

Elaborado por: Melissa Pérez

---

# MODELO DE CLUSTERING

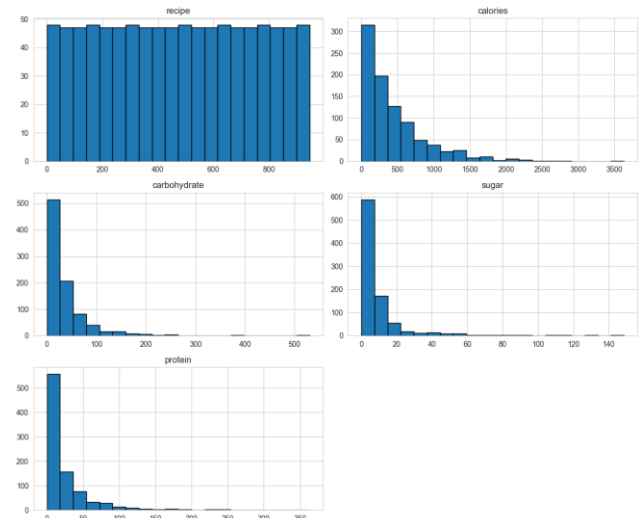
# ANÁLISIS EXPLORATORIO DE DATOS

Carga de datos y detección de valores nulos.

Distribución de variables mediante histogramas.

Los histogramas nos dan una idea de la distribución de cada variable. Por ejemplo, calories y protein están sesgadas a la derecha.

```
--- Información General del DataFrame ---  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 947 entries, 0 to 946  
Data columns (total 8 columns):  
#   Column          Non-Null Count  Dtype    
---  -  
0   recipe          947 non-null    int64    
1   calories        895 non-null    float64   
2   carbohydrate     895 non-null    float64   
3   sugar           895 non-null    float64   
4   protein         895 non-null    float64   
5   category        947 non-null    object    
6   servings         947 non-null    object    
7   high_traffic     574 non-null    object    
dtypes: float64(4), int64(1), object(3)  
memory usage: 59.3+ KB
```



# PREPROCESAMIENTO DE DATOS

Completar valores faltantes.

Estandarización de características.

Detección y tratamiento de valores atípicos.

```
df_clean = df.copy()
df_clean.dropna(inplace=True)
df_clean = pd.get_dummies(df_clean, columns=['category'], drop_first=True)
df_clean.drop(['recipe', 'high_traffic'], axis=1, inplace=True)
df_clean['servings'] = df_clean['servings'].astype(str).str.extract('(\d+)').astype(int)

scaler = StandardScaler()
X_scaled = scaler.fit_transform(df_clean)

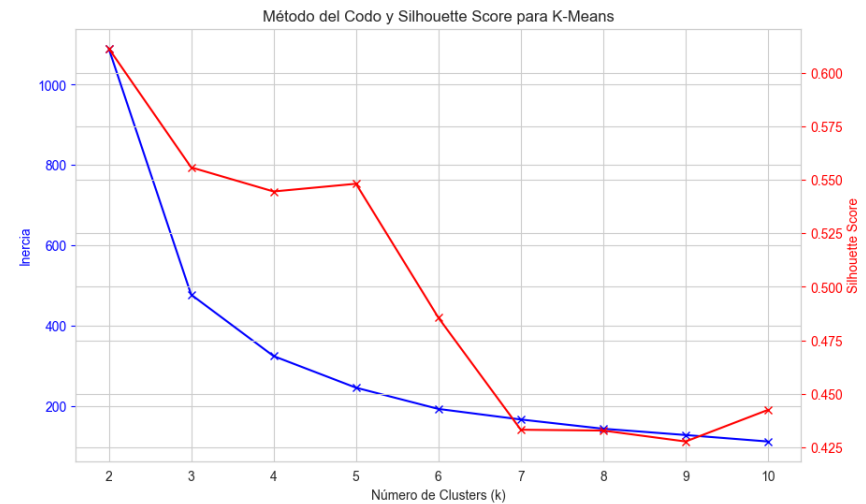
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)
```

# CLUSTERING K-MEANS

Elección de  $k = 4$  clusters.

Proceso iterativo de asignación y actualización de centroides.

