Evidencia 2: Redes bayesianas caso continuo

AUTHORS
Alejandra Velasco Zárate A01635453
José Antonio Juárez Pacheco A00572186
Jose Carlos Yamuni Contreras A01740285
Juan Manuel Hernández Solano A00572208
Mayra Sarahí De Luna Castillo A01635774

PUBLISHED
August 24, 2023

Abstract

Este estudio se enfoca en la aplicación de Redes Bayesianas Gaussianas (RBG) para modelar la Enfermedad Crónica Renal (ECR), considerando enfoques de modelado lineal y no paramétrico. La comparación de modelos lineales y no paramétricos busca determinar cuál se ajusta mejor a los datos. Mediante entrevistas con expertos en salud, se busca mejorar la estructura de las RBG con el objetivo de abordar diversas hipótesis sobre la ECR y la identificación de factores de riesgo y la predicción de su progresión. Esta investigación integra RBG, conocimiento de expertos y análisis comparativo de modelos para lograr una comprensión más profunda de la ECR, con potenciales implicaciones para la toma de decisiones clínicas y preventivas.

Introducción

La Enfermedad Crónica Renal (ECR) representa un desafío significativo en el ámbito de la salud pública debido a su creciente incidencia y sus efectos debilitantes en la calidad de vida de los pacientes. Es en estas situaciones donde las Redes Bayesianas son de mucho provecho para comprender y predecir los factores que contribuyen a esta enfermedad. Las Redes Bayesianas Gaussianas son una herramienta poderosa que permite capturar relaciones complejas entre variables y modelar la incertidumbre en los datos. Este trabajo se centra en la aplicación de Redes Bayesianas Gaussianas (RBG) para el modelado de la ECR. Se explorará la comparación entre dos enfoques de modelado: uno basado

en una red lineal y otro en una red no paramétrica. Así mismo, se trabajará con expertos en el área de Salud para la construcción de las redes bayesianas y con eso, responder a diversas hipótesis relacionadas con ECR, buscando identificar aspectos claves de la enfermedad con base en las RBG.

Marco Teórico

Las Redes Bayesianas son modelos ampliamente utilizados para la representación de relaciones de dependencia condicional en datos multivariantes. Una Red Bayesiana es un gráfico que muestra variables con nodos en un conjunto de datos y la dependencias probabilísticas o condicionales entre ellas. Inicialmente, las Redes Bayesianas se definieron para un conjunto finito de variables aleatorias discretas de las que se conocían la distribución de probabilidad condicionada, dada la ocurrencia de sus nodo padres en el DAG. Aplicando conceptos básicos del cálculo de probabilidades, la obtención de probabilidades finales de interés $P(x_i|E)$, dado un conjunto de variables discretas se ha demostrado en la evidencia anterior, pero si se busca ampliar este concepto en variables continuas se necesita indagar en este concepto. Un tipo específico de redes bayesianas es la Red Bayesiana Gaussiana, que se utiliza para modelar relaciones probabilísticas entre variables continuas que se supone que siguen una distribución gaussiana (normal). Como las variables en este tipo de redes son modeladas como variables gaussianas, las relaciones de dependencia entre ellas se expresan en términos de media y covarianza. En estas redes se tienen variables $X = \{X_1, \ldots, X_n\}$ con una distribución normal multivariante que tiene la forma:

$$f(x) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\}$$

donde μ es el vector de medias de dimensión n , Σ es la matriz de covarianzas definida positiva de dimensión $n \times n$, $|\Sigma|$ es la determinante de la matriz de covarianzas, $(x-\mu)^T$ es el vector transpuesto de $(x-\mu)$. La matriz de presición es Σ^{-1} . Además de la distribución gaussiana, las Redes Bayesianas Gaussianas tienen la distribución de probabilidad conjunta de la red, que es el producto de elementos P, que son las funciones de densidad condicionada dada ocurrencias de los nodos padres en una DAG, $f(x_i|pa(X_i))$. Por lo tanto la probabilidad conjunta queda de esta forma:

$$f(x) = \prod_{i=1}^{n} f(x_i|pa(X_i))$$

Se describe la distribución condicionada de cada una de las variables de la red, dada la ocurrencia de sus nodos padres en el DAG, como una distribución normal univariante tal que:

$$f(x_i|pa(X_i)) \sim N(\mu_i + \sum_{j=1}^{i-1} \beta_{ij}(x_j - \mu_j), v_i)$$

donde β_{ij} con $j < i$ es el coeficiente de regresión de X_j en la regresión de X_i de sus nodos podres y v_i es la varianza condicionada de X_i dado sus nodos padres en el DAG (García, 2007). Esta es la teoría detrás de las Redes Bayesianas Gaussianas.

Existen varias tipos dentro de la Red Bayesiana Gaussiana, pero las utilizadas en el proyecto son las Redes Bayesianas Gaussianas Lineales y las Redes Bayesianas Gaussianas No Lineales (Modelos no paramétricos).

