Data preprocessing

AUTHORS
Mayra Sarahí de Luna Castillo A01635774
Juan Manuel Hernández Solano A00572208
Alejandra Velasco Zárate A01635453
José Antonio Juárez Pacheco A0057218
José Carlos Yamuni Contreras A01740285

PUBLISHED
August 16, 2023

Abstract

Introduction

Bayesian networks provide a visual representation for a set of random variables and the relationships between them. The structure of these networks allows us to specify the joint probability function of the variables as the product of conditional probability functions. The main difference between these models is found in their arcs since they are directed and represent conditional dependence between variables. The objective of this work is to use Bayesian networks to make inferences about some hypotheses and obtain the probability that they are true or not, in order to know the relationships and dependencies that exist between variables and/or events.

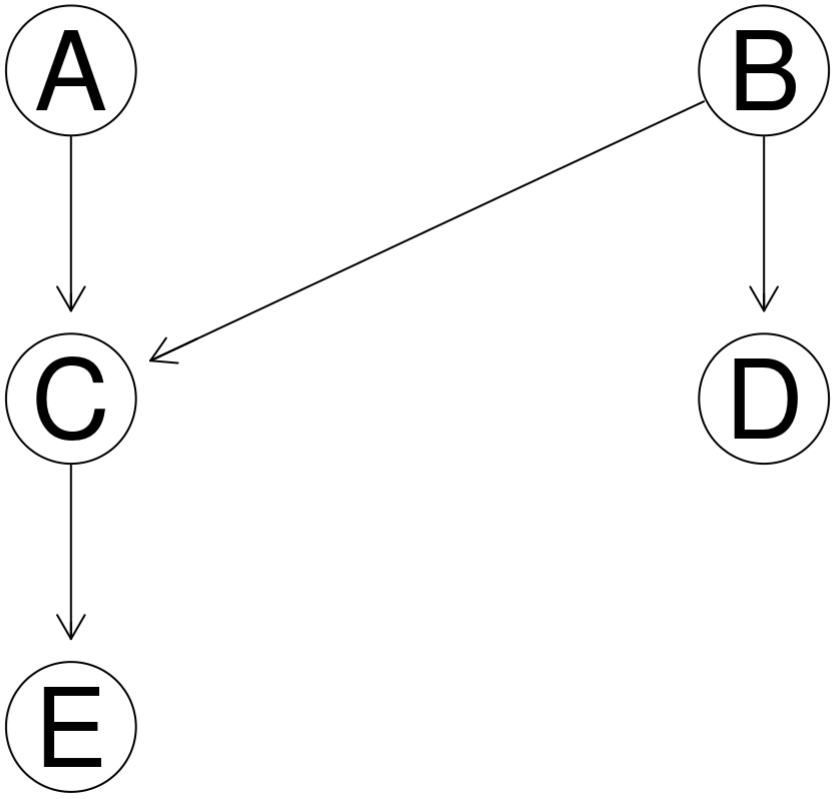
Theoretical framework

Bayesian networks, also known as probabilistic causal networks, are statistical tools that represent a set of associated uncertainties based on the conditional independence relationships established between them. They are directed acyclic graphs in which each node represents a random variable that has an associated conditional probability function (Santiesteban, 2012).

The probabilistic model is described by a directed acyclic graph (DAG), where the vertices of the graph that represent the variables are called nodes. These nodes are represented as circles that contain the name of the variable inside and the connections between nodes are called arcs. These arcs have an arrow ending, which indicates the dependency between variables. The node where the arc originates is called the parent, while the node where the arc ends is called the children. Nodes that can be reached from other nodes are called descendants. Nodes that lead a path to a specific node are called ancestors. The main point of Bayesian Networks is to allow probabilistic inference to be made.

```
Loading required namespace: Rgraphviz
```

DAG Ejemplo



En esta DAG los nodos padre son A y B, el nodo hijo es el E. C y E son descendientes de A y A y C son ancestros de E. En una red bayesiana no hay bucles ni ciclos, ya que ningún nodo puede ser su propio antepasado o descendiente.

Distribuciones de probabilidad conjunta

La probabilidad conjunta es la probabilidad de que una serie de eventos sucedan simultáneamente. La probabilidad conjunta de varias variables se puede calcular a partir del producto de probabilidades indivudales de los nodos.

$$P(X_1, \ldots, X_n) = \Pi_{i=1}^{n} P(X_i \mid parents(X_i))$$

En el ejemplo propuesto , la distribución conjunta de probabilidad es:

$$P(A, B, C, D, E) = P(A)P(B)P(C \mid A, B)P(D \mid B)P(E \mid C)$$

Si un nodo no tiene un padre, como el nodo A, su distribución de probabilidad se describe como incondicional. De lo contrario, la distribución de probabilidad local del nodo está condicionada a otros nodos (Wolf et al., 2019).

Teorema de Bayes

El Teorema de Bayes parte de una situación en la que es posible conocer las probabilidades de que ocurran una serie de sucesos A_i . Se tiene un evento B cuya ocurrencia proporciona información, ya que las probabilidades de que ocurra B son distintas si el suceso A_i sucede.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Donde $P(A)$ es la probabilidad a priori, $P(B \mid A)$ es la probabilidad condicional, $P(B)$ es la probabilidad total y el resultado $P(A \mid B)$ la probabilidad a posteriori.

Esta es la teoría detrás de las redes bayesianas.

