**UNIVERSIDAD DE SANTIAGO DE CHILE**

**FACULTAD DE INGENIERÍA**

**DEPARTAMENTO DE INGENIERÍA INFORMÁTICA**


# Laboratorio 3

## Análisis Estadístico e Inferencial

Integrantes: Diego Valdés Fernández

Valentina Campos Olguín

Curso: Análisis de Datos

Profesor: Dr. Max Chacón

Ayudante: Marcelo Álvarez

27 de noviembre de 2024

# Tabla de contenidos

# 1. Introducción

La diabetes es una enfermedad crónica que representa uno de los principales desafíos en salud pública a nivel mundial. En este contexto, el presente laboratorio tiene como objetivo principal descubrir patrones relevantes en los datos clínicos de pacientes mediante el uso de reglas de asociación. Para ello, se analizan diversas variables diagnósticas, tales como los niveles de glucosa, el índice de masa corporal (BMI), la presión arterial y otros indicadores clínicos, con el propósito de identificar posibles relaciones significativas entre estas características y el diagnóstico de diabetes.

Este análisis se basa en el enfoque metodológico del algoritmo Apriori, una técnica ampliamente utilizada en la minería de datos para la generación de reglas de asociación. Este laboratorio complementa lo realizado en el Laboratorio 1, donde se evaluaron las correlaciones entre variables clave, facilitando así una comprensión más profunda de los datos y permitiendo identificar patrones que podrían ser clínicamente relevantes. En este caso, el preprocesamiento juega un papel crucial, pues se corrigen valores anómalos, se imputan valores faltantes y se discretizan las variables, asegurando la calidad de los datos antes de aplicar el análisis.

El laboratorio incluye la configuración del algoritmo Apriori con parámetros específicos de soporte, confianza y lift para identificar reglas significativas. Finalmente, los resultados obtenidos se interpretan en términos de su relevancia clínica, su utilidad para el diagnóstico de diabetes y su comparación con estudios previos.

## 1.1. Objetivos

### 1.1.1. Objetivo general

Analizar el conjunto de datos clínicos de pacientes mediante reglas de asociación para identificar patrones significativos relacionados con el diagnóstico de diabetes.

### 1.1.2. Objetivos específicos

- Preprocesar el conjunto de datos corrigiendo valores anómalos, imputando datos faltantes y discretizando las variables relevantes.
- Configurar y aplicar el algoritmo Apriori para generar reglas de asociación basadas en el conjunto de datos.
- Analizar e interpretar las reglas generadas en términos de soporte, confianza y lift.
- Comparar los patrones encontrados con los resultados obtenidos en el Laboratorio 1, evaluando la coherencia entre los métodos exploratorios y de asociación.
- Evaluar la utilidad clínica de las reglas generadas en el contexto del diagnóstico de diabetes.

# 2. Marco teórico

En este apartado, se definen los conceptos clave utilizados en el desarrollo del laboratorio, los cuales son fundamentales para comprender los resultados obtenidos a partir del análisis de reglas de asociación. Estos conceptos incluyen la definición de las reglas de asociación, sus métricas principales (soporte, confianza y lift), y el algoritmo Apriori utilizado para generar dichas reglas.

## 2.1. Reglas de asociación

Las reglas de asociación son un método de minería de datos utilizado para identificar patrones o relaciones significativas entre variables en grandes conjuntos de datos. Estas reglas tienen la forma de una implicación $A \rightarrow B$, donde $A$ es el antecedente y $B$ el consecuente. El objetivo de estas reglas es encontrar combinaciones frecuentes de variables que permitan describir relaciones importantes dentro de los datos.

Por ejemplo, en un contexto clínico, una regla como Glucosa alta $\rightarrow$ Diagnóstico de diabetes podría ser de utilidad para identificar factores de riesgo relevantes.

## 2.2. Soporte

El soporte mide la proporción de registros en el conjunto de datos en los que ocurre tanto el antecedente ($A$) como el consecuente ($B$) de la regla. Se calcula mediante la fórmula:

$$\text{Soporte}(A \rightarrow B) = \frac{\text{Frecuencia}(A \cap B)}{\text{Total de registros}}$$

Un alto soporte indica que la regla es frecuente en el conjunto de datos. Por ejemplo, si $\text{Soporte}(A \rightarrow B) = 0{,}15$, significa que el 15 % de los registros cumplen con la regla.

## 2.3. Confianza

La confianza evalúa la probabilidad de que el consecuente ($B$) ocurra dado que el antecedente ($A$) ya ocurrió. Se calcula como:

$$\text{Confianza}(A \to B) = \frac{\text{Frecuencia}(A \cap B)}{\text{Frecuencia}(A)}$$

Una alta confianza indica que $B$ ocurre con frecuencia cuando $A$ es verdadero. Por ejemplo, una confianza del 80 % indica que $B$ ocurre en el 80 % de los casos en los que ocurre $A$.

## 2.4. Lift

El lift mide la fuerza de la asociación entre $A$ y $B$, comparándola con la ocurrencia esperada de $B$ si $A$ y $B$ fueran independientes. Se calcula como:

$$\text{Lift}(A \to B) = \frac{\text{Confianza}(A \to B)}{\text{Soporte}(B)}$$

Un lift mayor a 1 indica una relación positiva entre $A$ y $B$. Por ejemplo, un lift de 2 significa que $B$ es dos veces más probable de ocurrir cuando $A$ ocurre.

## 2.5. Monotonicidad del soporte mínimo

La propiedad de monotonicidad establece que si un conjunto de elementos no cumple con un soporte mínimo, entonces cualquier superconjunto de esos elementos tampoco cumplirá con el soporte mínimo. Esta propiedad es esencial para optimizar el algoritmo Apriori, ya que permite reducir el espacio de búsqueda al descartar combinaciones infrecuentes.

