

INTELIGENCIA ARTIFICIAL

Tarrea 1

Autor:

Joaquín Fernández

Profesor:

Martín Gutiérrez

Ayudante:

Nicolás Araya

3ndice

1. Introducci3n	2
1.1. Contexto y Dataset	2
2. Actividad y Supuestos	3
2.1. Desarrollo	3
2.2. Consultas de Inferencia	6
2.2.1. Consulta n3mero 1	6
2.2.2. Consulta n3mero 2	8
2.3. Consulta n3	11
2.4. Consulta n34	12
2.5. Consulta n35	14
3. Enlaces y Google Collab	16

1. Introducci3n

Para esta actividad de tarea se busca usar un *dataset* y dar posibles soluciones en el 3mbito de determinar inferencias y el tipo asociado a estas, dependiendo de una red Bayesiana que genera una estructura que abarcan todos los posibles tramos o caminos asociados al suceso en cuesti3n. Para abordar dicho problema es necesario usar un entorno de “Linux” en la plataforma de “Google Collab” junto a herramientas de Python y usando las librer3as “Matplotlib” y “Bnlearn”. De esta forma se tendr3 que estructurar, analizar, caracterizar par3metros y elementos de una red Bayesiana.

1.1. Contexto y Dataset

La Diabetes se encuentra entre las enfermedades cr3nicas m3s prevalentes en los Estados Unidos, esta afecta a millones de estadounidenses cada a3o y ejerce una carga financiera significativa en la econom3a. La enfermedad de la Diabetes es una enfermedad cr3nica grave en la que las personas pierden la capacidad de regular eficazmente los niveles de glucosa en la sangre y puede reducir la calidad de vida y la esperanza de vida.

Las complicaciones como la enfermedad cardiaca, la p3rdida de la visi3n, la amputaci3n de las extremidades inferiores y la enfermedad renal se encuentran asociadas con niveles cr3nicos altos de az3car en el torrente sangu3neo para las personas que padecen esta dolencia. De esta forma, el diagn3stico temprano puede conducir a cambio en el estilo de vida y aplicar as3 un tratamiento eficaz, lo que hace conveniente el uso de modelos predictivos de riesgo de diabetes en herramientas de suma importancia para estudios, p3blico y funcionarios del 3rea de salud.

2. Actividad y Supuestos

Para este apartado se dará a conocer los enunciados a tratar y para desarrollar el modelo probabilístico.

1. Usando la librería `bnlearn` (<https://pypi.org/project/bnlearn/>), y un *dataset* de su elección, aprenda la estructura de una red bayesiana en función de esos datos y caracterice los parámetros y elementos de la red obtenida.
2. Enuncie y efectúe cinco consultas de inferencia sobre la red construida. Deberá documentar cómo se reorganizan los parámetros para cada uno de los casos.

2.1. Desarrollo

Los parámetros pertenecientes a la red Bayesiana describen distintas condiciones y catalogan, según estados, el tipo de vida que posee una persona “x”; para así determinar predicciones asociadas no únicamente a la diabetes, sino más bien a una variedad de dolencias, malos hábitos y mal cuidado autogestionado que conducen a desequilibrios mentales, sustanciales y físicos.

Este “dataset” es un conjunto de datos limpio de 253,680 respuestas a la encuesta BRFSS2015 de los CDC. La variable objetiva `Diabetes_012` tiene 3 clases: 0 es sin diabetes o solo durante el embarazo, 1 es para prediabetes y 2 es para diabetes. Hay un desequilibrio de clases en este conjunto de datos. Este conjunto de datos tiene 21 variables con sus respectivas características.

- `MentHlth`: Escala en función de días de padecimientos y trastornos psicológicos; siendo 1 la más leve y 30 el más extremo.
- `PhysHlth`: Salud física en función duración en días, esto incluye enfermedades, desgarros, fracturas y otros malestares. Siendo 1 lo más leve y 30 el caso más extremo.
- `GenHlth`: salud asociada a la genética de la persona con los valores asociados 1 excelente, 2 muy buena, 3 buena, 4 con algunos defectos y 5 muy mala.
- `NoDocbcCost`: Si es que no haya asistido a consultas medicas por el tema del costo, en caso que sí 1 por el contrario 0.
- `HeartDiseaseorAttack`: En caso de padecer enfermedades al corazón y posibles infartos, sí 1 y no 0.
- `BMI`: Índice de masa corporal o indicador de gordura, si el índice es menor a 18.5 se encuentra dentro del rango de peso insuficiente, si el IMC es entre 18.5 y 24.9 se encuentra dentro del rango de peso normal o saludable, si el IMC se encuentra entre 25.0 y 29.9 se encuentra la persona con sobrepeso y el último caso en que el individuo tenga un IMC de 30.0 o superior ha de tener obesidad.
- `DiffWalk`: Si es que tiene dificultades al momento de subir escaleras o caminar 1, caso contrario 0.
- `Sex`: 0 femenino y 1 masculino.
- `Fruits`: En caso de que consuma frutas de forma reiterada 1 y caso contrario 0.
- `Veggies`: En caso de que consuma vegetales de forma reiterada 1 y caso contrario 0.
- `Education`: 1 no ha accedido a educación o ha cursado únicamente kínder, 2 haber cursado primaria, 3 que la persona haya asistido a HighSchool en grados de básica avanzados, 4 que haya terminado HighSchool (“la media”), 5 en Universidad primeros años y 6 egresando de enseñanza superior o universidad.

