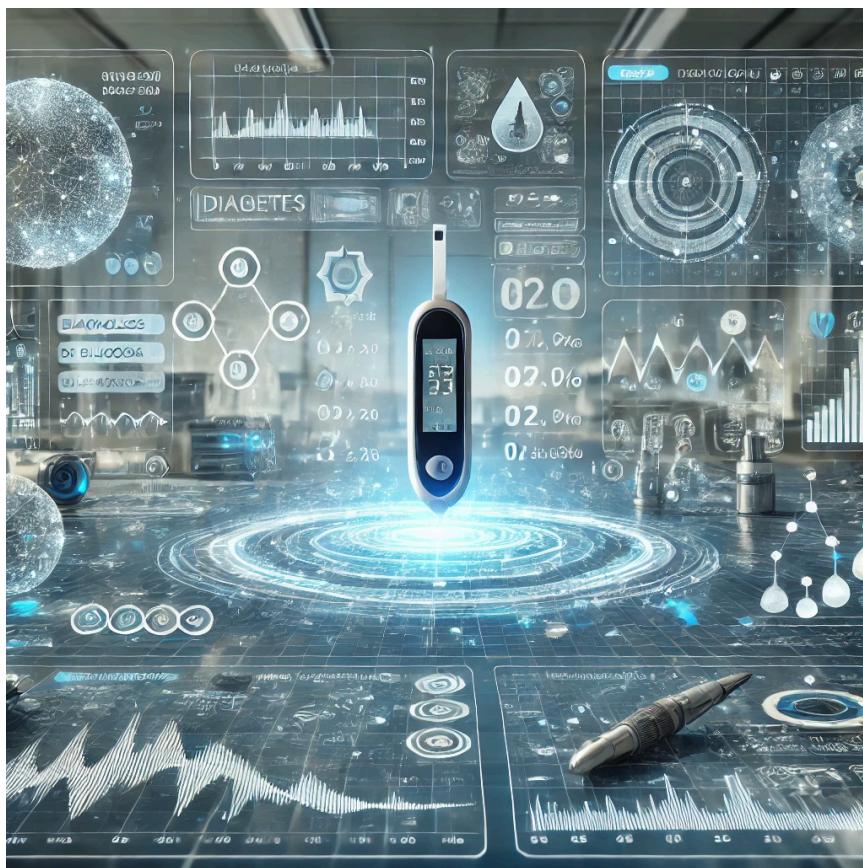

Sistema de Diagnóstico de la Diabetes Tipo 2 con Manejo de Incertidumbre

Razonamiento con Incertidumbre, 3º GRIA



Valeria Lagos Molins, 53819810V, vlagosmolins@gmail.com

Sara Porto Álvarez, 54153770V, saraportoalvarez@gmail.com

Índice

Resumen del problema.....	2
1. Introducción.....	3
1.1. Alcance del sistema propuesto.....	4
2. Planificación y seguimiento.....	5
3. Descripción de la arquitectura del sistema.....	10
3.1. Tecnologías utilizadas.....	10
3.2. Parámetros utilizados.....	11
3.3. Redes Bayesianas.....	13
3.4. Diagrama de Clases.....	17
3.5. Diagrama de Flujo.....	20
4. Pruebas realizadas.....	22
5. Manual de Uso.....	26
5.1. Instalación del programa.....	26
5.2. Uso del programa.....	26
6. Conclusiones finales.....	27
6.1. Problemas encontrados.....	27
6.2. Posibles mejoras.....	28
6.3. Conclusiones.....	28
7. Bibliografía consultada.....	29

Resumen del problema

Se ha creado un sistema de diagnóstico de la diabetes tipo 2 con manejo de incertidumbre con Redes Bayesianas. Se trata de un sistema pensado para pacientes que sospechen o tengan la curiosidad de saber si tienen dicha enfermedad.

El sistema presenta un cuestionario preliminar al paciente, preguntando por síntomas y datos que puede saber de antemano sin necesidad de hacer pruebas médicas. Estas serían preguntas sobre su edad, peso, frecuencia con la que siente fatiga o hambre, entre otras. En base a las respuestas al cuestionario, se da a conocer la probabilidad de tener diabetes gracias a la inferencia con las ya mencionadas Redes Bayesianas.

En el caso de que el paciente presente una probabilidad significativa de tener diabetes en base a las respuestas del cuestionario, sería necesario recoger pruebas clínicas en exámenes médicos sobre el nivel de azúcar en sangre y la presión arterial; con el fin de aclarar las dudas sobre si el paciente tiene o no diabetes. Se añadirían las respuestas de estas mediciones clínicas y, conociendo todas las respuestas del paciente, se realiza una nueva inferencia y se muestra la probabilidad total de que el paciente tenga diabetes.

1. Introducción

La diabetes mellitus tipo 2 (DM2), o simplemente diabetes tipo 2 es el tipo más común de diabetes. Se trata de un trastorno metabólico que se caracteriza por provocar altos niveles de glucosa en sangre (hiperglucemia), causado por la resistencia a la insulina y una producción insuficiente de dicha hormona por parte del páncreas. La glucosa no consigue entrar correctamente en las células dada la falta de insulina, lo que resulta en niveles peligrosamente altos de azúcar en sangre.

A diferencia de la diabetes tipo 1, que implica deficiencia total de insulina y se detecta en edades tempranas, la DM2 generalmente se manifiesta en la edad adulta. Otro tipo de diabetes es la diabetes gestacional, cuyo diagnóstico se detecta durante el embarazo. El desarrollo de la diabetes tipo 2 se relaciona con factores como el estilo de vida o la edad. Según el *FID Atlas de la Diabetes* (2021), el 10,5% de la población adulta (20-79 años) vive con diabetes, y casi la mitad desconoce que la padece.

Y es que, en los últimos años, la diabetes (concretamente la diabetes tipo 2) se ha convertido en una de las principales causas de enfermedades crónicas a nivel global, y la falta de detección temprana puede llevar a graves complicaciones. Una detección a tiempo de la diabetes tipo 2 es crucial para reducir al máximo los riesgos asociados a la enfermedad y comenzar cuanto antes un tratamiento que mejore la calidad y esperanza de vida del paciente diabético.

Es por ello que este proyecto tiene como objetivo desarrollar un sistema de diagnóstico de la diabetes tipo 2 incorporando el manejo de la incertidumbre. El sistema se basa en una serie de parámetros y en la utilización de Redes Bayesianas para realizar las inferencias necesarias para el diagnóstico.

Las Redes Bayesianas son un modelo gráfico probabilístico que representa las relaciones de dependencia condicional entre nodos. Cada nodo de la red es una variable (los parámetros elegidos y la variable diabetes), y las relaciones se representan mediante arcos dirigidos entre los nodos. Cada nodo tendrá asociada una tabla de distribución de probabilidad que ayuda a modelar la incertidumbre. En nuestro caso, para el diagnóstico de la diabetes, las Redes Bayesianas integrarán los diversos síntomas y factores de riesgo para realizar inferencias en torno a la enfermedad.

1.1. Alcance del sistema propuesto

Dado el desafío que implica el diagnóstico de la diabetes, fue necesario acotar el alcance de nuestro proyecto, que se enfoca únicamente en el diagnóstico de la diabetes tipo 2. Esto es debido a que la diabetes tipo 2 difiere de manera significativa de los otros tipos de diabetes, la diabetes tipo 1 y la gestacional, tanto en síntomas como en factores de riesgo y momento de detección.

Escogimos enfocar el diagnóstico en la diabetes tipo 2 debido a que es el tipo más prevalente a nivel global, manejable y prevenible si se detecta a tiempo. Por ello es que vimos la utilidad y necesidad de fijarnos en la DM2, ayudando a conocer la probabilidad de tener diabetes de manera sencilla; potencialmente ayudando a reducir el impacto de la enfermedad.

Principalmente, el sistema se trata de un cuestionario informal pensado para que el paciente pueda contestar incluso desde su propia casa. A partir de las respuestas al cuestionario, da a conocer al paciente la probabilidad que tiene de padecer diabetes. Si esta probabilidad superase un umbral determinado, se procedería a la recogida de datos clínicos. Se volvería a hacer una inferencia en base a las Redes Bayesianas y las respuestas y resultados del paciente, dando a conocer la probabilidad final de que el paciente padezca diabetes.

Adicionalmente y presentando una opción más orientada hacia desarrolladores del sistema o médicos que lo utilicen, se permite introducir los datos correspondientes a las respuestas del cuestionario directamente a través de un CSV que analice múltiples casos simultáneamente. Para el doctor, es una opción que puede permitir llevar un registro histórico de los datos de un paciente en concreto, o realizar inferencias para más de una persona con un único clic. Para el desarrollador, es útil para hacer pruebas con el fin de corroborar información y hacer revisiones periódicas que permitan actualizar valores de probabilidades.

