

**Análisis de métodos de razonamiento e incertidumbre** (Grupo. 301)

Dependencia entre factores de riesgo de la Diabetes descrita mediante Redes Bayesianas

**Docente:**

Daniel Otero Fadul

**Alumnos:**

Ricardo Kaleb Flores Alfonso | A01198716

Iván Alejandro López Valenzuela | A01284875

David Alejandro Acuña Orozco | A00571187

Raúl Correa Ocañas | A01722401

Carol Jatziry Rendon Guerrero | A01425341

# **1. Problematización**

La diabetes es una enfermedad crónica caracterizada por mantener un nivel alto de azúcar en sangre. En México, se ha convertido en la tercera causa principal de muerte, afectando a más de 12 millones de personas y cobrando la vida de más de cien mil personas al año. (Secretaria de Salud, 2021)

A pesar de esto, la diabetes a menudo pasa desapercibida en sus etapas iniciales, sobre todo en el caso de la diabetes tipo 2, donde los síntomas pueden ser imperceptibles durante años (WHO, 2024). Esta dificultad para dar un diagnóstico certero y muchas veces acertado provoca que una alta cantidad de personas no tengan un tratamiento adecuado hasta que la enfermedad ha avanzado significativamente, lo que incrementa la probabilidad de sufrir complicaciones en órganos vitales.

# **2. Enfoque**

De manera que los médicos cuenten con un sistema de apoyo para identificar pacientes con alta probabilidad de padecer diabetes, se propone implementar una red bayesiana. Se plantea que en esta se identifique dicha probabilidad a través de los valores en cierto rango de los siguientes factores de riesgo en el padecimiento de la diabetes: cantidad de embarazos, glucosa en sangre, presión arterial, que tan ancha se tiene la piel, resistencia a la insulina, índice de masa corporal, la función de probabilidad de tener diabetes basado en la edad y el historial familiar, edad y el sobrepeso.

Este sistema no reemplaza a un médico, sino que permite hacer una evaluación rápida de la evidencia recolectada acerca de un paciente de manera que, a través de la información del médico y el sistema, se tome una decisión en conjunto tomando en cuenta la mayor cantidad de información posible sobre estos factores, sus efectos y sus relaciones.

**3. Propósito**

El objetivo de este trabajo consiste en construir una red Bayesiana que, mediante su estructura, infiera las probabilidades de que distintos factores de riesgo relacionados con la diabetes e incluso el mismo diagnóstico de este padecimiento estén presentes y bajo qué niveles dados hechos ya determinados sobre estos, permitiendo así entender los efectos que el grado o condición de cada uno puede provocar sobre el resto de los factores con los que guardan dependencia.

**4. Información**

La información utilizada para ser analizada mediante esta estructura proviene del conjunto de datos “diabetes-dataset” que se encuentra disponible en la plataforma Kaggle, el cual contiene los registros recabados de 2000 pacientes en cuanto a mediciones de las distintas variables (factores de riesgo enlistados en la sección de Enfoque) y finalmente el diagnóstico definitivo de la enfermedad. De todos los pacientes, se determinó que 684 padecen diabetes mientras que este no es el caso para 1316 de ellos. Se puede acceder al conjunto de datos original en [este enlace](https://www.kaggle.com/datasets/vikasukani/diabetes-data-set).

Además, con la finalidad de mejorar la manejabilidad y comprensión de la información junto a las distribuciones de probabilidad en la implementación de la red Bayesiana, la experimentación y el análisis de resultados, todas las columnas (a excepción de Outcome y Overweight que son variables binarias) debieron ser discretizadas de manera que el valor original fuera sustituido por el valor del rango intercuartílico en el que se encuentra, ya fuera de 0 a 3, a manera de que la columna se convirtiera en una de tipo categórica y así concluir con el proceso de preparación de los datos.

De igual forma, parte de la información relevante para llevar a cabo el propósito es el recurso de visualización previa mediante un grafo que muestra la estructura a seguir por la red Bayesiana que se implementa, con una descripción clara de las conexiones entre nodos (cada variable o columna) y su relaciones del tipo padre-hijo que al denotar dependencia entre sí, más adelante influyen en cómo se comportan las distribuciones de probabilidad.

# **5. Razonamiento**

Siguiendo con el proceso de codificación y profundizando en la preparación de los datos, este es relevante dado que para el desarrollo de un modelo probabilístico basado en redes bayesianas se requiere que las variables sean discretas. Por lo tanto, se implementó un proceso de discretización con una función sobre el conjunto de datos para convertir las variables continuas en categorías, esto por los principios del modelado probabilístico discreto. La función de discretización llamada “discretize” categoriza de la siguiente forma

* 0: Valores menores al primer cuartil (25%).
* 1: Valores entre el primer y el segundo cuartil (25%-50%).
* 2: Valores entre el segundo y el tercer cuartil (50%-75%).
* 3: Valores mayores o iguales al tercer cuartil (75%-100%).

Por lo tanto, en esta preparación se simplifican las variables, lo que es útil para identificar patrones y relaciones en los datos. Por ejemplo, para la variable “Sobrepeso” (Overweight), que fue creada con base en el índice de masa corporal (BMI), la discretización permitió observar que 1694 personas están clasificadas como con sobrepeso (1) y 306 no lo están (0), lo que puede significar que esta variable podría ser un predictor importante para otras condiciones, como la presencia de diabetes.

