UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

TECNOLOGÍAS DE LA INFORMACIÓN

EXTRACCIÓN DE CONOCIMIENTO EN BASES DE DATOS

V.2. ELABORACIÓN DE GRÁFICAS

IDGS91N

PRESENTA:

REGINA CHÁVEZ TAMAYO - 6521110019

DOCENTE:

LUIS ENRIQUE MASCOTE CANO

Chihuahua, Chih., 30 de noviembre de 2025

Resumen Ejecutivo

El presente reporte documenta el desarrollo de una aplicación interactiva de visualización basada en los resultados del proyecto de la Unidad 4, el cual consistió en aplicar K-means para agrupar muestras del dataset Iris a partir de sus características numéricas. El objetivo principal de esta nueva etapa fue complementar dicho análisis mediante una capa visual robusta, clara y orientada a la comunicación efectiva de insights relevantes.

Para ello, se desarrolló una aplicación utilizando Streamlit + Plotly, que integra cuatro visualizaciones principales:

- Un histograma para mostrar la distribución de los valores de petal length.
- Un scatter plot para analizar la relación entre petal length y petal width.
- Un gráfico de pastel que muestra la proporción de los clusters encontrados.
- Una visualización PCA en 2D que permite observar la separación entre clusters en un espacio reducido.

Entre los hallazgos más importantes se identificó que Setosa forma un cluster perfectamente separado, mientras que Versicolor y Virginica presentan una mayor superposición en el espacio PCA. La aplicación facilita una interpretación visual clara que complementa el análisis matemático realizado previamente.

Introducción

El análisis no supervisado permite descubrir patrones, estructuras internas y agrupaciones naturales dentro de un conjunto de datos. En la Unidad 4 se aplicó el algoritmo K-means, obteniendo clusters que requerían una etapa adicional de interpretación visual para ser comprendidos en su totalidad.

La visualización es fundamental porque permite:

- Comprender distribuciones y relaciones entre las variables.

- Identificar patrones y separar grupos visualmente.
- Comunicar hallazgos a audiencias no técnicas.
- Detectar anomalías, tendencias o separaciones no evidentes numéricamente.

Objetivo del proyecto:

Desarrollar una aplicación de visualización profesional que permita explorar e interpretar los resultados del clustering realizado en la Unidad 4.

Objetivos específicos:

- Implementar al menos tres tipos diferentes de gráficas.
- Presentar visualizaciones interactivas con diseño profesional.
- Integrar una narrativa visual que apoye el proceso de extracción de conocimiento.

Metodología de Visualización

Herramientas utilizadas

Herramienta principal: Streamlit

- **Versión:** 1.x
- **Motivo de elección:** permite crear aplicaciones web fácilmente y con excelente integración con Python.
- **Ventajas:** interactividad, facilidad de despliegue, interfaz profesional.

Biblioteca principal: Plotly Express

- **Motivo de uso:** genera visualizaciones interactivas y de alta calidad sin esfuerzo adicional.
- **Tipos de gráficas soportadas:** scatter, histogramas, pie charts, mapas, líneas, boxplots, etc.

Otras tecnologías complementarias

- Scikit-learn para K-means y PCA
- Pandas para manipulación de datos
- StandardScaler para normalización

Proceso de desarrollo

Preparación de los datos

- Se utilizó el dataset Iris integrado en Plotly.
- Se escalaron las 4 variables numéricas.
- Se aplicó K-means con k=3.
- Se realizó PCA para obtener dos componentes principales.

Diseño de las visualizaciones

Se diseñaron bocetos iniciales basados en los tipos de gráficos solicitados:

1. Distribución - histograma
2. Relación - scatter plot
3. Proporción - pie chart
4. Extra (puntos adicionales) - PCA 2D cluster visualization

Implementación técnica

- Desarrollo realizado en un solo archivo app.py.
- Gráficas generadas con Plotly Express.
- Interfaz web construida con Streamlit.

Pruebas y ajustes

- Se verificó escalado correcto.
- Se compararon colores y se ajustaron para consistencia.

- Se probaron diferentes tamaños y resoluciones de gráficos.

