Navegación

Ir a la sección:

- Exploración Inicial
- Limpieza y Validación
- Indicadores y Documentación
- EDA Avanzado & Dashboards
- Modelado de Machine Learning

🏥 Análisis y Calidad de Datos de Pacientes de Hospital

Esta aplicación realiza un análisis exhaustivo de la calidad de los datos de pacientes, seguido de procesos de limpieza, validación, generación de KPIs, EDA avanzado y un modelo de Machine Learning.

5. 🧠 Modelado de Machine Learning: Agrupación de Pacientes (Clustering)

Identificación de segmentos de pacientes con características similares utilizando K-Means.

Preparación de Datos para ML y Selección de Características

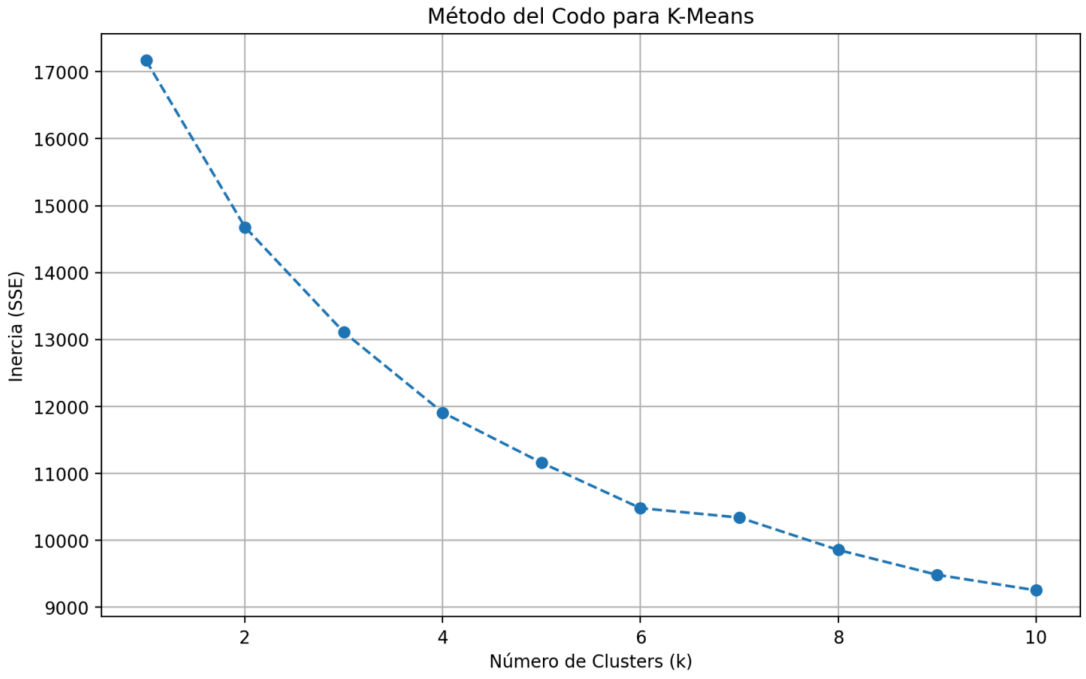
Selecciona las características para el clustering:

edad × sexo × ciudad × tipo_sangre × presion_arterial… × presion_arterial… ×

Datos preprocesados y escalados para el modelo de clustering (dimensiones): `(3318, 18)`

Justificación: Las características numéricas se escalan para igualar su contribución. Las categóricas se convierten a formato numérico (One-Hot Encoding) para que el algoritmo K-Means pueda procesarlas.

Determinación del Número Óptimo de Clusters (Método del Codo)



El **Método del Codo** ayuda a determinar el número óptimo de clusters (`k`). Se busca el punto en el gráfico donde la inercia (suma de cuadrados dentro del cluster) disminuye significativamente, formando una "rodilla" o "codo".

Configuración del Modelo K-Means

Selecciona el número de clusters (k):

3

2 8

Modelo K-Means entrenado con **3** clusters.