Inferencia

A partir de una red ya construida, y dados los valores concretos de algunas variables de una instancia, podrían tratar de estimarse los valores de otras variables de la misma instancia aplicando razonamiento probabilístico. El razonamiento probabilístico sobre las redes bayesianas consiste en propagar los efectos de las evidencias (variables conocidas) a través de la red para conocer las probabilidades a posteriori de las variables desconocidas. De esta manera se puede determinar un valor estimado para dichas variables en función de los valores de probabilidad obtenidos (Santiesteban, 2012).

Con la metodología se puede ver la creación y aplicación de la DAG con ciertos datos para responder unas preguntas.

Metodología

1. Lectura y análisis de los datos

Importación de librerías necesarias para redes bayesianas.

```
library(bnlearn)
```

```
#if (!requireNamespace("BiocManager", quietly = TRUE))
#  install.packages("BiocManager")
#BiocManager::install()
#BiocManager::install(c("graph", "Rgraphviz"))
```

Lectura de la base de datos final, con las variables necesarias para responder las queries establecidas.

```
data <- read.csv("data.csv")
head(data)
```

```
  Residencia   Sexo         Edad    Educacion      ing_fam          Transporte
1     grande  mujer adulto_mayor     primaria bajo ingreso           Automóvil
2     grande  mujer adulto_mayor preparatoria alto ingreso           Automóvil
3    pequeño  mujer adulto_mayor preparatoria bajo ingreso           Automóvil
```

```
4     grande hombre adulto_mayor     primaria bajo ingreso Tren urbano (Metro)
5    pequeño  mujer adulto_mayor     primaria bajo ingreso   Colectivo (Combi)
6     grande  mujer adulto_joven preparatoria alto ingreso                Taxi
  Eficiencia Seguridad          Ocupación
1  eficiente    seguro vendedor_ambulante
2  eficiente  inseguro               jefe
3  eficiente    seguro           empleado
4  eficiente  inseguro        comerciante
5  eficiente  inseguro        comerciante
6  eficiente  inseguro               jefe
```

Dimensión de la base de datos.

```
dim(data)
```

```
[1] 1191    9
```

Verificar datos faltantes

```
sum(is.na(data))
```

```
[1] 0
```

Conversión de variables a factor para el método MLE

```
data$Residencia<-as.factor(data$Residencia)
data$Sexo<-as.factor(data$Sexo)
data$Edad<-as.factor(data$Edad)
data$Educacion<-as.factor(data$Educacion)
data$ing_fam<-as.factor(data$ing_fam)
data$Transporte<-as.factor(data$Transporte)
data$Eficiencia<-as.factor(data$Eficiencia)
data$Seguridad<-as.factor(data$Seguridad)
data$Ocupación<-as.factor(data$Ocupación)
```

2. Creación de las DAGs

DAG 1

```
DAG<-empty.graph(nodes = c("Edad", "Sexo", "ing_fam", "Educacion", "Ocupación", "Residenc
```

Creación de relación y nodo entre variables

```
arc.set<-matrix(c("Edad", "Educacion",
                  "Sexo", "Educacion",
                  "ing_fam", "Educacion",
                  "Educacion", "Ocupación",
                  "Educacion", "Residencia",
```

```
                  "Ocupación", "Transporte",
                  "Residencia", "Transporte",
                  "Transporte", "Eficiencia",
                  "Transporte", "Seguridad"), byrow = TRUE, ncol = 2,
                dimnames = list(NULL, c("from", "to")))
arc.set
```

```
     from         to
[1,] "Edad"       "Educacion"
[2,] "Sexo"       "Educacion"
[3,] "ing_fam"    "Educacion"
[4,] "Educacion"  "Ocupación"
[5,] "Educacion"  "Residencia"
[6,] "Ocupación"  "Transporte"
[7,] "Residencia" "Transporte"
[8,] "Transporte" "Eficiencia"
[9,] "Transporte" "Seguridad"
```

