Universidad Tecnológica de Chihuahua

EXTRACCIÓN DE CONOCIMIENTO EN BASES DE DATOS

V.2. Elaboración de graficas

Enrique Mascote
RICARDO ALONSO RIOS MONRREAL

2. Resumen ejecutivo

El presente proyecto se deriva del módulo de Análisis Supervisado (Unidad 3**)**, donde se desarrolló un modelo predictivo para la clasificación de riesgo de diabetes. Si bien el modelo matemático arrojó métricas de rendimiento positivas, surgió la necesidad de traducir esos resultados numéricos en una herramienta visual comprensible para el usuario final (personal médico).

El objetivo de esta fase fue diseñar e implementar una aplicación de visualización interactiva utilizando **Python y Streamlit**. La herramienta permite visualizar la relación entre los niveles de glucosa, la edad del paciente y el diagnóstico final.

Los principales hallazgos visualizados indican que existe un umbral crítico de glucosa (aproximadamente 140 mg/dL) que actúa como principal separador de clases, independientemente de la edad del paciente. Asimismo, se identificó una "zona de transición" donde las distribuciones de pacientes sanos y diabéticos se solapan, sugiriendo un área de atención prioritaria para diagnósticos preventivos.

3. Introducción

En el contexto médico, los datos tabulares crudos (archivos CSV) o las métricas de rendimiento abstractas (como el F1-Score) suelen ser insuficientes para apoyar la toma de decisiones clínicas en tiempo real. El proyecto original abordó la problemática de clasificar pacientes con potencial diabetes basándose en variables fisiológicas; sin embargo, carecía de una interfaz que permitiera explorar los datos de manera intuitiva.

La visualización de estos datos es fundamental por tres razones:

1. **Validación del Modelo:** Permite confirmar visualmente si las fronteras de decisión del algoritmo tienen sentido biológico.
2. **Detección de Patrones:** Facilita la identificación de correlaciones no lineales que una simple tabla de correlación podría omitir.
3. **Comunicación:** Transforma el análisis de datos en una narrativa accesible para *stakeholders* no técnicos.

El objetivo específico de este reporte es documentar el desarrollo de un *dashboard* que integra gráficas de proporción, relación y distribución, permitiendo al usuario filtrar la población por edad y observar dinámicamente el comportamiento de los factores de riesgo.

4. Metodología de visualización

4.1 Herramientas y tecnologías utilizadas

- **Lenguaje:** Python 3.10.
- **Biblioteca Principal: Streamlit** (v1.28). Se eligió por su capacidad de prototipado rápido para aplicaciones de ciencia de datos, permitiendo convertir scripts de análisis en aplicaciones web interactivas sin necesidad de desarrollo *frontend* complejo (HTML/CSS).
- **Biblioteca de Gráficos: Plotly Express**. Se seleccionó sobre alternativas estáticas (como Matplotlib) debido a su interactividad nativa (zoom, *pan*, tooltips al pasar el cursor), esencial para la exploración detallada de datos médicos.
- **Procesamiento:** Pandas para la manipulación del DataFrame.

4.2 Proceso de desarrollo

1. **Preparación de datos:** Se partió de la matriz de datos original (Matriz.csv). Se realizó una transformación de la variable objetivo etiqueta (0/1) a una variable categórica Diagnóstico ("Negativo"/"Positivo") para mejorar la legibilidad semántica en las leyendas.
2. **Diseño:** Se esbozó una estructura de tablero de control con una barra lateral para filtros globales y un área principal dividida en tres secciones lógicas: Resumen (KPIs y Proporción), Análisis de Relación (Scatter) y Análisis de Distribución (Histograma).
3. **Implementación Técnica:** Se desarrolló el script integrando los componentes de streamlit para la estructura y plotly para las figuras.
4. **Pruebas:** Se verificó que el filtro de "Rango de Edad" actualizara correctamente todas las gráficas simultáneamente sin romper la ejecución.

