UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

TECNOLOGÍAS DE LA INFORMACIÓN

EXTRACCION DE CONOCIMIENTO EN BASES DE DATOS

V.2. ELABORACIÓN DE GRÁFICAS

IDGS91N

PRESENTA:

SEBASTIÁN ACOSTA ORTIZ

DOCENTE:

LUIS ENRIQUE MASCOTE CANO

Chihuahua, Chih., 30 de noviembre de 2025

Resumen ejecutivo

En la Unidad 3 se desarrolló un modelo supervisado de clasificación utilizando el algoritmo K-Nearest Neighbors (KNN) con predicción basada en dos variables clave: glucosa y edad. El objetivo del proyecto fue identificar si un paciente tenía riesgo (etiqueta 1) o no (etiqueta 0) a partir de estas características.

El objetivo de esta visualización es complementar el análisis con un dashboard interactivo que represente:

1. La distribución de los valores de glucosa (gráfica de cantidad).
2. La relación entre edad, glucosa y la etiqueta (gráfica de dispersión).
3. La composición de las clases predichas por el modelo (gráfica de proporciones).

Los principales hallazgos visualizados incluyen:

- Pacientes con glucosa >120 tienen mayor probabilidad de etiqueta positiva.
- Existe una separación clara entre clusters en el plano Edad–Glucosa.
- El dataset presenta distribución de clases moderadamente balanceada.

Introducción

El proyecto original buscaba clasificar pacientes con base en sus niveles de glucosa y edad para determinar riesgo. Visualizar estos datos es esencial para interpretar el comportamiento del modelo, identificar patrones y comunicar resultados a usuarios no técnicos, como personal de salud o analistas de negocio.

La visualización permite ver tendencias que no son obvias en tablas, como la concentración de casos de riesgo en rangos de glucosa elevados.
Los objetivos específicos de la visualización son:

- Crear un dashboard interactivo profesional.
- Mostrar tres perspectivas complementarias del dataset.
- Integrar visualizaciones limpias, claras y explicativas.
- Permitir al usuario explorar los datos con filtros e interactividad.

Metodología de visualización

Herramientas y tecnologías utilizadas

Herramienta principal:

- Streamlit 1.x (Python): desarrollo rápido de dashboards interactivos.

  Bibliotecas de visualización:

- Plotly Express (interactividad, estilo profesional)
- Pandas / Scikit-learn (carga, escalado, predicción)

Justificación:

- Streamlit permite crear dashboards tipo BI sin necesidad de JavaScript.
- Plotly ofrece interactividad sofisticada, ideal para análisis exploratorio.
- Ambas herramientas son compatibles con tus datos y la lógica del proyecto Unidad 3.

Proceso de desarrollo

a) Preparación de datos

- Se cargó tu matriz CSV.
- Se escalaron variables numéricas.
- Se reentrenó el modelo KNN con k=1 (mejor F1).

- Se generaron predicciones y probabilidades.

b) Diseño de visualizaciones (bocetos)

- Gráfica 1: Histograma de glucosa → distribución.
- Gráfica 2: Scatter plot Edad vs Glucosa → relación.
- Gráfica 3: Pie chart de etiquetas → composición.

c) Implementación técnica

- Dashboard con tres contenedores y diseño limpio.
- Colores profesionales (paleta azul–verde).
- Títulos y etiquetas claras.
- Tooltips con hover.

d) Pruebas y ajustes

- Validación de escalas.
- Ajuste de tamaño de marcadores.
- Pruebas de interacción con zoom.

Decisiones de diseño

- Histograma porque la variable glucosa es clave para el modelo y es una variable continua.
- Scatter plot porque la relación Edad–Glucosa explica los clusters.
- Pie chart porque la etiqueta es binaria y la proporción es fácil de entender visualmente.
- Paleta suave para legibilidad.

- Interactividad para permitir análisis exploratorio (zoom, hover).