Nuestro sistema tiene por lo tanto dos modalidades de uso, la de diagnóstico de un paciente (con el cuestionario y posibles mediciones clínicas) y la de introducción de los datos mediante un fichero CSV para médicos y desarrolladores. Por otra parte, el sistema se estructura en dos fases principales.

La fase de autoevaluación del paciente es la fase inicial correspondiente al cuestionario básico. Se evalúan una serie de factores de riesgo y síntomas que el propio paciente puede observar y conocer. La fase clínica para la confirmación de los resultados implica la recogida de datos clínicos del paciente, siendo necesario acudir a un centro médico para proporcionar un resultado más preciso. Cada fase tendrá asociada por lo tanto una Red Bayesiana con la que hacer la inferencia del diagnóstico.

2. Planificación y seguimiento

Para el desarrollo del proyecto, en líneas generales hemos seguido una metodología en cascada. Delimitamos bien las fases: documentación y toma de requisitos, diseño, implementación, pruebas y defensa.

Las tareas se planificaron de acuerdo con las fechas de entrega y el calendario de las prácticas; por lo que se ha hecho una gestión por hitos. Organizamos las semanas de manera que contuviesen tareas enfocadas a los objetivos intermedios de finalizar cada una de las fases del enfoque *waterfall* que implementamos.

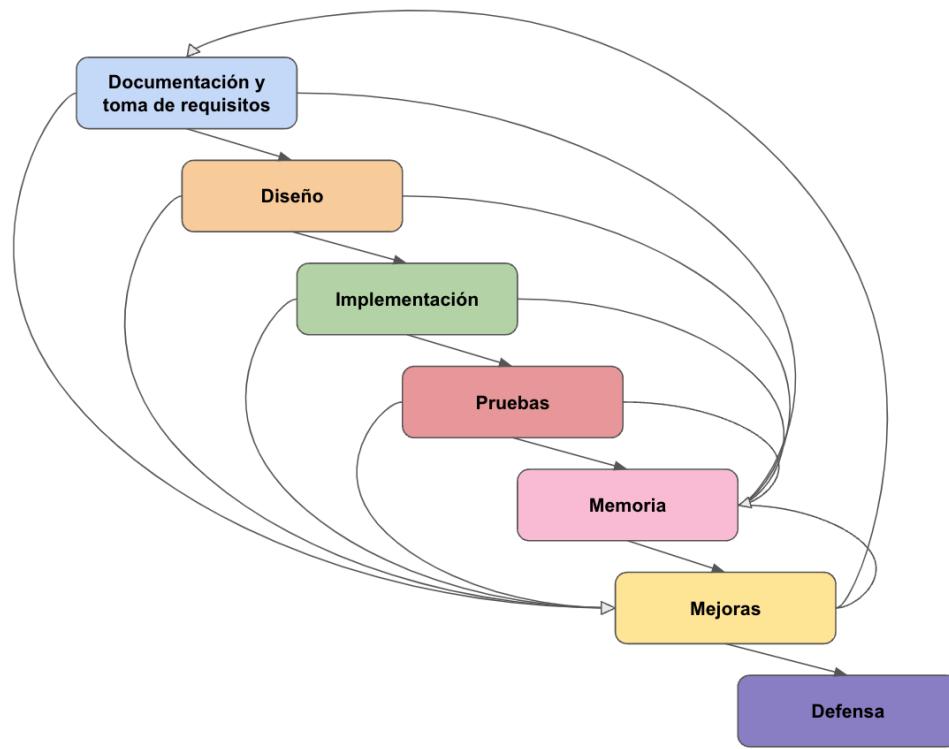
Adicionalmente, implementamos algunas pinceladas de metodología ágil. Las fases requerían una revisión y búsqueda de mejoras de las fases anteriores; así como la redacción de la memoria de manera paralela para realizar un seguimiento de cada una de las fases. Con ello, modificamos detalles de fases previas mientras desarrollábamos fases posteriores. Por ejemplo, modificando requisitos durante la implementación; o detalles de la implementación durante la fase de pruebas. Así que necesitamos definir dos fases adicionales, que son la memoria y las mejoras.

Las **fases**, por lo tanto, serían las siguientes:

- **Documentación inicial y toma de requisitos:** Recolectar requisitos, investigar tecnología, documentar conceptos. Dentro de esta fase también podemos incluir la planificación de tareas del proyecto.
- **Diseño:** Diseño del modelo de datos (parámetros y variables utilizadas), definición estructura de la red bayesiana, diseño de la arquitectura del programa, diseño de las relaciones entre parámetros y probabilidad de tener diabetes, con relaciones en CPDs.
- **Implementación:** Desarrollo completo del código según el diseño establecido.
- **Pruebas:** Realizar pruebas del funcionamiento del sistema, corrigiendo posibles errores en el código.
- **Memoria:** Desarrollo de la estructura de la memoria, manual de usuario y redacción de la memoria final. Esta fase se realizará en paralelo a todas las fases del proyecto, pero también tendrá su propia fase.
- **Mejoras:** Incluye tanto posibles imprevistos y retrasos en otras fases como las revisiones continuas del proyecto durante su desarrollo.
- **Defensa y entrega:** Preparación de la defensa y entrega final. También puede resumirse como despliegue final del proyecto.

Todo ello se ve reflejado en el siguiente gráfico sobre la metodología seguida. En él, las flechas rectas con relleno negro reflejan el flujo del enfoque en cascada; mientras que las flechas curvas con relleno blanco son las correspondientes a las pinceladas del enfoque ágil que acabamos de nombrar.

Metodología seguida:



Para la planificación consideramos el día de la práctica (en nuestro caso, coincidiendo en lunes) como el inicio de cada semana, y el día anterior a la siguiente práctica como el final de la semana. Así, comenzaremos con la semana 1 el día 23 de septiembre y acabaremos el día de la defensa (el 9 de diciembre).

Destacar asimismo que la semana 7 tendrá una duración de 2 semanas (del día 4 al día 18 de noviembre, ya que el día 11 no hay clase de prácticas). Además y como ya se mencionó, la memoria del proyecto se irá desarrollando en paralelo a todas las fases, y también se irán haciendo mejoras continuas cada semana.

Diagrama de Gantt:

Defensa												
Mejoras												
Memoria												
Pruebas												
Implementación												
Diseño												
Documentación y requisitos												
	Semana 1	Semana 2	Semana 3	Semana 4	Semana 5	Semana 6	Semana 7	Semana 8	Semana 9	Semana 10	Semana 11	

Utilizamos la siguiente tabla para realizar un seguimiento del desarrollo del proyecto, junto con las entregas a realizar. La última columna fue cambiando de rojo a verde por semanas, marcando las tareas que se iban terminando. Se estimaron también las horas previstas para cada una de las tareas

Tabla de seguimiento:

Semana	Tarea	Descripción de la tarea específica de cada semana	Tiempo estimado	¿Hecho?
1	Documentación inicial	Recolectar requisitos, Investigar tecnologías, Documentar conceptos.	2h	
2	Documentación inicial	Recolectar requisitos, Investigar tecnologías, Documentar conceptos	2h	
	Planificación	Planificar tareas del proyecto	1h	
	Diseño	Diseño del modelo de datos (parámetros y variables utilizadas), definición estructura de la red bayesiana, diseño de la arquitectura del programa	2 - 4h	
3	Entrega planificación: 7/10/24	Entrega y revisión de lo planificado	30 minutos	
	Diseño	Continuación de la semana anterior, con más detalle	3h	
4	Diseño	Diseño de las relaciones entre parámetros y probabilidad de tener diabetes, con relaciones en CPDs	3h	
5	Diseño	Continuación de la semana anterior, puliendo detalles	2h	
6	Entrega arquitectura: 28/10/24	Entrega y revisión de la arquitectura	30 minutos	
	Implementación	Desarrollo completo del código según el diseño establecido	5h	
7	Implementación	Continuación del desarrollo del código	6h	
	Pruebas	Realizar pruebas del	6h	

		funcionamiento del sistema, corrigiendo posibles errores en el código		
8	Entrega código parcial: 18/11/24	Entrega y revisión del código parcial	30 minutos	
	Pruebas	Realizar pruebas del funcionamiento del sistema, corrigiendo posibles errores en el código	5h	
	Estructura memoria	Desarrollo de la estructura de la memoria	30 minutos	
9	Entrega estructura de la memoria: 25/11/24	Entrega y revisión de la estructura de la memoria	15 minutos	
	Documentación de usuario	Desarrollo de un manual de uso	45 minutos	
	Memoria	Redacción de la memoria final del proyecto	4h	
10	Mejoras del proyecto	Posibles imprevistos y retrasos en otras fases	4h	
	Preparación de la defensa	Preparación de la defensa del proyecto	2,5h	
11-12	Entrega final: 18/12/24	Entrega y revisión final del proyecto	45 min	

Los tiempos, en general, se siguieron de acuerdo a lo estimado. Exceptuando algún apartado como la memoria y la preparación de la defensa, cuyos tiempos se excedieron más de lo pensado y utilizaron el tiempo reservado mejoras del proyecto. Sin embargo, ese tiempo adicional ya se había pensado para ese tipo de casos: cubrir posibles retrasos o imprevistos.