- PhysActivity: Si es que ha realizado actividad f3sica en los 3ltimos 30 d3as, un 1 caso de que este no haya hecho ejercicio 0.
- Income: Salario asociado a las personas tratadas o estudiadas con 1, menor a \$10.000, 5 menor a \$35.000 y 8 con ingresos mayores o iguales a \$75.000.
- HvyAlcoholConsump: Consume de forma frecuente alcohol 1 y en caso de que no lo haga 0.
- Smoker: En caso de que el usuario frecuente fumar con la notaci3n de 1 y caso de que no lo haga 0.
- HighBP: 0 para denotar que no posee una alta presi3n cardiaca, caso contrario de tener el valor 1 que significa que el individuo si lo padece.
- Stroke: Si alguna vez la persona haya tenido alg3n derrame cerebral.
- HighChol: 0 para denotar que no tiene colesterol alto en el torrente sangu3neo, opuesto al tratarse de tener el valor de 1.
- CholCheck: Esta columna act3a como un binario respecto a los valores 0 si es que no se ha hecho revisiones de colesterol tras 5 a3os, caso contrario del 1.
- AnyHealthCare: Si es que el sujeto se ha tratado o est3 inscrito en alg3n servicio de salud, 0 no y 1 s3.
- Age: 1 entre 18 y 24 a3os, 9 desde 60 a 64 a3os y 13 para gente con edad mayor o igual a 80.

```

1 import bnlearn as bn
2 import pandas as pd #lee el dataset
3
4
5 #df = pd.read_csv("clean_data.csv") ## abre el .csv como Dataframe
6 df = pd.read_csv("diabetes_012_health_indicators_BRFSS2015.csv")
7 print(df.shape) #tama o de dataframe
8
9 ##creacion de red bayesiana ##
10 model = bn.structure_learning.fit(df.iloc[3000:])
11 #model = bn.independence_test(model, df) #prueba de independencia
12
13 G = bn.plot(model)

