Instituto Tecnológico de Estudios Superiores de Monterrey

Evidencia 2. Artículo de investigación PBL2

Análisis de métodos de razonamiento e incertidumbre (Gpo 201)

Docente: Dr. Daniel Otero Fadul

Diego Armando Mijares Ledezma **- A01722421**

Pedro Soto Juárez - **A00837560**

Alexei Carrillo Acosta **- A01285424**

Marcos Renato Aquino Garcia - **A00835576**

Mauricio Octavio Valencia Gonzalez - **A01234397**

8 de octubre de 2024

Problematización

La diabetes es una de las enfermedades que a más personas afecta en el mundo. Según la OPS (2024) “Se estima que 62 millones de personas en las Américas viven con Diabetes Mellitus (DM) tipo 2.” Más preocupante aún es que se estima que entre el 30-40% de las personas con diabetes están sin diagnosticar, es por esto que es importante desarrollar nuevas formas de diagnosticar esta enfermedad, lo más rápido y simple posible (Organización Panamericana de la Salud, 2024).

El doctor Caro (2020) señala que la diabetes puede pasar desapercibida en las etapas iniciales de la enfermedad, pero algunas personas pueden pasar incluso hasta 10 años sin ser diagnosticadas: “...puede repercutir en su estado de salud, ya que los elevados niveles de glucosa en sangre pueden dar lugar a una serie de complicaciones importantes en diferentes órganos,” afirma el doctor Caro. Es importante señalar que al menos en el caso de la diabetes tipo II, es tratada y controlada principalmente con medidas higiénico-dietéticas, por lo que se debe diagnosticar la enfermedad en cada caso, lo más pronto posible, para que la calidad de vida del paciente no sufra a raíz de la enfermedad (Ministerio de Salud Pública y Bienestar Social & Fariña, 2020).

Enfoque

Se busca poder diagnosticar la diabetes a través del uso de una red bayesiana. La red bayesiana que se usará, está basada en la red mostrada en el artículo “Using Bayesian Network for the Prediction and Diagnosis of Diabetes” (Mohammadi et al., 2015). El objetivo de la red será calcular la probabilidad de que una persona tenga diabetes dadas algunas de las variables de nuestro dataset. Lo que se hace es que se mapean todas las variables de acuerdo a las relaciones que tienen entre ellas dentro de la red, así como los valores que cada variable puede tener. Con esta información se obtiene la probabilidad de cada nodo en el grafo, que a su vez representan a cada una de las variables. Con todo esto, se pueden hacer las inferencias necesarias para determinar qué tan probable es que una persona tenga diabetes.

Propósito

El propósito de esta actividad es desarrollar un modelo de red bayesiana para diagnosticar diabetes utilizando un conjunto de datos clínicos. Este modelo no solo busca predecir la presencia de diabetes, sino también entender las relaciones probabilísticas entre diferentes factores de riesgo y síntomas asociados con la enfermedad.

La actividad está diseñada para guiar el proceso de construcción, implementación y uso de una red bayesiana. Esto implica trabajar con datos reales, procesarlos, discretizarlos y utilizarlos para crear un modelo probabilístico complejo.

Se espera para el final de la actividad, una comprensión profunda de la estructura y funcionamiento de las redes bayesianas. Desde saber representar relaciones causales y probabilísticas entre variables, hasta practicar la comprensión de la topología de una red bayesiana basada en factores relacionados con la diabetes. Se va llevar a cabo la manipulación de datos con pandas, y se practicará la discretización de variables continuas; paso crucial para muchos modelos probabilísticos. Se espera conllevar el cálculo de probabilidades condicionales a partir de estos datos, y su interpretación. También se practicará inferencia probabilística y visualización de datos y modelos con *networkx* y *matplotlib.* Finalmente, se busca comprender cómo diferentes factores (edad, IMC, historial familiar, etc.) se relacionan probabilísticamente con el diagnóstico de diabetes.