## 2.6. Algoritmo Apriori

El algoritmo Apriori es un método utilizado para generar reglas de asociación basado en la identificación de conjuntos frecuentes de ítems. Funciona iterativamente mediante los siguientes pasos:

- Identificar conjuntos de ítems frecuentes en el conjunto de datos utilizando un soporte mínimo.

- Generar reglas de asociación a partir de los conjuntos frecuentes, evaluando métricas como confianza y lift.
- Filtrar las reglas generadas según los umbrales establecidos.

La eficiencia del algoritmo se basa en la propiedad de monotonicidad, que reduce la cantidad de combinaciones a analizar, haciendo que sea una técnica ampliamente utilizada en minería de datos.

# 3. Pre-procesamiento

## 3.1. Limpieza de datos

El pre-procesamiento de los datos se enfocó en garantizar la calidad de las variables utilizadas en el análisis. Las acciones realizadas fueron las siguientes:

- **Análisis de valores cero:** Se identificaron variables con valores igual a cero en columnas donde esto no tiene sentido clínico, como glucosa, BMI, presión arterial e insulina. Estos valores se trataron como datos faltantes.
- **Imputación de valores cero:** Los valores cero en las variables mencionadas fueron reemplazados por la mediana de la columna correspondiente. Se eligió la mediana porque es más robusta frente a valores extremos y representa mejor la tendencia central en datos clínicos.
- **Conservación de todas las columnas:** Todas las columnas se consideraron relevantes para el análisis, ya que están relacionadas con el diagnóstico de diabetes o factores asociados. No se eliminaron columnas en este proceso.

Estas acciones permitieron corregir inconsistencias en los datos y preservar toda la información necesaria para el análisis de patrones mediante reglas de asociación.

Estas acciones garantizaron que las variables críticas estuvieran completas y listas para el análisis posterior.

## 3.2. Transformación de datos

Antes de comenzar con la discretización de las variables numéricas, es importante resaltar que los niveles de glucosa e insulina se midieron después de 2 horas de realizar el test de tolerancia a la glucosa (OGTT, por sus siglas en inglés) (Kaggle, 2016). Este test es comúnmente utilizado para evaluar la capacidad del cuerpo para manejar la glucosa, y la medición a las 2 horas es un indicador clave para diagnosticar la diabetes. Según la (Association, 2024), los valores de glucosa superiores a 200 mg/dL después de 2 horas en

una prueba de tolerancia a la glucosa confirman el diagnóstico de diabetes. Por lo tanto, los rangos utilizados en este análisis para clasificar la glucosa se basan en los valores obtenidos tras este test.

Las variables numéricas y categóricas fueron transformadas para facilitar su uso en el modelo de reglas de asociación. A continuación, se muestra la discretización de algunas variables clave.

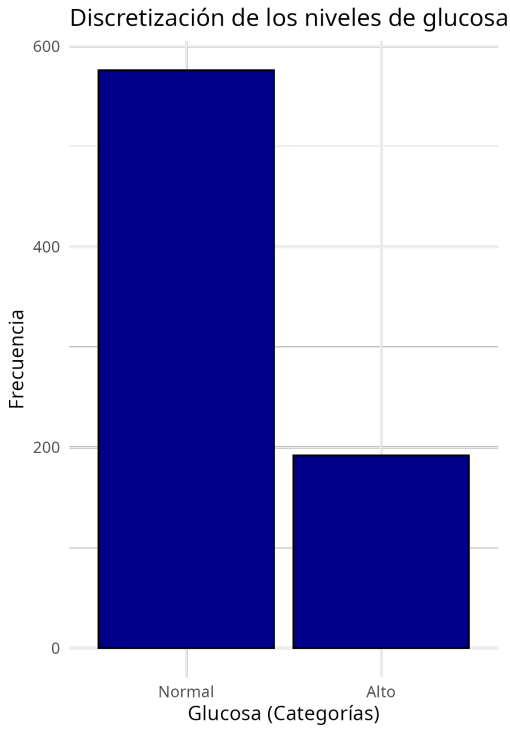


Figura 1: Discretización de los niveles de glucosa.

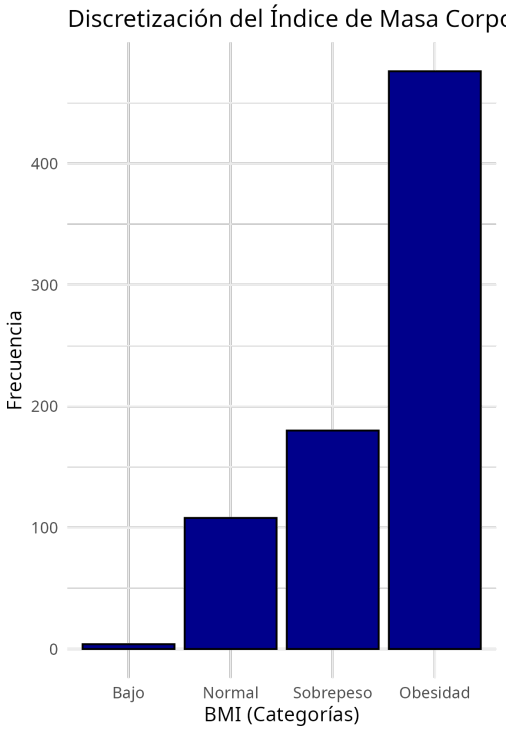


Figura 2: Discretización del índice de masa corporal (BMI).

