un entrenamiento por medio de validación cruzada y la creación de variables.