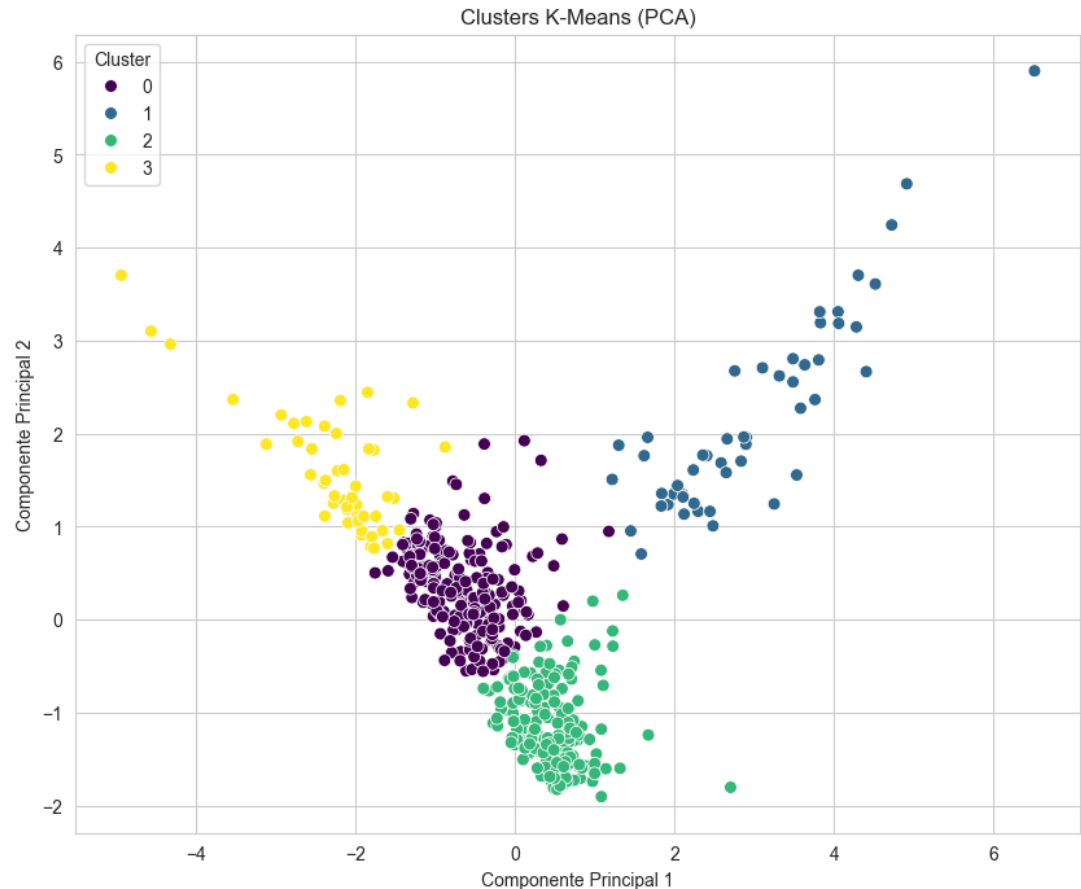
Visualización de clusters resultantes.