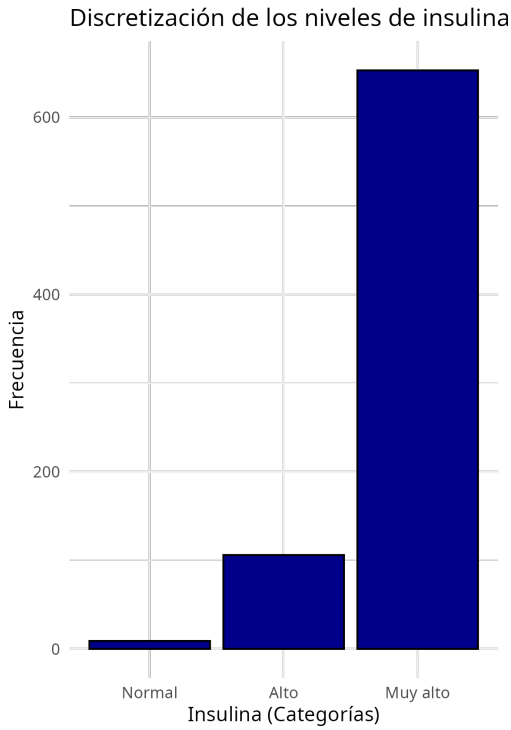


Figura 3: Discretización de los niveles de insulina.

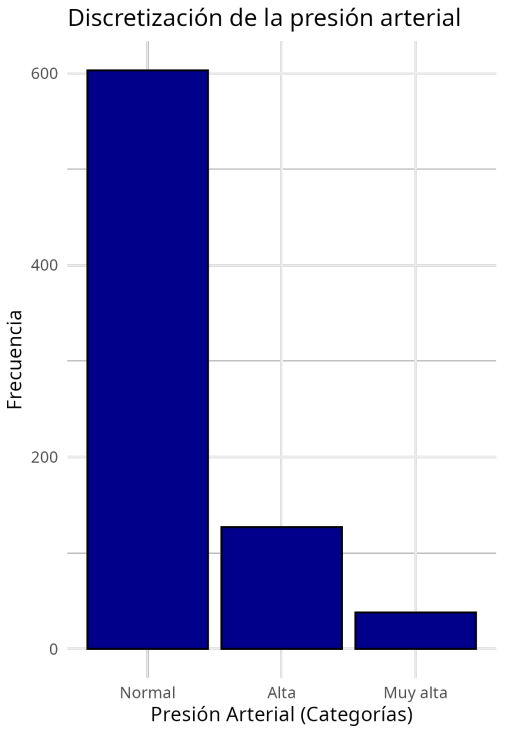


Figura 4: Discretización de la presión arterial.



Figura 5: Discretización de las variables clínicas clave.

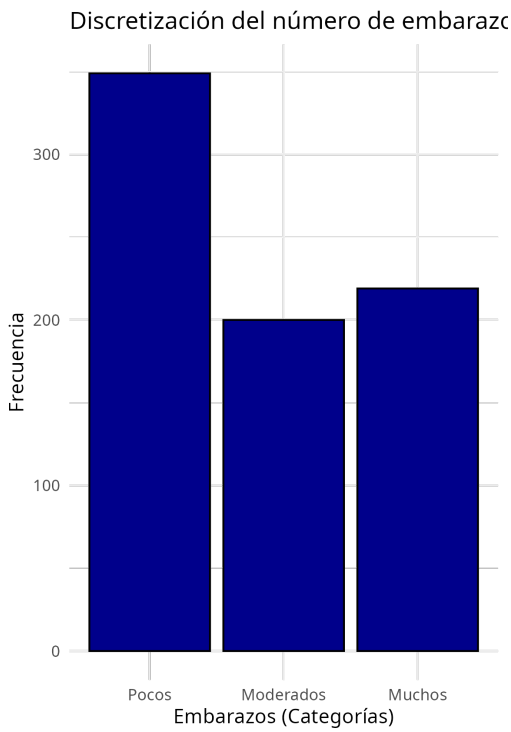


Figura 6: Discretización del número de embarazos.

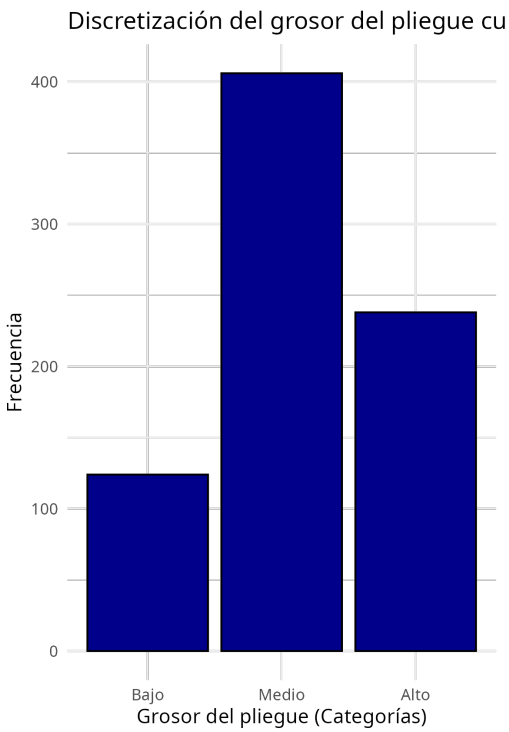


Figura 7: Discretización del grosor del pliegue cutáneo.

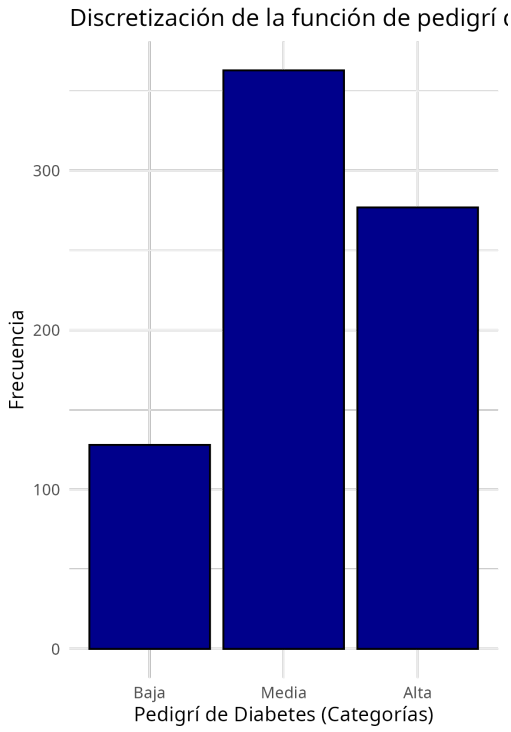


Figura 8: Discretización de la función de pedigrí de diabetes.