En general, la combinación de metodologías ágiles con el desarrollo en cascada del proyecto aseguraron el éxito del resultado del sistema final. Por lo que la planificación y la gestión por hitos fueron un punto clave para que el desarrollo del proyecto derivase en un resultado óptimo.

3. Descripción de la arquitectura del sistema

3.1. Tecnologías utilizadas

Para el correcto desarrollo e implementación del sistema, han sido necesarias una serie de tecnologías y herramientas. La más importante y básica de ellas es **Python**, el lenguaje de programación elegido para el desarrollo del proyecto. El resto de dependencias y tecnologías giran en torno a Python, ya que son las librerías y módulos utilizados para garantizar las funcionalidades necesarias.

- La librería **pgmpy** es una librería externa que sirve para construir y trabajar con modelos gráficos probabilísticos como son las Redes Bayesianas o los Modelos Ocultos de Markov. En nuestro caso, utilizamos la librería para crear el modelo de Red Bayesiana y realizar las inferencias correspondientes sobre los datos de respuesta de los pacientes.

Las **Redes Bayesianas** son la herramienta principal en torno a la cual gira nuestro sistema de diagnóstico. Este modelo probabilístico en forma de grafo que representa las dependencias condicionales entre nodos ha servido para indicar las relaciones entre la diabetes, sus síntomas y sus factores de riesgo. Aunque su implementación concreta se explicará más a fondo en el apartado correspondiente a Redes Bayesianas, la librería pgmpy nos facilitó la creación de la red y las inferencias necesarias para el diagnóstico (ya que las inferencias vienen implementadas en la propia librería, y solamente tuvimos que hacer el diseño de las redes).

- El módulo **time** es una librería estándar de Python que hemos utilizado únicamente para hacer una simulación de recogida de datos, con el fin de que el usuario lea bien los prints informativos antes de pasar a fases posteriores del diagnóstico. Realmente no tiene mayor trascendencia, ya que el uso de time es puntual y no tiene ningún impacto significativo en la funcionalidad principal del sistema.
- La librería externa **pandas** fue esencial para analizar y manipular datos de CSV. Proporciona una estructura de datos concreta, el *dataframe*, que permite trabajar con los datos de dichos ficheros. En nuestro caso, la hemos utilizado en dos partes del proyecto.

Por una parte, para leer los datos de ficheros en la funcionalidad CSV del sistema; es decir, para hacer inferencias de una batería de datos guardados en un fichero CSV. Se tienen que leer los datos y manipular el fichero para poder mostrar la probabilidad de tener diabetes de cada una de las filas (cada fila se corresponde con una instancia con respuestas del cuestionario).

Por otro lado, en la Red Bayesiana, la tabla de probabilidades condicionadas del nodo Diabetes era demasiado grande y complicada de interpretar y/o manipular si se escribía directamente en el código. Esto se debe a que el nodo diabetes tiene demasiados padres, lo que genera una CPD demasiado grande. Es por ello que decidimos utilizar un fichero CSV para guardar las CPDs (probabilidades condicionadas) de tener diabetes dados los factores de riesgo. Para leer las probabilidades del fichero e interpretarlas correctamente, utilizamos la librería pandas.

3.2. Parámetros utilizados

Se han seleccionado una serie de parámetros que, tras investigar sobre el diagnóstico de diabetes, hemos considerado como parámetros de interés. Se nombran a continuación por cada parámetro, el nombre de dicho parámetro relacionado con la diabetes, la pregunta que irá asociada a dicho parámetro (preguntas que se harán a la hora de hacer el cuestionario al paciente para hacer el diagnóstico), y las posibles respuestas que se esperan.

Estas respuestas están acotadas, ya que muchos de los parámetros que seleccionamos son continuos, y se ven reflejados en la columna “Rango”. Así, si por ejemplo una persona de 70 años responde a la pregunta de cuál es su edad, el número asociado a este parámetro será el 1 (al ser igual o mayor de 45 años).

Medición	Pregunta	Rango
Edad	¿Cuál es su edad?	0 → Joven: 0 - 44 años 1 → Mayor: 45+ años
Peso (kg)	¿Cuál es su peso? (en kg)	FLOAT
Altura (cm)	¿Cuál es su altura? (en cm)	FLOAT
Índice de Masa Corporal (IMC)	NO HAY PREGUNTA, se calcula en el código	0 → BajoPeso/Normal: <25 1 → Sobrepeso: 25 - 30 2 → Obesidad: 30+
Enfermedad o lesión grave del páncreas	¿Ha padecido alguna enfermedad o lesión grave del páncreas?	0 → No 1 → Sí
Antecedentes familiares	¿Hay antecedentes de diabetes en su familia (padre/madre)?	0 → No 1 → Sí
Veces que se va al baño	¿Con qué frecuencia siente ganas de orinar?	0 → Con poca frecuencia 1 → Con una frecuencia normal 2 → Con alta frecuencia
Sed	¿Con qué frecuencia siente ganas de beber agua?	0 → Con poca frecuencia 1 → Con una frecuencia normal 2 → Con alta frecuencia
Fatiga	¿Con qué frecuencia siente fatiga?	0 → Con poca frecuencia 1 → Con una frecuencia normal 2 → Con alta frecuencia
Hambre	¿Con qué frecuencia siente hambre?	0 → Con poca frecuencia 1 → Con una frecuencia normal 2 → Con alta frecuencia
Pérdida de peso inexplicada	¿Ha tenido alguna pérdida de peso inexplicada recientemente?	0 → No 1 → Sí

Enfermedades relacionadas (son consecuencia de la diabetes)	<p>¿Ha padecido recientemente o padece alguna de los siguientes?:</p> <ul style="list-style-type: none"> • <u>Retinopatía</u> diabética (visión borrosa, cuerpos flotantes en la vista, zonas de visión oscura y, en general, pérdida de capacidad visual) • <u>Patologías cardiovasculares</u> • <u>Apnea del sueño</u> • <u>Afecciones de la piel</u> (Acantosis nigricans o acantosis pigmentaria, Dermopatía diabética, Necrobiosis lipoídica, Bulosis diabética, Xantomatosi eruptiva, Infecciones cutáneas) 	0 → No 1 → Sí
Nivel de glucosa en sangre	¿Cuál es su nivel de glucosa en sangre en la medición realizada?	0 → Normal: menos de 100 mg/dL 1 → Prediabetes: 100 - 125 mg/dL 2 → Diabetes: +125 mg/dL
Presión sistólica	¿Cuál ha sido su medición de presión sistólica (número superior)? (en mmHg)	FLOAT
Presión diastólica	¿Cuál ha sido su medición de presión diastólica (número superior)? (en mmHg)	FLOAT
Presión sanguínea	NO HAY PREGUNTA, se calcula en el código	0 → Normal: sistólica < 130 Y diastólica < 80 1 → Hipertensión: sistólica +130 O diastólica +80

3.3. Redes Bayesianas

Nuestra implementación del diagnóstico de la diabetes se basa en dos redes bayesianas. En las dos redes aparece en el medio y como nexo, el nodo de "Diabetes". En ambas redes, se muestran encima del nodo "Diabetes" las causas que influyen en el riesgo de poder padecer diabetes; y, bajo el nodo "Diabetes", los síntomas comunes que podría padecer un paciente con esta enfermedad.