```

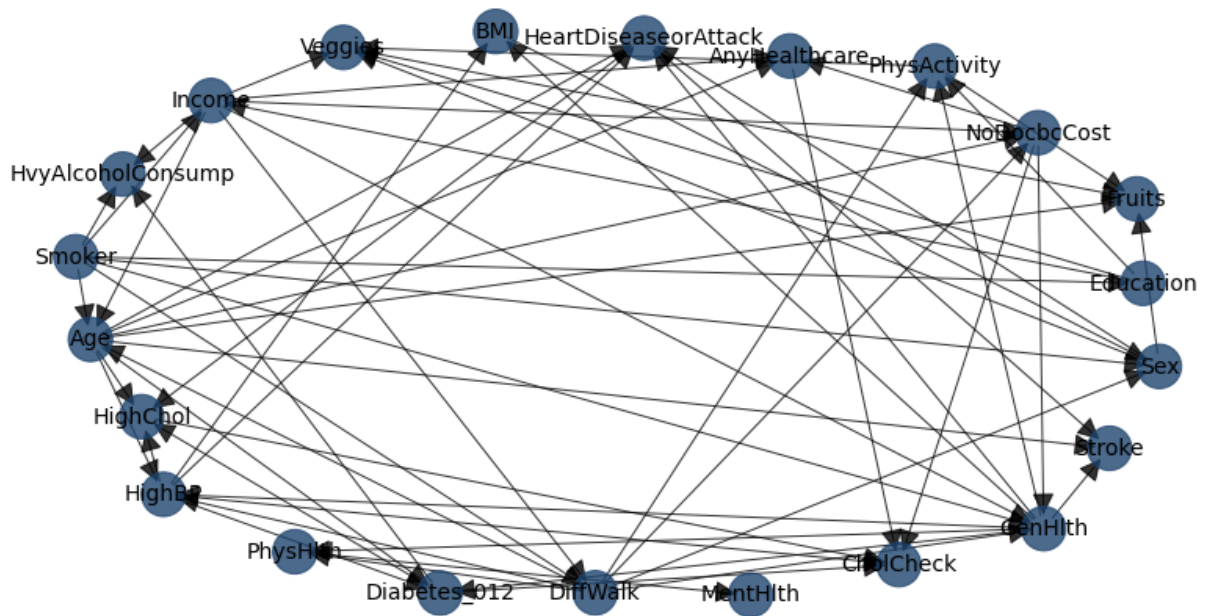


Figura 1: Estructura de Aprendizaje.

```

1 # Now we learn the parameters of the DAG using the df
2 model_update = bn.parameter_learning.fit(model, df)
3 Gg = bn.plot(model_update, interactive=True, params_interactive={'notebook':True})

```

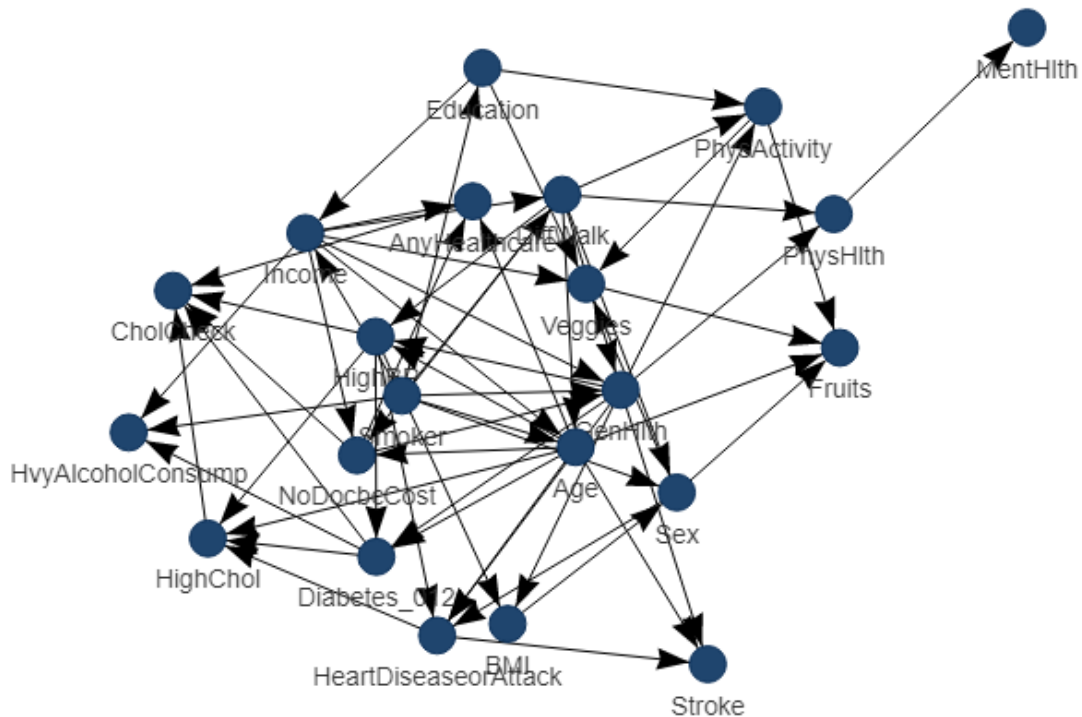


Figura 2: Red Bayesiana con parámetros.

2.2. Consultas de Inferencia

2.2.1. Consulta número 1

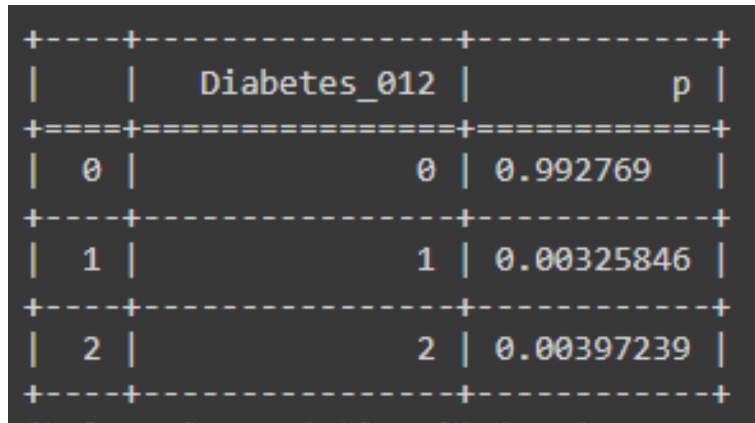
```

1 arr = [1,9,13]
2 for i in arr:
3     for j in range(1,6,1):
4         for k in range(0,2,1):
5             for h in range(0,2,1):
6                 for l in range(0,2,1):
7                     q_1 = bn.inference.fit(model_update, variables=['Diabetes_012'], evidence={'
                        GenHlth': j, 'HighBP':k, 'Age': i, 'Smoker':h, 'HvyAlcoholConsump':l})