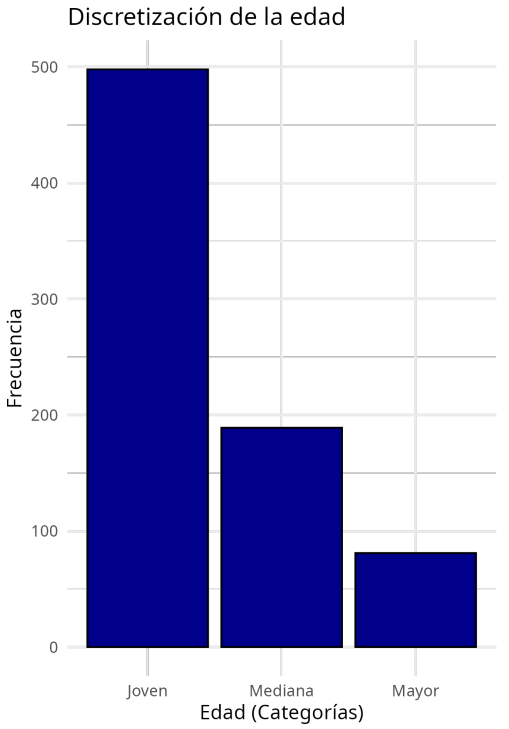


Figura 9: Discretización de la edad.



Figura 10: Discretización de otras variables clínicas.

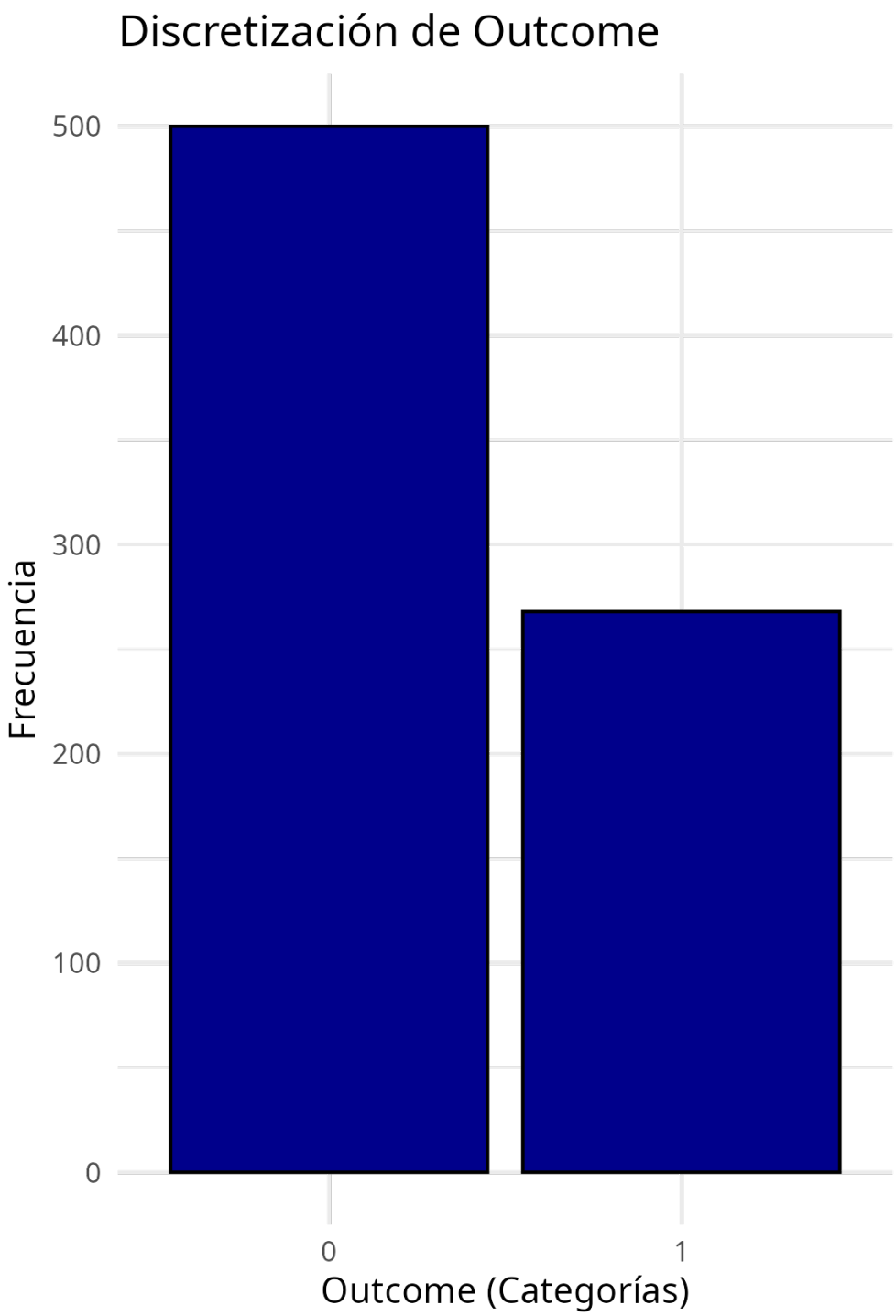


Figura 11: Discretización de Outcome (Diagnóstico de Diabetes).