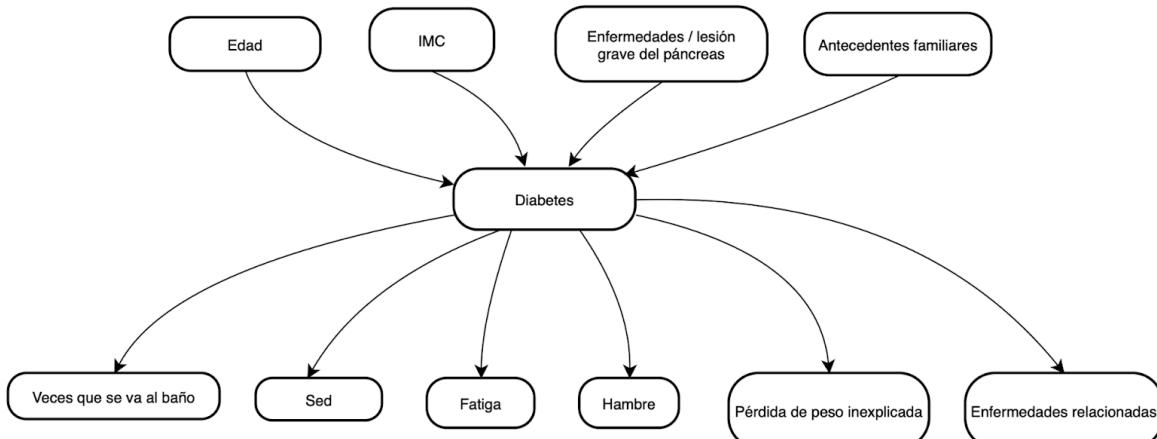
Por una parte, las causas son parámetros que aumentan la probabilidad de padecer diabetes (un IMC alto, la edad, etc.), mientras que los síntomas son manifestaciones que están presentes en aquellas personas con esta enfermedad (como la frecuencia de orinar o la pérdida de peso inexplicada, que aumenta en personas diabéticas). Las flechas que apuntan a "Diabetes" son factores que influyen en la probabilidad de tener diabetes. Por otra parte, las flechas que salen de "Diabetes" son síntomas que una vez se tiene la enfermedad, podrían manifestarse.

La primera Red Bayesiana ofrece un modelo básico de diagnóstico en el que, si los resultados indican una muy baja probabilidad de padecer diabetes (en nuestro caso menor del 30%), termina el diagnóstico. Esta red contiene los nodos de aquellos parámetros que forman parte de un cuestionario básico o inicial que responde el paciente que quiere saber si tiene diabetes. Es decir, son respuestas a preguntas sobre parámetros que conoce (como su edad, peso y altura) o sobre posibles síntomas (si se cansa con facilidad o tiene sed de manera frecuente). En resumen, el paciente podría responder a estas preguntas desde su casa y hacer un primer estudio sin necesidad de hacer mediciones médicas específicas.

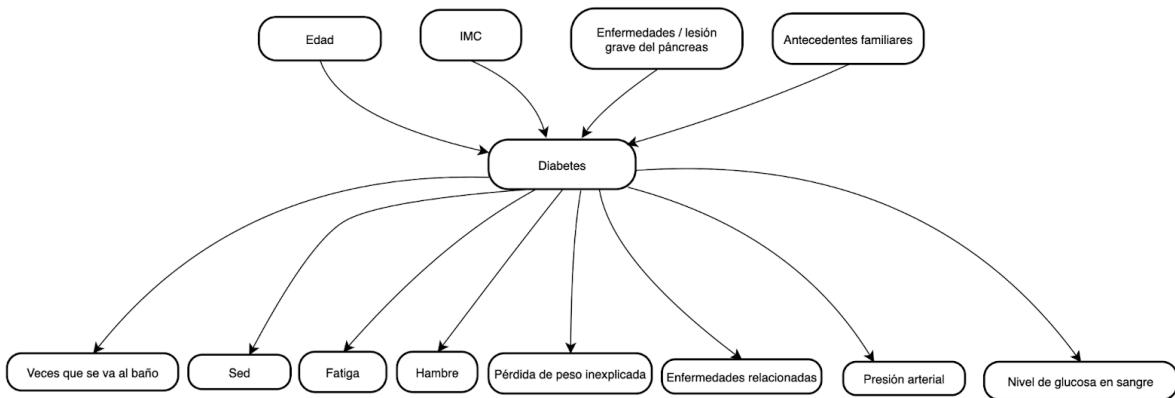
Si los resultados de la inferencia con la primera Red Bayesiana mostraran una probabilidad considerable de poder padecer diabetes (mayor al 30%, en este caso), se pasaría a implementar la **segunda Red Bayesiana**, que añadiría dos nodos con dos resultados clínicos de gran importancia (la presión arterial y los niveles de azúcar en sangre), así como sus relaciones con el nodo "Diabetes".

Esto, en el mundo real, sería el equivalente a que el médico mandase al paciente a hacerse pruebas clínicas tras detectar una posible diabetes en base al cuestionario inicial. Por lo que, además de las causas y síntomas anteriores, se añaden los resultados clínicos de presión arterial y nivel de glucosa en sangre, lo que permitiría tener un diagnóstico más claro.

Red Bayesiana 1:



Red Bayesiana 2:



Tablas CPDs de las Redes Bayesianas

A continuación, se detallan las tablas de **probabilidades** asociadas a cada uno de los nodos de las Redes Bayesianas. Se indican las CPDs de la Red Bayesiana 2, ya que son los mismos que los de la Red Bayesiana 1 pero añadiendo las tablas para la glucosa y la presión sanguínea.

Factores de riesgo:

Probabilidad *a priori* de Edad:

$P(\text{age}=0)$	$P(\text{age}=1)$
0.6	0.4

Probabilidad *a priori* de IMC:

$P(\text{bmi}=0)$	$P(\text{bmi}=1)$	$P(\text{bmi}=2)$
0.55	0.37	0.08

Probabilidad *a priori* de Enfermedad / Lesión grave del páncreas:

$P(\text{pancreas_diseases}=0)$	$P(\text{pancreas_diseases}=1)$
0.99	0.01

Probabilidad *a priori* de Antecedentes familiares:

$P(\text{family_history}=0)$	$P(\text{family_history}=1)$
0.8	0.2

Diabetes:**Probabilidad condicional de Diabetes:**

age	bmi	pancreas_diseases	family_history	P(diabetes=1)	P(diabetes=0)
0	0	0	0	0.98	0.02
0	0	0	1	0.95	0.05
0	0	1	0	0.93	0.07
0	0	1	1	0.88	0.12
0	1	0	0	0.90	0.10
0	1	0	1	0.85	0.15
0	1	1	0	0.83	0.17
0	1	1	1	0.75	0.25
0	2	0	0	0.80	0.20
0	2	0	1	0.70	0.30
0	2	1	0	0.65	0.35
0	2	1	1	0.50	0.50
1	0	0	0	0.90	0.10
1	0	0	1	0.85	0.15
1	0	1	0	0.80	0.20
1	0	1	1	0.70	0.30
1	1	0	0	0.75	0.25
1	1	0	1	0.65	0.35
1	1	1	0	0.60	0.40
1	1	1	1	0.45	0.55
1	2	0	0	0.50	0.50
1	2	0	1	0.35	0.65
1	2	1	0	0.30	0.70
1	2	1	1	0.20	0.80

Síntomas:**Probabilidad condicional de veces que se va al baño:**

diabetes	P(urinate_freq=0)	P(urinate_freq=1)	P(urinate_freq=2)
0	0.6	0.3	0.1
1	0.2	0.3	0.5

Probabilidad condicional de sed:

diabetes	P(thirst=0)	P(thirst=1)	P(thirst=2)
0	0.7	0.2	0.1
1	0.3	0.4	0.3

Probabilidad condicional de fatiga:

diabetes	P(fatigue=0)	P(fatigue=1)	P(fatigue=2)
0	0.5	0.3	0.2
1	0.2	0.3	0.5

Probabilidad condicional de hambre:

diabetes	P(hunger=0)	P(hunger=1)	P(hunger=2)
0	0.6	0.3	0.1
1	0.3	0.4	0.3

Probabilidad condicional de pérdida de peso inexplicada:

diabetes	P(weight_loss=0)	P(weight_loss=1)
0	0.8	0.2
1	0.3	0.7

Probabilidad condicional de enfermedades relacionadas:

diabetes	P(sympt_diseases=0)	P(sympt_diseases=1)
0	0.9	0.1
1	0.4	0.6

Y las correspondientes únicamente a la Red Bayesiana 2:

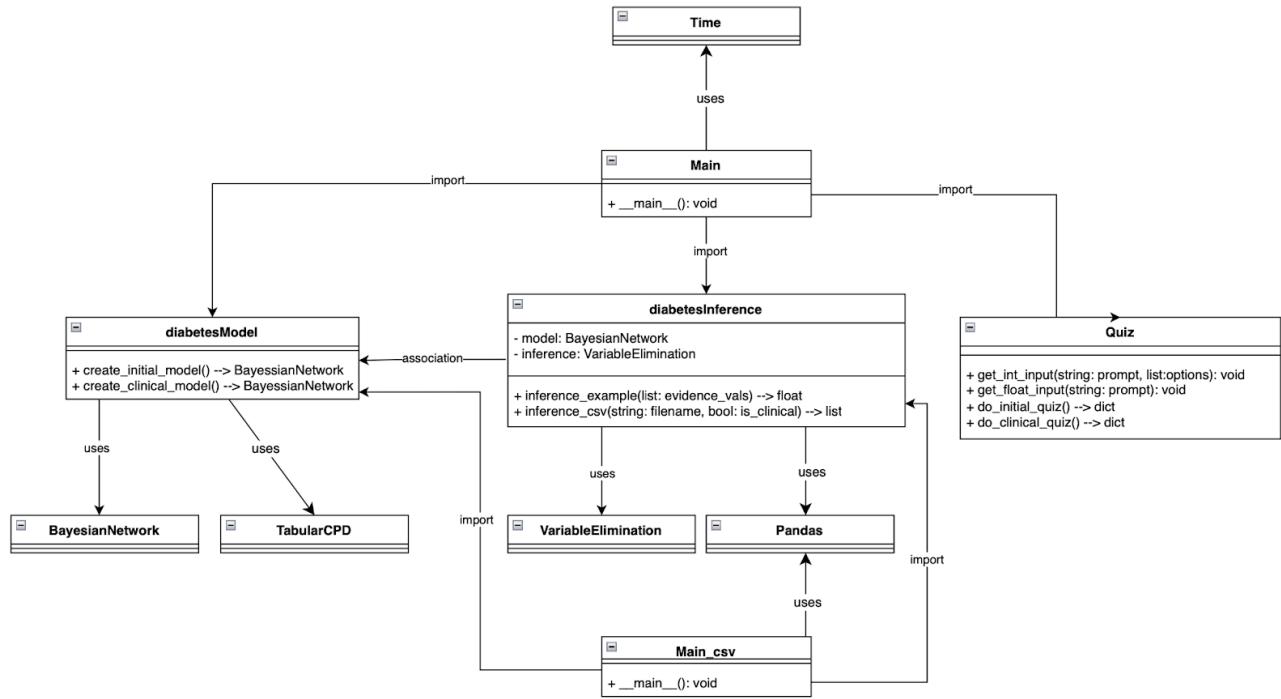
Probabilidad condicional de presión sanguínea:

diabetes	P(blood_pressure=0)	P(blood_pressure=1)	P(blood_pressure=2)
0	0.7	0.2	0.1
1	0.3	0.4	0.3

Probabilidad condicional de glucosa en sangre:

diabetes	P(glucose=0)	P(glucose=1)
0	0.7	0.3
1	0.4	0.6

3.4. Diagrama de Clases



En este diagrama, se representan todas las clases del programa de diagnóstico de diabetes. Hay que destacar por qué hay dos Main (el Main_csv y el Main). Y es que el Main_csv servirá esencialmente para hacer pruebas, ya que acepta un fichero de datos CSV. Esto será útil a la hora de enviar al modelo una batería de síntomas para comprobar que, efectivamente, la inferencia se esté haciendo de manera correcta. Por otro lado, el Main se utilizará para iniciar el programa que haga el cuestionario al paciente, y que a partir de ello pueda realizar la inferencia y determinar si padece o no diabetes.

Main es el controlador principal que inicializa y coordina la ejecución de los distintos componentes del sistema de diagnóstico de diabetes para un paciente. Su funcionamiento se detallarán en el apartado 3.5. correspondiente al Diagrama de Flujo. Utiliza las clases **diabetesModel**, **diabetesInference** y **Quiz**. Adicionalmente, utiliza la librería **Time** para poder hacer prints que simulen una recogida de datos.

Main_csv es el controlador principal del modo de ejecución con fichero CSV. En este caso, utilizará las clases **diabetesModel** y **diabetesInference** (no será necesario importar **Quiz** en este caso, ya que se recogerán los resultados directamente del CSV, y no de las respuestas de un paciente). Además, necesitará la librería **Pandas** para poder operar correctamente con los CSV, convirtiéndolos en *dataframes*. De nuevo, el funcionamiento concreto se detalla en el apartado 3.5. correspondiente al Diagrama de Flujo.,.

diabetesModel crea instancias de **BayesianNetwork** con ayuda de **TabularCPD** (ambas clases de la librería de pgmpy) para las distribuciones de probabilidades de los parámetros relacionados con la diabetes. Esto lo hace a través de dos funciones, correspondientes con las dos redes bayesianas explicadas en el apartado anterior.

- La función `create_initial_model()` se corresponde con la creación de la red bayesiana que representa a la primera Red Bayesiana (la que no contiene resultados clínicos). Crea el modelo, añade los nodos y arcos correspondientes, y por último añade las CPDs detalladas en el apartado de Redes Bayesianas (todas menos las de glucosa y presión sanguínea). La función devuelve el modelo (objeto BayesianNetwork).
- La función `create_clinical_model()` sería la encargada de añadir los nodos, arcos y CPDs faltantes para poder crear la segunda Red Bayesiana. Es decir, primero crea un modelo con `create_initial_model()`, para después añadir todo lo correspondiente a glucosa y presión sanguínea. De nuevo, devuelve el modelo (objeto BayesianNetwork).

diabetesInference es una clase que tiene como atributos el modelo (se asocia al modelo) y el objeto de clase VariableElimination, necesario para realizar inferencias. Las funciones de diabetesInference tendrán como objetivo dar una probabilidad de tener diabetes dados todos los síntomas y parámetros.

- La función de `inference_example(evidence_vals)` hace la inferencia sobre el caso del paciente. Es decir, es una función de inferencia para una única instancia. Toma como parámetro la lista de evidencias (de las respuestas del cuestionario), y evalúa si se trata de un caso de diagnóstico básico o si también hay resultados clínicos. Hace un diccionario combinando las claves con las respuesta y, con todo ello, realiza una inferencia en base al modelo asociado. Finalmente, devuelve la probabilidad de tener diabetes (en float).
- La función `inference_csv(filename, is_clinical)` hace la inferencia por cada una de las instancias en el fichero CSV indicado como filename, y devuelve una lista con todos los valores de las probabilidades de cada instancia (una lista de porcentajes). Esta función también toma como parámetro un booleano `is_clinical`, qué será verdadero cuando el fichero contenga datos clínicos; y falso cuando únicamente tenga datos del cuestionario inicial. Esto significa que `inference_csv` no podrá recibir un fichero que mezcle resultados clínicos con resultados del cuestionario inicial. Además, el resultado con las inferencias y la decisión de si se tiene o no diabetes se guardará en un nuevo CSV llamado filename original + `_solved.csv`.

Quiz es la clase encargada de lanzar los cuestionarios al usuario.

- El primer cuestionario se lanza con la función `do_initial_quiz()`, que almacena los resultados en un diccionario en el que las claves son los 10 parámetros y los valores el valor de cada uno de los parámetros en base a la respuesta del paciente (0, 1, 2, ...).
- En caso de ser necesario, la función `do_clinical_quiz()` lanza el segundo cuestionario. Este recolectaría las respuestas a los exámenes clínicos que se le hayan hecho al paciente; devolviendo de la misma forma un diccionario.

Para recoger los datos de forma eficiente y evitar errores por valores mal introducidos, dos funciones (`get_int_input` y `get_float_input`) serán las encargadas de lanzar un bucle que garantice que el usuario ingrese datos válidos. Ambas toman prompt como parámetro, que se corresponderá con la pregunta del cuestionario. La función `get_int_input(prompt, options)` recoge un número entero que se encuentre en la lista de options (opciones válidas).