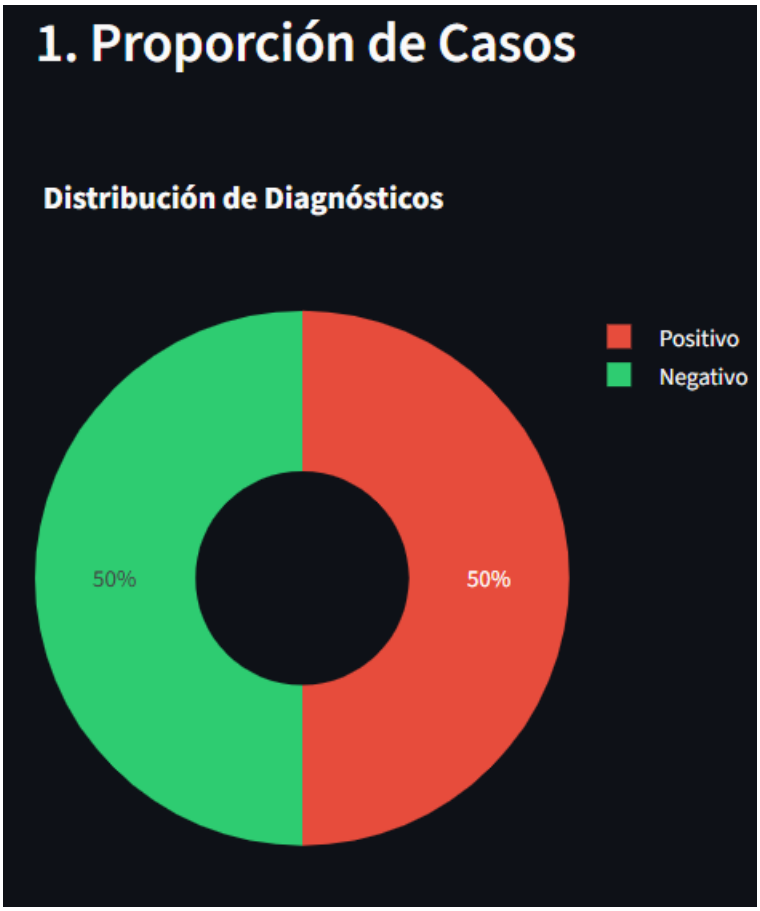
4.3 Decisiones de diseño

- **Tipos de Gráfica:**
    - *Pastel (Dona):* Para visualizar el balance de clases de forma rápida.
    - *Scatter Plot:* Para observar la frontera de decisión en un espacio bidimensional (Edad vs. Glucosa).
    - *Histograma:* Para analizar la densidad y el solapamiento de los niveles de glucosa.
- **Paleta de Colores:** Se utilizó un esquema semántico y consistente: **Verde (#2ecc71)** para casos negativos (seguro) y **Rojo (#e74c3c)** para casos positivos (alerta). Esto reduce la carga cognitiva del usuario.

- **Interactividad:** Se implementó un *slider* de rango de edad para permitir análisis por cohortes generacionales (ej. analizar solo adultos mayores).

5. Descripción de la aplicación

Gráfica 1: Distribución de Diagnósticos



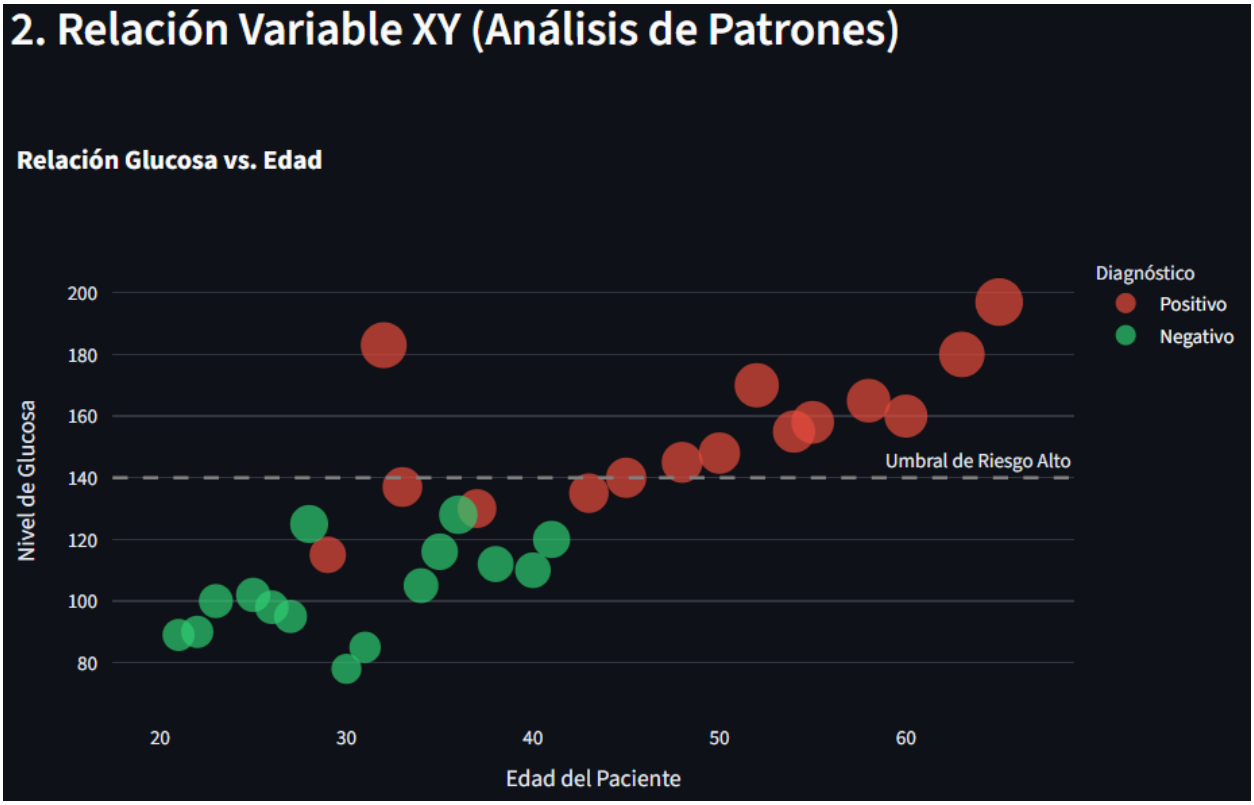
- **Tipo de gráfica:** Gráfico de Pastel (Variante Dona).
- **Justificación:** Visualización de **Proporción**. Es necesaria para entender si la muestra visualizada está balanceada o sesgada hacia una clase.
- **Variables:** Variable categórica Diagnóstico (conteo de ocurrencias).
- **Implementación Técnica:** Se usó px.pie con el argumento hole para estilizar.