Red Bayesiana Gaussiana Lineal

Este tipo de Red Bayesiana Gaussiana se enfoca en modelar relaciones lineales entre variables continuas utilizando distribuciones gaussianas. En este enfoque se asume que las relaciones entre variables puede ser aproximada de manera lineal. En esta red, todos los nodos siguen una distribución normal. Pero los nodos que no tiene padre tienen una distribución univariada normal marginal, explicada anteriormente. El efecto condicional de los nodos padres se describe como una combinación lineal aditiva en la media μ y no afecta la varianza σ^2 . Cada nodo tiene una varianza específica que no depende de sus padres y la distribución local de cada nodo puede representarse como un modelo lineal Gaussiano. Esta relación lineal entre variables, es decir, entre los nodos y sus padres, toma la forma de:

$$X = \beta_1 U_1 + \beta_2 U_2 + \ldots + \beta_n U_n + W_x$$

donde X es la variable, U_i son los padres de X, β son los coeficientes constantes y W representa el ruido gaussiano con media 0 (Sucar, s. f.).

Red Bayesiana Gaussiana No paramétrico (No Lineal)

A comparación de las redes lineales, estas permiten modelar relaciones

probabilísticas más complejas y no lineales entre variables continuas. Se combina la flexibilidad de las distribuciones gaussianas con la capacidad de capturar relaciones no lineales en los datos. Como en una Red Bayesiana, los arcos en el grafo indican dependencia probabilística entre variables, pero la diferencia clave radica en la naturaleza no lineal de las dependencias. La relación no lineal de una Red Bayesiana Gaussiana No Paramétrico tiene la forma de:

$$X = f_1(U_1) + f_2(U_2) + \ldots + f_n(U_n) + W_x$$

donde X es la variable, U_i son los nodos padres de X, $f_i(U_i)$ son funciones no lineales que transforman las variables de los nodos padres y W es el ruido (Córdoba, I.).

Metodología

1. Lectura de la Base de datos.

Como este proyecto consta de dos partes, hacer Redes bayesianas lineales y no lineales, se utilizarán dos bases de datos. Son las mismas, lo único que cambia es la variable 'Gravedad', ya que en los modelos lineales no tendrá ruido y en la de los modelos no paramétricos esta variable tendrá un poco de ruido, esto debido a la naturalidad del problema. La base de datos utilizada para este trabajo contiene información sobre diferentes marcadores para una muestra de pacientes con y sin Enfermedad Crónica Renal. Para fines del proyecto, la base de datos fue previamente limpiada, haciendo selección de variables continuas y rellenando los valores faltantes con imputación simple de la media. Cuenta con 397 filas y 12 columnas, las cuales son: 'Urea', 'Presion', 'Glucosa', 'Edad', 'Creatinina', 'Gravedad' (densidad de la urina), 'CelulasBlancas' (conteo de células blancas), 'Hemoglobina', 'Potasio', 'Sodio', 'CelulasRojas' (conteo de células rojas) y 'Hematrocito' (volumen celular).

Base de datos para los modelos lineales (sin ruido):

```
  Edad Presion Gravedad  Glucosa Urea Creatinina    Sodio
Potasio Hemoglobina
1   48      80    1.020 121.0000   36        1.2 137.5919
4.629126        15.4
2    7      50    1.020 148.0113   18        0.8 137.5919
```

```
4.629126        11.3
3   62      80    1.010 423.0000   53        1.8 137.5919
4.629126         9.6
4   48      70    1.005 117.0000   56        3.8 111.0000
2.500000        11.2
5   51      80    1.010 106.0000   26        1.4 137.5919
4.629126        11.6
6   60      90    1.015  74.0000   25        1.1 142.0000
3.200000        12.2
  Hematocrito CelulasBlancas CelulasRojas
1          44           7800     5.200000
2          38           6000     4.707435
3          31           7500     4.707435
4          32           6700     3.900000
5          35           7300     4.600000
6          39           7800     4.400000
```

Base de datos para modelos no paramétricos (con ruido):

```
  Edad Presion  Glucosa Urea Creatinina    Sodio  Potasio
Hemoglobina
1   48      80 121.0000   36        1.2 137.5919 4.629126
15.4
2    7      50 148.0113   18        0.8 137.5919 4.629126
11.3
3   62      80 423.0000   53        1.8 137.5919 4.629126
9.6
4   48      70 117.0000   56        3.8 111.0000 2.500000
11.2
5   51      80 106.0000   26        1.4 137.5919 4.629126
11.6
6   60      90  74.0000   25        1.1 142.0000 3.200000
12.2
  Hematocrito CelulasBlancas CelulasRojas gravedad_2
1          44           7800     5.200000   1.019752
2          38           6000     4.707435   1.020384
3          31           7500     4.707435   1.009931
4          32           6700     3.900000   1.004590
5          35           7300     4.600000   1.010434
6          39           7800     4.400000   1.014586
```

