Desarrollo de un dataset de un restaurante

1. Dataset trabajado en los exámenes (Descripción detallada)
El análisis de la API de Zomato es uno de los análisis más útiles para los amantes de la comida que quieren degustar las mejores cocinas de cada parte del mundo que se ajusten a su presupuesto. Este análisis también es para aquellos que quieren encontrar restaurantes que ofrezcan una buena relación calidad-precio en varias partes del país para las cocinas. Además, este análisis satisface las necesidades de las personas que se esfuerzan por obtener la mejor cocina del país y qué localidad de ese país ofrece esas cocinas con el mayor número de restaurantes incluyen:

Columnas principales:

1. **Restaurant ID**: Identificador único del restaurante.
2. **Restaurant Name**: Nombre del restaurante.
3. **Country Code**: Código del país donde se encuentra.
4. **City**: Ciudad del restaurante.
5. **Address**: Dirección completa.
6. **Locality**: Área o vecindario.
7. **Locality Verbose**: Descripción más detallada de la localidad.
8. **Longitude** y **Latitude**: Coordenadas geográficas del restaurante.
9. **Cuisines**: Tipo(s) de cocina ofrecida (ejemplo: Japonesa, Francesa).
10. **Average Cost for Two**: Costo promedio para dos personas.
11. **Currency**: Moneda utilizada para los precios.
12. **Has Table booking**: Indica si se puede reservar mesa ("Yes" o "No").
13. **Has Online delivery**: Indica si ofrece entrega en línea.
14. **Is delivering now**: Indica si están haciendo entregas en este momento.
15. **Switch to order menu**: Información sobre disponibilidad de menú para entrega.
16. **Price Range**: Rango de precios (1: Bajo, 4: Alto).
17. **Aggregate Rating**: Calificación promedio del restaurante.
18. **Rating Color**: Color asociado a la calificación.
19. **Rating Text**: Texto descriptivo de la calificación (ejemplo: "Excellent").
20. **Votes**: Número de votos o calificaciones dadas por los usuarios.

Ejemplo de los datos:

estaurant ID	Restaurant Name	City	Cuisines	Average Cost for Two	Aggregate Rating	Rating Text	Votes
6317637	Le Petit Souffle	Makati City	French, Japanese, Desserts	3	4.8	Excellent	314

6304287	Izakaya Kikufuji	Makati City	Japanese	3	4.5	Excellent	591
6300002	Heat - Edsa Shangri	Mandaluyong	Seafood, Asian, Indian	4	4.4	Very Good	270

Potencial de análisis:

- **Análisis de popularidad**: Identificar restaurantes mejor valorados.
- **Análisis geográfico**: Relación entre calificación y ubicación.
- **Preferencias culinarias**: Cocinas más populares por ciudad o país.
- **Patrones de consumo**: Relación entre costos promedio y votos.

El dataset contiene alrededor $9,551$ registros y 21 columnas.

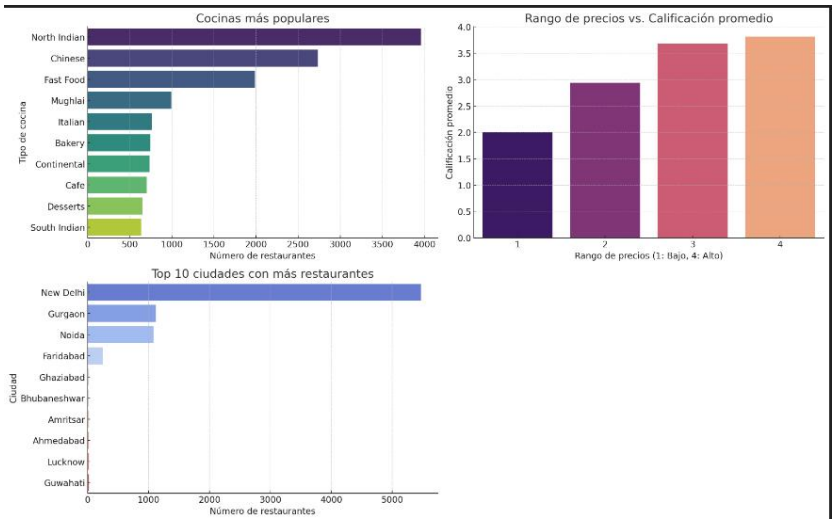
2. Objetivo de investigación

El objetivo es identificar patrones de consumo y factores que influyen en la satisfacción del cliente para:

- Optimizar el menú.
- Incrementar las ventas mediante ofertas personalizadas.
- Mejorar la experiencia del cliente.