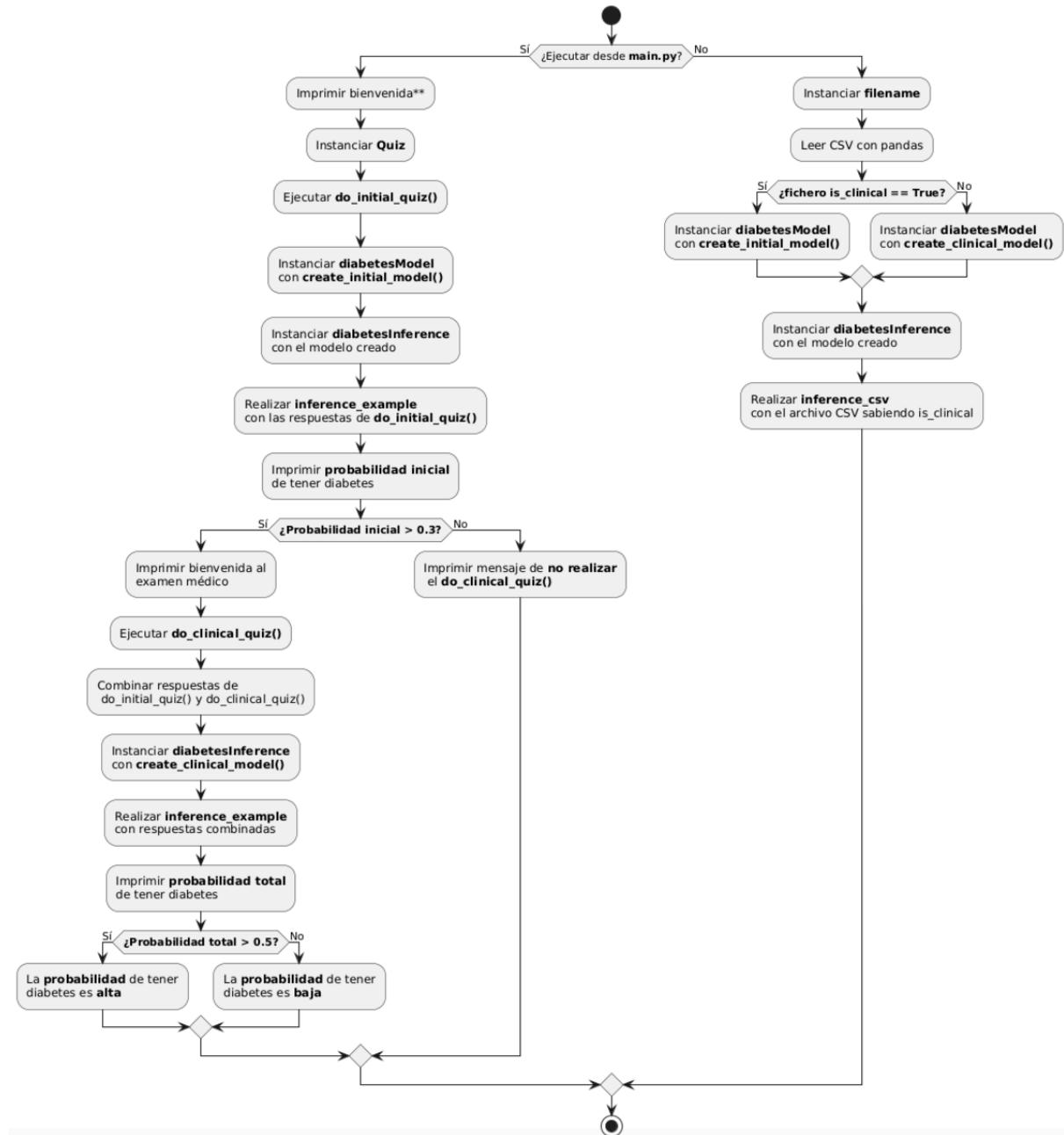
Por ejemplo, al recoger datos sobre la frecuencia de sed, la pregunta se lanza de la siguiente manera:

```
thirst = self.get_int_input("\n*****\n? ¿Con qué frecuencia siente  
ganas de beber agua? \n 0 = frecuencia baja\n 1 = frecuencia normal\n 2 =  
frecuencia alta\n>>>RESPUESTA FRECUENCIA BEBER: ", [0, 1, 2])
```

La prompt es la pregunta, y las opciones son 0, 1 y 2 (correspondientes con frecuencia baja, normal y alta de beber agua). Cualquier valor introducido por el usuario que se encuentre fuera de esos rangos, volverá a lanzar la pregunta hasta recoger un valor correcto.

En el caso de get_float_input(prompt) únicamente se comprueba que, en efecto, el valor introducido es un flotante introducido en el formato decimal adecuado (con un punto, y no con una coma).

3.5. Diagrama de Flujo



Para iniciar el programa tendremos dos opciones:

La primera será hacer un diagnóstico de diabetes a un paciente que llega con síntomas, ejecutando el **Main.py**. Primero, se imprime una bienvenida. Para continuar con el diagnóstico, se instancia una clase Quiz que usará su función `do_initial_quiz()` para recolectar las respuestas sobre los síntomas del paciente. Después, se instancia una clase `diabetesModel` con `create_initial_model()`; ya que aún estamos en la primera fase, la de los síntomas (y no la del estudio clínico).

Se pasa entonces a instanciar `diabetesInference` y ejecutar `inference_example`, con la lista de las respuestas del `do_initial_quiz()` como parámetro. A continuación se imprime la probabilidad inicial de padecer diabetes y, si esta es menor que 0.3 (que no es un valor muy alto, pero representa un riesgo de tener diabetes), no haría falta realizar el estudio clínico.

Si, por el contrario, la probabilidad ha resultado mayor a ese 0.3, se da la bienvenida al examen médico, y se ejecuta el `do_clinical_quiz()`, cuyas respuestas se combinan con las respuestas del cuestionario inicial. Se instancia un nuevo `diabetesInference` con `create_clinical_model()`. A continuación, se ejecuta `inference_example` con la lista de las respuestas combinadas como parámetro de la función. Finalmente, se imprime la probabilidad total de tener diabetes y se cierra el programa.

La segunda forma de iniciar el programa va más orientada a la fase de pruebas y a médicos que quieran utilizar la modalidad; y se ejecuta desde el **main_csv.py**. Primero, se instancia el fichero para el que queremos obtener las probabilidades de padecer diabetes; y se procede a leer dicho archivo CSV con ayuda de pandas.

A continuación, se comprueba si el número de atributos que contiene el fichero se asocia con un cuestionario completo (inicial + clínico) o únicamente con un cuestionario clínico. Dependiendo del resultado de esta comprobación, se instancia el `diabetesModel` con `create_initial_quiz()` o con `create_clinical_model()`. Después se pasa a instanciar `diabetesInference` con el modelo creado, y realizar `inference_csv` con el archivo CSV, y conociendo si se trata de un CSV con resultados clínicos o no.

4. Pruebas realizadas

Se han realizado una serie de pruebas que cubren diferentes casos de síntomas y factores de riesgo, para así comprobar que el modelo funciona correctamente.

Se han probado a introducir manualmente diversos tipos de respuestas al cuestionario para corroborar el buen funcionamiento del Quiz y el Main. A partir de dichas pruebas, comprobamos que se recogían bien las respuestas y se realizaba la inferencia tal y como debía. Además, se puso a prueba el introducir datos incorrectos para comprobar el buen funcionamiento de las funciones auxiliares del módulo Quiz.

También se comprobó que, introduciendo las mismas respuestas desde el cuestionario y desde el CSV, se recibían los mismos porcentajes de probabilidad de diabetes. Por ejemplo, para el siguiente caso en un cuestionario, se prueban a pasar los valores correspondientes por el Main_csv. La captura al CSV es el resultado que aparece en el fichero correspondiente a la inferencia resuelta.

En ambas se ve que se recibe como resultado de la inferencia un 76.9231% de probabilidad de tener diabetes, por lo que se concluye que la inferencia se realizó de manera correcta.

```
👋 Hola! Bienvenido a la prueba de diagnóstico para la diabetes
Comenzando test...

*****
? ¿Cuál es su edad?
>>RESPUESTA EDAD: 20

*****
? ¿Cuál es su peso? (en kg): 60

*****
? ¿Cuál es su altura? (en cm): 150

*****
Su IMC es de 26.67

*****
? ¿Ha padecido alguna enfermedad o lesión grave del páncreas?
0 = no
1 = sí
>>RESPUESTA ENFERMEDADES PÁNCREAS: 0

*****
? ¿Hay antecedentes de diabetes en su familia (padre/madre)?
0 = no
1 = sí
>>RESPUESTA ANTECEDENTES: 0

*****
? ¿Ha tenido alguna pérdida de peso inexplicada recientemente?
0 = no
1 = sí
>>RESPUESTA PÉRDIDA DE PESO: 0

*****
? ¿Ha padecido recientemente o padece alguna de los siguientes?:
*Visión borrosa, cuerpos flotantes en la vista, zonas de visión oscura o pérdida de la capacidad visual
*Patologías cardiovasculares
*Apnea del sueño
*Afecciones de la piel (Infecciones cutáneas, acantosis pigmentaria, necrobiosis lipoídica)

0 = no
1 = sí
>>RESPUESTA ENFERMEDADES: 1

-----
PROBABILIDAD DE TENER DIABETES:
76.9231 %

-----
```

age	bmi	pancreas_diseases	family_history	urinate_freq	thirst	fatigue	hunger	weight_loss	sympt_diseases	diabetes
0	1	0	0	2	1	1	1	0	1	76.9231 %

Para continuar con la prueba del correcto funcionamiento del cuestionario y el flujo del programa, seguimos con el mismo ejemplo del cuestionario. Este lanza el cuestionario clínico, ya que la probabilidad de tener diabetes es mayor a 0.3. A continuación realiza la inferencia e imprime la probabilidad total de tener diabetes, teniendo en cuenta todas las respuestas. De nuevo, podemos comprobar que realizando la inferencia con el CSV funciona de igual manera. La probabilidad final de tener diabetes es del 93.0233%, en ambos casos.

```
Los resultados del cuestionario indican que existe una probabilidad considerable de que tenga diabetes
```

```
. Es necesario hacer un examen médico.
```

```
Recogiendo muestras...
```

```
*****
```

```
💊 ¿Cuál es su nivel de glucosa en sangre en la medición realizada? (en mg/dL)  
>>>RESPUESTA GLUCOSA EN SANGRE: 120
```

```
*****
```

```
⌚ ¿Cuál ha sido su medición de presión sistólica (número superior)? (en mmHg): 120
```

```
⌚ ¿Cuál ha sido su medición de presión diastólica (número inferior)? (en mmHg): 90
```

----- PROBABILIDAD DE TENER DIABETES

(con el test clínico):

93.0233 %

age	bmi	pancreas_diseases	family_history	urinate_freq	thirst	fatigue	hunger	weight_loss	sympt_diseases	glucose	blood_pressure	diabetes
0	1	0	0	2	1	1	1	0	1	1	1	93.0233 %

Visto que el cuestionario funciona correctamente y da los mismos resultados que si se hace la inferencia con el CSV, pasamos a hacer pruebas más exhaustivas directamente desde los CSV que validen la correcta funcionalidad del modelo. No se detallan todas las pruebas hechas, pero sí algunas de las más significativas.

