**Analítica Computacional para la toma de Decisiones**

**Proyecto 1 – Predicción de enfermedades cardíacas**

**Integrantes: Nicolas Ricaurte y Juan Camilo Bello**

**Introducción**

El siguiente documento tiene como objeto el diseño de un proyecto que busca desarrollar e implementar un modelo de inferencia bayesiana basado en una base de datos del repositorio de la Universidad de California, con información sobre 303 pacientes. Para este caso, en particular, la base de datos solo contiene catorce parámetros. El prototipo desarrollado busca proporcionar a los pacientes información acerca del grado de urgencia con el que deben consultar asistencia médica con el fin de prevenir enfermedades cardiacas.

**Exploración de datos**

Inicialmente, después de descargar los datos del repositorio se procedió a limpiarlos, verificando los siguientes aspectos: Tipos de datos, caracteres no numéricos y valores nulos. Dentro de este proceso se resalta que en la columna correspondiente al número de arterias coloreadas por fluoroscopia (ca) y Gammagrafía cardíaca con talio (thal) se encontraron caracteres no numéricos con el signo “?”, por lo que se procedió a eliminar las filas con estos caracteres, debido a que no era posible utilizar esta información para inferir correctamente sobre las variables. Por otro lado, se modificó la variable de interés (num) la cual tiene valores superiores a 1, que indican la gravedad de la enfermedad, pero para efectos del proyecto nos interesa que sea binaria, indicando si se tiene una enfermedad cardiaca o no. Para esto se asumió que los pacientes con valores diferentes a 0, tienen una enfermedad cardiaca, es decir, toman el valor de 1. Al finalizar la limpieza de los datos, se cuenta con 297 pacientes.

Ahora bien, se continuo por encontrar estadísticas descriptivas de los datos para comprender la información con la que contamos y desarrollar inferencias apropiadas. Dividimos esta tarea en 3 partes: Visualización gráfica de los datos, investigación de valores teóricos, cuantificación de las estadísticas descriptivas.

**Visualización gráfica**

Realizamos un histograma, para identificar tendencias en los datos, verificar los valores que estos toman y lograr conceptualizar los valores teóricos que se deben encontrar para cada variable dentro del modelo.

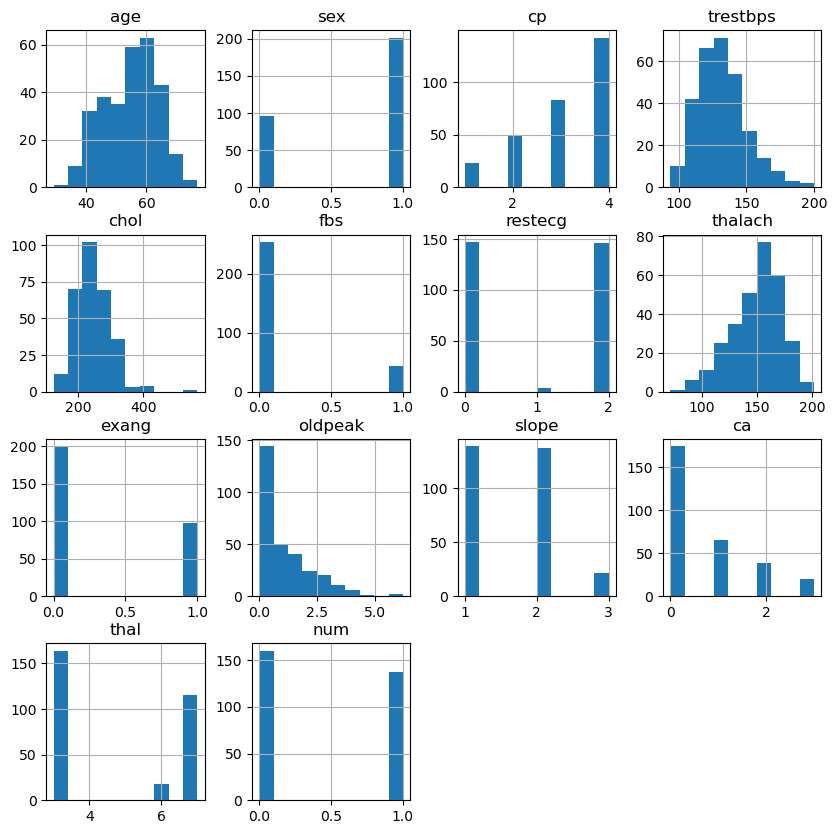


Ilustración 1

**Valores teóricos y Cuantificación de las Estadísticas Descriptivas**

Como se mencionó anteriormente el total de datos es 297 pacientes, usando los histogramas encontrados en la *ilustración 1* se pueden inferir las siguientes estadísticas descriptivas, como media, valor mínimo y máximo, como también, cantidad de pacientes dentro de los rangos teóricos para aquellas variables que los tengan.

**Edad**: La media es 54 años, donde 29 es el mínimo y 77 el máximo.

**Sexo**: Hay 201 hombres y 96 mujeres en la base de datos.

**Dolor torácico (cp):** Es una variable que se divide en 4 (1: Angina típica, 2: Angina atípica, 3: Dolor no relacionado con la angina, 4: Asintomático)). Hay 23 pacientes con angina tipo 1, 49 pacientes con angina tipo 2, 83 pacientes con angina tipo 3 y 142 pacientes con angina tipo 4.