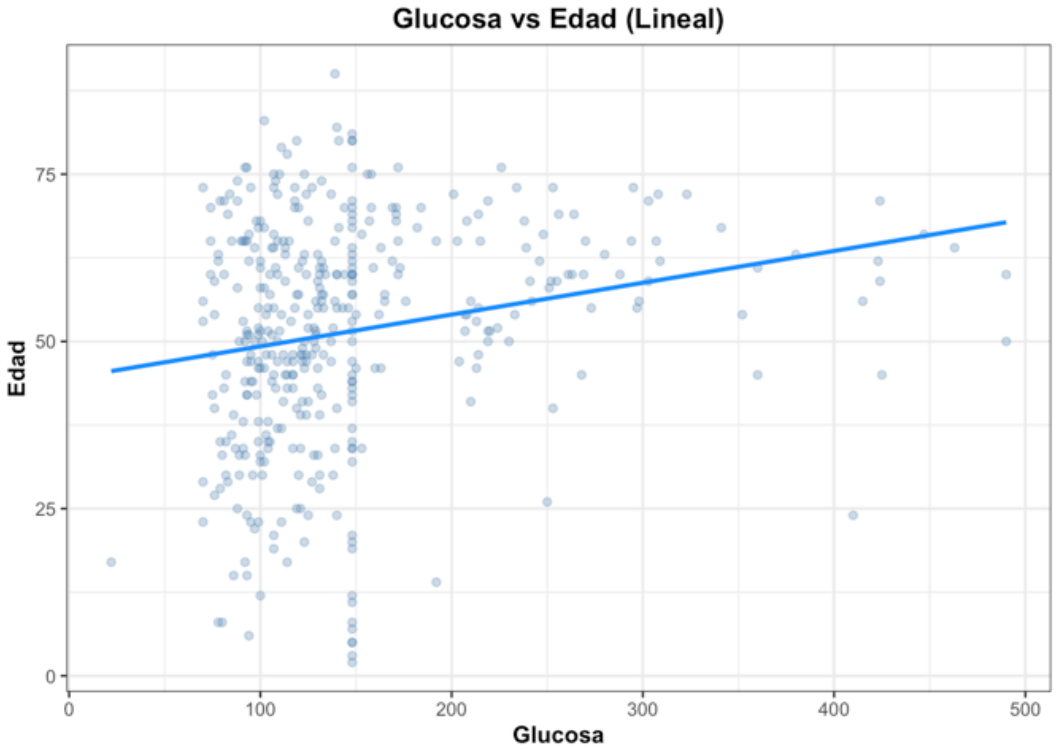
2. Análisis de los datos

Se debe analizar el comportamiento de la base de datos, para saber si siguen un comportamiento lineal o no lineal. La mejor manera de hacer esto es graficando los datos, pero solo es necesario graficarlo con una variable del conjunto de datos. Para estas visualizaciónes se utilizará con

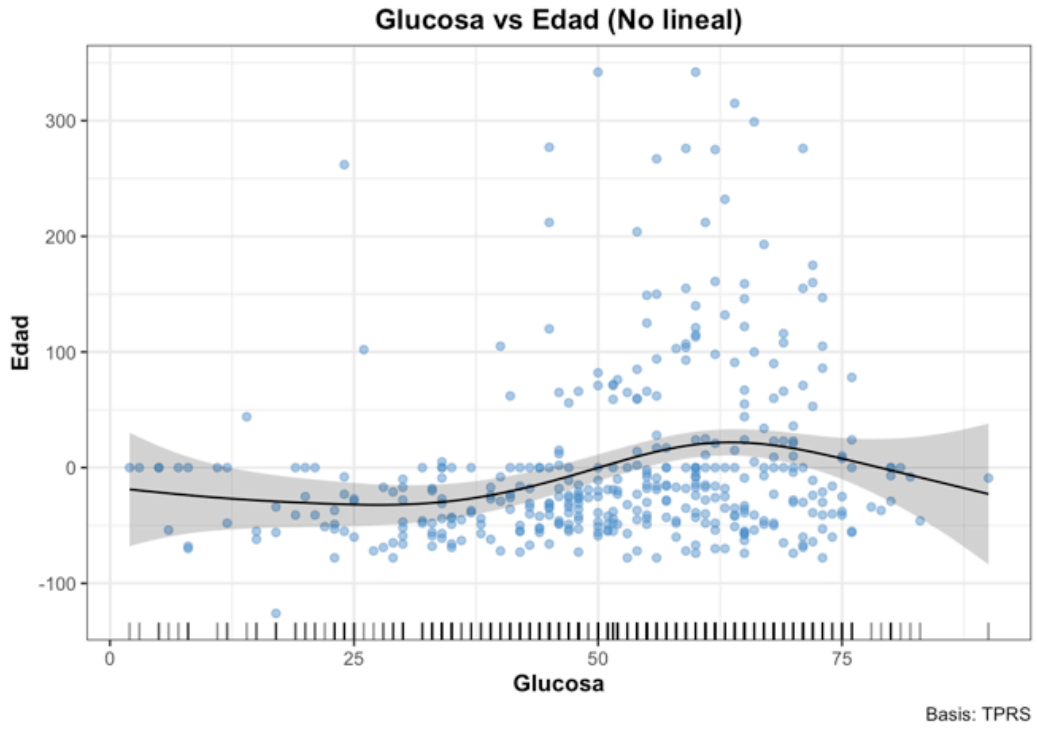
la variable ‘Glucosa’.

Modelo Lineal:

```
`geom_smooth()` using formula = 'y ~ x'
```



Modelo no paramétrico:



Claramente se puede observar que los datos tienen un comportamiento

no lineal, por lo que, las DAGs se tratarán como una Red Bayesiana Gaussiana No Paramétrica.

