Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

El histograma de la variable *Pregnancies* muestra la distribución de frecuencias del número de embarazos en la muestra, donde la mayoría de observaciones se concentra en valores bajos (entre 0 y 3), con una cola larga hacia valores más altos. Esto refleja una distribución asimétrica positiva, por lo que la mediana resulta más representativa que la media. En consecuencia, para la imputación o resumen estadístico se recomienda emplear medidas robustas como la mediana. Sin embargo, el histograma no permite distinguir entre embarazos válidos o erróneos ni muestra dependencia con otras variables, por lo que se complementa con el boxplot y una revisión de outliers.

Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

El boxplot de *Pregnancies* permite observar la mediana, los cuartiles y algunos valores por fuera de los bigotes, que podrían considerarse atípicos en el extremo superior. Estos registros con número de embarazos inusualmente alto pueden representar tanto casos reales poco frecuentes como errores de captura. Por ello, se sugiere aplicar el criterio del rango intercuartílico (IQR) para detectar outliers, revisar manualmente los registros extremos y conservar únicamente los casos válidos. Aunque el boxplot resume eficazmente la distribución, no aporta información sobre la causa del atípico ni su relación con la variable objetivo.

Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

En el histograma de *Glucose* se observa una distribución sesgada a la derecha, con una concentración en rangos bajos y moderados y una cola que se extiende hacia valores altos. Esta asimetría indica la presencia de valores extremos que influyen sobre la media, por lo cual la mediana describe mejor la tendencia central. Se recomienda la imputación por mediana en caso de valores nulos y aplicar un escalado robusto antes del modelado para mitigar el efecto de los outliers. El histograma no permite separar errores de medición de casos clínicos reales, por lo que se requiere validación adicional.

Gráfico

El contenido generado por IA puede ser incorrecto.

El boxplot de *Glucose* expone la dispersión de los valores y evidencia observaciones por encima del rango intercuartílico esperado. Algunos de estos valores extremos pueden corresponder a errores de medición o a pacientes con hiperglucemia. Por ello, es necesario validar los registros extremos y decidir si deben corregirse, truncarse o conservarse según su verosimilitud clínica. Esta visualización tampoco muestra si los outliers se concentran en determinados subgrupos, por lo que resulta útil cruzar la variable con otras, como edad o diagnóstico.

Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

En el histograma de *BloodPressure* se aprecia una distribución con moda en valores normales y algunos picos en rangos bajos, incluyendo registros con valores cercanos a cero. Estos valores atípicos bajos pueden deberse a errores de captura o a mediciones faltantes codificadas como cero. Se recomienda revisarlos y, de ser necesario, tratarlos como nulos e imputarlos con la mediana, además de aplicar un escalado posterior. El histograma no distingue entre tipos de presión ni condiciones de medida, por lo que debe revisarse el protocolo de recolección de datos.

Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

El boxplot de *BloodPressure* muestra registros extremos tanto por debajo como por encima de los bigotes. Estos casos pueden representar errores de entrada o valores poco probables fisiológicamente. Es conveniente aplicar el filtro IQR para marcarlos y revisarlos manualmente, corrigiendo errores de digitación antes de la imputación. La eliminación mecánica de outliers puede ser riesgosa, ya que algunos podrían tener relevancia clínica.

Gráfico

El contenido generado por IA puede ser incorrecto.

En el histograma de *SkinThickness* se observa una concentración en valores bajos a medios, con picos dispersos y una posible acumulación en ceros. Esta distribución sugiere que algunos ceros podrían corresponder a mediciones no realizadas. Se recomienda tratarlos como valores nulos e imputar con la mediana, además de escalar la variable antes del modelado.

Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

El boxplot de *SkinThickness* confirma la presencia de outliers en el extremo superior que aumentan la dispersión aparente; sin embargo, la mediana permanece estable. Se recomienda revisar manualmente los valores extremos y emplear un escalado robusto para reducir su impacto. El boxplot no indica si los outliers están concentrados por subgrupos, por lo que conviene segmentar por edad o sexo.

Tabla

El contenido generado por IA puede ser incorrecto.

El histograma de *Insulin* presenta una distribución altamente sesgada con una larga cola de valores altos y posible acumulación de ceros, lo que sugiere mediciones ausentes o codificadas incorrectamente. Dada su heterogeneidad, se recomienda considerar los ceros como nulos, imputar por mediana y escalar con un método robusto.

Gráfico, Escala de tiempo

El contenido generado por IA puede ser incorrecto.