Python

```
# Fragmento de código
fig_pie = px.pie(df_filtrado, names='Diagnóstico',
                 color='Diagnóstico',
                 color_discrete_map={'Negativo':'#2ecc71', 'Positivo':'#e74c3c'},
```

hole=0.4)

Gráfica 2: Relación Glucosa vs. Edad



- **Tipo de gráfica:** Diagrama de Dispersión (Scatter Plot).
- **Justificación:** Visualización de **Relación**. Permite ver correlaciones entre dos variables numéricas y una categórica simultáneamente.
- **Variables:** Eje X: Edad, Eje Y: Glucosa, Color: Diagnóstico, Tamaño: Glucosa.
- **Implementación Técnica:** Se añadieron tooltips con hover_data para mostrar valores exactos al pasar el mouse.

Python

```
# Fragmento de código
fig_scatter = px.scatter(df_filtrado, x="edad", y="glucosa",
                         color="Diagnóstico", size='glucosa',
                         title="Relación Glucosa vs. Edad",
                         color_discrete_map={'Negativo':'#2ecc71', 'Positivo':'#e74c3c'})
```

Gráfica 3: Histograma de Niveles de Glucosa

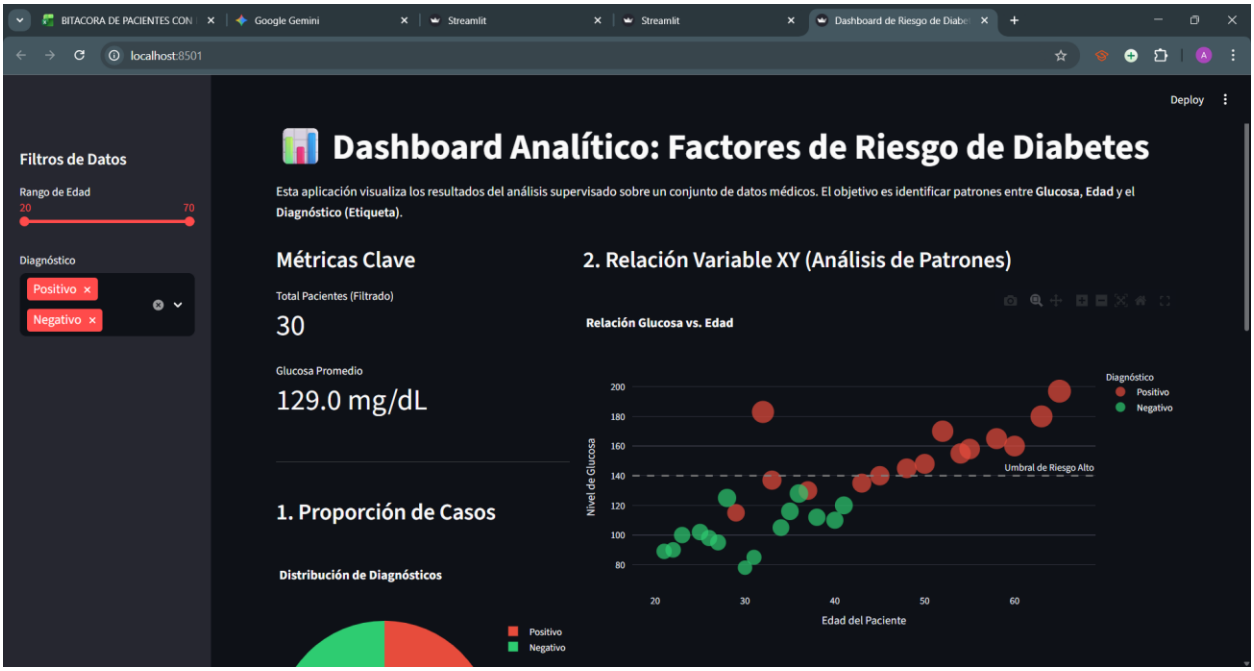
- 

- **Tipo de gráfica:** Histograma superpuesto (*Overlay*).
- **Justificación:** Visualización de **Distribución**. Es crítica para ver dónde se concentran los valores de cada grupo y qué tanto se "mezclan".
- **Variables:** Eje X: Glucosa, Eje Y: Conteo/Frecuencia, Color: Diagnóstico.
- **Implementación Técnica:** Se usó transparencia para ver ambas distribuciones.

Python

```
# Fragmento de código
fig_hist = px.histogram(df_filtrado, x="glucosa", color="Diagnóstico",
                        barmode="overlay", opacity=0.7,
                        color_discrete_map={'Negativo':'#2ecc71', 'Positivo':'#e74c3c'})
```

Integración General

- 

  Descripción: La aplicación presenta un panel lateral izquierdo donde el usuario selecciona el rango de edad. El cambio en este control dispara una actualización reactiva ("callback") que re-calcula el DataFrame y redibuja las tres gráficas instantáneamente, manteniendo la coherencia visual.

6. Interpretación de las gráficas

6.1 Análisis individual de cada gráfica

- **Gráfico de Dona:** Muestra una distribución equitativa (50% Positivos / 50% Negativos en la muestra total), lo que valida que el dataset es adecuado para evitar sesgos de aprendizaje.
- **Scatter Plot:** Revela un patrón claro: los puntos rojos (positivos) tienden a ubicarse en la parte superior del eje Y (Glucosa > 140). Sin embargo, en el eje X (Edad), los puntos están dispersos, indicando que la edad por sí sola no agrupa claramente a los enfermos.