Información

El código elaborado construye una red bayesiana para predecir el riesgo de diabetes. El primer paso es cargar un conjunto de datos y crear una nueva columna llamada *Overweight* para identificar si una persona tiene sobrepeso, lo cual se determina a partir del *BMI* (Body Mass Index). Si el *BMI* es mayor a 25, se clasifica como sobrepeso (valor 1), si no, se clasifica como no sobrepeso (valor 0). Se debe considerar que se discretizan todas las variables del dataset (excepto *Outcome* y *Overweight*) en función de sus cuartiles, lo que facilita el análisis probabilístico en la red.

La estructura de la red bayesiana se define mediante un diccionario llamado *graph*, que muestra las relaciones de dependencia entre las variables. Otro diccionario, *values*, almacena los

valores discretos que puede tomar cada variable. Esta estructura se utiliza para realizar predicciones sobre el riesgo de diabetes basándose en los datos ingresados.

La función *probabilities(df, node)* es clave para calcular las probabilidades de cada nodo. Si el nodo es independiente, es decir, no depende de otros nodos, se calculan las probabilidades marginales a partir de las frecuencias observadas en los datos. Si el nodo depende de otros nodos (es decir, tiene padres), se calculan las probabilidades condicionales considerando todas las combinaciones posibles de los valores de los nodos padres.

La función *tables(df)* organiza las probabilidades en un diccionario, donde cada nodo tiene asociada una tabla de probabilidades. Este diccionario facilita la construcción de la red, ya que permite acceder rápidamente a las probabilidades de cada nodo.

Los nodos de la red se crean utilizando la clase *bbnNode*, donde cada nodo representa una variable médica, como *overweight*, *age* o *glucose*. A cada nodo se le asignan tanto sus posibles valores como las probabilidades previamente calculadas. La estructura de la red se define con el objeto *bbn*, que agrega los nodos y las aristas que conectan las variables, indicando relaciones de dependencia causal entre ellas.

Para realizar inferencias probabilísticas, la red bayesiana se convierte en un árbol de unión mediante la función *InferenceController.apply(bbn)*. Esto permite calcular las probabilidades de ciertos nodos y actualizar estas probabilidades cuando se incorpora nueva información. Por otro lado, también permite visualizar la red bayesiana utilizando *networkx*, lo que muestra las conexiones entre variables como el sobrepeso, la glucosa y la presión arterial. Los nodos representan variables y las aristas describen dependencias causales entre ellas.

Por último, la función *print_probs()* muestra las probabilidades marginales actuales de cada nodo en la red, permitiendo observar las distribuciones de probabilidad antes y después de introducir evidencia. Para esto, se usa la función *evidence()*, que permite fijar valores específicos para ciertos nodos, como indicar si una persona tiene sobrepeso o un nivel específico de glucosa, y luego actualizar las probabilidades de los otros nodos en la red según esta información.

Resultados

Los valores marginales indican la frecuencia con la que cada categoría de un atributo particular aparece en los datos. Un ejemplo es que el 84.7% de los individuos son clasificados como con sobrepeso, lo que sugiere que la mayoría de personas en el set de datos tienen este factor de riesgo. Estas probabilidades ayudan a entender la distribución de los datos y la relevancia de cada uno en la predicción del diabetes.

Valor	0	1	2	3
Overweight	0.153	0.847	0	0
SkinThickness	0.2865	0.2245	0.2405	0.2485
BMI	0.25	0.2535	0.2555	0.241
Outcome	0.60497	0.39503	0	0
Insulin	0.48036	0.02313	0.24496	0.25155
Glucose	0.24411	0.23812	0.25016	0.2676
BloodPressure	0.24612	0.29071	0.24617	0.217
Diabetes Pedigree Function	0.251	0.2505	0.2495	0.249
Age	0.2915	0.2335	0.2295	0.2455
Pregnancies	0.3285	0.2395	0.2315	0.2005

Tabla 1.0: Distribución de las probabilidades marginales en distintos nodos

Se puede observar como la mayoría de los nodos se distribuyen equitativamente entre los cuartiles con algunas excepciones como la insulina, el outcome y el sobrepeso. Una vez con estas probabilidades marginales, se realizaron evidencias los cuales son casos de personas ficticias con ciertos atributos de los nodos para ver sus probabilidades de diagnóstico de cada nodo.

