Predicción de Enfermedades Cardíacas

Universidad de los Andes, Bogotá, Colombia
Analítica Computacional para la Toma de Decisiones

Santiago González Montealegre	s.gonzales35	202012274
Juliana Carolina Cárdenas Barragán	jc.cardenasb1	202011683

Fecha de presentación: Marzo 14 de 2023

Tabla de contenido

1 Análisis Exploratorio de Datos

Se explorarán los datos para Enfermedades Cardíacas, el cual es compuesto por 14 atributos para los que se cuenta con 303 observaciones. De estos 14 atributos, 5 son variables cuantitativas (numéricas) y 9 son variables categóricas. A continuación, se exploran estas variables.

1.1 Variables

1.1.1 Numéricas

- Edad (`age`): la edad del paciente medida en años.
- Presión Arterial (`trestbps`): presión arterial en reposo al entrar al hospital medida en mm Hg.
- Colesterol Sérico (`chol`): el colesterol sérico medido en mg/dl.
- Frecuencia cardiaca máxima (`thalach`): la frecuencia cardíaca máxima alcanzada por el paciente.
- Depresión Segmento ST (`oldpeak`): depresión en el segmento ST inducida por el ejercicio relativo al cansancio.

1.1.2 Categóricas

- Sexo (`sex`): Sexo del paciente.
  - 1: Masculino
  - 0: Femenino
- Tipo Dolor Pecho (`cp`): el tipo de dolor en el pecho del paciente.

        - 1: Angina típica
        - 2: Angina atípica
        - 3: Dolor no anginal
        - 4: Asintomático
    - Azúcar en Sangre (`fbs`): el azúcar en sangre en ayunas es mayor a 120 mm Hg
        - 1: Verdadero
        - 0: Falso
    - Resultados Electrocardiográficos Reposo (`restecg`):
        - 0: Normal
        - 1: Anomalía en la onda ST
        - 2: Hipertrofia ventricular izquierda
    - Angina Inducida Ejercicio (`exang`): si la angina es inducida por hacer ejercicio
        - 1: Verdadero
        - 0: Falso
    - Pendiente Segmento ST Ejercicio (`slope`): el tipo de pendiente del segmento ST en el nivel pico de ejercicio.
        - 1: Ascendente
        - 2: Plana
        - 3: Descendente
    - Número Vasos Coloreados (`ca`): el número de vasos coloreados por fluoroscopia.
        - 0: 0 vasos coloreados
        - 1: 1 vasos coloreados
        - 2: 2 vasos coloreados
        - 3: 3 vasos coloreados
    - Talasemia (`thal`): nivel de presencia de la enfermedad talasemia.
        - 3: Normal
        - 6: Defecto fijo – daño permanente en el corazón.
        - 7: Defecto reversible – un área del corazón no recibe suficiente sangre, pero se puede restaurar.
    - Presencia Enfermedad Cardíaca (`heartdis`):
        - 0: Ausencia de enfermedad
        - 1: Presencia de enfermedad

1.2 Análisis Univariado

Una vez se tienen claras las variables y la categoría a la cual pertenece cada una, se procede a realizar un análisis de las distribuciones de estas. Sin embargo, para algunas columnas (`ca` y `thal`) se hallan registros con datos nulos o faltantes; dado que ambas variables que presentan faltantes son categóricas se decide utilizar la moda como mecanismo para llenar estos datos faltantes, con el fin de evitar tener que eliminar estas filas.

Una vez resuelto el problema de los valores nulos, se procede analizar cada una de las variables para las que tenemos datos, dividiendo este análisis en variables Numéricas y Categóricas. El código y las gráficas generadas se pueden observar en el repositorio en la carpeta `Soportes` en el archivo `EDA_Univariate.ipynb` (1).

1.2.1 Variables Numéricas

El análisis para las variables numéricas consistirá en cuatro gráficas: un histograma que se realizará con el número de clases definido; $k = \sqrt{n}$, una gráfica de densidad ajustada a la distribución, un diagrama de caja y un diagrama de violín.

- **Edad**

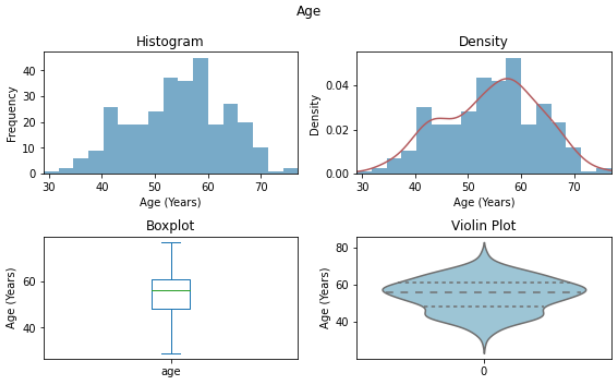

Figura 1. Univariado age

Los datos de edad no se encuentran uniformemente distribuidos, presenta cierta forma de campana alrededor de los 50 y los 60 años, además, se presenta otro pico alrededor de los 40 años. El dato mínimo de la edad es de 29 años y hasta los 40 años no hay gran cantidad de datos por lo que se puede concluir que no se tienen datos de población joven. El dato máximo de la edad es de 77.