Primero, destacamos que se han hecho pruebas sobre **datos reales** de personas con y sin diabetes, y los resultados fueron los esperados. La persona con diabetes dio un alto porcentaje en la probabilidad total, y la persona sin diabetes dio una probabilidad muy baja.

Primero, para los tests normales:

tag	age	bmi	pancreas_diseases	family_history	urinate_freq	thirst	fatigue	hunger	weight_loss	sympt_diseases	diabetes_prob	DIABETES
caso_diabetico	1	0	0	1	2	2	2	0	1	0	83.7321 %	SI
caso_no_diabetico	1	1	0	0	0	1	1	0	0	0	1.8182 %	NO

Y luego, para los tests clínicos:

tag	age	bmi	pancreas_diseases	family_history	urinate_freq	thirst	fatigue	hunger	weight_loss	sympt_diseases	glucose	blood_pressure	diabetes_prob	DIABETES
caso_diabetico	1	0	0	1	2	2	2	0	1	0	1	0	85.4701 %	SI
caso_no_diabetico	1	1	0	0	0	1	1	0	0	0	1	0	2.0725 %	NO

También quisimos probar el **impacto del test clínico** sobre otro caso. Para un caso que recomendaba hacer el test clínico y después terminó dando negativo en la probabilidad completa de tener diabetes.

Primero, el test normal dio una alta probabilidad de padecer diabetes:

tag	age	bmi	pancreas_diseases	family_history	urinate_freq	thirst	fatigue	hunger	weight_loss	symp_t_diseases	diabetes_prob	DIABETES
si_luego_no	0	1	0	0	2	1	1	1	0	1	76.9231 %	SI

Y después, se esclarecieron las dudas realizando pruebas clínicas, y el caso terminó dando negativo en el test:

tag	age	bmi	pancreas_diseases	family_history	urinate_freq	thirst	fatigue	hunger	weight_loss	symp_t_diseases	glucose	blood_pressure	diabetes_prob	DIABETES
si_luego_no	0	1	0	0	2	1	1	1	0	1	0	0	44.9438 %	NO

Para la **batería de pruebas**, hay dos ficheros, que se encuentran en la carpeta test del directorio raíz del proyecto: normal.csv y clinico.csv. Ambos tienen su contraparte resuelto, normal_solved.csv y clinico_solved.csv (respectivamente). Las pruebas estarán en **normal.csv** y **clinico.csv** (resueltas en **normal_solved.csv** y **clinico_solved.csv**):

test1. Prueba para que de que tiene diabetes, en función de criterios subjetivos o de “sentido común”. **Se espera un resultado positivo.**

test2. Prueba para que de que no tiene diabetes, en función de criterios subjetivos o de “sentido común”. **Se espera un resultado negativo.**

test3. Pruebas con únicamente probabilidades altas en los síntomas. **Se espera un resultado positivo.**

test4. Pruebas con únicamente probabilidades altas en los factores de riesgo. **Se espera un resultado positivo.**

test5. Prueba para una persona joven, con IMC normal, sin antecedentes familiares de diabetes, que no haya sufrido ninguna enfermedad o lesión grave de páncreas y sin ningún síntoma de padecer diabetes. **Se espera que el resultado de esta prueba sea negativo.**

test6. Prueba que evalúa a una persona mayor, con IMC normal, sin antecedentes familiares, con fatiga y sed normales, que no haya tenido síntomas graves ni una pérdida de peso reciente, aunque sí que tiene una frecuencia alta de micción. **Se espera un resultado negativo.**

test7. Esta prueba evalúa a una persona de 70 años, con sobrepeso, sin cambio de peso inexplicable, sin enfermedades pasadas y sin antecedentes familiares, aunque con sed y fatiga altas. **Se espera un resultado negativo.**

test8. Esta prueba evalúa a un niño con un IMC bajo y con antecedentes familiares de diabetes, sin síntomas; pero que ha sufrido una bajada de peso inexplicable recientemente. **Se espera que el resultado de esta prueba sea negativo.**

test9. Esta prueba evalúa a una persona joven, IMC normal, pero con antecedentes familiares y que haya tenido una lesión grave en el páncreas. La persona tiene una frecuencia alta en orinar, sed y hambre; y una baja frecuencia de fatiga normal. La persona no tiene más síntomas de diabetes. Como el resultado de esto puede dar **riesgo**, se realizan a mayores unas pruebas clínicas que dan con nivel de glucosa en prediabetes y presión arterial normal. **Se espera que el resultado de esta prueba de positivo.**

test10. La prueba evalúa el caso anterior, pero variando la edad a mayor. Se espera un resultado de **riesgo** del caso inicial, y un **caso positivo en la prueba total**. Sin embargo, en este caso las probabilidades serán más altas que en el caso anterior.

test11. Esta prueba evalúa a una persona joven, IMC alto, y con antecedentes familiares y que haya tenido una lesión grave en el páncreas. La persona tiene una frecuencia alta en orinar, sed y hambre; aunque una baja frecuencia de fatiga. Como el resultado de esto puede dar **riesgo**, se realizan a mayores unas pruebas clínicas que dan con nivel de glucosa en prediabetes y presión arterial alta. **Se espera que el resultado de esta prueba de positivo.**

test12. Esta prueba evalúa a una persona mayor con IMC en sobrepeso. Tiene una baja frecuencia de hambre, pero tiene alta frecuencia de fatiga, sed y de micción. Además, ve cuerpos flotantes. El cuestionario inicial indicaría un **riesgo** de padecer diabetes. Se realiza un análisis clínico para comprobar los valores de presión arterial y nivel en sangre de glucosa, para verificar si tiene diabetes partiendo de los síntomas dados. El resultado de estos análisis de hipertensión y nivel alto de glucosa en sangre. **Se espera que el resultado de esta prueba de positivo.**

Por último, en **z_more_normal.csv** y **z_more_clinical.csv** se ven los resultados de otras pruebas realizadas. Incluyen el máximo y mínimo de probabilidad posible, el resultado al variar un único atributo y otras pruebas variadas que comprobaron el funcionamiento de la funcionalidad CSV.

Prácticamente todas las pruebas han ido según lo esperado, aunque ha habido alguna excepción. Por ejemplo, para el test4, se esperaba una probabilidad positiva y finalmente se obtuvo una probabilidad del 1.8692 %. Esto indica que en nuestro modelo los síntomas son los que marcan la diferencia y producen un mayor impacto sobre la probabilidad final. Esto se podrá mejorar refinando las probabilidades.

5. Manual de Uso

A continuación se detallan los pasos de instalación y uso del programa de diagnóstico de diabetes para su uso en un sistema operativo Debian/Ubuntu. El programa funcionará en otros sistemas operativos como Windows o MacOS, pero la instalación de dependencias deberá hacerse manualmente y los pasos detallados podrán variar de los de Linux.

5.1. Instalación del programa

1. Descomprimir el zip del proyecto y entrar en la carpeta “**diabetes_diagnose**”.
2. Abrir una terminal en Linux en ese mismo directorio.
3. Ejecutar el comando “**chmod +x dependencies.sh**” para dar permisos de ejecución al instalador de dependencias.
4. Ejecutar el comando “**sudo bash dependencies.sh**” para instalar todas las dependencias de manera automática.

5.2. Uso del programa

Como se ha indicado previamente, hay dos posibles formas de ejecutar el programa:

Ejecución para un paciente, con cuestionario:

1. Abrir la terminal en el directorio “**diabetes_diagnose**”.
2. Ejecutar el comando “**python main.py**”, que iniciará el cuestionario.
3. Leer las preguntas planteadas con atención y responder por línea de comandos donde se indica.
4. Al finalizar, el cuestionario indicará la probabilidad total de padecer diabetes, por lo que el paciente deberá consultar con su doctor según los resultados finales.