3. Investigación previa: Enfermedad Crónica Renal

En primera estancia, para proponer una estructura de una DAG se necesita tener contexto sobre la Enfermedad Crónica Renal y como las variables que se tienen en la base de datos dependen de padecerla o no. La Enfermedad Crónica Renal (ERC) es un padecimiento médico en el cual los riñones pierden su funcionalidad. Los riñones desempeñan un papel muy importante en el cuerpo humano: eliminan desechos y toxinas del cuerpo, regulan el equilibrio de líquidos y electrolitos, producen hormonas importantes para la presión arterial y la formación de glóbulos rojos (Lorenzo y Luis, 2022). La gravedad de la ERC se ha clasificado en 5 categorías dependiendo del estado del paciente, durante las dos primeras etapas las funciones del cuerpo siguen trabajando de manera adecuada , pero cuando alcanza una etapa avanzada en el cuerpo se acumulan niveles peligrosos de líquidos, electrolitos y desechos (González, 2011).

La presión arterial (fuerza ejercida por la sangre contra las paredes de las arterias) suele aumentar con la edad, debido a que con el tiempo, las arterias se pueden volver menos elásticas e incrementar la resistencia al flujo sanguíneo (‘Presión arterial alta’, s.f.). A medida que las personas envejecen tienden a crear resistencia a la insulina lo que puede resultar en niveles elevados de glucosa (azúcar simple que da energía al cuerpo) , y mientras más edad tenga un paciente la función de los riñones se ve comprometida y su dieta suele cambiar en la cantidad de proteína que consumen lo que afecta directamente los niveles de la urea en sangre (parámetro bioquímico utilizado para evaluar la función renal que se produce principalmente en el hígado y se excreta a través de los riñones) (López, 2022). Los glóbulos blancos son células sanguíneas que protegen al organismo contra infecciones y enfermedades, cuando la presión arterial es alta, el cuerpo puede llegar a tener respuestas inflamatorias, lo que hace que los glóbulos blancos estén más activos y acumularse en áreas dañadas.

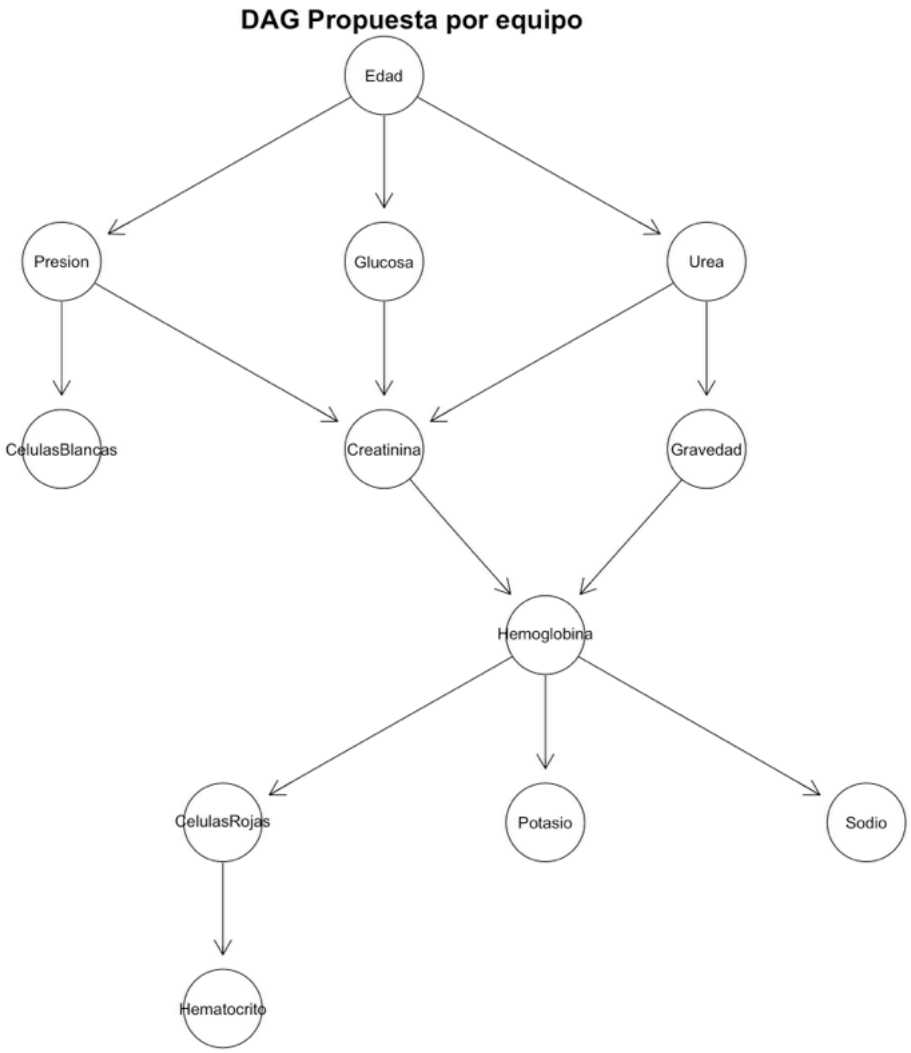
La creatinina es un compuesto químico orgánico que se encuentra en el músculo y se forma como resultado de la degradación natural de la creatina (sustancia utilizada por los músculos para obtener energía). La creatinina se excreta principalmente a través de los riñones y se elimina del cuerpo a través de la orina, cuando el cuerpo tiene niveles elevados de