Resultados del Agrupamiento

Características Promedio por Cluster (en escala original para numéricas)

Cluster	edad	presion_arterial_sistolica	presion_arterial_di
0	26.8181	133.8013	
1	59.4402	136.1108	
2	48.5744	131.5045	

Distribución de Características Categóricas por Cluster

Distribución de 'sexo' por Cluster:

cluster	Female	Male
0	47.72%	
1	51.41%	
2	51.07%	

Distribución de 'ciudad' por Cluster:

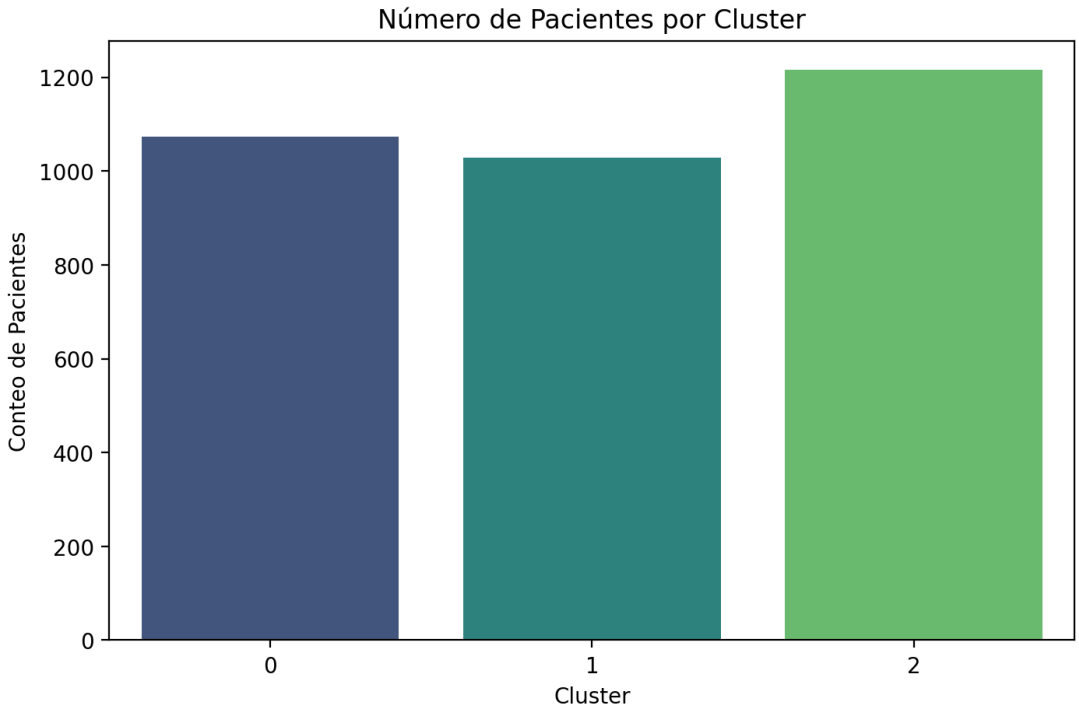
cluster	Barranquilla	Bogotá	Bucaramanga	Cali
0	19.11%	20.97%	20.50%	
1	19.53%	19.14%	20.31%	
2	19.33%	19.74%	22.37%	

Distribución de 'tipo_sangre' por Cluster:

cluster	A+	A-	AB+	AB-	B+	B-
0	14.26%	12.12%	12.58%	13.98%	11.00%	12.30%
1	11.18%	13.70%	11.86%	13.31%	13.12%	12.24%
2	12.17%	12.34%	11.76%	12.99%	12.75%	11.35%

Estos valores representan el centro de cada cluster para características numéricas y la distribución de categorías para las categóricas, ayudando a interpretar lo que define a cada grupo de pacientes.