El boxplot de *Insulin* refuerza esta observación al mostrar amplia dispersión y múltiples valores extremos en el percentil superior. Se sugiere mantener los registros válidos, corregir errores evidentes y considerar una transformación logarítmica si el modelo lo requiere, aunque esto puede afectar la interpretabilidad clínica.

Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

El histograma de *BMI* muestra una mayor concentración de valores en rangos bajos a moderados, con una cola hacia valores asociados a obesidad y algunos casos en cero, posiblemente por mediciones faltantes. La distribución es coherente con poblaciones mixtas, por lo que la variable debe mantenerse por su relevancia clínica, pudiendo escalarse o categorizarse según el análisis.

Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

El boxplot de *BMI* confirma una baja proporción de valores fuera de los whiskers, lo que sugiere pocos outliers. Estos deben revisarse y, si es necesario, corregirse con criterio clínico.

Gráfico

El contenido generado por IA puede ser incorrecto.

El histograma de *DiabetesPedigreeFunction* muestra una concentración en valores bajos con una cola hacia valores altos, reflejando una asimetría positiva. La mayoría de individuos tiene baja predisposición genética, con pocos casos de riesgo alto. Se recomienda mantenerla como variable continua y escalarla para mejorar su comparabilidad.

Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

En el boxplot de *DiabetesPedigreeFunction* se observan varios outliers en el extremo superior, aunque la dispersión central es reducida. Los valores altos deben conservarse, pues pueden representar información genética relevante.

Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

El histograma de *Age* evidencia una distribución centrada en adultos jóvenes y mayores, con menor frecuencia en edades extremas. Representa una heterogeneidad etaria esperada y potencialmente correlacionada con otras variables clínicas.

Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

El boxplot de *Age* muestra una dispersión moderada y pocos outliers, principalmente en edades extremas, que deben conservarse tras verificar su validez.

Tabla

El contenido generado por IA puede ser incorrecto.

En el histograma de *Outcome*, variable binaria que representa la presencia o ausencia de la condición, se observa un desbalance de clases, con una predominancia de la clase negativa. Esto implica que en el modelado deberán considerarse métricas robustas como F1 o AUC, y técnicas de balanceo si se justifica.

Gráfico

El contenido generado por IA puede ser incorrecto.

El boxplot de *Outcome*, aunque menos informativo por la naturaleza binaria de la variable, confirma el desbalance y no aporta información adicional más allá de la separación entre clases.

Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

El histograma de *StressLevel* revela una distribución sesgada hacia niveles bajos y moderados, con pocos valores altos. Esta heterogeneidad sugiere que el estrés podría correlacionar con variables de estilo de vida, como *Lifestyle* o *SmokingStatus*.

Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

El boxplot de *StressLevel* muestra algunos valores extremos que amplían la varianza; se recomienda mantenerlos si son registros válidos y escalar con *RobustScaler* para mitigar su impacto.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

En los diagramas de frecuencia, la variable *Lifestyle* muestra predominancia de estilos sedentarios y activos, reflejando la distribución esperada en la población estudiada. Para su codificación se recomienda *One-Hot Encoding*, agrupando categorías poco frecuentes bajo “Otros” para evitar dispersión en el espacio de características.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

En *DietType*, predominan las dietas bajas en carbohidratos, mientras que las balanceadas son menos frecuentes; esta variable se considera relevante para el estado metabólico y se recomienda también su codificación binaria, consolidando categorías raras.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

El diagrama de frecuencia de *SmokingStatus* indica una mayoría de no fumadores, con presencia menor de fumadores activos. Esta variable representa un factor de riesgo importante y debe incluirse codificada mediante *One-Hot Encoding*.

Gráfico

El contenido generado por IA puede ser incorrecto.

En el caso de *FamilyStory*, la mayoría de individuos reporta historia familiar de la condición, lo que la convierte en un indicador útil de predisposición genética. Dado su carácter binario, puede mantenerse tal cual y considerarse en interacción con *DiabetesPedigreeFunction*.

Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

Finalmente, la matriz de correlación de variables numéricas calculada mediante el coeficiente de Pearson muestra correlaciones moderadas entre *Insulin* y *SkinThickness*, así como entre *Age* y *Pregnancies*, sin pares con correlaciones mayores a 0.6 en valor absoluto. Esto indica baja colinealidad, por lo que todas las variables pueden incluirse en los modelos predictivos. Sin embargo, es importante reconocer que la correlación solo mide relaciones lineales y puede verse afectada por outliers o ruido, por lo que debe complementarse con otros análisis de dependencia.