- **Presión Arterial en Reposo**

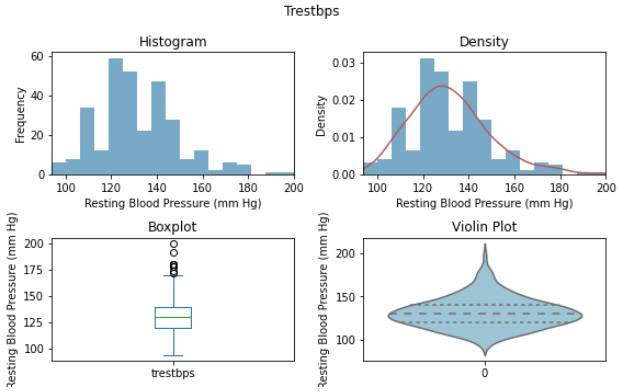

Figura 2. Univariado trestbps

Los datos de la presión de la sangre en reposo no se encuentran uniformemente distribuidos. Se presenta cierta forma de campana alrededor de 130 mm Hg. Esta forma de campana no es simétrica ya que tiende a estar cargada al lado derecho de la media. El dato mínimo es de 94 y el dato máximo es de 200.

- **Colesterol Sérico**

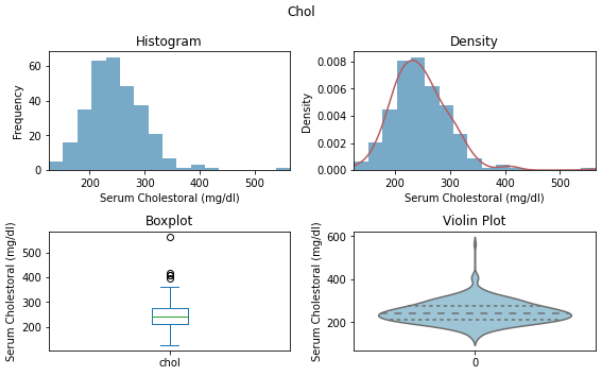

Figura 3. Univariado chol

Los datos del colesterol sérico no se encuentran uniformemente distribuidos. Se presenta cierta forma de campana alrededor de 220 mm Hg, aunque no es de forma simétrica ya que presenta cierta carga hacia el lado derecho de la media. El dato mínimo es de 126 y el dato máximo es de 564.

- **Frecuencia Cardíaca Máxima**

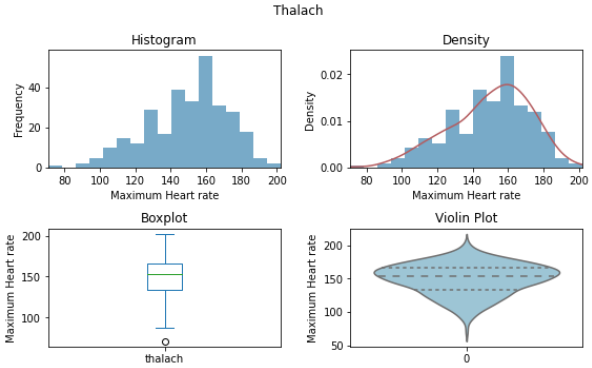

Figura 4. Univariado thalach

Los datos de la frecuencia cardíaca máxima no se encuentran uniformemente distribuidos. Se presenta cierta forma de campana alrededor de 160. Esta forma de campana es asimétrica, ya que el lado izquierdo de la media se encuentra cargado. El dato mínimo es de 71 y el dato máximo es de 202.

- **Depresión Segmento ST**

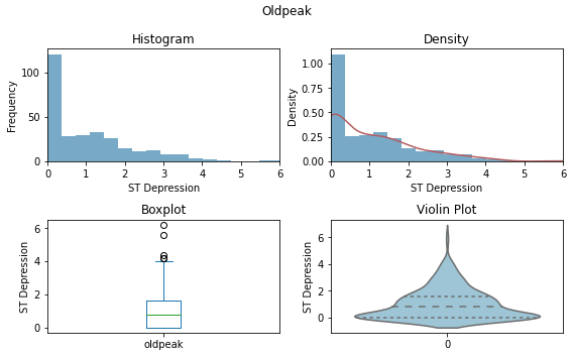

Figura 5. Univariado oldpeak

Los datos de la presión de la sangre en reposo no se encuentran uniformemente distribuidos. Se presenta cierta similitud con la distribución exponencial, aunque no llega a ajustarse. El dato mínimo es de 0 y el dato máximo es de 6.2.

1.2.2 Variables Categóricas

Para las variables categóricas se realiza un análisis a partir de las cuentas para cada uno de los valores que puede tomar la variable categórica, esto permitirá saber la distribución de la variable y ver si los datos se encuentran distribuidos uniformemente o si hay prelación por ciertos atributos de la variable.