En el primer caso, únicamente alimentamos la evidencia de que la persona tiene sobrepeso, y su tabla de probabilidades queda de la siguiente manera:

Valor	0	1	2	3
Overweight	0	1	0	0
SkinThickness	0.25266	0.18713	0.27273	0.28749
BMI	0.11452	0.29929	0.30165	0.28453
Outcome	0.57119	0.42881	0	0
Insulin	0.48187	0.02257	0.23920	0.25636
Glucose	0.23399	0.23438	0.25154	0.28009
BloodPressure	0.21251	0.29648	0.25122	0.23979
Diabetes Pedigree Function	0.251	0.2505	0.2495	0.249
Age	0.2915	0.2335	0.2295	0.2455
Pregnancies	0.3285	0.2395	0.2315	0.2005

Tabla 1.1: Distribución de las probabilidades del caso 1

Se puede observar como la probabilidad de que la persona tenga sobrepeso es de 1 ya que esta fue la evidencia que se decidió probar, una vez con esto, al compararlo con la tabla 1.0 de valores marginales, se puede observar que las probabilidades de los nodos de Diabetes Pedigree Function, Age y Pregnancies fueron las únicas que mantuvieron sus probabilidades, lo que indica que estas no dependen del sobrepeso. Podemos observar que la probabilidad de tener diabetes aumentó a 0.42, esto nos indica que el tener sobrepeso aumenta tus probabilidades de tener diabetes.

En el caso 2, se decidió hacer el caso de un atleta de alto rendimiento, el cual tiene bajo BMI, baja presión de sangre, baja insulina y alto grosor de piel. Su tabla de probabilidades quedaría de la siguiente manera:

Valor	0	1	2	3

Overweight	1	0	0	0
SkinThickness	0	0	0	1
BMI	1	0	0	0
Outcome	0.87805	0.12195	0	0
Insulin	1	0	0	0
Glucose	0.32593	0.26839	0.23902	0.16666
BloodPressure	1	0	0	0
Diabetes Pedigree Function	0.25146	0.25096	0.24996	0.24763
Age	0	1	0	0
Pregnancies	0.33119	0.24146	0.23157	0.19579

Tabla 1.2: Distribución de las probabilidades del caso 2

En este caso se puede observar la distribución de probabilidades del caso 2, podemos ver que al ser un atleta saludable, el *outcome* qué es la probabilidad de tener o no tener diabetes, es de 0.87 a favor de que no tenga diabetes, lo cual es mucho mayor (y mejor) comparado con las probabilidades marginales. Al ser varias variables las que cambiamos, estas influyen en todas las demás características no cambiadas.

Para el caso 3, se usó una persona de media edad con prediabetes, es decir alta glucosa, BMI medio, y un poco presión alta. Esta es su tabla de probabilidades:

Valor	**0**	**1**	**2**	**3**
Overweight	0	1	0	0
SkinThickness	0	0	0	1
BMI	0	0	1	0
Outcome	0.62378	0.37622	0	0
Insulin	0	1	0	0

Glucose	0	0	1	0
BloodPressure	0	0	1	0
Diabetes Pedigree Function	0.28616	0.24422	0.23726	0.23236
Age	0	0	1	0
Pregnancies	0.28715	0.25455	0.25881	0.19949

Tabla 1.3: Distribución de las probabilidades del caso 3

Con estas probabilidades, y las características predefinidas mencionadas anteriormente, se puede observar como la probabilidad de tener diabetes disminuye ligeramente a comparación con las probabilidades marginales, lo que nos indica que una persona con estas características tiene un 37.6% de probabilidad de tener diabetes.

Dentro del código anexado en la entrega, se pueden observar más casos - estos incluyen los siguientes: Persona de edad avanzada con salud bien controlada pero con una alta función pedigrí de diabetes, mujer embarazada de veintitantos años con diabetes gestacional (IMC alto, insulina alta, glucosa alta), y adolescente con resistencia a la insulina (IMC alto, insulina alta, glucosa normal, edad temprana). Favor de acudir al código en caso de desear explorar la distribución de probabilidades en la red bayesiana con las evidencias respectivas a cada uno de estos casos adicionales.