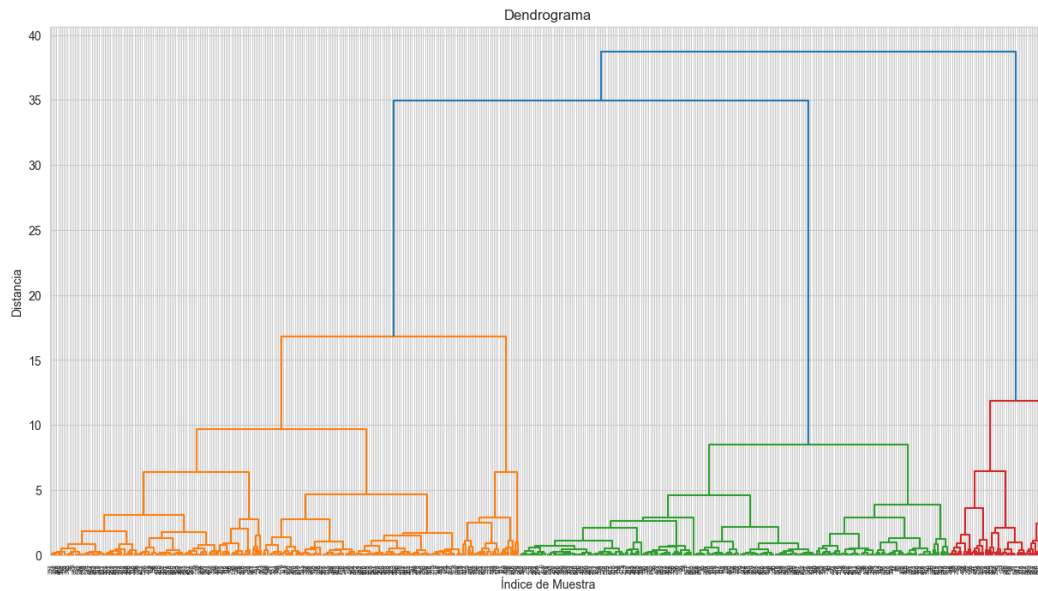
# PCA PARA REDUCCIÓN DE DIMENSIONALIDAD

Aplicación de PCA.

En modelado de datos, **PCA (Análisis de Componentes Principales)** es una técnica de reducción de dimensionalidad que busca transformar un conjunto de variables posiblemente correlacionadas en un nuevo conjunto más pequeño de variables no correlacionadas (componentes principales), manteniendo la mayor parte de la variabilidad original.

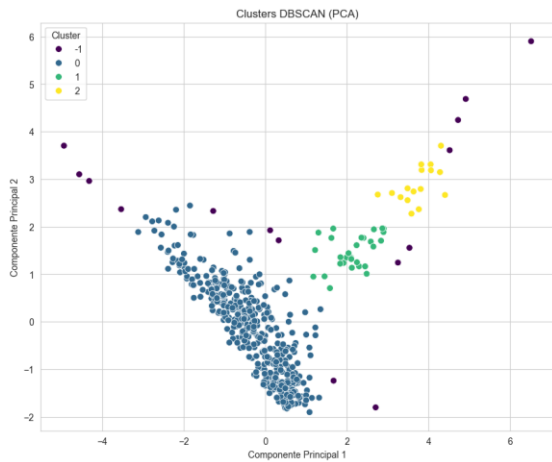


# CLUSTERING JERÁRQUICO



Método aglomerativo.

Interpretación del  
dendrograma.

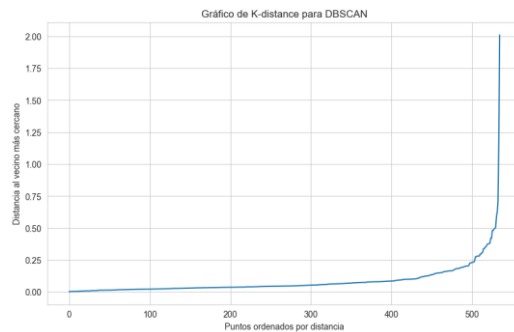


# CLUSTERING DBSCAN

Método basado en densidad.

Parámetros `eps` y `min_samples`.

Limitaciones observadas en este dataset.



Métricas de validación: Silueta y Davies-Bouldin.

Interpretación de los valores obtenidos para cada modelo.


	Silhouette Score	Calinski-Harabasz	Davies-Bouldin
K-Means	0.544559	828.083951	0.591387
Jerárquico	0.544583	783.051941	0.566152
DBSCAN	0.496556	116.160375	2.232661

# EVALUACIÓN DE MODELOS




# RESULTADOS Y COMPARACIÓN

K-Means con  
mejor  
puntuación de  
Silueta.



Rendimiento  
similar del  
modelo  
jerárquico.



DBSCAN no  
adecuado para  
este conjunto de  
datos.

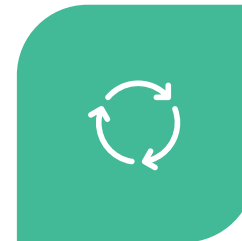
## CONCLUSIONES Y RECOMENDACIONES



SELECCIÓN DEL MODELO  
K-MEANS CON 4  
CLUSTERS.



APLICACIONES  
PRÁCTICAS E INSIGHTS  
DE NEGOCIO.



PASOS FUTUROS: AJUSTE  
DE PARÁMETROS Y  
VALIDACIÓN ADICIONAL.