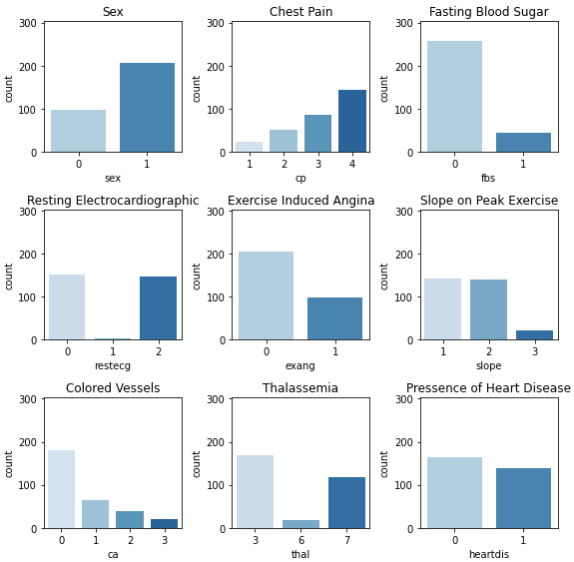

Figura 6. Univariado Categóricas

Se puede observar que hay más datos de personas con sexo masculino. Mayormente personas asintomáticas respecto al dolor de pecho. Con nivel de sangre en ayunas principalmente menor a 120 mm Hg. Los resultados electrocardiográficos en reposo son casi la mitad normales, casi otra mitad de hipertrofia ventricular izquierda, pero con pocos datos para anomalías en la onda. La mayor parte de las personas no sufren de angina inducida por ejercicio. La pendiente del segmento ST en el pico de ejercicio es casi equiprobale que sea ascendente o plana, mientras que sea descendente es muy poco probable. Más de la mitad de los datos son de personas con 0 vasos coloreados por fluoroscopia. Más de la mitad de los datos no tienen talasemia, seguido por talasemia con efecto reversible y con pocos datos de talasemia de defecto fijo. Finalmente, con distribución casi equiprobable hay o no presencia de enfermedades cardíacas en los datos.

1.3 Análisis Multivariado

Para hacer el análisis de varias variables en conjunto con el fin de identificar correlacion se divide en 3 partes: variables categóricas contra categóricas, variables categóricas contra numéricas y variables numéricas contra numéricas. El código y las gráficas generadas se pueden observar en el repositorio en la carpeta `Soportes` en el archivo `EDA_Multivariate.ipynb` (1).

1.3.1 Categóricas vs. Categóricas

Para el análisis multivariado de las variables categóricas contra las variables categóricas se realizaron countplots con el fin de identificar si los distintos valores de la variable categórica afectan o se ven afectados por las demás variables categóricas.

- Sexo

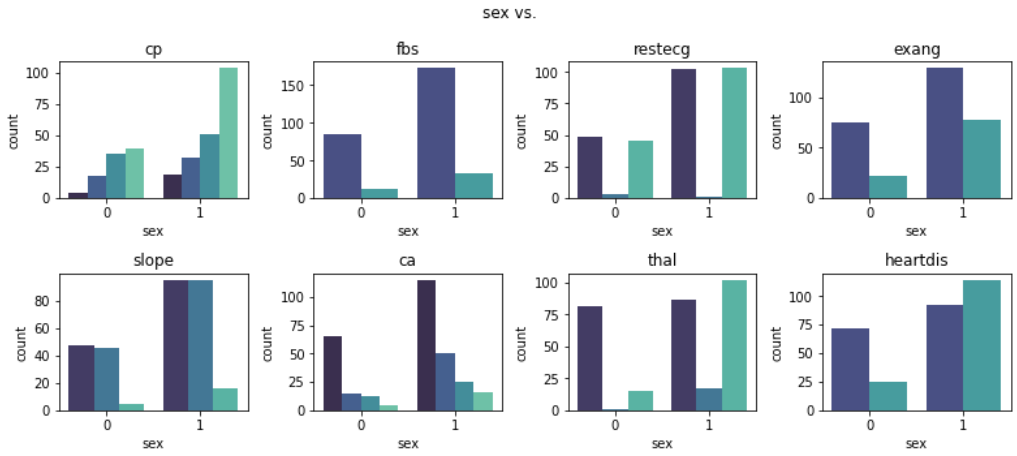

Figura 7. Multivariado categóricas sex

El sexo si afecta o se ve afectado por las variables categóricas tal y heartdis. Esto dado que hay una diferencia en la distribución de estas variables categóricas si el sexo es femenino a si es masculino.