**Presión sistólica (trestbps) (mmHg)**: Los valores teóricos encontrados son: Baja: <90, Normal: <120, prehipertensión: <130, hipertensión grado 1: <140, hipertensión grado 2: <180+. La media se encuentra en 131.69 mmHg y los rangos para los pacientes con respecto a los valores teóricos son: Baja: 0 pacientes, Normal: 97 pacientes, prehipertensión: 71 pacientes, hipertensión grado 1: 63 pacientes, hipertensión grado 2: 66 pacientes.

**Niveles de cholesterol (chol) (mg/dl):** (Prueba de sangre) Los valores teóricos son: Sano: <200, En riesgo: <239, Peligroso: >240. La media se encuentra en 247.35 mg/dl y los rangos para los pacientes con respecto a los valores teóricos son: Sano: 49 pacientes, en riesgo: 93 pacientes, peligroso: 155 pacientes.

**Glucemia en ayunas (fbs) (mg/dl):** (Prueba de sangre) Es una variable que se divide en 1 si es mayor a 120 mg/dl y 0 de lo contrario. Como es una variable binaria se encuentra una cantidad de pacientes con fbs menor a 120 mg/dl de 254 pacientes y 25 pacientes la tienen mayor.

**Resultados electrocardiográficos en reposo (restecg):** (Se necesita hacer un electrocardiograma)Es una variable categórica dividida de la siguiente forma: 0: normal, 1: Tener anomalías en la onda ST-T (inversión de la onda T y/o elevación o depresión del ST >0,05 mVi, 2: hipertrofia ventricular izquierda probable o definida según los criterios de Estes. Hay 147 pacientes con restecg normal, 4 con restecg con anormalidad ST-T y 146 con restecg con hipertrofia ventricular izquierda.

**Frecuencia cardíaca máxima (thalach) (bpm):** El valor teórico para esta variable siguiente el siguiente procedimiento: Normal: 220 - Edad >= Thalach. Con esto en mente, la media se encuentra en 149.6 bpm, el mínimo es de 71 y máximo 202. Hay 232 pacientes con frecuencia cardiaca normal y 65 anormal.

**Angina inducida por ejercicio (exang):** Es una variable binaria que toma los valores 1 si el paciente tiene angina inducida por ejercicio y 0 si no. Hay 200 pacientes con angina inducida por ejercicio y 97 sin angina inducida por ejercicio.

**Depresión del segmento ST (ECG) (oldpeak):** El valor mínimo de la variable es 1.05, el valor mínimo es 0 y el máximo es 6.2. De esta forma se realizaron los siguientes rangos para categorizar a nuestros pacientes: Entre 0-1: 174, entre 1-2: 73, entre 2-3: 30, entre 3-4:15, entre 4-6.2: 5.

**Pendiente del segmento ST (ECG) (slope):** Es una variable categórica que toma los valores: 1: Plano, 2: Inclinado, 3: Pronunciado. Se encontró que 139 pacientes tienen una pendiente plana, 137 inclinada y 21 pronunciada.

**Número de arterias coloreadas por fluoroscopia (ca):** Esta es una variable categórica que toma los valores 0,1,2 y 3 indicando la cantidad de arterias coloreadas por fluoroscopia. Los resultados de los pacientes fueron los siguientes: Hay 174 personas con ca = 0, 65 personas con ca = 1, 38 personas con ca = 2, 20 personas con ca = 3.

**Gammagrafía cardíaca con talio (thal):** Es una variable categórica que toma los valores 3,7 y 6, donde los resultados de dicha caracterización en la muestra de pacientes son: 3: 164, 7: 115, 6: 18.

**Variable objetivo (num):** Es una variable binaria que muestra si el paciente tiene o no enfermedad tomando los valores de 1 y 0 respectivamente. Encontrando que de la muestra de pacientes 160 no tienen enfermedad y el restante 137 sí.

Con el objetivo de encontrar entre qué variables o grupos de variables existe la posibilidad de que ocurra una interacción, visualizamos el siguiente gráfico de correlación donde los valores negativos indican que mientras una variable aumenta, la otra disminuye y viceversa (inversamente proporcional), y los valores positivos indican que mientras una variable aumenta, la otra también aumenta (directamente proporcional), donde:

* Cuanto más cerca de 1(Claro), más correlación hay entre las variables
* Cuanto más cerca de -1, menos correlación hay entre las variables
* Cuanto más cerca de 0(Oscuro), no hay correlación entre las variables