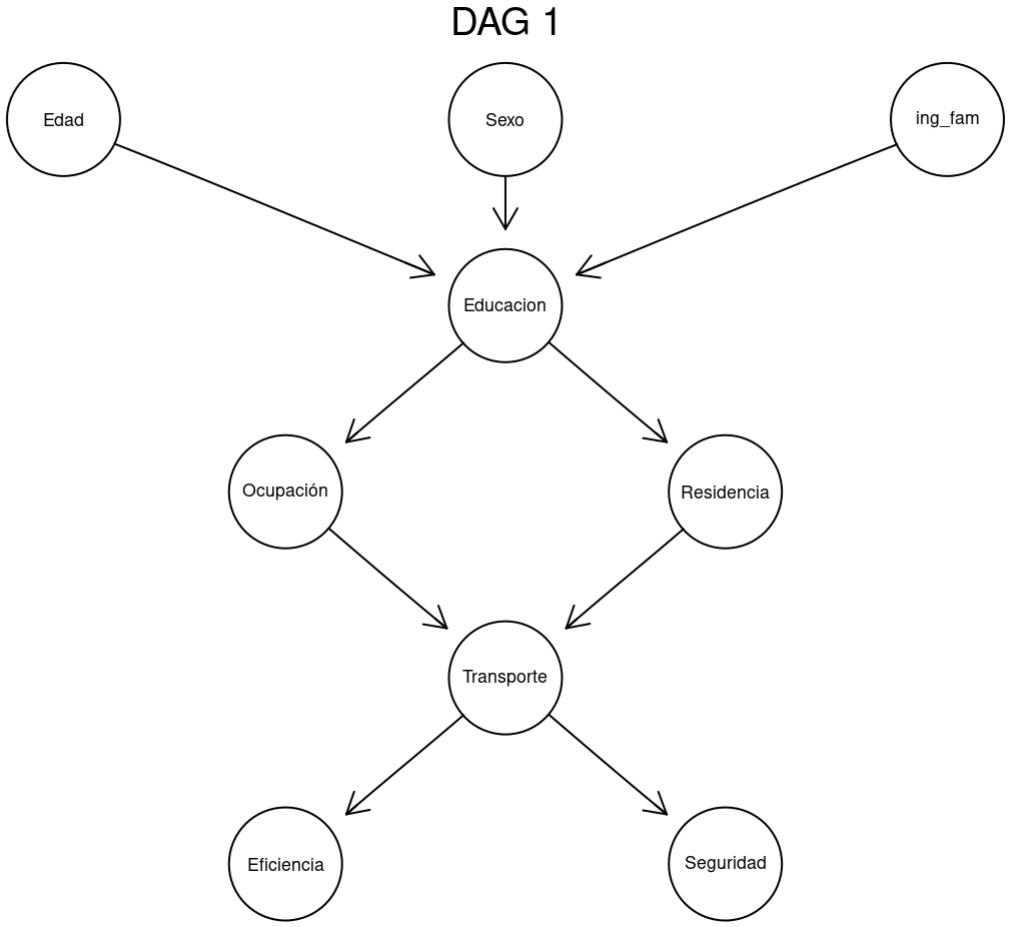
Implementación de los nodos a la DAG 1

```
arcs(DAG)<-arc.set
DAG
```

```
  Random/Generated Bayesian network

  model:
   [Edad][Sexo][ing_fam][Educacion|Edad:Sexo:ing_fam][Ocupación|Educacion]
   [Residencia|Educacion][Transporte|Ocupación:Residencia]
   [Eficiencia|Transporte][Seguridad|Transporte]
  nodes:                                9
  arcs:                                 9
    undirected arcs:                    0
    directed arcs:                      9
  average markov blanket size:          2.89
  average neighbourhood size:           2.00
  average branching factor:             1.00

  generation algorithm:                 Empty
```

Visualización de la DAG 1

```
graphviz.plot(DAG, main = "DAG 1")
```



La primera DAG propuesta consta de 3 nodos padres: ‘Edad’, ‘Sexo’ e ‘Ingreso familiar’, la razón de esto es que edad y sexo son características íntrinsicas del ser humano, es decir, hacen referencia a la naturaleza del ser humano, por lo que no dependen de ningún factor externo fuera de los atributos humanos. El ingreso familiar se consideró como nodo padre porque por razones estructurales de la sociedad y la economía, la educación depende del ingreso familiar. Las familias con ingresos más altos generalmente tienen más recursos disponibles para invertir en la educación de calidad de sus hijos. Por otro lado, las familias de bajos ingresos pueden tener dificultades para costear estos recursos, esto se debe a las desigualdades socioeconómicas y las limitaciones de acceso a empleos bien remunerados y esto puede perpetuar un ciclo intergeneracional de desventaja (Torres, 2020). La educación juega un papel crucial en la determinación de las oportunidades laborales y el éxito profesional de una persona, es decir, la ocupación laboral del individuo. Para ascender en la jerarquía laboral y acceder a roles de mayor responsabilidad y remuneración, a menudo se requiere una educación continua y el desarrollo de habilidades adicionales. Las personas con educación superior pueden tener más oportunidades de avanzar en sus carreras que los que no cuentan con educación (‘La educación en México y su influencia en la ocupación’, s.f.). Por otro lado, tanto la residencia como la ocupación que se tiene pueden influir en el tipo de transporte que se utiliza diariamente, ya se por distintos factores como: la distancia al trabajo, los costos y valores personales pueden influir al momento de optar por vehículos privados, transporte público, bicicletas u otras alternativas. Por último, la eficiencia y seguridad del transporte dependen del transporte más utilizado y preferido, ya que estos atributos están directamente relacionados a el medio de transporte. Bajo estos argumentos se obtuvo la primera propuesta para la DAG.

Estimación de parámetros para la DAG 1

```
bn.mle<-bn.fit(DAG, data = data, method = "mle")
```