- Tipo Dolor de Pecho

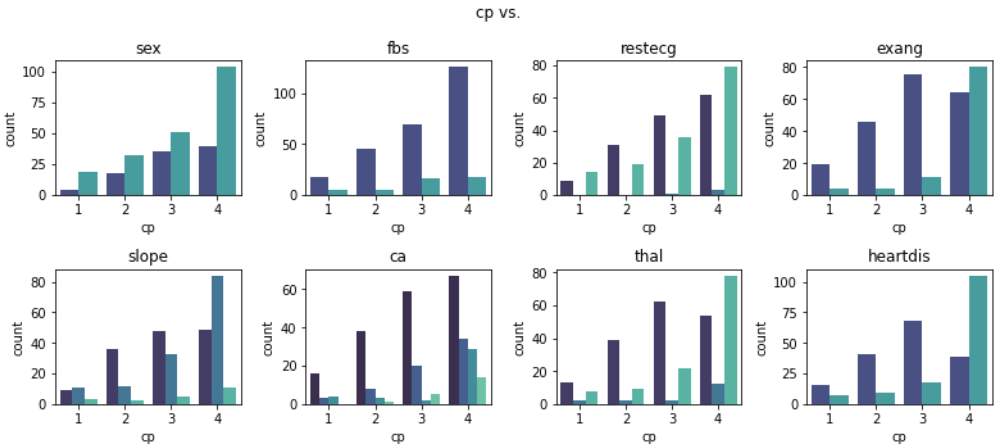

Figura 8. Multivariado categórico cp

El tipo de dolor de pecho si afecta o se ve afectado por las variables categóricas exang y heartdis. Esto dado que hay una diferencia en la distribución de estas variables categóricas si el dolor es anginal típico, atípico, no anginal o asintomático.

- Azúcar en sangre

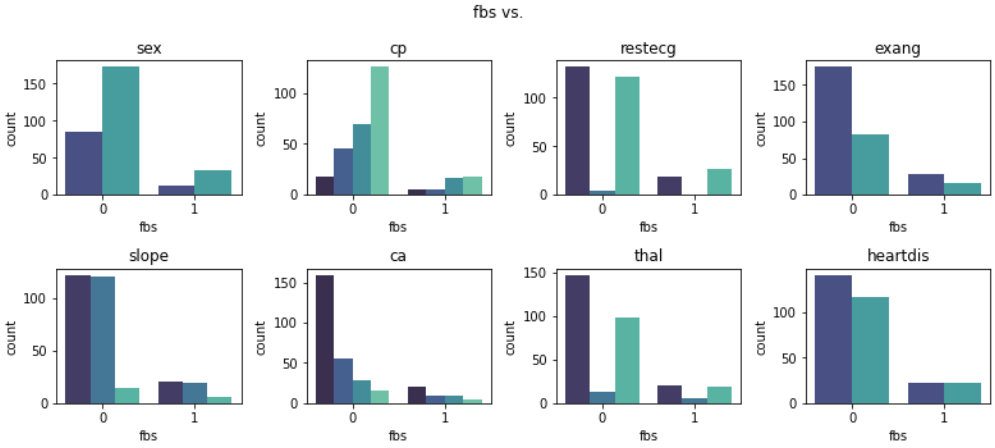

Figura 9. Multivariado categórico fbs

Si el azúcar en sangre en ayunas es mayor a 120 mm Hg no afecta o se ve afectado por las variables categóricas. Esto dado que hay no una diferencia en la distribución de estas variables categóricas si es mayor o no el azúcar.

- Resultados electrocardiográficos reposo

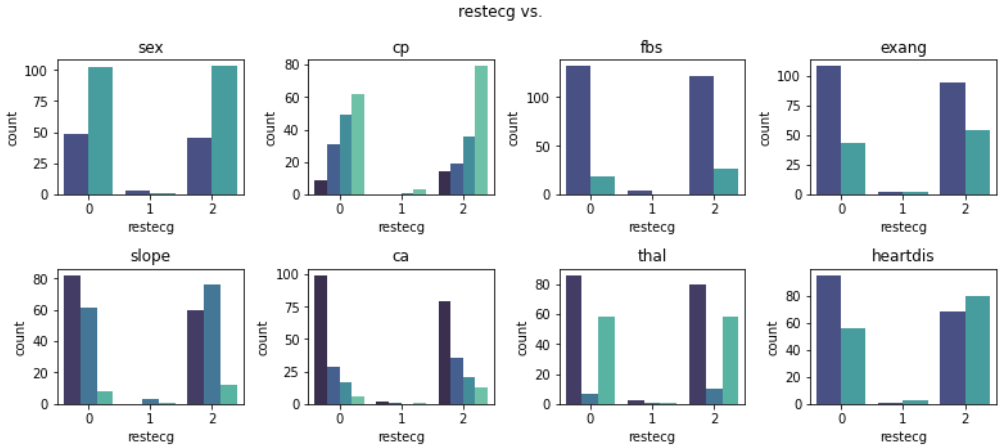

Figura 10. Multivariado categórico restecg

El resultado electrocardiográfico en reposo no afecta o se ve afectado por las variables categóricas. Esto dado que hay no hay diferencia en la distribución de las variables categóricas dado el resultado del examen.