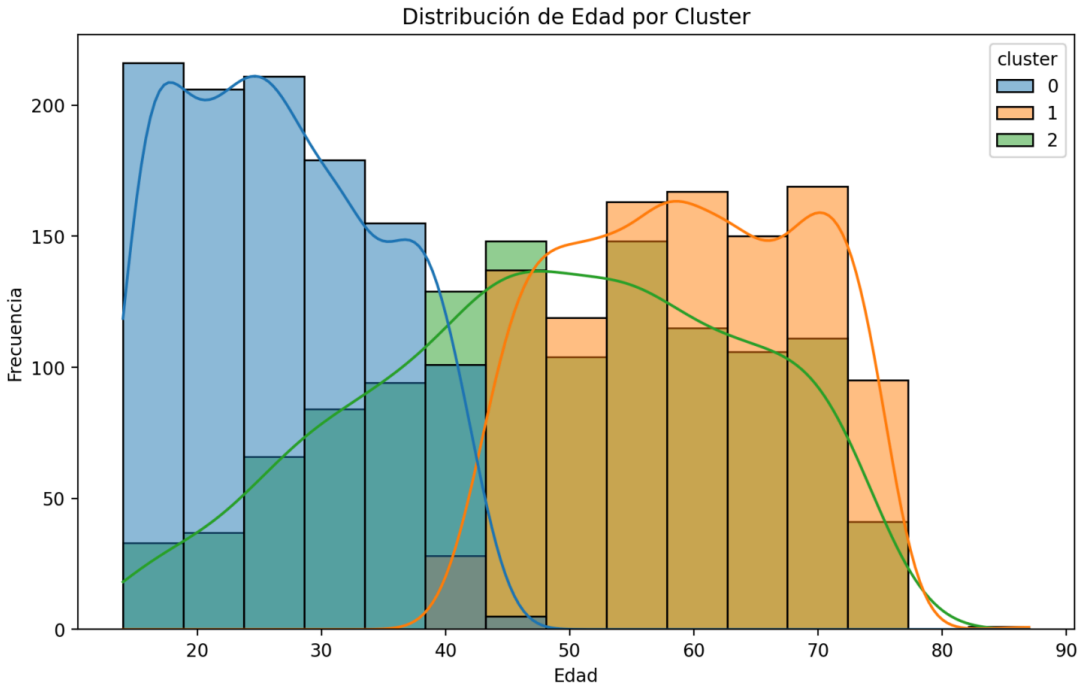
Conteo de Pacientes por Cluster



cluster	Conteo
0	
1	
2	

Este gráfico de barras apiladas muestra cuántos pacientes fueron asignados a cada cluster.

Visualización de Clusters (Distribución de Edad por Cluster)



Este histograma superpuesto muestra cómo se distribuyen las edades dentro de cada cluster, ayudando a entender los perfiles de edad de cada grupo.

Métricas de Evaluación del Clustering

Silhouette Score

0.12

El **Silhouette Score** mide cuán similar es un objeto a su propio cluster (cohesión) en comparación con otros clusters (separación).

- Un valor cercano a +1 indica que el objeto está bien agrupado.
- Un valor cercano a 0 indica que el objeto está en la frontera entre dos clusters.
- Un valor cercano a -1 indica que el objeto ha sido asignado al cluster incorrecto. Un score alto sugiere una buena separación de los clusters.

Davies-Bouldin Index

2.08

El **Davies-Bouldin Index** mide la similitud promedio entre cada cluster y el cluster más similar, donde la similitud es la relación entre la distancia dentro del cluster y la distancia entre clusters. Un valor más bajo indica un mejor agrupamiento.

Calinski-Harabasz Index

513.47

El **Calinski-Harabasz Index** (también conocido como Variance Ratio Criterion) es la relación entre la dispersión inter-cluster y la dispersión intra-cluster. Un valor más alto generalmente corresponde a modelos con clusters mejor definidos.

Inercia (SSE)

13108.45

La **Inercia** (Sum of Squared Errors - SSE) es la suma de las distancias cuadradas de cada punto a su centroide asignado. Un valor más bajo generalmente indica clusters más compactos.

Interpretación y Aplicaciones:

Basado en las características seleccionadas, el modelo K-Means ha identificado **3** grupos distintos de pacientes. La interpretación de estos grupos dependerá de los valores promedio y las distribuciones de las características dentro de cada cluster.

Aplicaciones potenciales:

- **Marketing y Comunicación Personalizada:** Enviar información relevante sobre prevención o programas de salud específicos para cada grupo de pacientes.
- **Gestión de Recursos Hospitalarios:** Anticipar las necesidades de ciertos grupos de pacientes (ej., especialidades pediátricas para el cluster joven, geriatría para el cluster mayor, o recursos para pacientes con ciertas condiciones).
- **Investigación Clínica:** Estudiar patrones de enfermedades o tratamientos que sean más prevalentes en un segmento de pacientes particular.
- **Recomendación de Tratamientos/Alertas:** Aunque un modelo de clasificación es más directo para esto, el clustering puede sentar las bases. Por ejemplo, si un nuevo paciente cae en un cluster específico, el sistema podría sugerir alertas de salud comunes para ese grupo o tratamientos que han demostrado ser efectivos para pacientes similares.