Comprobación del método de máxima verosimilitud (MLE) con probabilidad condicional

```
bn.mle$Educacion
```

```
  Parameters of node Educacion (multinomial distribution)

Conditional probability table:

, , Sexo = , ing_fam = alto ingreso

              Edad
Educacion      adulto_joven adulto_mayor joven
  ninguno       0.000000000  0.000000000
  preescolar    0.000000000  0.000000000
  preparatoria  0.000000000  0.000000000
  primaria      0.000000000  0.000000000
  profesional   0.000000000  0.000000000
  secundaria    1.000000000  1.000000000

, , Sexo = hombre, ing_fam = alto ingreso

              Edad
Educacion      adulto_joven adulto_mayor joven
  ninguno       0.083333333  0.062500000
  preescolar    0.000000000  0.000000000
  preparatoria  0.333333333  0.291666667
  primaria      0.000000000  0.208333333
  profesional   0.083333333  0.104166667
  secundaria    0.500000000  0.333333333

, , Sexo = mujer, ing_fam = alto ingreso

              Edad
Educacion      adulto_joven adulto_mayor joven
  ninguno       0.000000000  0.023622047
  preescolar    0.016393443  0.007874016
  preparatoria  0.368852459  0.236220472
  primaria      0.073770492  0.230971129
  profesional   0.106557377  0.083989501
  secundaria    0.434426230  0.417322835

, , Sexo = , ing_fam = bajo ingreso

              Edad
Educacion      adulto_joven adulto_mayor       joven
  ninguno       0.000000000  0.000000000 0.000000000
```

```
  preescolar    0.000000000  0.000000000 0.000000000
  preparatoria  0.000000000  0.000000000 0.000000000
  primaria      0.000000000  0.000000000 0.000000000
  profesional   0.000000000  0.000000000 0.000000000
  secundaria    1.000000000  1.000000000 1.000000000

, , Sexo = hombre, ing_fam = bajo ingreso

              Edad
Educacion      adulto_joven adulto_mayor joven
  ninguno       0.000000000  0.051282051
  preescolar    0.000000000  0.000000000
  preparatoria  0.076923077  0.333333333
  primaria      0.000000000  0.051282051
  profesional   0.000000000  0.025641026
  secundaria    0.923076923  0.538461538

, , Sexo = mujer, ing_fam = bajo ingreso

              Edad
Educacion      adulto_joven adulto_mayor       joven
  ninguno       0.016949153  0.073490814 0.000000000
  preescolar    0.000000000  0.018372703 0.000000000
  preparatoria  0.254237288  0.149606299 0.000000000
  primaria      0.093220339  0.322834646 0.000000000
  profesional   0.067796610  0.062992126 0.000000000
  secundaria    0.567796610  0.372703412 1.000000000
```

Estructura de la DAG 1

```
arc.strength(DAG, data = data, criterion = "x2")
```

```
        from         to     strength
1       Edad  Educacion 1.433130e-01
2       Sexo  Educacion 1.738102e-07
3    ing_fam  Educacion 2.661524e-01
4  Educacion  Ocupación 2.652852e-16
5  Educacion Residencia 5.871046e-12
6  Ocupación Transporte 7.572555e-01
7 Residencia Transporte 1.009214e-02
8 Transporte Eficiencia 2.282981e-20
9 Transporte  Seguridad 3.256080e-06
```

DAG 2

```
DAG2<-empty.graph(nodes = c("Edad", "Sexo", "Educacion", "Ocupación","ing_fam", "Residenc
```

Creación de relación y nodo entre variables de la DAG 2

```
arc.set2<-matrix(c("Edad", "Educacion",
                   "Sexo", "Educacion",
                   "Educacion", "Ocupación",
                   "Educacion", "Residencia",
                   "Ocupación", "ing_fam",
                   "Residencia", "ing_fam",
                   "Residencia", "Eficiencia",
                   "Eficiencia", "Seguridad",
                   "Eficiencia", "Transporte"), byrow = TRUE, ncol = 2,
                 dimnames = list(NULL, c("from", "to")))
arc.set2
```

```
     from         to
[1,] "Edad"       "Educacion"
[2,] "Sexo"       "Educacion"
[3,] "Educacion"  "Ocupación"
[4,] "Educacion"  "Residencia"
[5,] "Ocupación"  "ing_fam"
[6,] "Residencia" "ing_fam"
[7,] "Residencia" "Eficiencia"
[8,] "Eficiencia" "Seguridad"
[9,] "Eficiencia" "Transporte"
```

Implementación de los nodos a la DAG 2

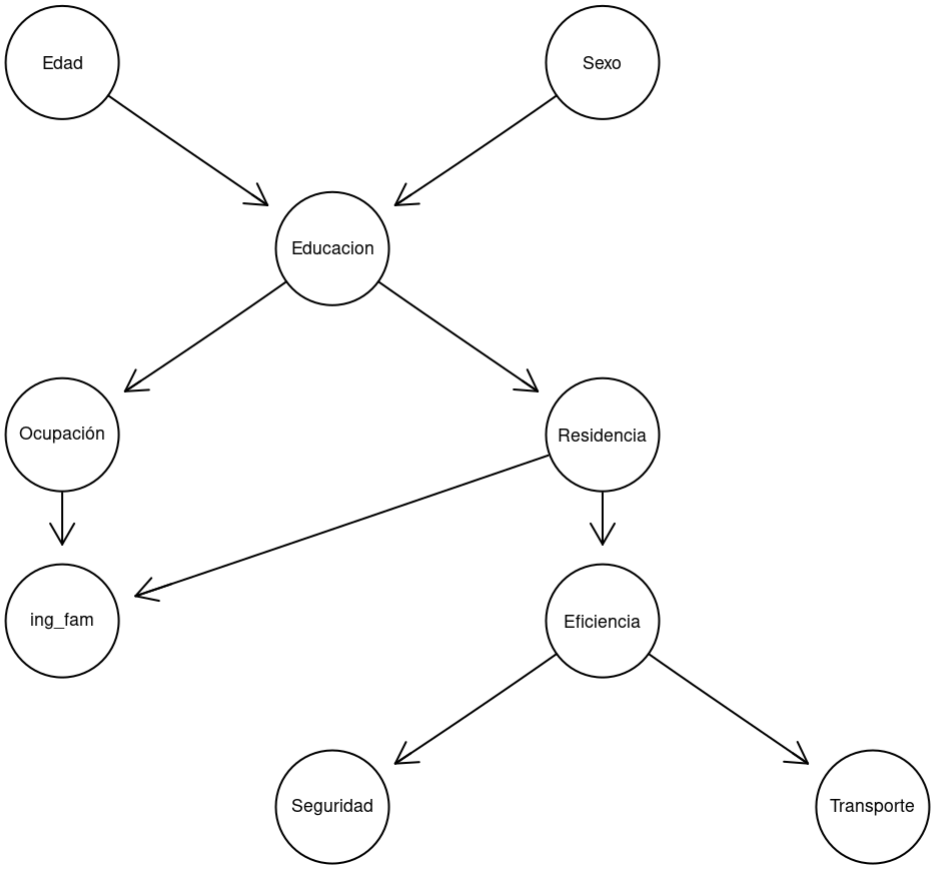
```
arcs(DAG2)<-arc.set2
DAG2
```

```
  Random/Generated Bayesian network

  model:
   [Edad][Sexo][Educacion|Edad:Sexo][Ocupación|Educacion][Residencia|Educacion]
   [ing_fam|Ocupación:Residencia][Eficiencia|Residencia][Seguridad|Eficiencia]
   [Transporte|Eficiencia]
  nodes:                                9
  arcs:                                 9
    undirected arcs:                    0
    directed arcs:                      9
  average markov blanket size:          2.44
  average neighbourhood size:           2.00
  average branching factor:             1.00

  generation algorithm:                 Empty
```