```

Para esta consulta se da uso de un supuesto de independencia en serie, puesto que en este caso el nodo Diabetes.012 se recorre de forma ascendente, mediante una cota que incluye una gran cantidad de nodos, además se incluye un nodo no perteneciente a la relación que actuará como podneración y/o producto en la probabilidad; de esta forma se obtienen la probabilidad de que tenga Diabetes dado que su salud genética sea buena o mala, su presión sanguínea si es elevada o no, su edad, si es que fuma también y si también consume alcohol. De esta forma, con esta consulta puede hacerse un seguimiento de la persona respecto a su rutina y hábitos en su vida. Además, se itera todas las posibles soluciones dado todos los estados asociados a las variables y parámetros contenidos en la inferencia.

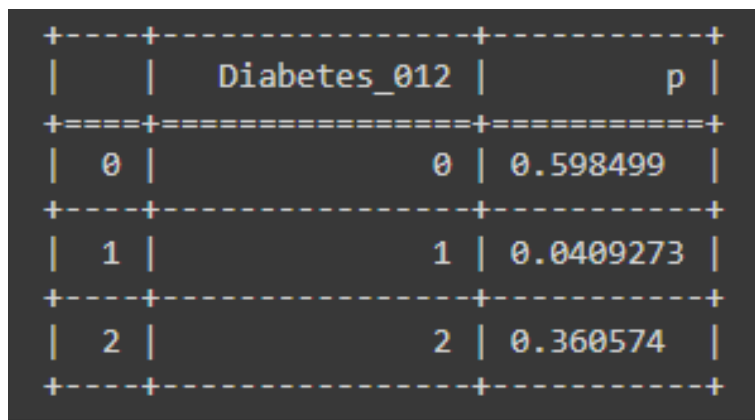
Por ejemplo, el primer output:



	Diabetes_012	p
0	0	0.992769
1	1	0.00325846
2	2	0.00397239

Figura 3: Ejemplo output inferencia 1.

Este indica la probabilidad de que tenga diabetes dado que su salud genética no es buena, que no tenga alta presión sanguínea, que su edad se encuentra entre 18 y 24 años, que no fume y que no consuma alcohol. Por lo que se puede ver que lo más posible es que no tenga diabetes; Ahora bien, pasando a ver otro caso:



	Diabetes_012	p
0	0	0.598499
1	1	0.0409273
2	2	0.360574

Figura 4: Ejemplo 2 output inferencia 1.

La probabilidad aumenta según como varían todos los parámetros, en algunos casos, las iteraciones se mantienen iguales o el cambio no es muy notorio; pero al momento que iteren todas las evidencias se puede ver como afecta según el incremento en la cifra asociada a los diagnósticos obtenidos.