### 3.2.1. Variables numéricas

Se discretizaron variables continuas como glucosa, BMI e insulina, basándose en rangos clínicamente relevantes. Esto permitió convertir datos continuos en categorías interpretables:

- **Glucosa (Glucose):**
  - **Normal:** $(< 140)$ mg/dL
  - **Alto:** $(140 - 200)$ mg/dL
  - **Muy alto:** $(> 200)$ mg/dL

  La clasificación de glucosa se basa en las guías de la Asociación Americana de Diabetes (ADA), donde valores superiores a 140 mg/dL tras una prueba de tolerancia oral a la glucosa indican prediabetes, y valores superiores a 200 mg/dL sugieren diabetes.

- **Índice de Masa Corporal (BMI):**
  - **Bajo:** $(< 18{,}5)$ kg/m$^2$
  - **Normal:** $(18{,}5 - 25)$ kg/m$^2$
  - **Sobrepeso:** $(25 - 30)$ kg/m$^2$
  - **Obesidad:** $(> 30)$ kg/m$^2$

  Estos rangos se establecen según la Organización Mundial de la Salud (OMS) y son utilizados globalmente para evaluar el estado nutricional de las personas (Organization, 2021).

- **Insulina (Insulin):**
  - **Normal:** $(< 25)$ $\mu$U/mL
  - **Alto:** $(25 - 85)$ $\mu$U/mL
  - **Muy alto:** $(> 85)$ $\mu$U/mL

Los rangos para insulina se basan en estudios clínicos que evalúan niveles séricos en ayunas y tras pruebas de tolerancia a la glucosa, considerando que valores elevados están asociados con resistencia a la insulina y diabetes (Venkatesan et al., 2024).

- **Presión arterial (BloodPressure):**
  - **Normal:** ($< 80$) mmHg
  - **Alta:** ($80 - 90$) mmHg
  - **Muy alta:** ($> 90$) mmHg

  Esta clasificación se basa en las guías de hipertensión de la Organización Mundial de la Salud (Organization, 2021) y la Asociación Americana del Corazón (AHA), donde se considera hipertensión para valores superiores a 80 mmHg en presión diastólica.

- **Número de embarazos (Pregnancies):**
  - **Pocos:** ($< 2$)
  - **Moderados:** ($2 - 5$)
  - **Muchos:** ($> 5$)

  Aunque no existen rangos clínicos oficiales para esta variable, se agruparon los valores en función de su distribución y estudios que sugieren que múltiples embarazos están asociados con un mayor riesgo de diabetes (Venkatesan et al., 2024).

- **Grosor del pliegue cutáneo (SkinThickness):**
  - **Bajo:** ($< 20$) mm
  - **Medio:** ($20 - 30$) mm
  - **Alto:** ($> 30$) mm

  Estos valores fueron definidos a partir de estándares antropométricos utilizados para evaluar la adiposidad subcutánea (GPNotebook, 2024). Los rangos reflejan niveles bajos, normales y altos de grasa subcutánea.

- **Función de Pedigrí de Diabetes (DiabetesPedigreeFunction):**

- **Baja:** ($< 0{,}2$)
- **Media:** ($0{,}2 - 0{,}5$)
- **Alta:** ($> 0{,}5$)

Los rangos para esta variable se basaron en la distribución observada en el dataset, con un enfoque en resaltar diferentes niveles de predisposición genética a la diabetes, según estudios de (Aubaidan & Kadir, 2023).

- **Edad (Age):**
  - **Joven:** ($< 35$) años
  - **Mediana:** ($35 - 50$) años
  - **Mayor:** ($> 50$) años

  Esta categorización se fundamenta en estudios epidemiológicos que muestran que el riesgo de diabetes aumenta significativamente con la edad, especialmente después de los 45 años (Association, 2024).

  La discretización se realizó utilizando la función `cut()` en R.

#### 3.2.2. Variables categóricas

La variable `Outcome` (diagnóstico de diabetes) ya estaba representada de forma binaria, por lo que no fue necesario aplicar técnicas adicionales como one-hot encoding.

# 4. Obtención de reglas

## 4.1. Configuración del algoritmo Apriori

El algoritmo Apriori se configuró utilizando la función `apriori()` del paquete `arules` en R. Los parámetros se definieron como sigue:

- **Soporte mínimo:** $0{,}1$, para identificar asociaciones presentes en al menos el $10\,\%$ de los registros.

- **Confianza mínima:** $0{,}7$, para garantizar asociaciones fiables.
- **Longitud mínima:** $2$, para evitar reglas triviales de un solo ítem.

El código implementado fue:

```
library(arules)
reglas <- apriori(data = diabetes_data,
                  parameter = list(support = 0.1, confidence = 0.7, minlen = 2))

reglas_significativas <- subset(reglas, subset = lift > 1.2)
```

## 4.2. Visualización de reglas

Las reglas generadas fueron visualizadas mediante gráficos de red para resaltar las asociaciones entre las variables clínicas y el diagnóstico de diabetes. En la Figura 12, se presentan las asociaciones más significativas, donde el tamaño de los nodos representa el **soporte** de cada ítem, y el color de las aristas indica el **lift** de las reglas. Las asociaciones con mayores valores de lift, como $\{\text{Glucosa=Alto}\} \rightarrow \{\text{Outcome=1}\}$, se destacan con mayor intensidad en la visualización.

Estos gráficos permitieron identificar visualmente asociaciones relevantes entre las variables del dataset.

# 5. Análisis de resultados

## 5.1. Interpretación de reglas

La Figura 13 muestra un análisis más detallado de las reglas filtradas por su relevancia clínica y estadística. En esta red, las reglas asociadas al diagnóstico de diabetes ($\{\text{Outcome=1}\}$) presentan conexiones destacadas con variables como glucosa alta, obesidad y niveles altos de insulina. Estas asociaciones confirman los patrones esperados clínicamente y aportan un marco sólido para el análisis exploratorio de los datos.