Visualización de la DAG 2

```
graphviz.plot(DAG2, main = "DAG 2")
```

DAG 2



Los cambios realizados en la segunda propuesta de la DAG fue que el nodo de ‘Ingreso familiar’ ya no es nodo padre, sino nodo descendiente de ‘Ocupación’ y ‘Residencia’. Este cambio fue porque el ingreso familiar tiende a depender de la ocupación y la residencia debido a las interacciones complejas entre factores económicos, sociales y geográficos, en primer lugar porque la ocupación suele ser la primera fuente de ingreso debido a las renumeraciones en el mercado laboral. En segundo, la residencia puede influir en el ingreso debido a factores como el costo de vida, las oportunidades laborales disponibles en un área geográfica y la presencia de industrias específicas (Agualongo y Garcés, 2020). Otro cambio fue que la eficiencia del transporte público y privado depende de la residencia, ya que la movilización o el trasporte público es una de los sectores gubernamentales que más financiamiento y planeación requieren. Depende de la zona geográfica y las características de la residencia, es el presupuesto que se tendrá para la movilización y planeación de calles, de la vía pública y de los medios de transporte. Todo esto recae directamente en le eficiencia del transporte público y privado (Calvillo y Moncada, 2008). El último cambio fue que la seguridad y el transporte más utilizado/preferido depende de la eficiencia, la razón de este cambio va mucho de la mano con la razón de que la eficiencia depende de la residencia. La eficiencia del transporte está relacionada con la rapidez y la comodidad con la que las personas pueden desplazarse de un lugar a otro. Si un medio de transporte es eficiente, es más probable que las personas lo prefieran, ya que les permite ahorrar tiempo y viajar de manera más cómoda. Así mismo, un sistema de transporte eficiente suele estar respaldado por una planificación cuidadosa de rutas y horarios. Esto puede conducir a rutas más seguras que evitan áreas peligrosas o congestionadas, reduciendo así el riesgo de accidentes y situaciones peligrosas. Es por estas razones, que la eficiencia esta directamente ligada a la elección de transporte y la seguridad de la misma.

Estimación de parámetros para la DAG 2

```
bn.mle2<-bn.fit(DAG2, data = data, method = "mle")
```

Comprobación del método de máxima verosimilitud (MLE) con probabilidad condicional

```
bn.mle2$Educacion
```

```
  Parameters of node Educacion (multinomial distribution)

Conditional probability table:

, , Sexo =

              Edad
Educacion      adulto_joven adulto_mayor       joven
  ninguno       0.000000000  0.000000000 0.000000000
  preescolar    0.000000000  0.000000000 0.000000000
  preparatoria  0.000000000  0.000000000 0.000000000
  primaria      0.000000000  0.000000000 0.000000000
  profesional   0.000000000  0.000000000 0.000000000
  secundaria    1.000000000  1.000000000 1.000000000

, , Sexo = hombre

              Edad
Educacion      adulto_joven adulto_mayor joven
  ninguno       0.040000000  0.057471264
  preescolar    0.000000000  0.000000000
  preparatoria  0.200000000  0.310344828
  primaria      0.000000000  0.137931034
  profesional   0.040000000  0.068965517
  secundaria    0.720000000  0.425287356

, , Sexo = mujer

              Edad
Educacion      adulto_joven adulto_mayor       joven
  ninguno       0.008333333  0.048556430 0.000000000
  preescolar    0.008333333  0.013123360 0.000000000
  preparatoria  0.312500000  0.192913386 0.000000000
  primaria      0.083333333  0.276902887 0.000000000
  profesional   0.087500000  0.073490814 0.000000000
  secundaria    0.500000000  0.395013123 1.000000000
```

Estructura de la DAG 2

```
arc.strength(DAG2, data = data, criterion = "x2")
```

```
        from         to     strength
1       Edad  Educacion 2.824048e-04
2       Sexo  Educacion 2.362316e-11
3  Educacion  Ocupación 2.652852e-16
4  Educacion Residencia 5.871046e-12
5  Ocupación    ing_fam 2.014204e-06
6 Residencia    ing_fam 3.271055e-08
7 Residencia Eficiencia 2.652738e-10
8 Eficiencia  Seguridad 1.119869e-56
9 Eficiencia Transporte 2.282981e-20
```

3. Evaluación del rendimiento de las DAGs

Criterios basado en la verosimilitud para probar que tan bueno son los DAGs

Bayesian Information Criterion (BIC)

DAG 1

```
score(DAG, data = data, type = "bic")
```

```
[1] -10714.4
```

DAG 2

```
score(DAG2, data = data, type = "bic")
```

```
[1] -10042.96
```

Mientras más grande sea el BIC, mejor será el modelo. DAGs con scores más altos ajustan mejor a los datos.