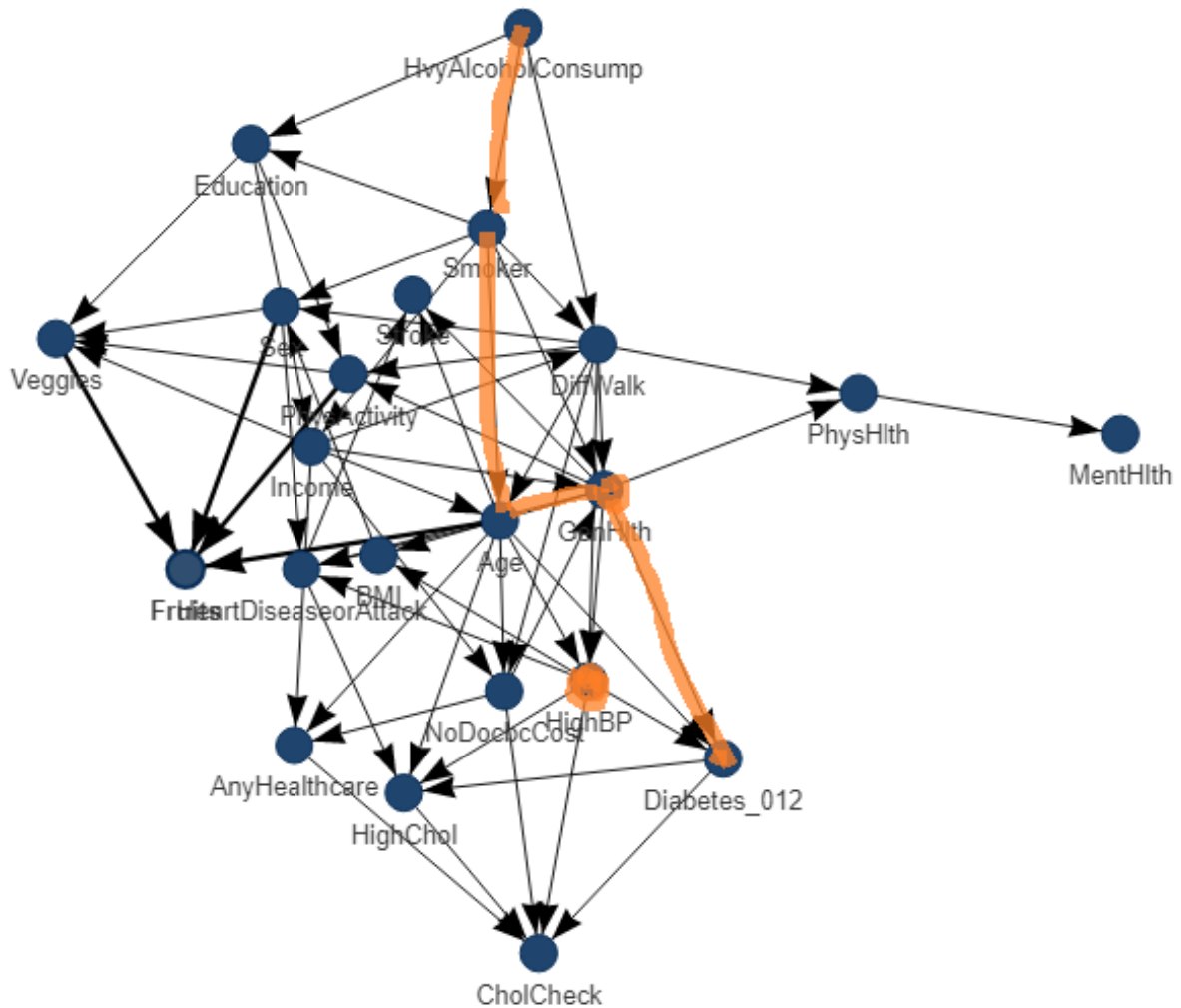


Figura 5: Recorrido inferencia n° 1.

2.2.2. Consulta número 2

```

1 for i in range(0,3,1):
2   q_n = bn.inference.fit(model_update, variables=['HighChol','CholCheck'], evidence={'
      Diabetes_012':i,'GenHlth':5,'NoDocbcCost':1,'Income':1,'Education':1,'Smoker':1,'
      HvyAlcoholConsump':1})

```

Para esta inferencia actúa una relación de probabilidad independiente divergente, puesto que se tiene la independencia de los nodos o parámetros “HighChol” y “CholCheck”. La forma de interpretar dicha inferencia sería la probabilidad que tenga si o no tenga colesterol alto y que se haga revisiones, dado que tenga diabetes o no, que tenga mala genética, que no acuda a ir al doctor por el costo, que los ingresos sean mínimos, que su educación sea solo hasta kínder, que fume y que beba alcohol.

	HighChol	CholCheck		p
0	0	0		0.0608124
1	1	0		0.0342741
2	0	1		0.501385
3	1	1		0.403528
[bnlearn] >Variable Elimination..				
Finding Elimination Order: : 100%				
Eliminating: DiffWalk: 100%				
	HighChol	CholCheck		p
0	0	0		0.0706882
1	1	0		0.069657
2	0	1		0.331447
3	1	1		0.528208
[bnlearn] >Variable Elimination..				
Finding Elimination Order: : 100%				
Eliminating: DiffWalk: 100%				
	HighChol	CholCheck		p
0	0	0		0.0246485
1	1	0		0.0270027
2	0	1		0.285792
3	1	1		0.662557

Figura 6: Output inferencia n° 2.

Mediante los resultados de la inferencia se puede ver que la probabilidad aumenta en función de las iteraciones asociadas a la diabetes, siendo el primer recuadro no tener diabetes, el segundo pre diabético y el tercero diabético. Además, según la combinación puede apreciarse una relación directa e inversamente proporcional; ya que teniendo el caso de no tener el colesterol alto y que se haya hecho una revisión de su estado de colesterol, dado que no es diabético, pasa desde un 50 %, cuando es pre diabético 33 % y si es diabético un 28 %. Caso contrario de que tenga el colesterol alto y que se haya revisado el estado de colesterol. Además, puede verse una distribución en la relación de tipo Gaussiana, ya que se forman una campana por iteración. La representación vendría a ser similar a este ejemplo:

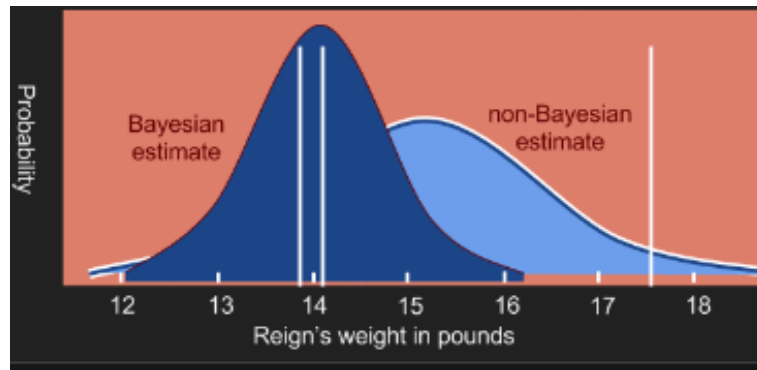


Figura 7: Distribuci3n de probabilidad Gaussiana inferencia n3 2.

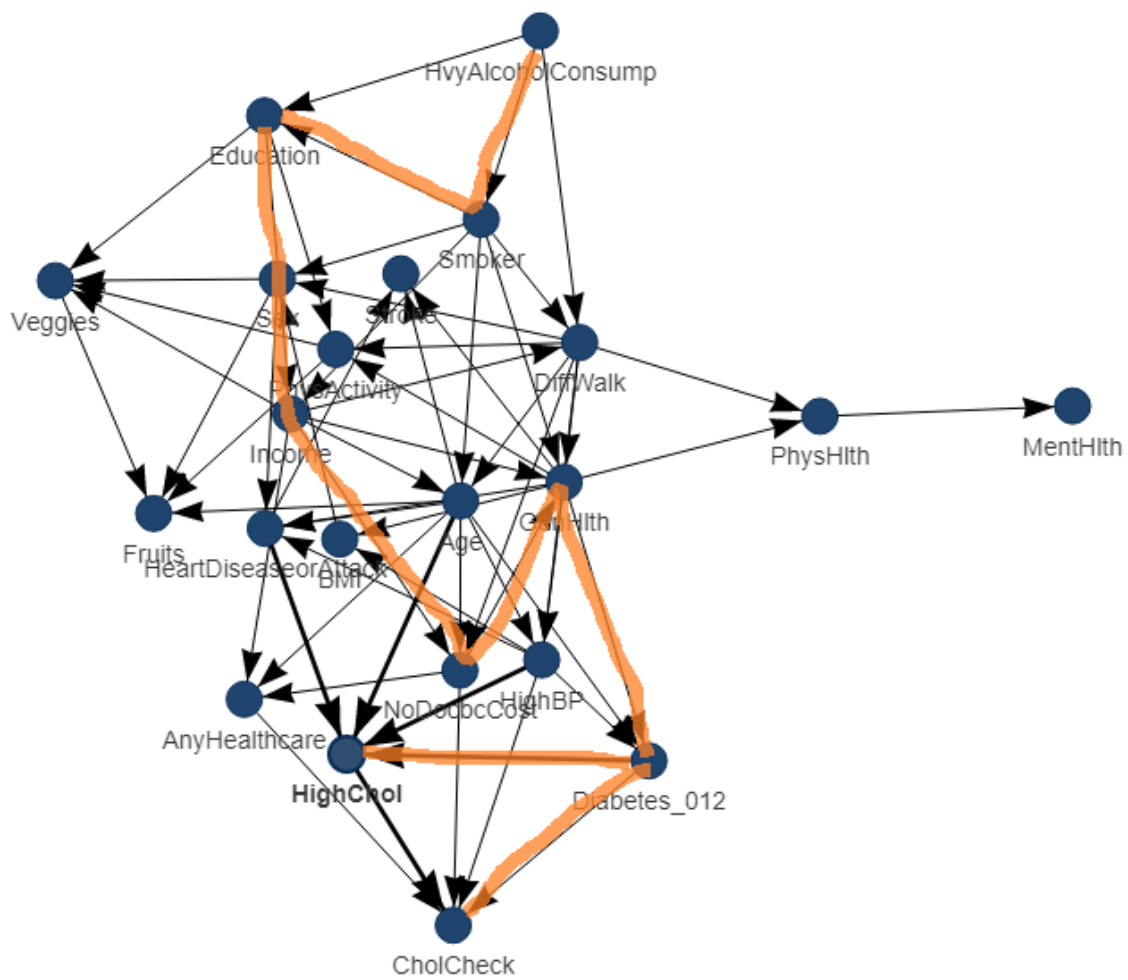


Figura 8: Recorrido inferencia n3 2.