Objetivos del análisis:

1. **Cocinas más populares**: Identificar los tipos de cocina con mayor número de restaurantes.
2. **Relación entre rango de precios y calificaciones**: ¿Los restaurantes más caros tienden a tener mejores calificaciones?
3. **Ciudades con más restaurantes**: Top 10 de ciudades con mayor número de restaurantes en el dataset.
4. **Calificaciones y votos**: Relación entre calificación promedio y cantidad de votos recibidos.



Resultados del análisis

Cocinas más populares:

1. Las cocinas más frecuentes son aquellas como India, China, Continental, e Italiana.
2. Esto sugiere que los restaurantes con estas especialidades tienen mayor representación en el dataset.

Relación entre rango de precios y calificaciones:

Existe una tendencia a que los restaurantes de precios más altos (rango 4) obtengan calificaciones promedio más altas que los de rango bajo.

Ciudades con más restaurantes:

Las ciudades con mayor representación incluyen Makati City, Baguio City, y Mandaluyong, destacando ciertas áreas urbanas como hubs gastronómicos.

Relación entre calificación y votos:

Hay una **correlación positiva débil** (0.31) entre la calificación promedio y los votos, indicando que los restaurantes mejor valorados tienden a recibir más votos, pero otros factores también pueden influir.

3. Proceso básico de análisis de datos

a. Preprocesamiento

Validación de datos

- Revisar registros faltantes y valores atípicos en variables clave (e.g., precios, calificaciones).

Transformaciones necesarias

- Normalizar las columnas de precios y propinas para un análisis uniforme.
- Codificar variables categóricas como género y método de pago.

Balanceo de datos

- Revisar la distribución de calificaciones. Si hay sesgo significativo, aplicar técnicas como submuestreo o sobremuestreo.

b. Selección del clasificador

Clasificadores elegidos (con justificación)

1. **Árboles de decisión supervisados**: Adecuados para clasificar satisfacción según múltiples variables independientes (edad, frecuencia, gasto, etc.).
2. **K-means (no supervisado)**: Útil para segmentar clientes y entender sus patrones de consumo.

El algoritmo supervisado permite predecir satisfacción, mientras que el no supervisado identifica grupos de clientes.

Referencia técnica:

- Breiman, L. (2001). *Random forests*. Machine Learning Journal. ISBN: 9780387848570
- MacQueen, J. (1967). *Some Methods for Classification and Analysis of Multivariate Observations*. Proceedings of the Fifth Berkeley Symposium. DOI: 10.2307/2334205

c. Primera ejecución

1. **Confiabilidad**: Medir el rendimiento inicial del modelo.
2. **Matriz de confusión**: Analizar falsos positivos y negativos.
3. **Splits:**
    - Académico: 80% entrenamiento, 20% prueba.
    - Investigación: 50% entrenamiento, 50% prueba.

d. Aplicar PCA

Reducir dimensionalidad del dataset. Probar con 12, 10, 9, 5 y 3 columnas para evaluar si los resultados mejoran.

Fundamentos de PCA (Álgebra lineal):

- PCA encuentra las componentes principales al calcular los *autovalores* y *autovectores* de la matriz de covarianza de los datos.
- Los autovalores más altos corresponden a direcciones con mayor varianza.

e. Proceso no supervisado

Aplicar K-means para segmentar clientes en clusters según patrones de consumo. Evaluar los resultados sin usar la variable "satisfacción".

5. N-reinas con análisis combinatorio y simulado-recocido

Explicación del algoritmo:

- **Enfoque combinatorio:** Enumerar todas las configuraciones posibles y verificar cuáles son soluciones válidas.
- **Simulado-recocido:**
    1. Inicializar con una solución aleatoria.
    2. Realizar cambios incrementales.
    3. Aceptar soluciones subóptimas con cierta probabilidad para evitar mínimos locales.
    4. Disminuir gradualmente la probabilidad de aceptar soluciones peores.

Justificación: Simulado-recocido es eficiente para problemas combinatorios complejos como este.