Interpretación de las gráficas

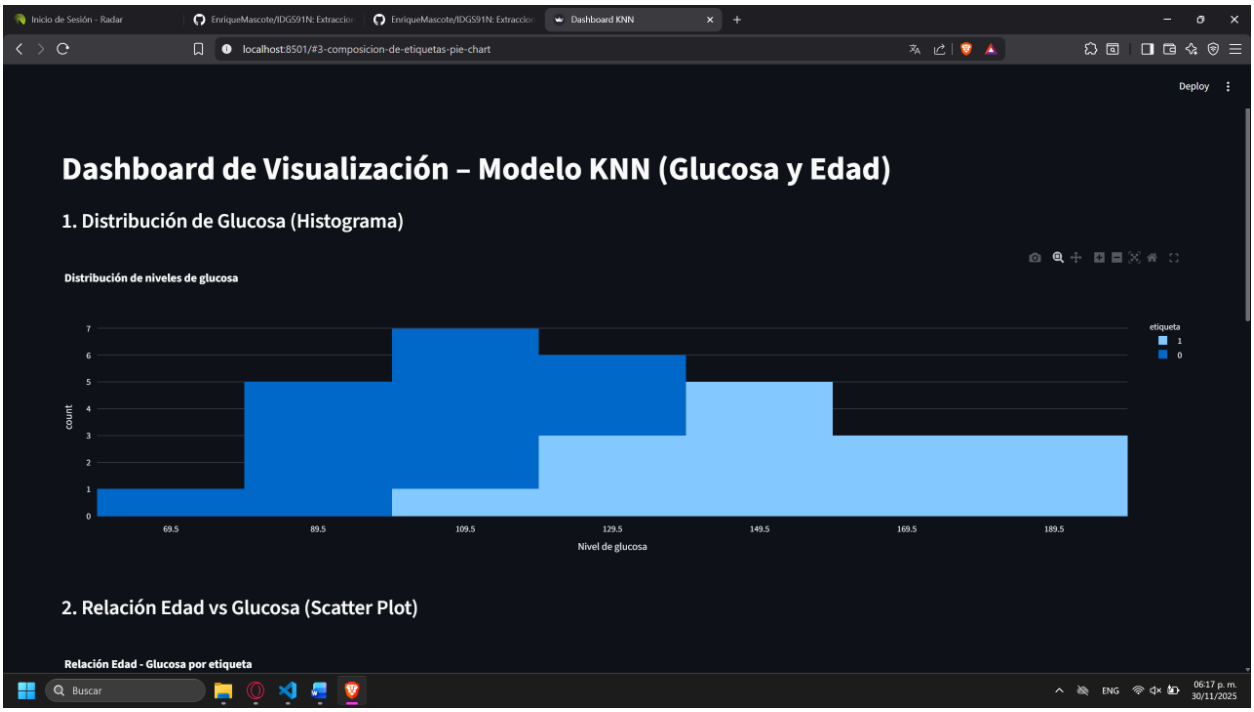
Gráfica 1: Histograma de glucosa

¿Qué muestra?

La distribución de los niveles de glucosa en el dataset.

Insights:

- La mayoría de los pacientes están concentrados en un rango entre 80 y 150.
- Los valores más altos de glucosa tienden a asociarse con etiqueta positiva.