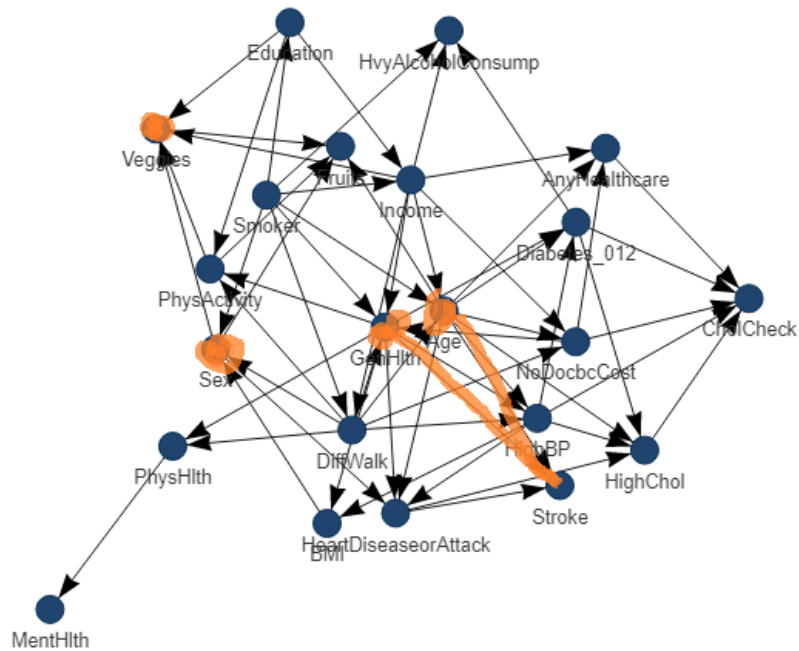


Figura 10: Recorrido inferencia n° 3.

2.4. Consulta n°4

```

1 for i in range(0,2,1):
2     for j in range(1,6,1):
3         q_2 = bn.inference.fit(model_update, variables=['HvyAlcoholConsump','Smoker'], evidence
                                ={'DiffWalk': i, 'Education':j, 'HighBP':1, 'HeartDiseaseorAttack':1, 'Stroke':1})

```

Nuevamente, se buscó trabajar en una inferencia de tipo divergente unida con una serial en la misma, de esta forma se buscaba crear un entorno de probabilidades híbrido para observar así su comportamiento.

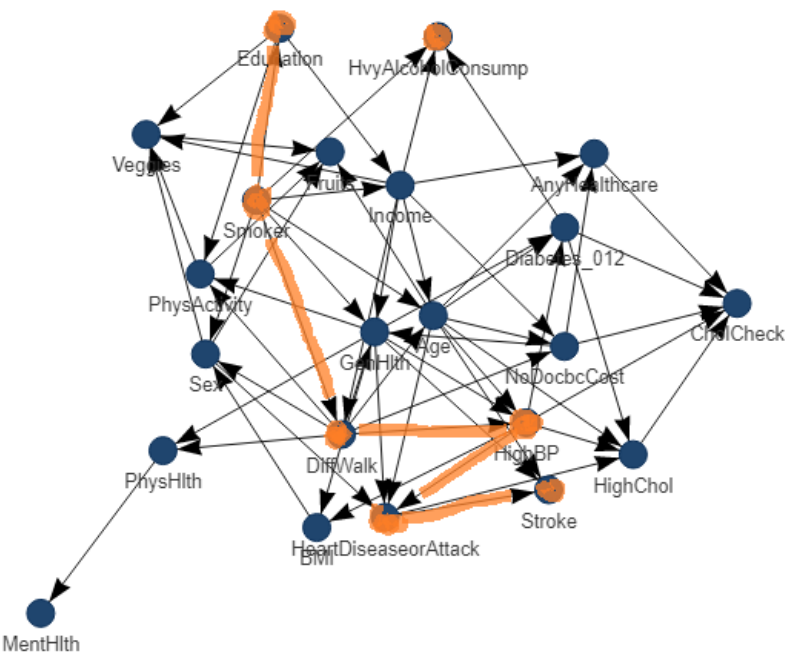


Figura 11: Recorrido inferencia n3 4.

	HvyAlcoholConsump	Smoker	p
0	0	0	0.514338
1	1	0	0.0134322
2	0	1	0.442247
3	1	1	0.0299834

[bnlearn] >Variable Elimination..
Finding Elimination Order: : 100%
Eliminating: GenHlth: 100%

	HvyAlcoholConsump	Smoker	p
0	0	0	0.47606
1	1	0	0.0118284
2	0	1	0.482353
3	1	1	0.0297583

[bnlearn] >Variable Elimination..
Finding Elimination Order: : 100%
Eliminating: GenHlth: 100%

	HvyAlcoholConsump	Smoker	p
0	0	0	0.343985
1	1	0	0.00870907
2	0	1	0.608317
3	1	1	0.0389887

Figura 12: Output inferencia n° 4.