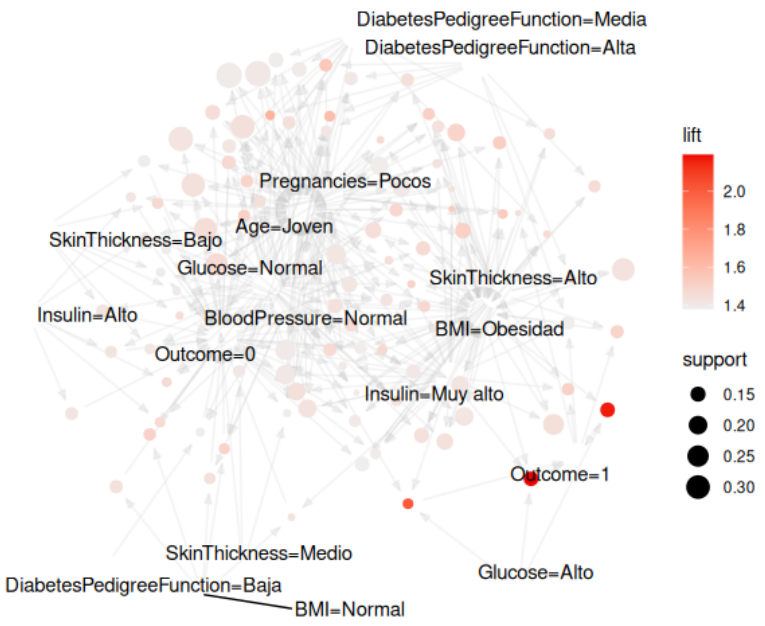


Figura 12: Red de reglas significativas generadas mediante el algoritmo Apriori.

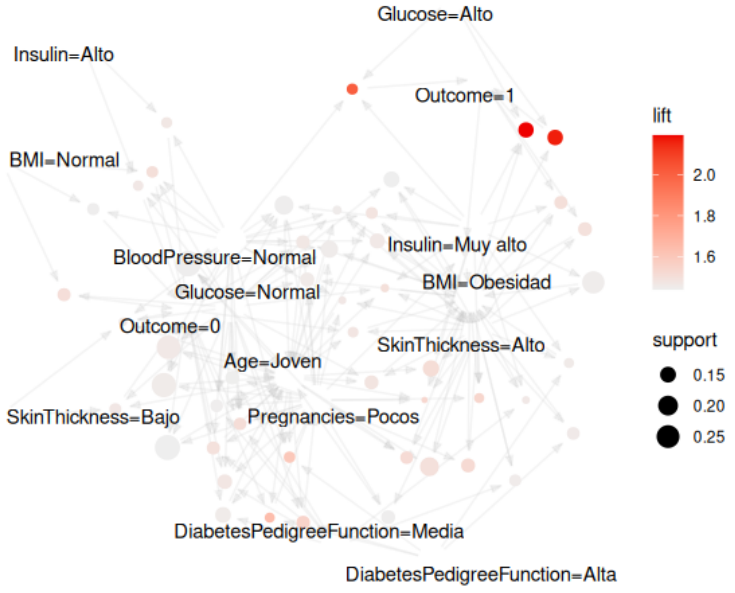


Figura 13: Red de reglas filtradas por relevancia clínica y estadística.

Se identificaron reglas con soporte, confianza y lift significativos. A continuación, se presentan las reglas más destacadas junto con su interpretación:

- **Regla 1:** {Glucosa=Alto, Insulina=Muy alto, BMI=Obesidad} $\rightarrow$ {Outcome=1}

  - **Soporte:** $14{,}5\,\%$
  - **Confianza:** $76{,}6\,\%$
  - **Lift:** $2{,}19$

  **Interpretación:** Las pacientes con niveles altos de glucosa, insulina muy alta y obesidad tienen $2{,}19$ veces más probabilidades de ser diagnosticadas con diabetes en comparación con lo esperado por azar.

- **Regla 2:** {Glucosa=Alto, BMI=Obesidad} $\rightarrow$ {Outcome=1}

  - **Soporte:** $14{,}6\,\%$
  - **Confianza:** $75{,}7\,\%$
  - **Lift:** $2{,}17$

  **Interpretación:** Las pacientes con niveles altos de glucosa y obesidad tienen $2{,}17$ veces más probabilidades de ser diagnosticadas con diabetes.

- **Regla 3:** {Glucosa=Alto, Presión arterial=Normal, Insulina=Muy alto} $\rightarrow$ {Outcome=1}

  - **Soporte:** $11{,}3\,\%$
  - **Confianza:** $70{,}2\,\%$
  - **Lift:** $2{,}01$

  **Interpretación:** Las pacientes con glucosa alta, presión arterial normal e insulina muy alta tienen $2{,}01$ veces más probabilidades de ser diagnosticadas con diabetes.

- **Regla 4:** {Embarazos=Pocos, Grosor del pliegue cutáneo=Alto, Insulina=Muy alto} $\rightarrow$ {BMI=Obesidad}

- **Soporte:** $10{,}8\,\%$
- **Confianza:** $95{,}4\,\%$
- **Lift:** $1{,}54$

**Interpretación:** Las pacientes con pocos embarazos, grosor del pliegue cutáneo alto e insulina muy alta tienen $1{,}54$ veces más probabilidades de presentar obesidad.

- **Regla 5:** {Glucosa=Normal, DiabetesPedigreeFunction=Alta, Edad=Joven} $\rightarrow$ {Embarazos=Pocos}
  - **Soporte:** $12{,}8\,\%$
  - **Confianza:** $70{,}5\,\%$
  - **Lift:** $1{,}55$

  **Interpretación:** Las pacientes jóvenes con glucosa normal y una alta predisposición genética a la diabetes tienen $1{,}55$ veces más probabilidades de haber tenido pocos embarazos.