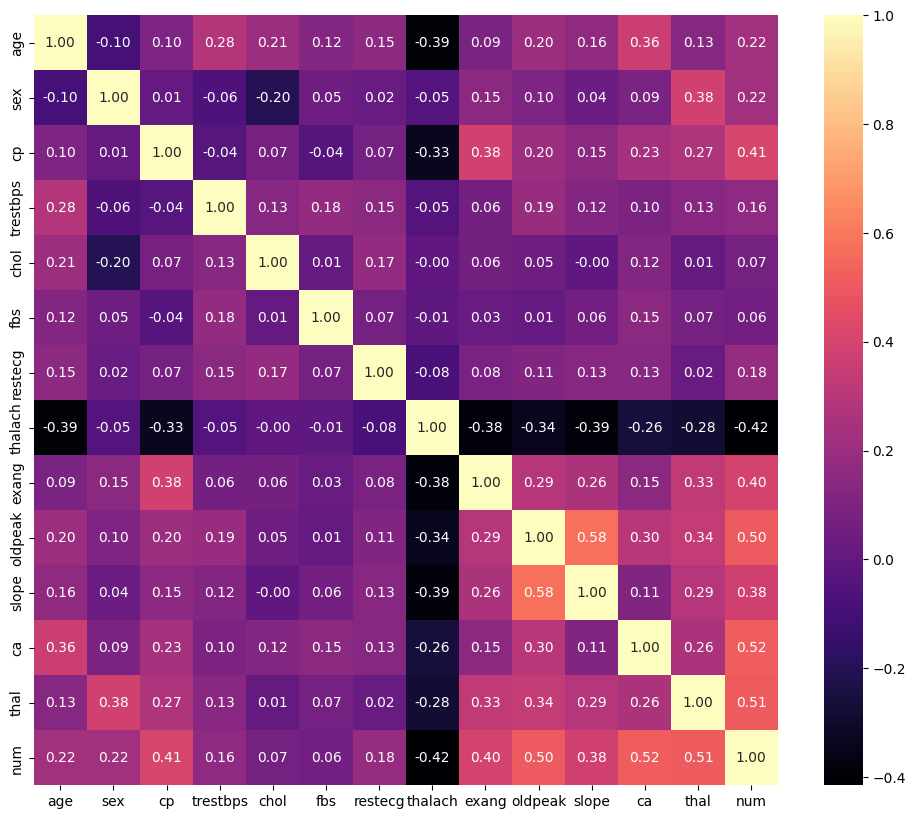


Ilustración 2

Las variables que tienen correlación perfecta positiva son las que están en la diagonal, con esto en mente podemos ver que algunas de las variables que tienen mayor correlación en la siguiente tabla:

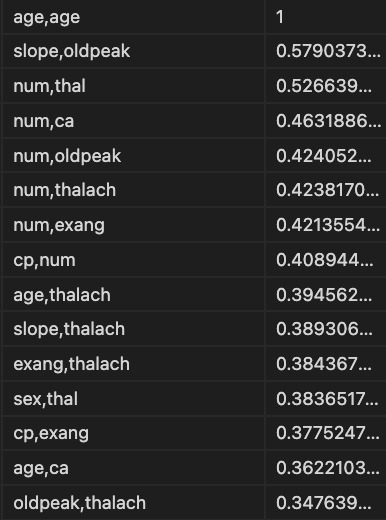


Ilustración 3

En este punto de la exploración podemos asumir que existe una relación entre el número de arterias coloreadas por fluoroscopia (no obstruidas) CA, y la presencia de una enfermedad cardiaca. La angina inducida EXANG por ejercicio y la angina en reposo CP están relacionadas entre sí y a su vez están relacionadas con cambios en la información suministrada por electrocardiograma OLDPEAK. Por último, el máximo ritmo cardiaco alcanzado THALACH, está ligado a la pendiente y las inversiones de las ondas ST en el electrocardiograma.

**Construcción del grafo**

Para iniciar con la definición del grafo a utilizar en nuestro proyecto, determinamos que los usuarios finales de nuestro aplicativo serian pacientes que, al introducir sus datos, puedan saber si es prioritario que reciban asistencia médica y se realicen exámenes orientados a la identificación de problemas cardiacos.

Con el fin de dirigir nuestra investigación entrevistamos a Alfonso Ricaurte Riveros, médico cirujano de la Universidad Nacional, quien después de comprender los alcances de nuestro proyecto, nos recomendó no tener en cuenta las variables obtenidas mediante los electrocardiogramas pues los pacientes actualmente no cuentan con una forma eficaz de auto realizar dicho examen. Dicho esto, el entrevistado procedió a graficar las siguientes relaciones causales en conjunto con nosotros:

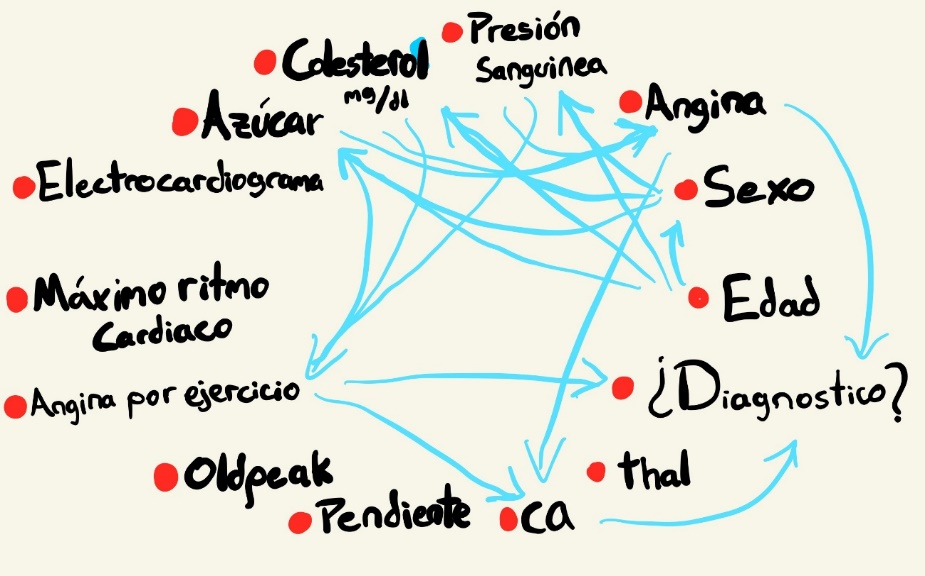


Ilustración 4

La anterior información nos proporcionó los insumos suficientes para empezar a buscar relaciones preexistentes entre las variables y así justificar su conexión. A continuación, revisaremos el porqué de cada una de las relaciones de nuestra red bayesiana:

Schematic

Description automatically generated with low confidence

Ilustración 5

Si bien el sexo puede no tener una relación muy directa con la existencia de enfermedades cardiacas, el sexo puede convertirse en un factor relevante para las mujeres mayores de 55 años aproximadamente. En esta edad terminan los ciclos fértiles de la mujer y este cambio hormonal puede generar una gran variedad de complicaciones medicas incluyendo problemas cardiacos (Vieira & Medrado, 2021, p2), (ilustración 5).