Akaike Information Criterion (AIC)

DAG 1

```
score(DAG, data = data, type = "aic")
```

```
[1] -9964.728
```

DAG 2

```
score(DAG2, data = data, type = "aic")
```

```
[1] -9735.463
```

Después de analizar los resultados de los métodos de rendimiento BIC y AIC para las 2 redes bayesianas propuestas anteriormente, se puede observar que ambos valores de las métricas son mayores en la

segunda DAG. Esto significa que la DAG 2 ajusta de una mejor manera los datos del trabajo y permiten tener una mejor aproximación a las probabilidades e hipótesis planteadas. Esta DAG número 2 se va a comparar con la DAG propuesta por Hill-Climbing para ver cual es la mejor y así poder hacer los queries.

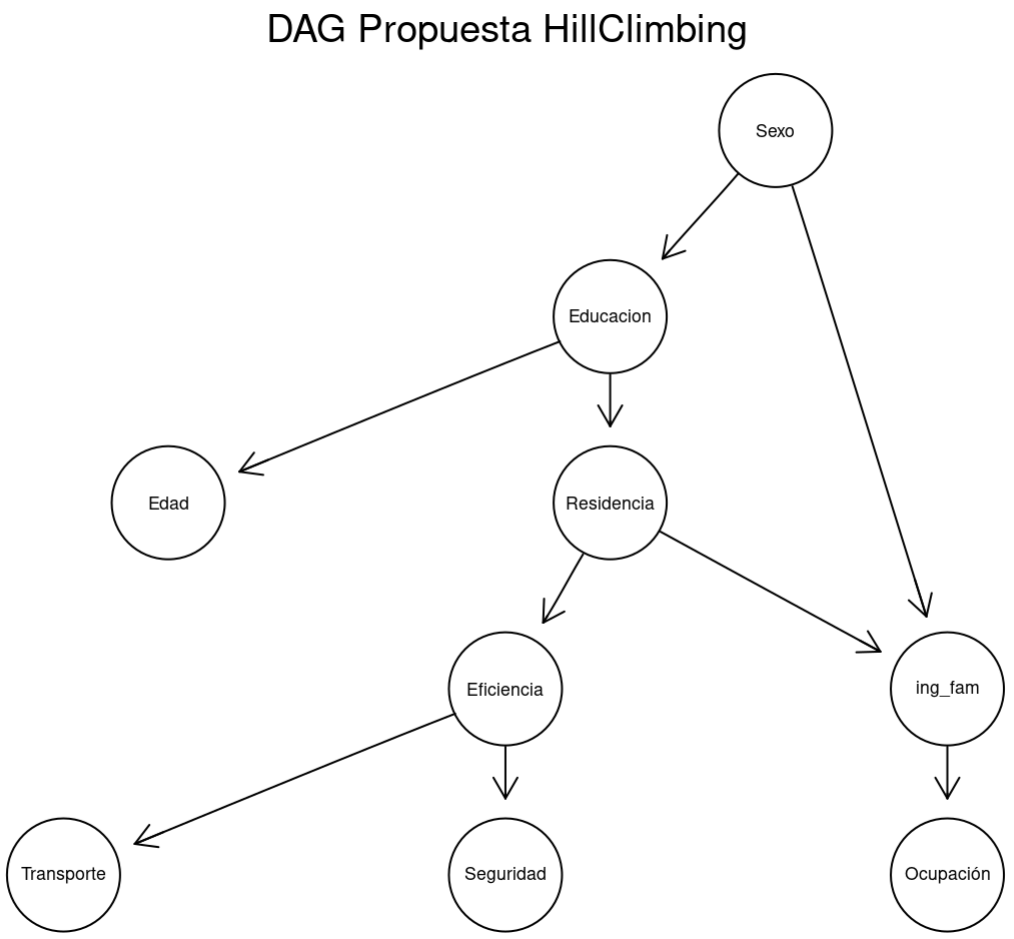
4. Optimización de la DAG seleccionada con Hill-Climbing (HC)

```
best_DAG<-hc(data)
modelstring(best_DAG)
```

```
[1] "[Sexo][Educacion|Sexo][Residencia|Educacion][Edad|Educacion]
[ing_fam|Residencia:Sexo][Eficiencia|Residencia][Transporte|Eficiencia]
[Seguridad|Eficiencia][Ocupación|ing_fam]"
```

Visualización de la nueva DAG

```
graphviz.plot(best_DAG, main = "DAG Propuesta HillClimbing")
```



Esta DAG propuesta por la función Hill-Climbing no tiene mucho sentido ya que se pueden observar nodos en los que la relación entre variables no llegan a ser coherentes. Por un lado, decir que la edad depende de la educación sería una afirmación ilógica porque la edad es una característica intrínseca y natural del tiempo transcurrido desde el nacimiento de una persona, mientras que la educación es un proceso que implica la adquisición de conocimientos, habilidades y experiencias a lo largo de la vida.