glucosa, presión arterial y urea en sangre, la creatinina se ve afectada y comienza a tener indicios de problemas renales (‘Prueba de creatinina’, s. f.). Así mismo, mientras más alto tenga una persona los niveles de urea, la personas es más propensa a la deshidratación lo que puede provocar que la gravedad específica en la orina se altere.

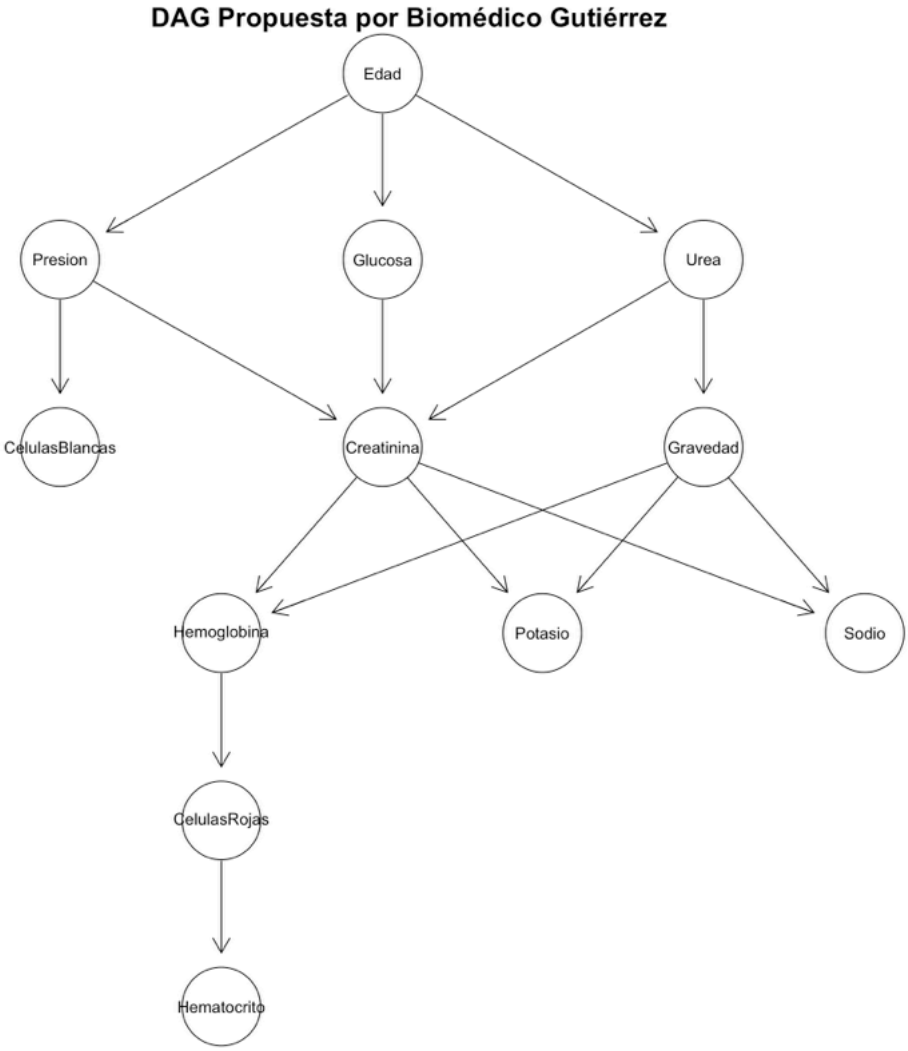
La hemoglobina es el componente principal de los glóbulos rojos y están formados por una proteína llamada hemo, cuya función es fijar el oxígeno para intercambiar el dióxido de carbono. Si los riñones no funcionan correctamente los niveles de creatinina se ven afectados y el cuerpo puede tener dificultades en la hormona de eritropoyetina que transporta la hemoglobina, lo mismo pasa cuando el cuerpo tiene gravedad específica de la orina anormal, lo que indica problemas renales y no se produce eritropoyetina suficiente (Standford Medicine, s. f.). La hemoglobina a su vez contienen potasio en su interior y el equilibrio de este es crucial para mantener la función celular normal, incluida la función de los glóbulos rojos. Los glóbulos rojos también están involucrados en la regulación de los niveles de sodio en el cuerpo debido a su influencia en el equilibrio ácido-base. Por último, el hematocrito indica la proporción de volumen que ocupa los glóbulos rojos en la sangre, cuando los glóbulos rojos aumentan el hematocrito igual y de la misma manera cuando los glóbulos disminuyen (MedicalNewsToday, s.f.).