Diagram, schematic

Description automatically generated

Ilustración 6

Según una investigación conducida para la revista facultad de salud en el 2015, la menopausia puede generar cambios en el metabolismo lo que ha demostrado reflejar alteraciones abruptas en el azúcar, la presión arterial y sobre todo en los niveles de colesterol (Pinzon & Celemin, 2015, p49), (ilustración 6).

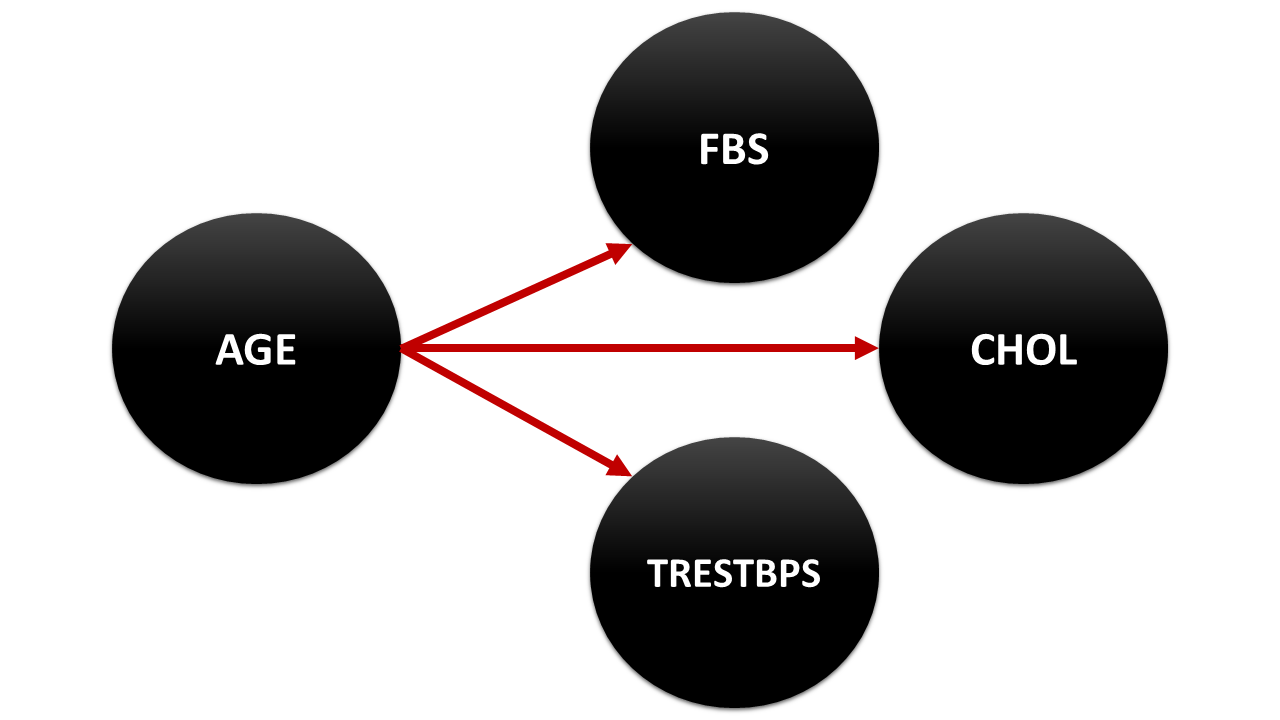


Ilustración 7

Teniendo en cuenta lo anterior, es importante mencionar que, si bien la influencia de la variable sexo en otros factores está condicionada a la existencia de la edad como parámetro en el estudio, la variable edad tiene una relación independiente con los signos vitales básicos tenidos en cuenta (azúcar, presión y colesterol). Según el libro de la salud cardiovascular, una investigación promovida por el banco BBVA, *“El nivel de colesterol en la sangre está determinado por factores genéticos y ambientales que incluyen: la edad, el sexo y el peso corporal”* (Cachofeiro, 2009, p135). En el capítulo 13 del mismo libro, explican cómo esto también puede conducir a problemas relacionados con la presión arterial (TRESTBPS). Por último, estudios realizados por el instituto nacional de ciencias médicas y nutrición de México afirman la edad puede, por si sola, influir en la manera en la que el cuerpo maneja la glucosa (Lerman & Rosales, 2010), por lo que trazamos también esta relación entre ambas variables (ilustración 7).