Estas 2 nociones son conceptos diferentes y no están vinculadas en términos de causalidad directa. Por otro lado, establecer que la ocupación depende del ingreso familiar podría ser incoherente porque son 2 conceptos diferentes que generalmente no están directamente relacionados en términos de causa y efecto. La ocupación se refiere al trabajo, profesión o actividad que una persona realiza para ganarse la vida, mientras que el ingreso familiar se refiere a la cantidad de dinero que una familia gana de diversas fuentes. Si bien el ingreso familiar puede influir en las decisiones de carrera de un individuo, no determina completamente la ocupación que elijan. Es por estas 2 razones, que se asenta la conclusión que la DAG propuesta de Hill-Climbing no tiene fundamentos lógicos y racionales (Benno, 1985).

5. Evaluación del rendimiento de la óptima DAG

Bayesian Information Criterion (BIC)

```
score(best_DAG, data = data, type = "bic")
```

```
[1] -9943.853
```

Akaike Information Criterion (AIC)

```
score(best_DAG, data = data, type = "aic")
```

```
[1] -9738.01
```

Se puede observar que en la métrica AIC el resultado de la DAG número 2 es ligeramente mejor que el de la DAG propuesta por la función hill-climbing. Aún así, la función hill-climbing tiene un resultado mejor en la métrica BIC con respecto a la DAG número 2 propuesta al inicio. Ambos DAG son buenos, sin embargo, la métrica BIC suele tener más peso que la métrica AIC. Es por eso, que la DAG propuesta por la función hill-climbing es una mejor estructura.

Por la misma razón de que la DAG propuesta por la función de Hill-Climbing carece de razonamiento lógico y porque las métricas BIC y AIC de ambas DAGs tienen valores cercanos, se utilizará la DAG 2 para realizar las preguntas de hipótesis.

Aplicación

1. Impresión de diccionario

Se utiliza el diccionario para tener las variables como referencia para resolver las queries.

```
unique(data$Residencia)
```

```
[1] grande  pequeño
Levels: grande pequeño
```

```
unique(data$Sexo)
```

```
[1] mujer  hombre
Levels:  hombre mujer
```

```
unique(data$Edad)
```

```
[1] adulto_mayor adulto_joven joven
Levels: adulto_joven adulto_mayor joven
```

```
unique(data$Educacion)
```

```
[1] primaria     preparatoria secundaria   profesional  ninguno
[6] preescolar
Levels: ninguno preescolar preparatoria primaria profesional secundaria
```

```
unique(data$ing_fam)
```

```
[1] bajo ingreso alto ingreso
Levels: alto ingreso bajo ingreso
```

```
unique(data$Transporte)
```

```
 [1] Automóvil            Tren urbano (Metro)  Colectivo (Combi)
 [4] Taxi                 Camión               Mototaxi
 [7] Autobús foráneo      BRT                  Motocicleta
[10] Bicicleta            Animal               Transporte eléctrico
[13] Avión                Tren                 Patineta
15 Levels: Animal Autobús foráneo Automóvil Avión Bicicleta BRT ... Tren urbano (Metro)
```

```
unique(data$Eficiencia)
```

```
[1] eficiente   ineficiente
Levels: eficiente ineficiente
```

```
unique(data$Seguridad)
```

```
[1] seguro   inseguro
Levels: inseguro seguro
```

```
unique(data$Ocupación)
```

```
[1] vendedor_ambulante jefe               empleado           comerciante
[5] servidor
Levels: comerciante empleado jefe servidor vendedor_ambulante
```

Entrenar la DAG con los datos para responder las queries.

```
bn<-bn.fit(DAG2, data = data)
```

1.- Queremos saber si el transporte público en ciudades grandes es más eficiente que en ciudades pequeñas.

Probabilidad de eficiencia transporte público para ciudades grandes:

```
cpquery(bn, event = (Eficiencia == "eficiente") , evidence = (Residencia == "grande"), n
```

```
[1] 0.5675451
```

Probabilidad de eficiencia transporte público para ciudades pequeñas:

```
cpquery(bn, event = (Eficiencia == "eficiente") , evidence = (Residencia == "pequeño"), n
```

```
[1] 0.7630648
```

Se puede ver que hay más posibilidades de que el transporte público sea más eficiente en localidades pequeñas comparada a localidades grandes.

2.- ¿Qué probabilidad hay de que una persona viaje en tren, dado que sea vendedor ambulante?

En este ejemplo se utiliza la variable de tren urbano porque este es más concurrido que el tren ferrocarril:

```
cpquery(bn, event = (Transporte == "Tren urbano (Metro)") , evidence = (Ocupación == "ven
```

```
[1] 0.06529723
```

La probabilidad de que una persona viaje en tren dado que es vendedor ambulante es de 7%.

3.- ¿Quiénes son más probables a sentirse seguros en el transporte público, los hombres con estudios universitarios o las mujeres con estudios universitarios?

Probabilidad de que hombres con estudios universitarios se sientan seguros en transporte:

```
cpquery(bn, event = (Seguridad == "seguro") , evidence = ((Sexo == "hombre") & (Educacion
```

```
[1] 0.4186243
```

Probabilidad de que mujeres con estudios universitarios se sientan seguros en transporte:

```
cpquery(bn, event = (Seguridad == "seguro") , evidence = ((Sexo == "mujer") & (Educacion
```

```
[1] 0.424031
```

Se puede decir que es más probable que un hombre con estudios universitarios se sienta más seguro en el transporte público que una mujer con los mismos estudios.