Ejecución para médicos o desarrolladores, con fichero CSV:

1. Crear o seleccionar el fichero CSV del que queremos conocer la inferencia.
2. Abrir un editor de código y, en “**main_csv.py**” de la carpeta “**diabetes_diagnose**”, descomentar o escribir la línea correspondiente al CSV del archivo del que se quiera conocer la inferencia.
3. Abrir la terminal en el directorio “**diabetes_diagnose**”.
4. Ejecutar el comando “**python main_csv.py**”.
5. Se imprimirá un mensaje de dónde se encuentra el fichero con la inferencia resuelta.

NOTA: En el caso de no poder ejecutar el programa correctamente, se recomienda instalar todas las dependencias manualmente en un entorno.

6. Conclusiones finales

6.1. Problemas encontrados

Durante el desarrollo del proyecto se han encontrado una serie desafíos que hemos tenido que superar para poder llevar a cabo el sistema de diagnóstico. Los primeros problemas los encontramos durante la fase de recogida de requisitos e información necesarios para hacer el trabajo.

Inicialmente, el sistema estaba pensado para poder ser capaz de diagnosticar la diabetes Tipo 1 y la Tipo 2; por lo que toda la información que buscábamos era en relación con ambas. Después de investigar y ver la dificultad a la que conllevaría el problema, decidimos que lo mejor sería hacer el diagnóstico únicamente para la diabetes Tipo 2. Este enfoque resultó más práctico, ya que la diabetes tipo 2 es la más común, tiene síntomas más claros y cuantificables.

Previo a acotar el alcance del problema, también considerábamos una cantidad demasiado grande de parámetros (síntomas y factores de riesgo). Por una parte, algunos de ellos eran síntomas propios de la diabetes Tipo 1, por lo que los descartamos una vez descartamos el diagnóstico de ese tipo de diabetes. Pero a la hora de diseñar las Redes Bayesianas, tuvimos que sacrificar unos cuantos factores de riesgo y relaciones en la red entre varios nodos.

Tener un número tan elevado de factores de riesgo, con una cardinalidad tan alta (al principio considerábamos hasta 5 o 6 valores posibles por cada nodo) conllevaba una complejidad computacional demasiado elevada; ya que había demasiados posibles valores para demasiados padres para el nodo Diabetes, lo que significaba hacer una tabla de probabilidad condicionada demasiado grande. Ello también repercutía en el sentido de que había que buscar las probabilidades correspondientes a todas las posibles combinaciones.

Para mejorar la legibilidad y comprensión, decidimos hacer el fichero que contuviese la CPD de Diabetes. Inicialmente, antes de eliminar los factores de riesgo innecesarios así como sus posibles valores, teníamos más de 11.000 líneas con las distintas combinaciones de probabilidad. Finalmente y tras sacrificar los que menos impacto tenían en el diagnóstico, tenemos 24 posibles combinaciones en el fichero.

Otro de los problemas a los que nos tuvimos que enfrentar fue el de la asignación de probabilidades. Tratamos de buscar los datos en internet de fuentes fiables. Aún así, no toda la información deseada estaba disponible, y mucha de ella era contradictoria. Por lo que tuvimos que asignar algunas probabilidades según nuestros propios criterios.

En la implementación del diseño, el mayor problema encontrado es el de el tiempo que tarda en comenzar la ejecución la primera vez que se ejecuta el programa. Realmente aún no sabemos muy bien a qué se debe esto. Una de nuestras hipótesis fue que quizás sería por no cerrar los ficheros, pero esto no es así; ya que manejamos los ficheros con pandas.

6.2. Posibles mejoras

El sistema, aunque presenta una muy buena solución al problema planteado, tiene algunos aspectos que dan pie a posibles mejoras a largo o corto plazo.

Uno de ellos es el de poder incluir más factores clínicos, síntomas y factores de riesgo que puedan influir en el diagnóstico. Esto permitiría obtener resultados más precisos y exhaustivos, además de aumentar la sensibilidad del modelo al detectar casos potenciales.

Otra posible mejora sería que ampliar la capacidad del sistema y además de detectar el riesgo de padecer la diabetes de tipo 2, también detectase la de tipo 1 o gestacional, añadiendo otros parámetros específicos y probabilidades.

También se podrían revisar y actualizar las tablas de probabilidades condicionadas utilizadas en el modelo, incorporando datos más recientes, específicos y comprobados por médicos, para así mejorar la precisión del diagnóstico (ya que la mayoría de las probabilidades están fundamentadas en estimaciones aproximadas).

En esta línea, también se podría ajustar el modelo para que sea capaz de analizar datos específicos de diferentes grupos de personas y zonas geográficas. Esto ayudaría a mejorar la utilidad del sistema en poblaciones diversas y a generar diagnósticos más precisos adaptados a cada contexto.

Además, se puede diseñar una interfaz más atractiva para los usuarios, de forma que facilite la interacción con el sistema; así como implementar comprobaciones adicionales para verificar que se introducen correctamente los datos de los CSV (ya que actualmente solo se lanza un error)

Para mejoras a más largo plazo, se podría considerar añadir simulaciones de casos que permitan predecir cómo podría evolucionar la diabetes en un paciente si no se toman medidas. En un sistema real de diagnóstico, estas simulaciones podrían ayudar a planificar estrategias preventivas a largo plazo.

6.3. Conclusiones

Dado el gran desafío que suponía hacer un sistema de diagnóstico de la diabetes, hemos conseguido unos resultados bastante favorables que resuelven el problema tal y como queríamos y que ofrecen una funcionalidad interesante.

A pesar de ello, el sistema da pie a posibles mejoras, especialmente en cuanto a la precisión de las probabilidades. Se podrían añadir nodos y refinar probabilidades para mejorar el diagnóstico, que necesitarían una revisión periódica para mantener actualizado el sistema a las últimas investigaciones y avances médicos. Así como implementar una interfaz más amigable con el usuario.

Con todo, se ha conseguido que el sistema tenga la eficacia esperada, y las pruebas hechas demuestran un buen comportamiento a la hora de hacer el diagnóstico. Con un diseño modular y comprensible, se implementan correctamente las Redes Bayesianas; con el punto a mayores de complejidad que supone el hacer el diagnóstico en dos pasos.

7. Bibliografía consultada

PGMPY. (n.d.). pgmpy: Probabilistic graphical models using Python. Recuperado de <https://pgmpy.org/>

BNlearn. (n.d.). Diabetes: Discrete very large dataset. Recuperado de <https://www.bnlearn.com/bnrepository/discrete-verylarge.html#diabetes>

Veritas Intercontinental. (2021, 15 de noviembre). ¿La diabetes es hereditaria? Veritas Intercontinental. Recuperado de <https://www.veritasint.com/blog/es/la-diabetes-es-hereditaria/>

National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). (n.d.). ¿Qué es la diabetes tipo 2? Recuperado de <https://www.niddk.nih.gov/health-information/informacion-de-la-salud/diabetes/informacion-general/que-es/diabetes-tipo-2>

National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). (n.d.). Síntomas y causas de la diabetes tipo 2. Recuperado de <https://www.niddk.nih.gov/health-information/informacion-de-la-salud/diabetes/informacion-general/sintomas-causas>

Vivo Labs. (2023, 12 de abril). 12 enfermedades asociadas a la diabetes. Recuperado de <https://vivolabs.es/12-enfermedades-asociadas-a-la-diabetes/>

Mayo Clinic. (2023, 19 de septiembre). Diabetes: síntomas y causas. Recuperado de <https://www.mayoclinic.org/es/diseases-conditions/diabetes/symptoms-causes/syc-20371444>

MedlinePlus. (s. f.). Diabetes mellitus tipo 2. Recuperado el 6 de diciembre de 2024, de <https://medlineplus.gov/spanish/ency/article/000313.htm>

Wikipedia. (2024). Diabetes mellitus tipo 2. En Wikipedia, la enciclopedia libre. Recuperado el 6 de diciembre de 2024, de https://es.wikipedia.org/wiki/Diabetes_mellitus_tipo_2

International Diabetes Federation (IDF). (s. f.). Diabetes: Datos y cifras. Recuperado el 6 de diciembre de 2024, de <https://idf.org/es/about-diabetes/diabetes-facts-figures/>

Fernández Lanza, S. (s. f.). Apuntes del tema 3: Razonamiento con incertidumbre. Universidad de Vigo. Documento del Grado en Inteligencia Artificial. Recuperado el 6 de diciembre de 2024.