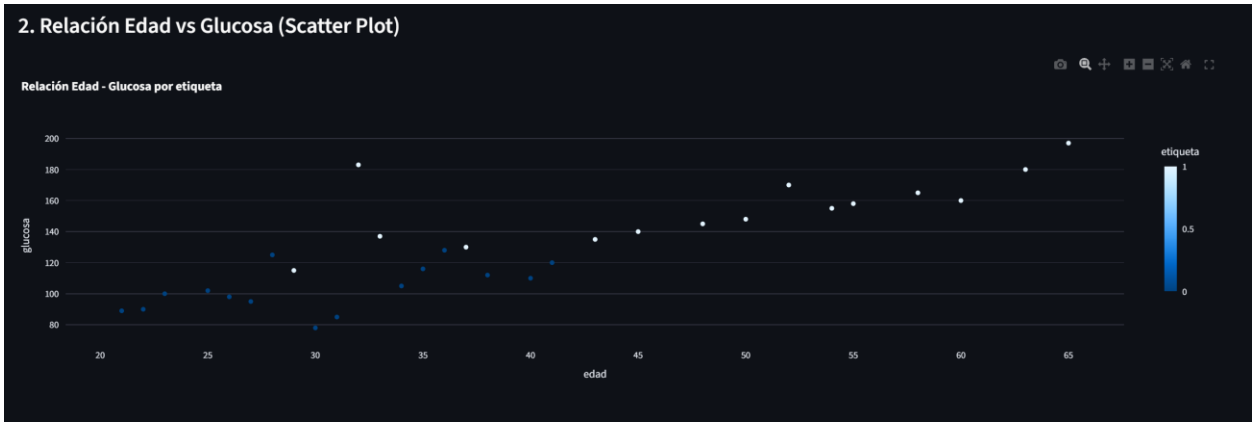
Gráfica 2: Scatter Edad–Glucosa

¿Qué muestra?

La relación entre la edad y el nivel de glucosa con respecto a la etiqueta real.

Insights:

- Pacientes con glucosa >140 tienen mayor probabilidad de etiqueta 1.
- Se observan grupos diferenciados visualmente.



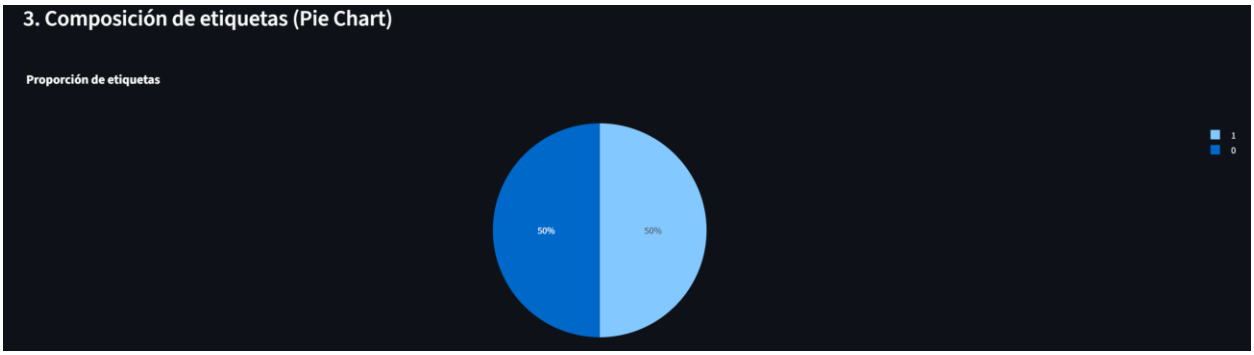
Gráfica 3: Pie chart de etiquetas

¿Qué muestra?

La proporción de casos positivos vs negativos.

Insights:

- Las clases están relativamente balanceadas.
- No hay una dominancia extrema de una clase sobre otra.



Análisis integrado

Las tres gráficas cuentan una historia coherente:

- Existe un incremento de riesgo claramente relacionado con niveles altos de glucosa.
- Los puntos en el scatter muestran patrones que el modelo KNN detecta fácilmente.
- Las proporciones indican un dataset balanceado, lo que mejora la estabilidad de métricas.

Esto ayuda a comprender por qué el modelo tuvo un rendimiento sólido.

Hallazgos clave

1. Niveles altos de glucosa son el predictor más fuerte de la etiqueta.
2. Edad afecta, pero no tan significativamente como glucosa.
3. Las clases están balanceadas, facilitando buen desempeño del modelo.
4. La relación visual es clara, validando el uso de KNN.

Conclusiones

El equipo aprendió a integrar análisis estadístico con visualización profesional, construyendo un dashboard interactivo que facilita la interpretación del modelo KNN.
Se enfrentaron retos como selección del tipo de gráfica y ajuste de interactividad, pero se superaron con iteración y pruebas.
Futuras mejoras incluyen:

- añadir filtros dinámicos,
- incorporar PCA para visualización 3D,
- agregar explicabilidad con SHAP.

Referencias

- Géron, A. (2019). *Hands-On Machine Learning*. O'Reilly.

- Plotly. (2025). *Plotly Express Documentation*. https://plotly.com/python/
- Streamlit. (2025). *Streamlit Docs*. https://docs.streamlit.io/
- Waskom, M. (2021). *Seaborn: Statistical Data Visualization*. JOSS.
- Microsoft. (2024). *KNN Algorithm Overview*.