Decisiones de diseño

Selección del tipo de gráfica

- Histograma - mejor opción para mostrar distribución.
- Scatter plot - ideal para relaciones XY.
- Pie chart - comunica proporciones rápidamente.
- PCA - permite observar agrupamientos con claridad.

Esquema de colores

- Se usaron paletas profesionales Plotly Set1, Set2 y Pastel.
- Colores diferenciados por cluster.

Interactividad utilizada

- Tooltips (datos flotantes)
- Zoom
- Movimiento dinámico
- Hover para ver valores exactos

Descripción de la aplicación

Gráfica 1 — Histograma (Distribución)

- **Tipo:** Histograma
- **Variables:** petal_length y cluster
- **Justificación:** Permite ver la forma de la distribución dentro de cada cluster.
- **Código:**

Histograma

```
1 fig_hist = px.histogram(df, x="petal_length", nbins=20, color="cluster")
```

Gráfica 2 — Scatter Plot (Relación)

- **Variables:** petal_length, petal_width, cluster
- **Justificación:** Identifica separación de clusters en dos dimensiones reales.
- **Código:**

Scatter Plot

```
1 fig_scatter = px.scatter(df, x="petal_length", y="petal_width", color="cluster")
```

Gráfica 3 — Pie Chart (Proporciones)

- **Variables:** cluster (frecuencias)
- **Justificación:** Muestra balance entre clusters.
- **Código:**

Pie Chart

```
1 fig_pie = px.pie(cluster_counts, values="count", names="cluster")
```

Gráfica Extra — PCA 2D

- **Justificación:** Demuestra separación entre clusters.
- **Código:**

PCA 2D

```
fig_pca = px.scatter(df, x="PC1", y="PC2", color="cluster")
```

Interfaz general

- Aplicación organizada en secciones
- Navegación vertical simple
- Gráficas con excelente resolución

Interpretación de las Gráficas

Análisis individual

Histograma

- Setosa presenta valores más pequeños de petal_length.
- Virginica tiene los valores más altos.
- Se observa clara separación entre clusters.

Scatter Plot

- Setosa aparece como un grupo compacto y aislado.
- Versicolor y Virginica tienen una ligera superposición.

Pie Chart

- Los clusters están casi balanceados (aprox. 33% cada uno).
- Esto confirma que K-means asignó muestras equitativamente.

PCA 2D

- Setosa es perfectamente separable.
- Virginica y Versicolor comparten espacio pero mantienen cierta estructura.

Análisis integrado

Las gráficas juntas cuentan una historia clara:

- Existe una variable clave (petal_length) que separa naturalmente los clusters.
- Las relaciones entre variables confirman lo bien que funciona K-means en este dataset.
- El PCA hace evidente que Setosa es el cluster más distinto.

Relación con la extracción de conocimiento

- La visualización permitió comprender por qué K-means funciona bien: la separación es real y evidente.
- Las gráficas apoyan decisiones basadas en observación visual y no solo en métricas.
- Ayudan a explicar clustering a un público no técnico.

Hallazgos clave

1. Setosa es completamente separable.
2. Virginica y Versicolor tienen superposición parcial.
3. La variable petal_length define la mayor parte de la separación.
4. El PCA conserva más del 95% de la variabilidad.
5. Los clusters son casi perfectamente balanceados.

Conclusiones

El equipo aprendió a integrar visualizaciones profesionales con análisis no supervisado, utilizando bibliotecas especializadas y buenas prácticas de diseño. Los principales retos

fueron: garantizar la interactividad, el ajuste de color y la reducción de ruido visual. Como mejora futura se propone integrar filtros dinámicos, paneles de comparación y más opciones de clustering como DBSCAN.

Referencias

Plotly. (2024). *Plotly Python Documentation*. https://plotly.com/python/

Streamlit. (2024). *Streamlit Documentation*. https://docs.streamlit.io

Scikit-learn. (2024). *Clustering: K-means Algorithm*. https://scikit-learn.org/stable/modules/clustering.html

IBM. (2024). *What is Data Visualization?*. https://www.ibm.com/topics/data-visualization

Google Developers. (2024). *Data Visualization Guide*. https://developers.google.com/datavis