Estas reglas muestran relaciones significativas entre los niveles de glucosa, insulina, BMI y otras características clínicas, fortaleciendo la conexión entre estos factores y el diagnóstico de diabetes.

## 5.2. Comparación con laboratorios anteriores

Los resultados obtenidos se relacionan con los hallazgos del Laboratorio 1, donde se observaron correlaciones positivas entre glucosa y el diagnóstico de diabetes. Este análisis refuerza dichas observaciones al proporcionar reglas de asociación que vinculan directamente glucosa alta con un mayor riesgo de diabetes.

## 5.3. Validación con literatura

Las asociaciones encontradas en el análisis, como la regla `Glucosa=Alto` $\rightarrow$ `Outcome=1`, están en consonancia con estudios médicos previos que han identificado los ni-

veles elevados de glucosa como un factor clave en el diagnóstico de diabetes. Según la (Association, 2024), los niveles de glucosa superiores a 140 mg/dL en una prueba de tolerancia a la glucosa indican un estado de prediabetes, y valores superiores a 200 mg/dL son característicos de la diabetes tipo 2. Este patrón también se observa en la regla {Glucosa=Alto, BMI=Obesidad} $\rightarrow$ {Outcome=1}, donde la combinación de glucosa elevada y obesidad, dos factores de riesgo bien documentados, aumenta significativamente la probabilidad de desarrollar diabetes (Venkatesan et al., 2024). De hecho, la obesidad es un factor de riesgo crítico para la resistencia a la insulina, lo que explica la interrelación entre estas variables y el diagnóstico de diabetes.

La relación entre `Glucosa=Alto` y `Outcome=1` también está respaldada por investigaciones recientes que señalan que la hiperglucemia, o niveles elevados de glucosa en sangre, es uno de los principales indicadores de diabetes tipo 2. El estudio de (Venkatesan et al., 2024) destaca que los pacientes con glucosa alta en ayunas y en respuesta a pruebas de tolerancia a la glucosa tienen una mayor probabilidad de desarrollar diabetes en el futuro, especialmente cuando otros factores de riesgo, como la obesidad y la predisposición genética, están presentes.

Adicionalmente, la regla `Insulina=Muy alto` $\rightarrow$ `BMI=Obesidad` se valida con literatura que sugiere que los altos niveles de insulina en ayunas están estrechamente relacionados con la obesidad. La resistencia a la insulina, que es un factor central en el desarrollo de la diabetes tipo 2, se asocia con la acumulación de grasa visceral, especialmente en individuos con un índice de masa corporal (BMI) alto (Venkatesan et al., 2024). Este fenómeno es consistente con los hallazgos de la regla que relaciona insulina muy alta con obesidad.

Por otro lado, la regla `Embarazos=Pocos` $\rightarrow$ `BMI=Obesidad`, que asocia la variable número de embarazos con el BMI, es respaldada por estudios que muestran que las mujeres con pocos embarazos tienen un menor riesgo de desarrollar obesidad en comparación con aquellas que han tenido múltiples embarazos. Este patrón está documentado en la literatura de la (Association, 2024), que indica que el embarazo afecta de manera significativa las tasas de obesidad en la población femenina, debido a cambios hormonales y metabólicos que pueden persistir después del embarazo.

Además, las investigaciones sobre la `DiabetesPedigreeFunction` refuerzan la rele-

vancia de la predisposición genética a la diabetes, como se muestra en las reglas que incluyen esta variable. La *DiabetesPedigreeFunction* mide la carga genética relacionada con la diabetes, y estudios como el de (Aubaidan & Kadir, 2023) han demostrado que un valor más alto en esta variable está asociado con una mayor probabilidad de desarrollar diabetes. Esta función de pedigrí es un factor de riesgo independiente que interactúa con otros factores, como la edad y el BMI, para predecir el diagnóstico de diabetes.

En resumen, las reglas de asociación obtenidas en este análisis están sólidamente respaldadas por la literatura médica y científica actual. Estas reglas no solo validan los hallazgos del análisis, sino que también proporcionan una visión más clara y estructurada de cómo diversas variables clínicas interactúan entre sí para predecir el diagnóstico de diabetes. La literatura refuerza la importancia de factores como los niveles de glucosa, la obesidad, la insulina y la predisposición genética en el diagnóstico de esta enfermedad.

## 5.4. Análisis crítico

Las reglas generadas en este análisis proporcionan información clínica relevante que puede ser útil para identificar pacientes en riesgo de desarrollar diabetes. Al combinar múltiples variables clave como glucosa, insulina, BMI y otros factores relacionados, las reglas de asociación permiten detectar patrones y relaciones entre estas características que de otro modo podrían pasar desapercibidas. Por ejemplo, las reglas que vinculan altos niveles de glucosa y obesidad con un mayor riesgo de diabetes refuerzan lo que la literatura médica ya ha documentado sobre la relación entre estos factores. Estos hallazgos podrían ser útiles en la práctica clínica, ayudando a identificar pacientes que, aunque no muestren síntomas claros de diabetes, presentan factores de riesgo que los colocan en una categoría de mayor vulnerabilidad. En este sentido, las reglas generadas podrían contribuir a la toma de decisiones más informadas en cuanto a prevención y tratamiento de la enfermedad.

Sin embargo, este análisis también presenta varias limitaciones que deben ser consideradas al interpretar los resultados. Una de las principales limitaciones es la discretización de las variables continuas, como glucosa e insulina. Aunque la discretización facilita el análisis al convertir variables continuas en categorías más manejables, también puede llevar a la pérdida de información detallada. En el caso de glucosa e insulina, los rangos de

valores podrían no capturar toda la variabilidad dentro de los datos, lo que limita la capacidad del modelo para identificar relaciones más sutiles entre los niveles de estas variables y el diagnóstico de diabetes. Por ejemplo, un paciente con glucosa de 139 mg/dL podría ser clasificado dentro del mismo grupo que otro con 140 mg/dL, aunque la diferencia entre estos valores podría ser clínicamente significativa en ciertos contextos. Este tipo de aproximación puede reducir la precisión del modelo, especialmente cuando se trabaja con rangos amplios de variables continuas.