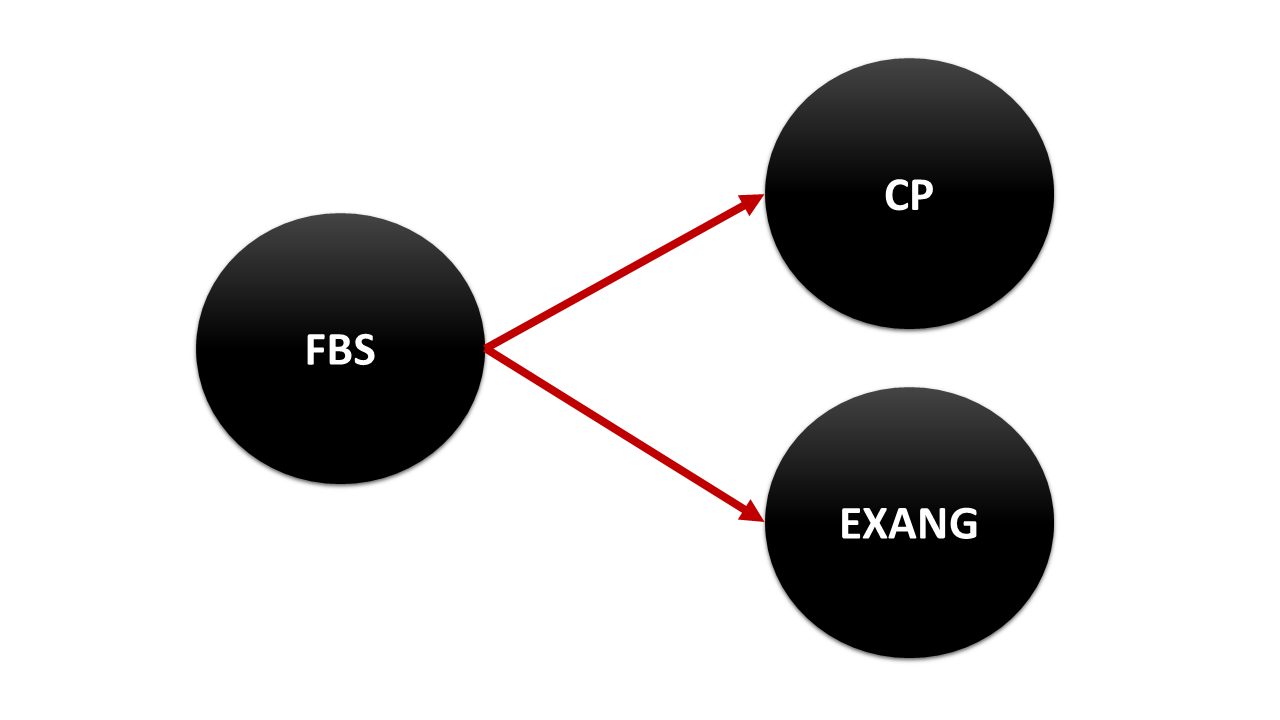


Ilustración 8

Según el Instituto Nacional de Diabetes y Enfermedades Digestivas y de Riñón de Los Estados Unidos - NIDDK, altos niveles de azúcar en la sangre pueden causar problemas cardiacos ya que esta puede dañar los vasos sanguíneos y los nervios que los controlan. Así mismo, sugieren a los pacientes con angina que si experimentan síntomas relacionados con niveles extremos de azúcar busquen asistencia médica inmediatamente (Nanayakkara, Curtis, Heritier, 2021) (Diwivedi & Pandey, 2020). Es por esto que la variable FBS (Fasting Blood Sugar) se encuentra representada como padre de CP (Chest Pain o Angina) y EXANG (Angina inducida por ejercicio). (ilustración 8).

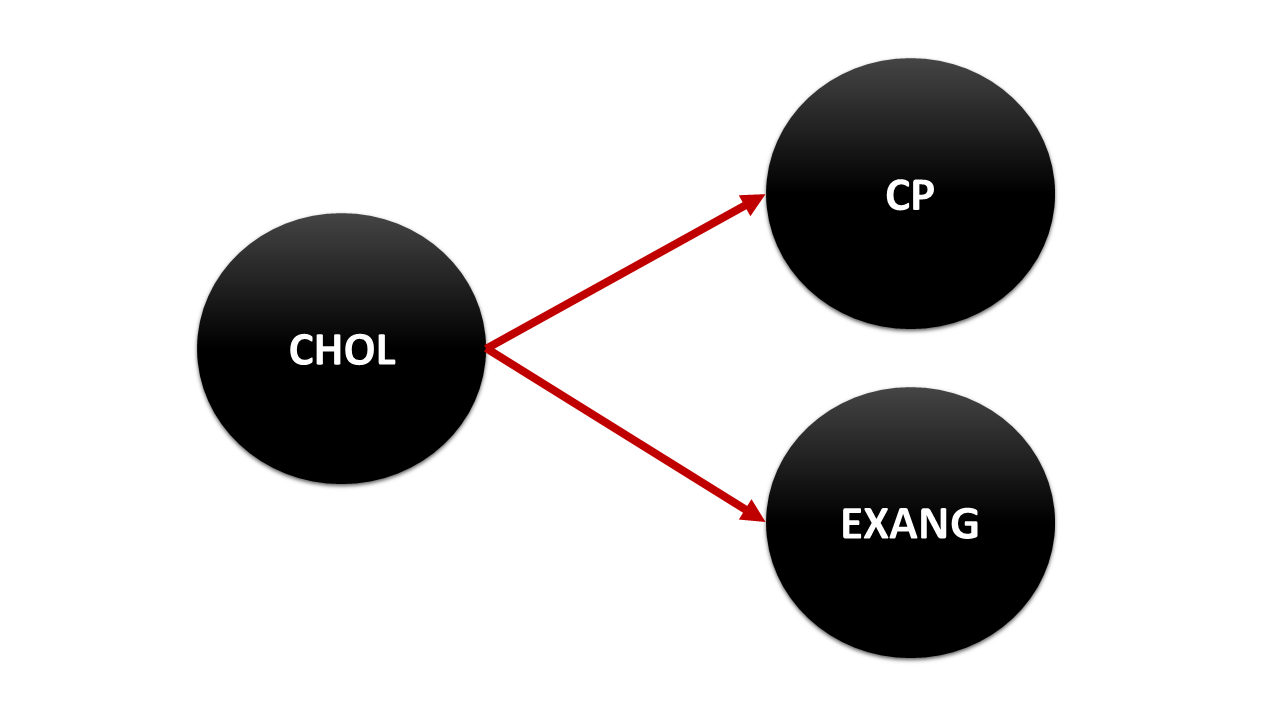


Ilustración 9

Según el libro de la Salud Cardiovascular los tratamientos para la angina de pecho incluyen el uso de fármacos, no obstante, sugieren a los pacientes que:

***“Nunca se hará suficiente hincapié en factores como el abandono del hábito fumador, el tratamiento de las cifras elevadas de lípidos —y en especial del colesterol, de la diabetes mellitus, de la hipertensión y de la obesidad—, así como en favorecer una actividad física moderada.”*** (Calvo & Nieto, 2009, p.256)

Por lo que tomamos a la variable CHOL(Colesterol) como padre para las variables CP(Angina) y EXANG (Angina inducida por ejercicio). (ilustración 9).

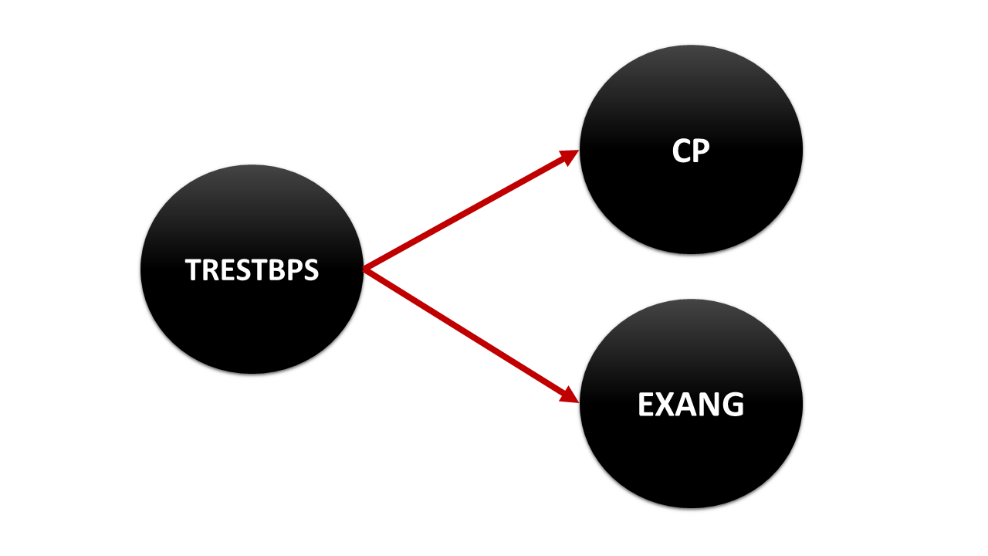


Ilustración 10

De acuerdo con el Instituto Nacional del Corazón, Pulmones y Sangre de los Estados Unidos - NIH, uno de los principales desencadenantes de la angina es la presión arterial, por lo que la variable TRESTBPS (Resting blood pressure) se encuentra representada como padre de las variables CP(Angina) y EXANG (Angina inducida por ejercicio). (ilustración 10).

Según el mismo instituto la angina está directamente relacionada con diversos problemas cardiacos por lo que así representamos estas variables como padres para la variable de interés de nuestro análisis causal:

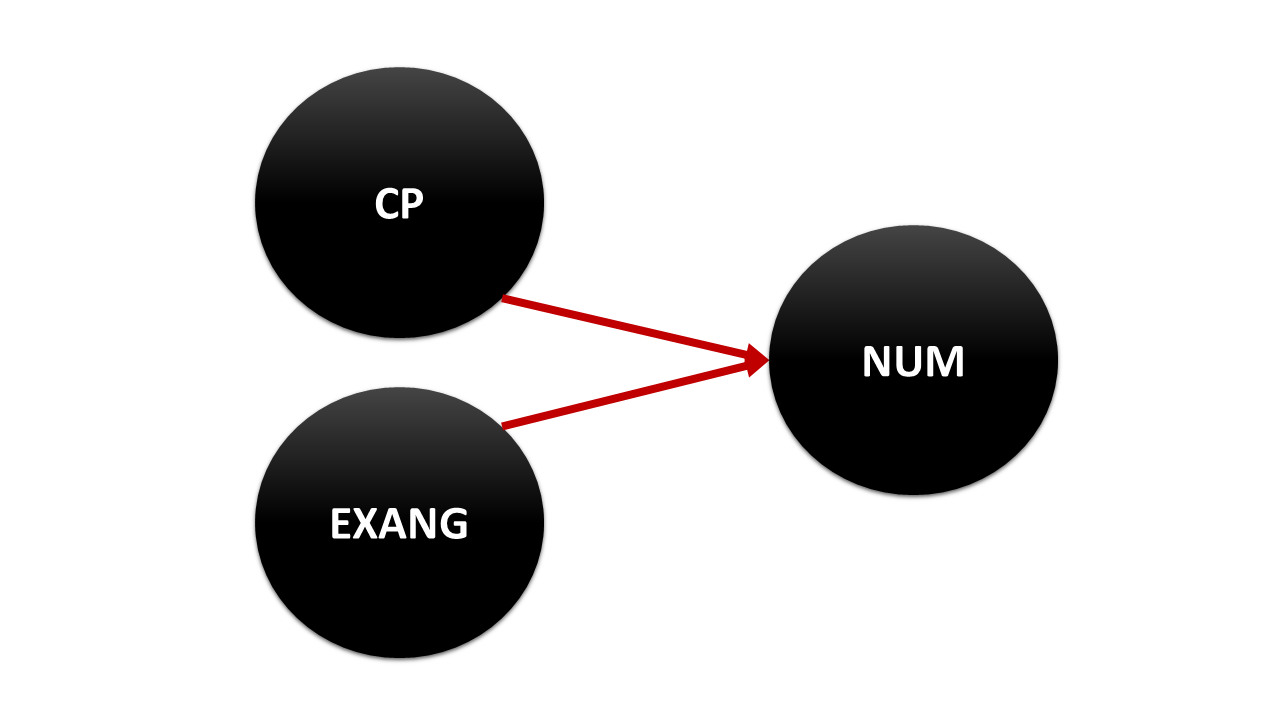


Ilustración 11

Como resultado de lo anterior, a continuación, se presenta la estructura de nuestra red bayesiana que será utilizada para realizar inferencias mediante análisis causal, tomando todas las variables como evidencia a excepción de NUM, la variable de interés:

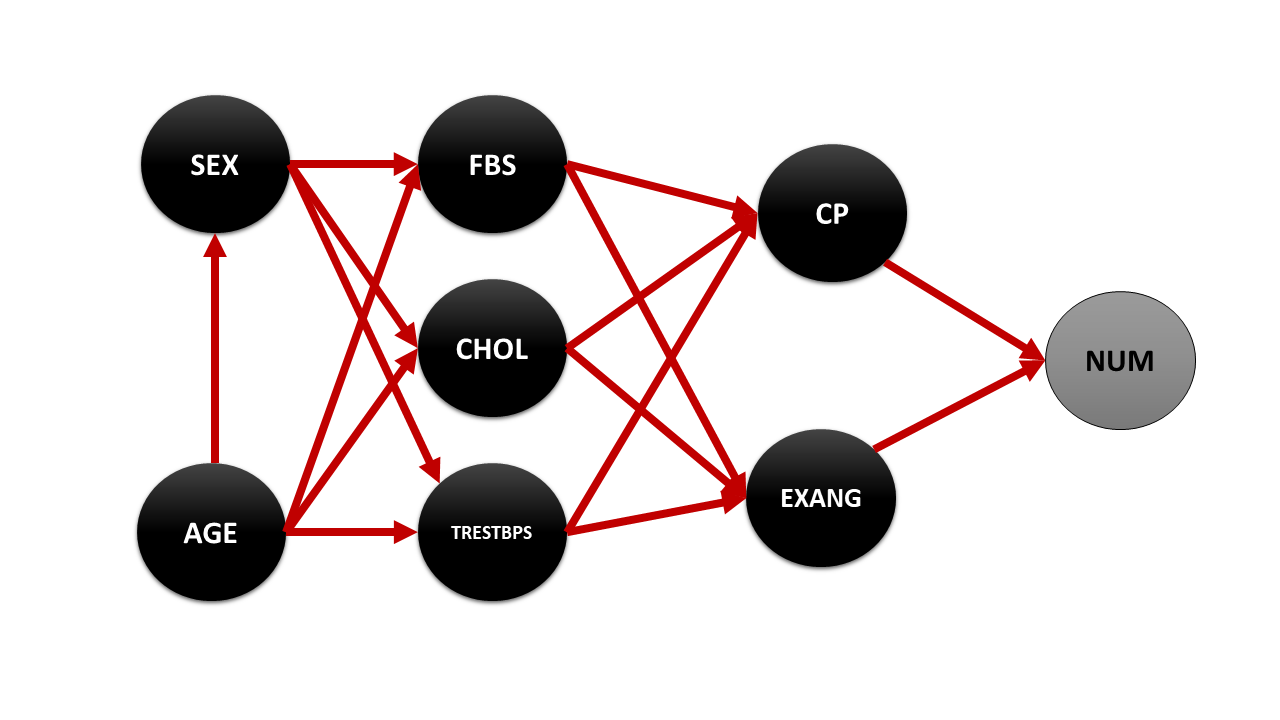
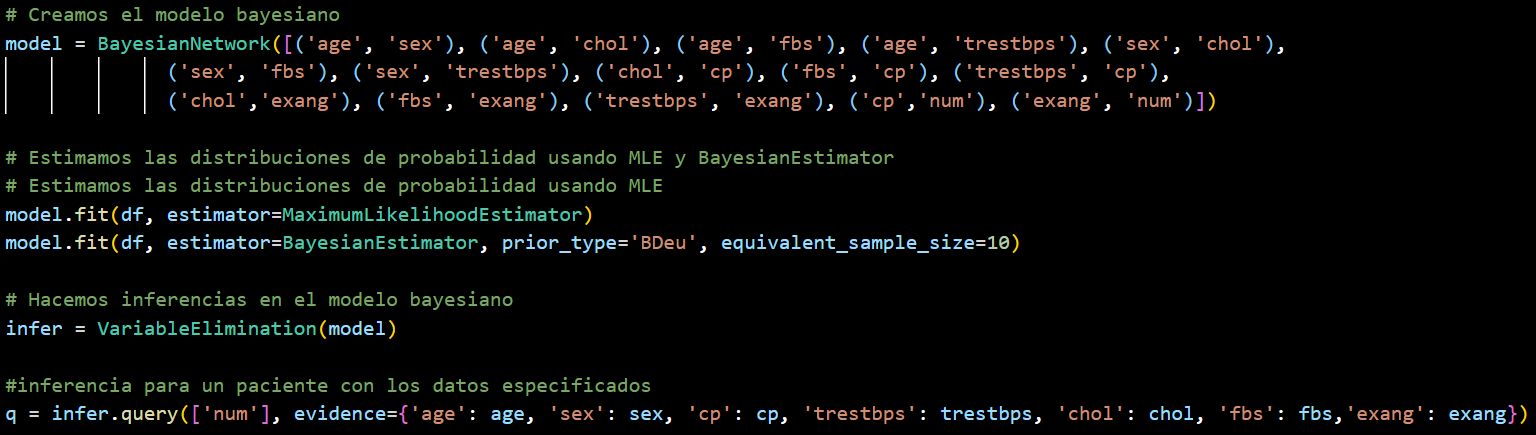