Conclusiones

A partir de los resultados obtenidos en las simulaciones con individuos de diferentes características, es evidente que ciertos factores tienen una mayor influencia en la probabilidad de que una persona desarrolle diabetes. Entre los más importantes destacan el índice de masa corporal (BMI) junto al sobrepeso, los niveles de glucosa y el antecedente familiar (Diabetes Pedigree Function).

En primer lugar, el sobrepeso (y alto BMI) tiene un impacto interesante. Al simular un individuo con sobrepeso y dejar el resto de las variables constantes, la probabilidad de que este

desarrolle diabetes aumenta un 3.3%, pasando de una probabilidad base de 39.5% a 42.88%. Este resultado sugiere que el sobrepeso es uno de los factores más influyentes por sí solo.

Por otro lado, también se observó que un individuo joven y saludable (con bajo BMI, baja insulina y baja presión arterial) tiene una probabilidad de sólo 12% de desarrollar diabetes. Este resultado es consistente con la literatura médica, que resalta la importancia de un estilo de vida saludable en la prevención de enfermedades metabólicas.

También, la glucosa alta juega un papel crucial. Un individuo de mediana edad con glucosa elevada, sobrepeso y presión arterial ligeramente alta tiene una probabilidad de diabetes apenas un 5.2% mayor que alguien que solo tiene sobrepeso. Esto sugiere que aunque estos factores adicionales aumentan el riesgo, el nivel de glucosa es uno de los indicadores clave por encima de otros factores.

Lo que resulta aún más alarmante es el impacto de la genética. Un adolescente con tres familiares diabéticos y sobrepeso tiene la mayor probabilidad de desarrollar diabetes en todas las pruebas, lo que refuerza la idea de que la combinación entre genética y sobrepeso es un fuerte predictor de la enfermedad. Estos resultados coinciden con la evidencia científica actual, que subraya la importancia de los antecedentes familiares y el estilo de vida en el riesgo de diabetes.

Finalmente, variables como la presión arterial y los niveles de insulina parecen tener un impacto menos directo, pero igualmente contribuyen cuando se combinan con otros factores de riesgo. Por ejemplo, un individuo con niveles normales de insulina y presión arterial, pero con glucosa alta, sigue presentando una alta probabilidad de desarrollar la enfermedad.

En conclusión, los resultados sugieren que los factores más determinantes para el desarrollo de la diabetes son el nivel de glucosa, el BMI junto al sobrepeso, y la función de antecedentes familiares (Diabetes Pedigree). Otros factores, como la edad, la insulina y la presión arterial, tienen influencia, pero su impacto es más pronunciado cuando se presentan en combinación con estos factores principales. El sobrepeso y la genética emergen como los principales indicadores del riesgo, por lo que la gestión del peso y la atención a los antecedentes familiares deben ser prioritarios en estrategias de prevención, según las inferencias hechas en este reporte con los resultados del código elaborado.

Referencias

Caro, J. (2023, May 12). *¿Es asintomática la diabetes tipo 2? - Vithas*. Vithas. https://vithas.es/consejo/diagnostico-tardio-de-la-diabetes-tipo-2-por-ser-asintomatica/

Ministerio de Salud Pública y Bienestar Social, & Fariña, F. (2020, November 14). *La diabetes puede cursar de forma asintomática*. Ministerio De Salud Pública Y Bienestar Social - Paraguay. https://www.mspbs.gov.py/portal/22134/la-diabetes-puede-cursar-de-forma-asintomatica.html

Mohammadi, M., Hosseini, M., & Tabatabaee, H. (2015). Using Bayesian Network for the Prediction and Diagnosis of Diabetes. *Bulletin of Environment, Pharmacology and Life Sciences*. http://www.bepls.com

Organización Panamericana de la Salud. (2024, September 3). *Diabetes*. OPS/OMS | Organización Panamericana De La Salud. https://www.paho.org/es/temas/diabetes