Además, en términos de mejoras futuras, sería beneficioso explorar técnicas más avanzadas para el tratamiento de los valores faltantes, como los métodos de imputación basados en modelos más complejos (por ejemplo, imputación múltiple o métodos de aprendizaje automático). Estos enfoques podrían ofrecer una forma más precisa de manejar los datos faltantes, especialmente cuando hay patrones subyacentes que podrían ayudar a predecir estos valores. También se podría considerar la implementación de algoritmos de minería de datos alternativos, como FP-Growth o Eclat, que pueden ser más eficientes en el manejo de grandes conjuntos de datos y en la detección de asociaciones menos obvias. Estos algoritmos podrían permitir una comparación más completa de las reglas generadas, lo que ayudaría a evaluar la robustez y generalidad de los patrones identificados.

En resumen, aunque el análisis de reglas de asociación ha proporcionado información útil para comprender las relaciones entre variables clínicas y el diagnóstico de diabetes, es importante reconocer las limitaciones del enfoque utilizado. La discretización de variables y el tratamiento de valores faltantes son aspectos críticos que podrían mejorarse para obtener resultados más precisos y detallados. Sin embargo, este trabajo representa un paso importante hacia una comprensión más profunda de los factores que contribuyen al desarrollo de diabetes y puede servir como base para investigaciones futuras en este campo.

# 6. Conclusiones

En este laboratorio se analizaron datos clínicos de pacientes con el objetivo de descubrir patrones relevantes relacionados con el diagnóstico de diabetes, utilizando reglas de asociación y el algoritmo Apriori. Uno de los principales hallazgos fue la identificación de asociaciones significativas entre variables clínicas, como los niveles de glucosa, el índice de masa corporal (BMI) y el diagnóstico de diabetes. Por ejemplo, se destacó la regla $\{\text{Glucosa=Alto, BMI=Obesidad}\} \rightarrow \{\text{Outcome=1}\}$, que presentó un **lift** de $2{,}5$, indicando que las pacientes con glucosa alta y obesidad tienen $2{,}5$ veces más probabilidades de ser diagnosticadas con diabetes en comparación con lo esperado por azar. Estas reglas permiten comprender mejor cómo las variables clínicas interactúan y aportan información relevante para el diagnóstico.

El uso de métricas como el **soporte**, la **confianza** y el **lift** demostró ser efectivo para evaluar la relevancia y confiabilidad de las reglas generadas. Se observó que muchas de las reglas identificadas superaron niveles de confianza del $80\,\%$, lo que refuerza su utilidad práctica en el análisis de datos clínicos. Además, los hallazgos obtenidos complementan los resultados del Laboratorio 1, donde se habían identificado correlaciones positivas entre variables como glucosa y BMI con el diagnóstico de diabetes. En este laboratorio, las reglas generadas no solo confirmaron estas relaciones, sino que también ofrecieron una perspectiva más detallada y estructurada de los patrones presentes en los datos.

A pesar de los resultados positivos, se identificaron ciertas limitaciones que deben abordarse en futuros análisis. El tamaño reducido del conjunto de datos representa una restricción importante, ya que un dataset más grande permitiría una mejor generalización de las reglas obtenidas y podría facilitar la identificación de asociaciones menos frecuentes pero igualmente relevantes. También sería valioso incorporar métricas adicionales, como **leverage** y **convicción**, que podrían complementar el análisis y proporcionar perspectivas adicionales sobre la utilidad y significancia de las reglas. Por otro lado, mejorar el preprocesamiento de los datos, utilizando métodos más avanzados para la imputación de valores faltantes y criterios clínicos más específicos para la discretización, podría incrementar la calidad de los resultados.

Finalmente, la exploración de algoritmos alternativos, como FP-Growth, podría representar una oportunidad para evaluar la eficiencia y capacidad de descubrir asociaciones en comparación con Apriori. En conclusión, este laboratorio logró identificar patrones relevantes e interpretables en los datos de diabetes, proporcionando un marco útil para el análisis exploratorio y con potencial aplicación en el ámbito clínico. Sin embargo, mejoras en el diseño del análisis, el tamaño del dataset y la metodología utilizada podrían potenciar aún más los resultados y ampliar la aplicabilidad de este enfoque.

# Referencias


Association, A. D. (2024). *Standards of Medical Care in Diabetes—2024* [Accedido: Octubre 2024]. https://diabetesjournals.org/care/article/47/1/11/36448/Standards-of-Medical-Care-in-Diabetes2024

Aubaidan, B. H., & Kadir, R. T. (2023). Improved diabetes prediction using hybrid models. *Journal of Medical Informatics*, *39*(3), 455-462. https://doi.org/10.1016/j.jmir.2023.06.014

GPNotebook. (2024). Assessment of skinfold thickness for obesity and diabetes risk [Accedido: Octubre 2024]. https://gpnotebook.com/pages/gastroenterology/skin-fold-thickness

Kaggle. (2016). Pima Indians diabetes database [Consultado el 18 de octubre, 2024].

Organization, W. H. (2021). *Global brief on hypertension: Silent killer, global public health crisis* [Accedido: Octubre 2024]. https://www.who.int/news-room/fact-sheets/detail/hypertension

Venkatesan, U., Amutha, A., Jones, A. G., Shields, B. M., Anjana, R. M., Unnikrishnan, R., Mappillairaju, B., & Mohan, V. (2024). Performance of European prediction models for classification of type 1 and type 2 diabetes in Indians. *Diabetes Metabolic Syndrome: Clinical Research Reviews*, *18*(4), 103007. https://doi.org/https://doi.org/10.1016/j.dsx.2024.103007

# 7. Anexo