4. DAG propuesta

Con la información encontrada, se pudo proponer la siguiente DAG:

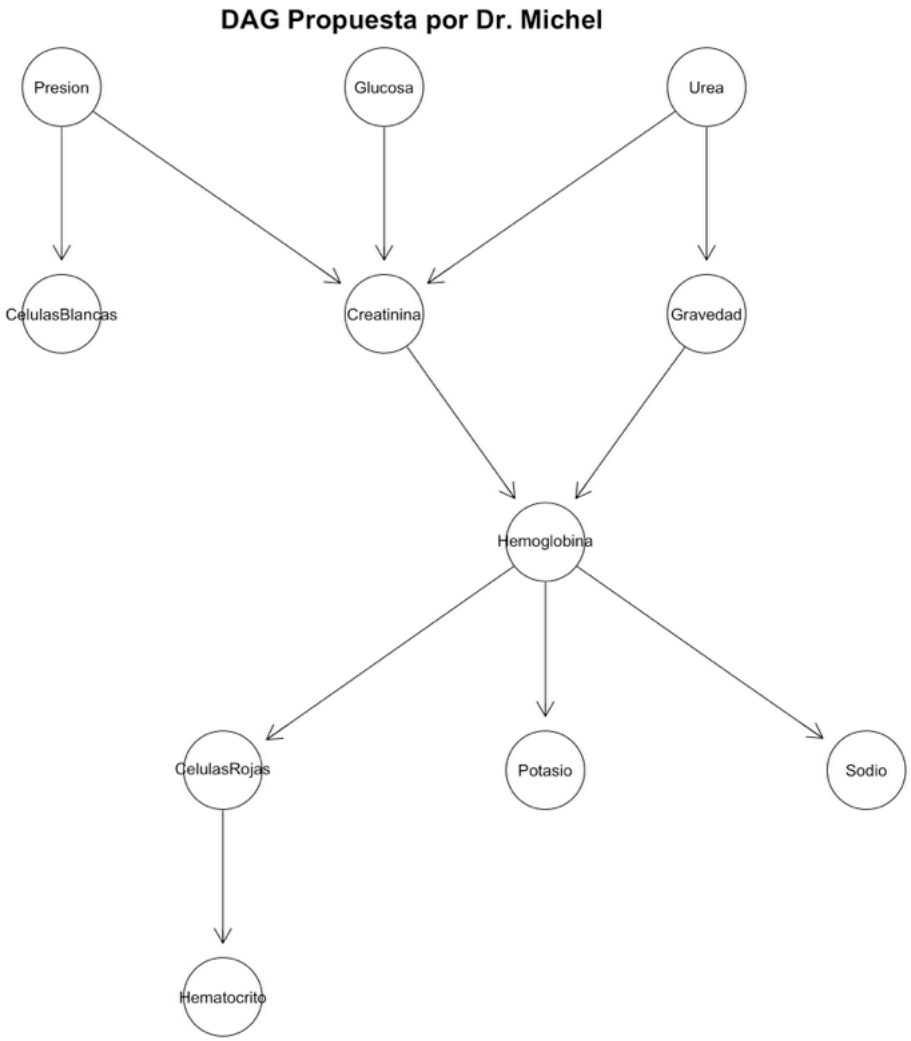


5. Modificación DAG: Biomédico Abel Gutiérrez



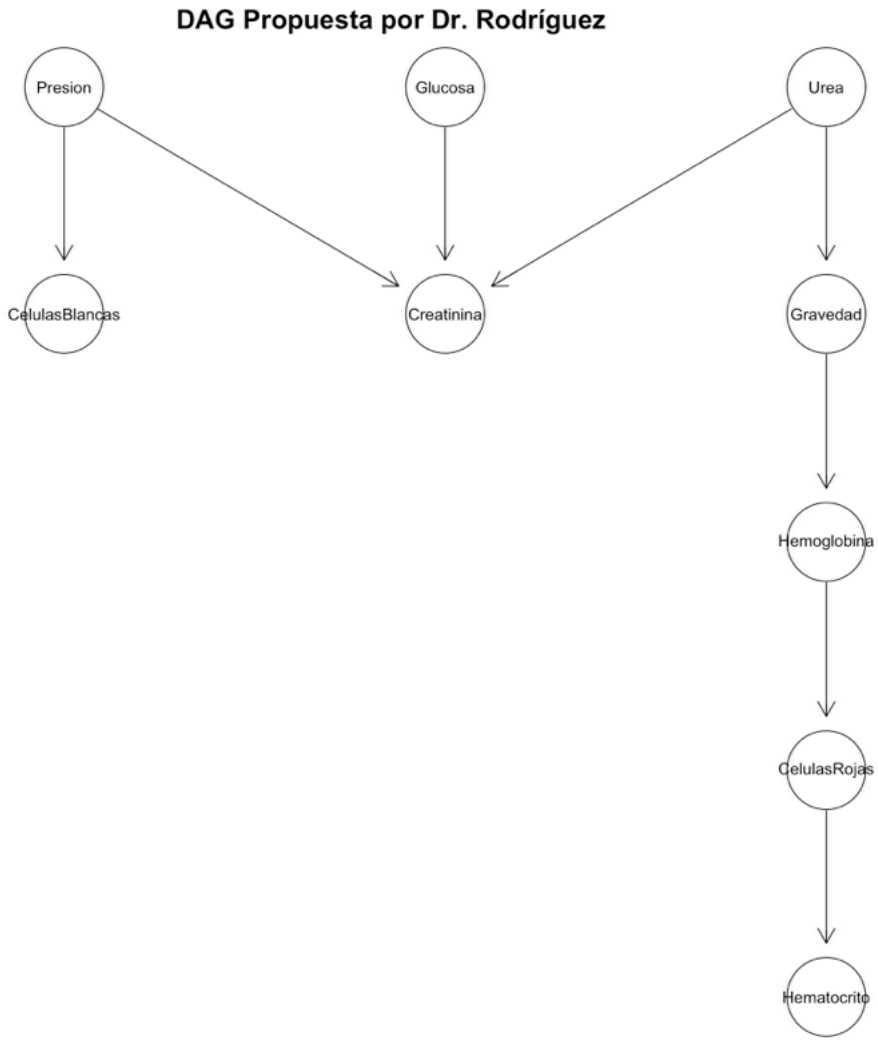
El DAG propuesto por el Maestro Biomédico Abel Gutiérrez es parecido al del equipo, los cambios realizados fue que el Potasio y Sodio no dependan de la Hemoglobina. En su lugar, el Potasio y Sodio dependieran de la Urea y de la Creatinina, al igual que la Hemoglobina.