- Angina inducida por ejercicio

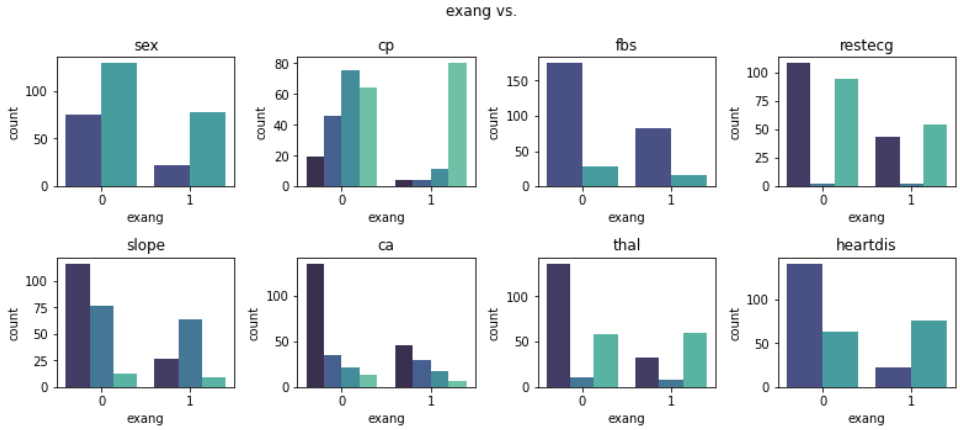

Figura 11. Multivariado categórico exang

Si la angina se produce por hacer ejercicio si afecta o se ve afectado por las variables categóricas thal y heartdis. Esto dado que hay una diferencia en la distribución de estas variables si la angina es producida o no.

- Pendiente segmento ST ejercicio

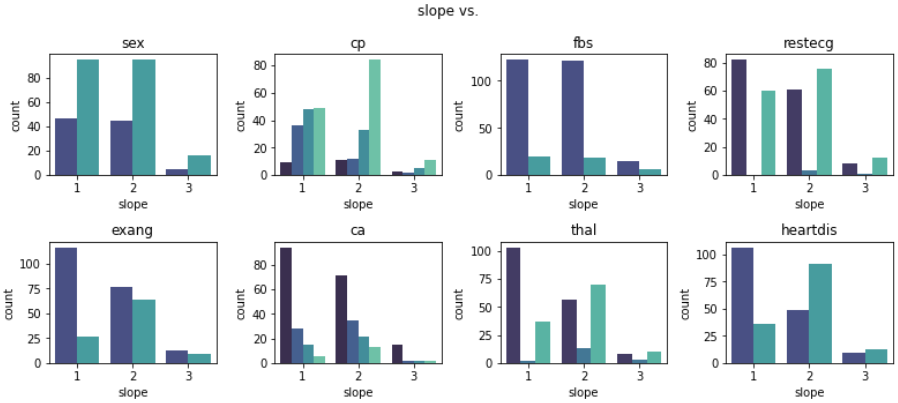

Figura 12. Multivariado categórico slope

La pendiente del segmento ST en el pico de ejercicio si afecta o se ve afectado por la variable categórica heartdis. Esto dado que hay una diferencia en la distribución de esta variable categórica si la pendiente cambia.

- Número vasos coloreados

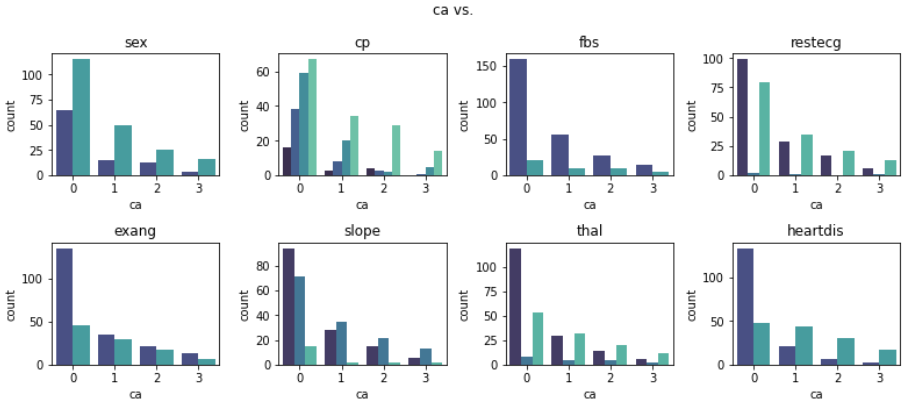

Figura 13. Multivariado categórico ca

El número de vasos coloreados por fluoroscopia si afecta o se ve afectado por la variable categórica heartdis. Esto dado que hay una diferencia en la distribución de esta variable categórica si el número cambia.