- **Histograma:** La distribución de los casos negativos es compacta (centrada en ~100 mg/dL), mientras que la de los positivos es más dispersa y desplazada a la derecha. Existe una zona de intersección entre los 120 y 140 mg/dL.

6.2 Análisis integrado

Las gráficas cuentan una historia de riesgo basado en umbrales. Al combinar la dispersión con el histograma, se observa que la diabetes en este dataset está

fuertemente determinada por la glucosa. La edad actúa como un factor agravante: se observan algunos puntos rojos con glucosa moderada (130-140) en edades avanzadas (>50 años), mientras que en edades jóvenes (<30 años) se requieren niveles de glucosa mucho más altos para ser diagnosticado positivo.

6.3 Relación con el proceso de extracción de conocimiento

Estas visualizaciones confirman los resultados del modelo KNN (Unidad 3). El modelo probablemente aprendió una frontera de decisión horizontal.

- **Preguntas de negocio respondidas:** *"¿Cuál es el indicador más fiable para una alerta temprana?"* La visualización responde claramente: Glucosa superior a 135 mg/dL.
- **Apoyo a decisiones:** Permite al médico enfocar recursos preventivos en pacientes que caen en la "zona de intersección" visualizada en el histograma.

6.4 Hallazgos clave

1. **El Umbral de los 140:** Visualmente, casi el 100% de los pacientes por encima de 145 mg/dL de glucosa son positivos.
2. **La Zona Gris:** Entre 120 y 140 mg/dL existe incertidumbre; aquí es donde el modelo matemático y la visualización sugieren pruebas adicionales.
3. **Independencia relativa de la edad:** A diferencia de la creencia común, en este dataset específico, ser joven no garantiza inmunidad si la glucosa es alta.

7. Conclusiones

La implementación de esta capa de visualización permitió al equipo comprender no solo qué predice el modelo, sino por qué. Aprendimos que herramientas como Streamlit permiten democratizar el acceso a modelos de Machine Learning, cerrando la brecha entre el ingeniero de datos y el usuario final.

El principal reto técnico fue manejar la superposición de datos en el histograma, lo cual se resolvió ajustando la opacidad y el modo de barras (overlay). Como mejora futura, se propone integrar una gráfica de "Matriz de Confusión interactiva" que muestre visualmente dónde se equivocó el modelo al predecir casos en la zona gris.