Después, en la construcción de la Red de Creencia Bayesiana (BBN) su construcción parte del cálculo de probabilidades, por lo que se crea una función para ello donde cada nodo representa una variable del sistema. Si un nodo no tiene relaciones parentales, se calculan sus probabilidades marginales directamente a partir de las frecuencias relativas en los datos. En caso de tener nodos padres, se obtienen probabilidades condicionales considerando todas las combinaciones posibles de valores de dichos padres, permitiendo capturar relaciones dependientes y complejas.

En cuanto a la red, esta se define mediante dos diccionarios clave: el primero especifica las relaciones entre nodos y el segundo los posibles estados discretizados de cada nodo. Las conexiones que varían como SkinThickness y BMI se ven afectadas con Overweight, mientras que este último influye en Outcome y BloodPressure. El nodo Outcome está influenciado por variables como Age, Diabetes, PedigreeFunction, Overweight y Pregnancies. Además, BloodPressure depende tanto de Overweight como de Outcome, mostrando cómo la red captura las interacciones entre factores biológicos y de estilo de vida, esto puede decir que un paciente con altos niveles de glucosa y un elevado índice de masa corporal tiene una probabilidad mayor de un diagnóstico positivo en Outcome.

Pasando a hablar de los métodos necesarios para hacer inferencias dados hechos determinados acerca de las variables, estos se logran concretar mediante una función de inferencia probabilística para un estado de un nodo específico de la red (según el de los parámetros que se ingrese), para que posteriormente con un ciclo que muestra el reajuste de probabilidades para cada categoría de cada factor de riesgo luego de que la función evidence() actua.

A manera de ejemplo, en el primer experimento, se incluyó la evidencia de que una persona presenta Overweight = 1 (sobrepeso). Al introducir esta información, se observó un cambio en la distribución de probabilidades, aumentó la probabilidad de que el Outcome (Resultado de diabetes) fuera 1 (diabetes)

De esta manera, resulta posible realizar más experimentos de este tipo que no estén limitados sólo a una evidencia debido a que, como en el ejemplo anterior, las redistribuciones de probabilidades pueden llevarse a cabo múltiples veces siguiendo la secuencia de ingreso de las evidencias, en los que finalmente los resultados puedan brindar indicios de las ya mencionadas relaciones y consecuencias que la presencia conocida de los factores de riesgo tengan sobre otros y principalmente, sobre el diagnóstico de la diabetes. Es justo mediante ir avanzando en estos experimentos que se pudieron ir identificando patrones y variables influyentes que sirvieron para conocer las combinaciones de evidencia más favorables para que el sistema tome mayor valor en el cumplimiento del objetivo.

# **6. Conclusiones**

En este trabajo, se logró implementar una red Bayesiana que modela las relaciones probabilísticas entre factores de riesgo de la diabetes. Después de repetidos experimentos, se identificaron relaciones significativas entre las variables. Según el peso de las relaciones, estas muestran cómo ciertos factores tienen un impacto más pronunciado en la probabilidad de diagnóstico. Un ejemplo concreto sugiere que las variables de sobrepeso y el IMC tienen una alta correlación con la probabilidad de que se tenga un diagnóstico de diabetes. Estos se vuelven aún más pronunciados al considerarse junto a niveles elevados de la variable de glucosa. También se encontró que el historial de familiares que padecen de diabetes no parece influir severamente en la probabilidad de diagnóstico de este mismo. Claro, también se encuentran relaciones triviales y esperadas en el funcionamiento del modelo, como lo es la relación entre la variable de resistencia a la insulina y el diagnóstico de diabetes.

Es importante mencionar que el tener estados discretos para cada variable simplifica la forma de cómo se interpretan las probabilidades dependientes. Por lo tanto, es aparente que las redes bayesianas son útiles para proporcionar una visión interactiva y completa de la interdependencia de factores que contribuyen al desarrollo de la diabetes. Modelos como estos no solo mejoran la comprensión de las relaciones entre factores de riesgo, sino también ofrecen una base para el análisis para un enfoque médico. Entre posibles mejoras y caminos por explorar, se podrían probar incluso más experimentos o probar con diferentes configuraciones de estados discretos.

**Referencias**

Secretaria de Salud. (2021, 5 octubre). *Diabetes en México.* Gobierno de México. Recuperado 21 de noviembre de 2024, de https://www.gob.mx/promosalud/acciones-y-programas/diabetes-en-mexico-284509

World Health Organization: WHO & World Health Organization: WHO. (2024, 14 noviembre). *Diabetes*. Recuperado de: https://www.who.int/es/news-room/fact-sheets/detail/diabetes

**A. Anexos**

Acceso al código ejecutado en [este enlace](https://drive.google.com/file/d/1OtrhPVZN_Z5dXcIrVod3F4-jMN9vWWW5/view?usp=sharing)

Acceso a la tabla de resultados de experimentos en [este enlace](https://docs.google.com/spreadsheets/d/1SjijYbw8VCqn3oT_3TLdDozPgfUnHgtn/edit?usp=sharing&ouid=101127852425430275285&rtpof=true&sd=true)