- Talasemia

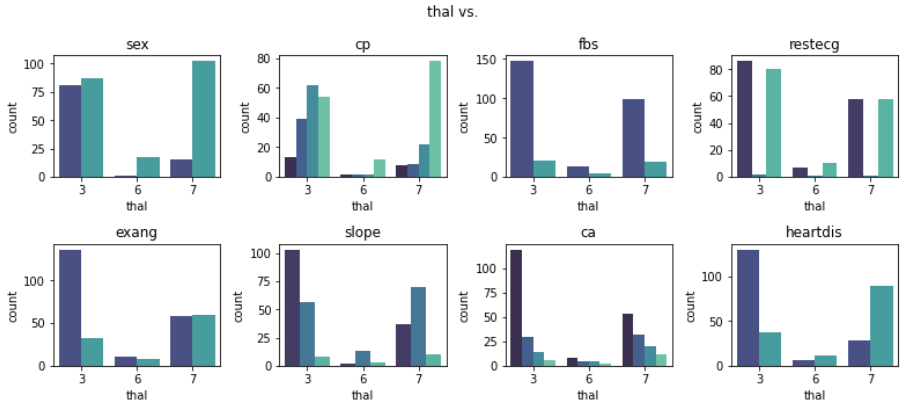

Figura 14. Multivariado categórico thal

La enfermedad talasemia si afecta o se ve afectado por las variables categóricas sex, cp, exang, slope, ca y heartdis. Esto dado que hay una diferencia en la distribución de estas variables categóricas según el tipo de talasemia.

- Presencia enfermedad cardíaca

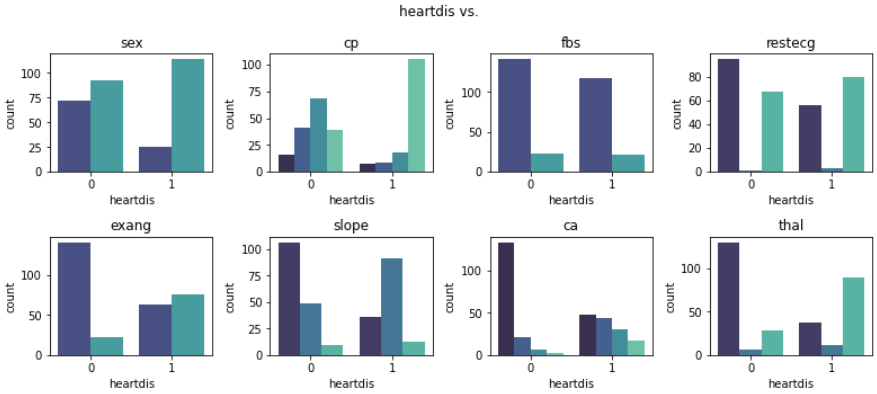

Figura 15. Multivariado categórico heartdis

La presencia de enfermedades cardiacas si afecta o se ve afectado por las variables categóricas sex, exang, slope, thal. Esto dado que hay una diferencia en la distribución de estas variables si la pendiente cambia.

1.3.2 Categóricas vs. Numéricas

Para el análisis multivariado de las variables categóricas contra las variables numéricas se realizaron diagramas de violín y de caja con el fin de identificar si los distintos valores de la variable categórica afectan o se ven afectados por la variable numérica.

- **Sexo**

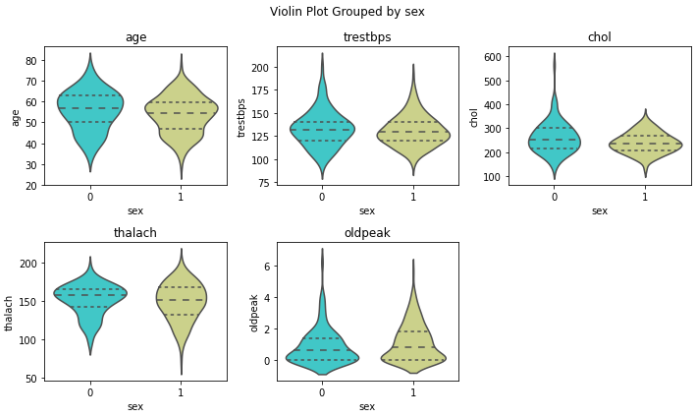

Figura 16. Multivariado sex

El sexo de la persona no afecta o se ve afectado por ninguna variable numérica. Esto se debe a que no hay diferencia en la distribución de las variables numéricas si la persona es de sexo femenino o masculino. Este fenómeno sucede a pesar de que se tienen más datos de personas de sexo masculino lo que reafirma que el sexo no afecta variables numéricas.

- **Tipo Dolor de Pecho**

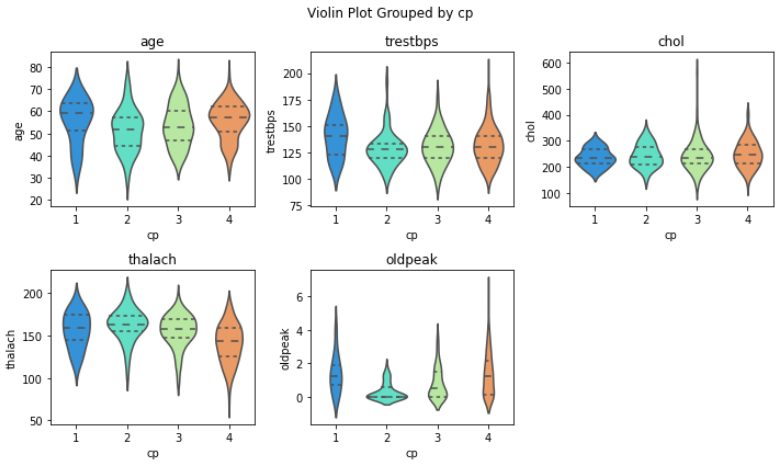

Figura 17. Multivariado cp

El tipo de dolor de pecho no afecta o se ve afectado por ninguna variable numérica. Esto se debe a que no hay diferencia en la distribución de las variables numéricas si el dolor de pecho es de angina típica, atípica, no anginal o sin dolor. Este fenómeno sucede a pesar de que se tienen más datos de personas sin dolor de pecho o asintomáticos.