4.- ¿Cómo influye el sexo de la persona en la elección del medio de transporte más utilizado, tomando en cuenta el nivel de ingreso familiar y la eficiencia del transporte público?

Primero, se debe encontrar el medio de transporte más utilizado.

```
freq_table <- table(data$Transporte)
most_common_name <- names(freq_table)[which.max(freq_table)]

print(paste("El medio de transporte más utilizado es:", most_common_name))
```

```
[1] "El medio de transporte más utilizado es: Automóvil"
```

Se sabe que para la pregunta 4, el automóvil es el medio de transporte más utilizado.

Generación de probabilidades para los diferentes casos que se puede presentar para hombre y mujer respectivamente

Hombre:

Probabilidad de que el ser hombre influya en la elección de automóvil como medio de transporte cuando el nivel de ingreso familiar es bajo y el transporte público es eficiente:

```
probh1 <- cpquery(bn, event = ((Transporte == "Automóvil") & (ing_fam == "bajo ingreso")
probh1
```

```
[1] 0.06880366
```

Probabilidad de que el ser hombre influya en la elección de automóvil como medio de transporte cuando el nivel de ingreso familiar es bajo y el transporte público es ineficiente:

```
probh2 <- cpquery(bn, event = ((Transporte == "Automóvil") & (ing_fam == "bajo ingreso")
probh2
```

```
[1] 0.08454695
```

Probabilidad de que el ser hombre influya en la elección de automóvil como medio de transporte cuando el nivel de ingreso familiar es alto y el transporte público es eficiente:

```
probh3 <- cpquery(bn, event = ((Transporte == "Automóvil") & (ing_fam == "alto ingreso")
probh3
```

```
[1] 0.06460112
```

Probabilidad de que el ser hombre influya en la elección de automóvil como medio de transporte cuando el nivel de ingreso familiar es alto y el transporte público es ineficiente:

```
probh4 <- cpquery(bn, event = ((Transporte == "Automóvil") & (ing_fam == "alto ingreso")
probh4
```

```
[1] 0.08800587
```

Mujer:

Probabilidad de que el ser mujer influya en la elección de automóvil como medio de transporte cuando el nivel de ingreso familiar es bajo y el transporte público es eficiente:

```
probm1 <- cpquery(bn, event = ((Transporte == "Automóvil") & (ing_fam == "bajo ingreso")
probm1
```

```
[1] 0.07072431
```

Probabilidad de que el ser mujer influya en la elección de automóvil como medio de transporte cuando el nivel de ingreso familiar es bajo y el transporte público es ineficiente:

```
probm2 <- cpquery(bn, event = ((Transporte == "Automóvil") & (ing_fam == "bajo ingreso")
probm2
```

```
[1] 0.08514148
```

Probabilidad de que el ser mujer influya en la elección de automóvil como medio de transporte cuando el nivel de ingreso familiar es alto y el transporte público es eficiente:

```
probm3 <- cpquery(bn, event = ((Transporte == "Automóvil") & (ing_fam == "alto ingreso")
probm3
```

```
[1] 0.0644893
```

Probabilidad de que el ser mujer influya en la elección de automóvil como medio de transporte cuando el nivel de ingreso familiar es alto y el transporte público es ineficiente:

```
probm4 <- cpquery(bn, event = ((Transporte == "Automóvil") & (ing_fam == "alto ingreso")
probm4
```

```
[1] 0.08877855
```

Con todas las probabilidades y sus combinaciones, se hace una suma de probabilidades y así responder la pregunta

Suma probabilidad Hombre:

```
probh <- probh1 + probh2 + probh3 + probh4
probh
```

```
[1] 0.3059576
```

Suma probabilidad mujer:

```
probm <- probm1 +probm2 + probm3 + probm4
probm
```

```
[1] 0.3091336
```

Con estos resultados se puede decir que la mujer es más probable a elegir el automóvil como medio de transporte más utilizado, tomando en cuenta su nivel de ingreso familiar y la eficiencia de este transporte público. Pero como la diferencia entre probabilidades es muy pequeña, se puede inferir que el sexo no influye en la elección del transporte público dado el ingreso familiar y eficiencia.

Conclusión

Referencias

Agualongo, D. y Garcés, A. (2020). El nivel socioeconómico como factor de influencia

en temas de salud y educación. Universidad de las Fuerzas Armadas Espe. [PDF]

Benno, S. (1985). Educación y dependencia: el papel de la educación comparada. UNESCO. [PDF]

Calvillo, A. y Moncada, G. (2008). Eficiencia del transporte público y privado. El consumidor. [PDF]

Education in Mexico and its influence on the occupation. (sf) Espinosa Yglesias Study Center. Retrieved from https://ceey.org.mx/la-educacion-en-mexico-y-su-influencia-en-la-trabajo/

Santiesteban, JC, Pérez, d. and Hernández, C. (2012). Definition of Bayesian Networks and their applications. Vinculando Magazine. https://vinculando.org/articulos/redes-bayesianas.html

Torres, G. and Ayala, E. (November 2020). Family income as a determinant of school attendance of young people in Mexico. Problems of development, 201. Retrieved from https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0301-70362020000200085

Wolf et. to the. (March 11, 2019). Dynamics and controls of chemical processes. [PDF]