Las probabilidades son en torno de la inferencia de la probabilidad de que consuma Alcohol y Fume; dado que tenga dificultades al caminar o no, su nivel de Educación, presión sanguínea, problemas al corazón y si la persona ha tenido accidentes cerebrovasculares. Se puede decir que las probabilidades cambian en torno a los diferentes estados que posee cada nodo del tramo en cuestión; lo que proporciona un visión no únicamente médica sino social de los sujetos observados.

2.5. Consulta n°5

```

1 arr = [1,9,13]
2 for i in arr:
3     for j in range(0,2,1):
4         for k in range(1,6,1):
5             q_t = bn.inference.fit(model_update, variables=['MentHlth'], evidence={'Age':i,
                PhysHlth':j,'GenHlth':k,'PhysActivity':1})

```

Esta inferencia es de tipo divergente y toma los tramos desde salud mental y actividad física; teniendo como nodo central el estado genético. Por lo que la inferencia se lee como la probabilidad de la salud mental dada su salud física, actividad física, estado genético y edad.

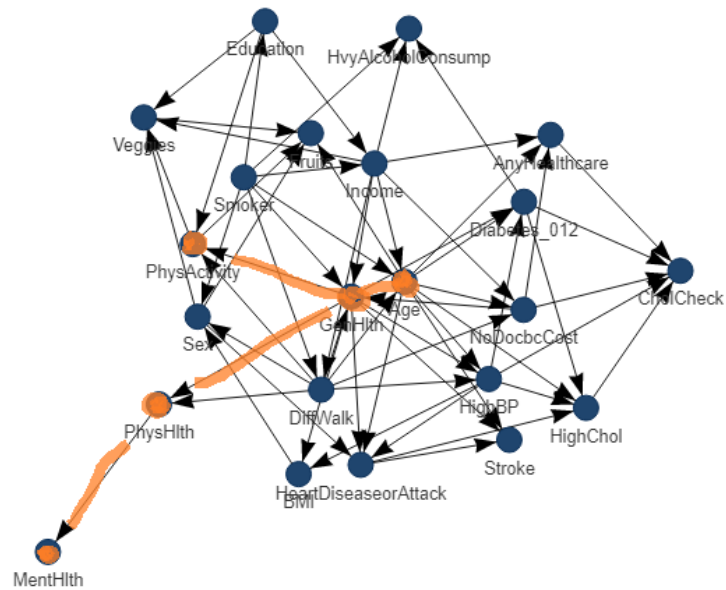


Figura 13: Recorrido inferencia n° 5.


```

+-----+-----+-----+
| | MentHlth | p |
+-----+-----+-----+
| 0 | 0 | 0.786998 |
+-----+-----+-----+
| 1 | 1 | 0.0280855 |
+-----+-----+-----+
| 2 | 2 | 0.0455575 |
+-----+-----+-----+
| 3 | 3 | 0.024275 |
+-----+-----+-----+
| 4 | 4 | 0.0117253 |
+-----+-----+-----+
[bnlearn] >Variable Elimination..
Finding Elimination Order: 0/0 [00:00<?, ?i/s]
0/0 [00:00<?, ?i/s]
+-----+-----+-----+
| | MentHlth | p |
+-----+-----+-----+
| 0 | 0 | 0.786998 |
+-----+-----+-----+
| 1 | 1 | 0.0280855 |
+-----+-----+-----+
| 2 | 2 | 0.0455575 |
+-----+-----+-----+
| 3 | 3 | 0.024275 |
+-----+-----+-----+
| 4 | 4 | 0.0117253 |
+-----+-----+-----+
[bnlearn] >Variable Elimination..
Finding Elimination Order: 0/0 [00:00<?, ?i/s]
0/0 [00:00<?, ?i/s]
+-----+-----+-----+
| | MentHlth | p |
+-----+-----+-----+
| 0 | 0 | 0.786998 |
+-----+-----+-----+
| 1 | 1 | 0.0280855 |
+-----+-----+-----+
| 2 | 2 | 0.0455575 |
+-----+-----+-----+
| 3 | 3 | 0.024275 |
+-----+-----+-----+
| 4 | 4 | 0.0117253 |
+-----+-----+-----+

```

Figura 14: Output inferencia n3 5.

3. Enlaces y Google Collab

<https://colab.research.google.com/drive/1nfCEr1DqkEK5odYN5221ORoYbTdKYMxM?usp=sharing>