- **Azúcar en Sangre**

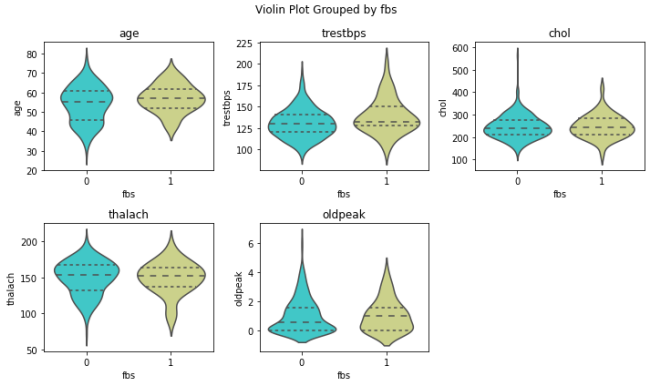

Figura 18. Multivariado fbs

Si el azúcar en sangre en ayunas es mayor a 120 mm Hg no afecta o se ve afectado por ninguna variable numérica. Esto se debe a que no hay diferencia en la distribución de las variables numéricas si es mayor o no el azúcar en sangre en ayunas a 120 mm Hg. Este fenómeno sucede a pesar de que se tienen más datos de personas con nivel de azúcar en sangre menor a 120 mm Hg.

- **Resultados Electrocardiográficos Reposo**

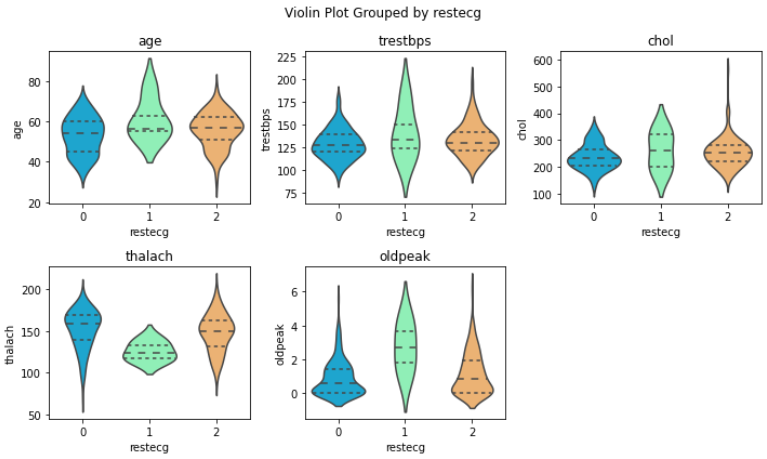

Figura 19. Multivariado restecg

Los resultados electrocardiográficos en reposo si afectan o se ven afectados por las variables numéricas thalach y oldpeak. Esto dado que hay una diferencia en la distribución de estas variables numéricas si los resultados son normales, si hay anomalía en la onda ST o si hay hipertrofia ventricular. Sin embargo, esta diferencia se da cuando los resultados son de anomalía en la onda ST y esto se puede deber a que se cuenta con muy pocos datos (4 de 303) para este valor de la categórica. Por lo tanto, toca tomar estos hallazgos con cuidado.

- **Angina Inducida Ejercicio**

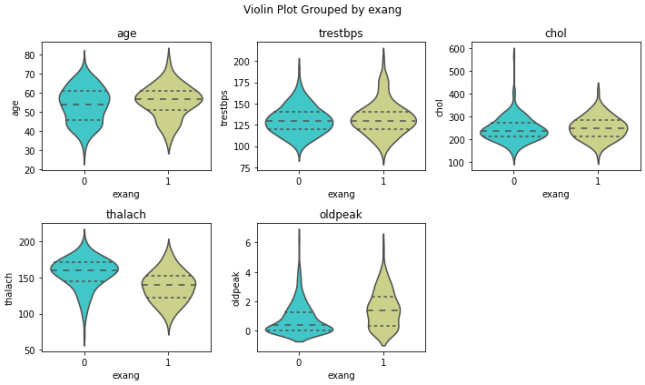

Figura 20. Multivariado exang

Si la angina es provocada o inducida por el ejercicio si afecta o se ve afectada por las variables numéricas thalach y oldpeak. Esto dado que hay una diferencia en la distribución de esta variable numérica si la angina es provocada por el ejercicio o no.