6. Modificación DAG: Doctor Jorge Michel



La moficación del Doctor con Especialidad en Nefrología Jorge Michel fue quitar el Edad como el nodo padre, ya que esta variable no condiciona la enfermedad. Comentó que pueden haber niños con ECR y adultos, por lo que no era una buena aseveración a la DAG propuesta por el equipo. Por lo que, los nodos padres de esta nueva DAG son Presión, Urea y Glucosa.

7. Modificación DAG: Doctor Sergio Rodríguez



8. Comparación de resultados: BIC y AIC

Teniendo todas las DAGs propuestos por los experto en el área de Salud, se puede hacer una comparación de lo valores de BIC y AIC para obtener el mejor modelo para responder a las hipótesis en la sección de Aplicación.

8.1 Tabla de modelos lineales

	BIC Score	AIC Score
Modelo Lineal Equipo	−15980.29	−15904.59
Modelo Lineal IMD GutiÃ©rrez	−15898.83	−15819.15
Modelo Lineal Dr. Michel	−14301.90	−14236.16
Modelo Lineal Dr. RodrÃguez	−11902.00	−11850.21

8.2 Tabla de modelos no paramétricos

	BIC Score	AIC Score
Modelo no paramétrico Equipo	−15940.91	−15779.04
Modelo no paramétrico IMD Gutiérrez	−15676.92	−15485.15

```
Modelo no paramétrico Dr. Michel    −14000.40 −13832.60
Modelo no paramétrico Dr. Rodríguez −11895.66 −11780.16
```

En general, los modelos no paramétricos tenían mejores valores en las métricas de BIC y AIC, que sustenta la hipótesis de que los datos no siguen un comportamiento lineal. De los modelos no paramétricos, el que tienes mejores puntajes y mejor rendimiento fue el del Doctor Sergio Rodríguez.

9. ¿Cómo incluir variables categóricas a la Red Bayesiana Gaussiana?

Según Scutari y Denis (2014) existe una manera de implementar las variables categóricas en las Redes Bayesianas Gaussianas y que se puede usar cualquier tipo de distribución. La manera de hacer esto es con los Métodos de Montecarlo basados en cadenas de Markov (MCMC) mediante la implementación de BUGS, especialmente JAGS (Just Another Gibbs Sampler). MCMC es una cadena construida que eventualemente converge a la distribución posterior y muestrea directamente esta probabilidad. JAGS es un software que programa cadenas de MCMC para modelos bayesianos y es un sucesor de BUGS, que es Bayesian inference using Gibbs sampling. Para implementar esto, primero se necesita definir el modelo. BUGS es un lenguaje declarativo y no un lenguaje de programación, por ejemplo, en R los operadores de asignación <- y ~ definen los valores y las distribuciones de cada nodo en la Red Bayesiana, que no pueden modificarse posteriormente. La sintaxis es muy concisa, una linea de código para cada nodo:

```
model {
  csup ~ dcat(sp);
  cdiam ~ dnorm(mu[csup], 1/sigma^2);
}
```

El primer nodo es categórico y el segundo numérico continuo. El primer nodo categórico sigue una distribución categórica (dcat) y assume valores $s1$ y $s2$. El argumento sp es un vector que provee las probabilidades de las dos categorías. El segundo nodo sigue una distribución normal. La dependencia entre los dos nodos se introduce por la presencia de $csup$ en los argumentos que definen la distribución de $cdiam$, exactamente igual que en una fórmula matemática. Esta similitud es una de las principales características de la codificación BUGS. La manera de adaptar BUGS a R, sería mediante JAGS y los paquetes existentes en R (Scutari y Denis, 2014). Y esta es una manera de cómo se pueden implementar

variables categóricas en una Red Bayesiana Gaussiana.