Ilustración 12

Como producto de este proyecto, esperamos que el paciente, a partir de datos como: edad, sexo, reconocimiento de dolor en el pecho, niveles de azúcar, colesterol y presión arterial (datos que según el gremio médico son algunos de los exámenes más comúnmente realizados); pueda reconocer el nivel de probabilidad de existencia o presencia de enfermedades cardiacas sin importar su grado, por lo que la variable NUM es tomada como binaria. Esto podría permitir que los pacientes accedan con mayor prioridad a los servicios de asistencia médica y que les realicen exámenes más específicos relacionados con enfermedades del corazón.

**La Aplicación**

Para el desarrollo del aplicativo, implementamos la red bayesiana, y ajustamos el modelo utilizando el estimador de máxima verosimilitud MLE. Posteriormente utilizamos estimación bayesiana con una distribución a-priori Dirichlet con un tamaño de 10 para la muestra. Lo anterior con el fin de realizar las inferencias pertinentes dada la evidencia. Aquí el código utilizado:



A continuación, se muestra una breve descripción de los elementos de la interfaz de la aplicación.

Graphical user interface, text, application

Description automatically generated

Los valores tomados para la edad **AGE** son números enteros positivos. Es importante mencionar que edades que están por fuera de los datos utilizados para el aprendizaje del modelo, no serán computadas por la aplicación. La variable de sexo **SEX** toma valores binarios asignados a ‘hombre’ y ‘mujer’. Las variables **CP** Y **EXANG**, son tomadas por el siguiente dropdown menú, el cual ofrece cuatro tipos de valores categóricos asignados a números enteros para **CP** y dos valores para **EXANG**. La presión arterial **TRESTBPS** es también un valor numérico que junto con el nivel de colesterol **CHOL**. Por último, la variable FBS es también un valor binario que registra si la persona se encuentra o no por encima de los niveles normales. El reporte para el paciente puede mostrarse, ocultarse y actualizarse utilizando el botón “CALCULAR”.

**Resultados de la aplicación**

La aplicación mostrara recomendaciones basadas en el estado de los signos vitales de los pacientes, de esta manera describirá qué acciones puede tomar el paciente según los signos que estén fuera de los valores normales. Dichas recomendaciones no tienen contraindicaciones notorias y están respaldadas por la información publicada en MedlinePlus quien es producida por la Biblioteca Nacional de Medicina de Estados Unidos NLM. A continuación, se muestra un ejemplo en el que todas las recomendaciones posibles están visibles:

****

Por último, mostramos como podría verse el reporte par aun paciente completamente saludable según los parámetros establecidos.

Chart

Description automatically generated

**Despliegue en la nube**

La aplicación fue desplegada utilizando una imagen de Linux en la siguiente instancia en la nube:



El código fue ejecutado desde dicha instancia la cual tiene una IP public address 54.144.133.47. En este caso corrimos la aplicación en el puerto 8070.

Text

Description automatically generated

**Bibliografía**

Cachofeiro, V. C., Dra. (2009). *libro de la salud cardiovascular*. Fundacion BBVA.

Calvo, E. C. M., Dr., & Nieto, A. N. S., Dr. (2009). *libro de la salud cardiovascular*. https://www.fbbva.es/microsites/salud\_cardio/mult/fbbva\_libroCorazon\_cap27.pdf

*Diabetes, Heart Disease, & Stroke*. (2023, 28 febrero). National Institute of Diabetes and Digestive and Kidney Diseases. https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/heart-disease-stroke

Dwivedi, M. D., & Pandey, A. R. P. (2020). Diabetes Mellitus and Its Treatment: An Overview. *Journal of Advancement in Pharmacology*, *1*(1). https://d1wqtxts1xzle7.cloudfront.net/64743286/Diabetes\_mellitus\_48\_58\_-libre.pdf

Lerman-Garber, I., & Rosales-Calderón, M. (2010). Cambios en la tolerancia a la glucosa en el anciano. *Revista De Investigacion Clinica*, *62*(4), 312-317.

Tovar, A. C., & Celemin, C. (2015). Menopausia y síndrome metabólico. *RFS revista facultad de salud universidad surcolombiana sede neiva*. https://doi.org/10.25054/rfs.v7i1.183

Vieira, v. s., & Medrado, B. M. (2021). Hormonas y Mujeres en la Menopausia. *Scielo Brazil*. https://doi.org/10.1590/1982

*What Is Angina?* (2022, 24 marzo). NHLBI, NIH. <https://www.nhlbi.nih.gov/health/angina>