Aplicación

Se analizó anteriormente las métricas para evaluar las diferentes propuestas de las DAGs y se obtuvo que la mejor DAG con el mejor rendimiento es la propuesta por el Doctor Sergio Rodríguez. Esta será la DAG que se usará para responder las hipótesis (queries).

Queries planteadas:

1. *¿Cuál es la probabilidad de que un indivduo tenga más de 11000 células blancas en la sangre dado que su presión diastólica esté entre 80 y 110?*

   La probabilidad de que el individuo tenga mas de 11,000 células blancas dado que su presión está en un rango de 80 a 110 es de 15.95%.

2. *Si una persona tiene urea en la sangre mayor a 40, ¿cuál es la probabilidad de que la densidad de su orina esté entre 1.005 y 1.015?*

   La probabilidad de que la persona tenga densidad en su orina entre 1.005 y 1.015 dado que la urea en su sangre es mayor a 40 es de 37.16%.

3. *¿Cuál es la probabilidad de que un paciente de enfermedad crónica renal tenga un volumen celular (hematocrito) entre 40 y 60, dado que su conteo de células rojas sea menor o igual a 4.3 y su hemoglobina en la sangre menor a 13?*

   Si un paciente con ECR tiene un conteo de celular menor o igual a 4.3 y su hemoglobina en la sangre menor a 13, la probabilidad de que tenga hematocrito entre 40 y 60 es de 94.16%.

4. *Si un paciente de enfermedad crónica renal tiene su presión entre 105 y 130, ¿cuál es la probabilidad de que tenga un conteo de células blancas menor a 3000?*

   La probabilidad de que el paciente con ECR tenga un conteo de células blancas menor a 3000 dado que su presión esté en un rango de 105 a 130 es de 1.53%.

Conclusión

Este trabajo comprueba la efectividad y utilidad de las Redes Bayesianas Gaussianas, ya que demuestra que son herramientas esenciales para el análisis de datos complejos que presentan tanto relaciones lineales como no lineales. Estas redes proporcionan una representación visual intuitiva de las relaciones dependientes entre variables, permitiendo modelar y cuantificar la incertidumbre de manera coherente. Su capacidad para modelar con precisión estas relaciones, junto con su enfoque en la inferencia probabilística, las convierte en una opción poderosa para la exploración y validación de hipótesis en diversas disciplinas, desde la ciencia de datos hasta la investigación científica y médica.

Referencias

Análisis de hematocrito. (s.f). *Mayo Clinic*. Recuperado de https://www.mayoclinic.org/es/tests-procedures/hematocrit/about/pac-20384728

Córdoba, I. (2015). Fusión de redes Bayesianas Gaussianas. *Universidad Politécnica de Madrid*. Recuperado de https://oa.upm.es/39091/1/TFM_CORDOBA_SANCHEZ_IRENE.pdf

Enfermedad crónica del riñón. (s.f.). *OPS*. Recuperado de https://www.paho.org/es/temas/enfermedad-cronica-rinon

Descripción general de la sangre y sus componentes. (s.f.). *Stanford Medicine*. Recuperado de https://www.stanfordchildrens.org/es/topic/default?id=overview-of-blood-and-blood-components-90-P05425

García, R. S. (2007). Análisis de Sensibilidad en Redes Bayesianas Gaussianas. *Universidad Complutense de Madrid*. Recuperado de https://docta.ucm.es/rest/api/core/bitstreams/d9376029-fba2-40aa-960f-b69d2f182e6d/content

González, O. (2011). Envejecimiento y función renal. *Mecanismos de predicción y progresión*. Recuperado de https://www.revistanefrologia.com/es-envejecimiento-funcion-renal-mecanismos-prediccion-articulo-X2013757511000284

López, G. (2022). Urea alta en análisis : ¿qué significa y cómo tratarla?.

Savia. Recuperado de https://www.saludsavia.com/contenidos-salud/articulos-especializados/urea-alta-en-analisis-que-significa-y-como-tratarla

Lorenzo, V. & Luis, D. (2022). *Enfermedad Renal Crónica*. Nefrología al día. Recuperado de https://www.nefrologiaaldia.org/es-articulo-enfermedad-renal-cronica-136

Niveles de hematocrito: Definición, niveles bajos, niveles altos, y más.(s.f). *MedicalNewsToday*. Recuperado de https://www.medicalnewstoday.com/articles/es/niveles-de-hematocrito

Presión arterial alta.(s.f.). *Medlineplus* .Recuperado de https://medlineplus.gov/spanish/highbloodpressure.html

Prueba de glucosa en sangre. (s.f.) *Medlineplus*. Recuperado de https://medlineplus.gov/spanish/pruebas-de-laboratorio/prueba-de-glucosa-en-la-sangre/

Prueba de creatinina. (s.f). *Medlineplus*. Recuperado de https://medlineplus.gov/spanish/pruebas-de-laboratorio/prueba-de-creatinina/

Scutari, M. y Denis J.B. (2014). *Bayesian Networks with Examples in R*. CRC Press.

Sucor, L. (s. f.). Redes Bayesianas: extensiones y aplicaciones. *INAOE*. Recuperado de https://ccc.inaoep.mx/~esucar/Clases-mgp/pgm13-rbeya-2012.pdf