- **Pendiente Segmento ST Ejercicio**

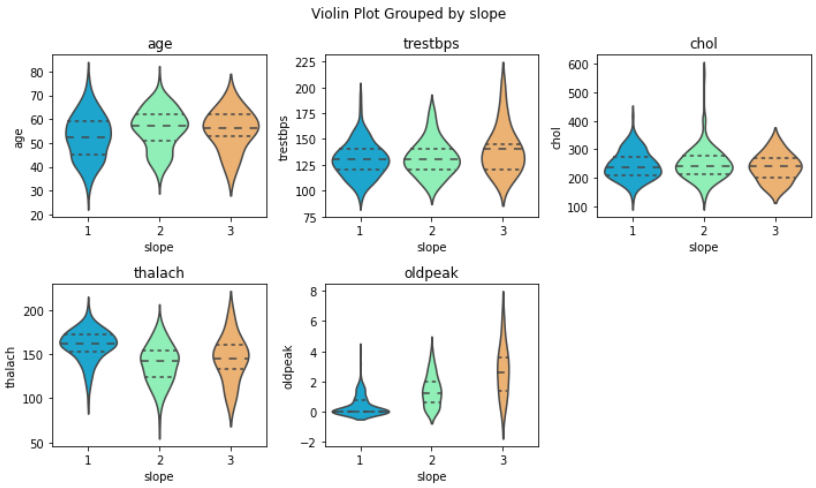

Figura 21. Multivariado slope

La pendiente del segmento ST en el ejercicio si afecta o se ve afectada por las variables numéricas thalach y oldpeak. Esto dado que hay una diferencia en la distribución de estas variables numéricas si la pendiente es ascendente, plana o descendente.

- **Número Vasos Coloreados**

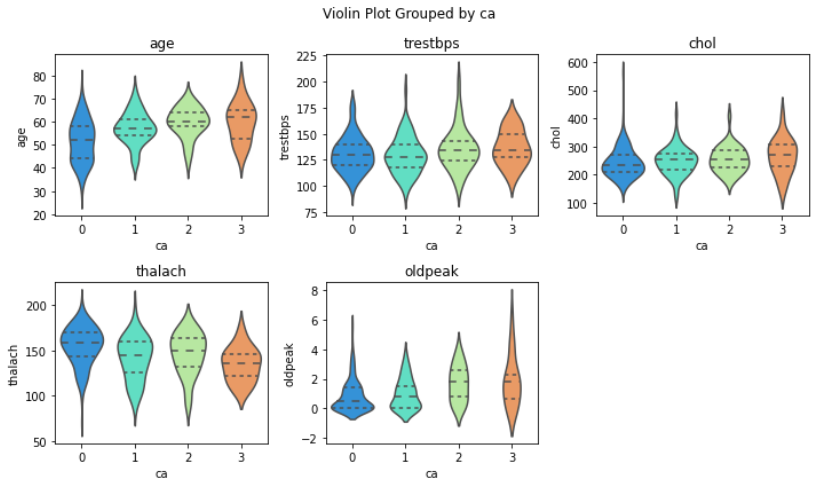

Figura 22. Multivariado ca

El número de vasos coloreados por fluoroscopia si afectan o se ven afectados por las variables numéricas age, thalach y oldpeak. Esto dado que hay una diferencia en la distribución de estas variables numéricas si el número de vasos es 0, 1, 2 o 3.

- **Talasemia**

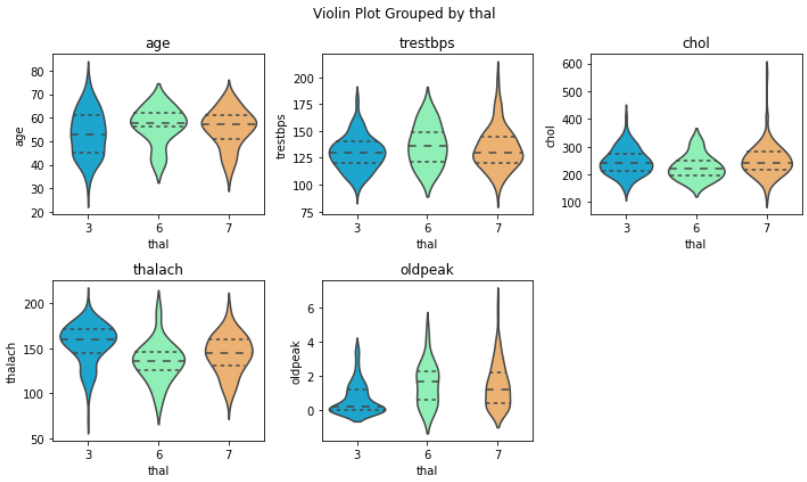

Figura 23. Multivariado thal

El nivel de presencia de Talasemia si afecta o se ve afectada por la variable numérica thalach. Esto dado que hay una diferencia en la distribución de esta variable numérica si es normal, con defecto fijo o con defecto reversible.

- **Presencia Enfermedad Cardíaca**

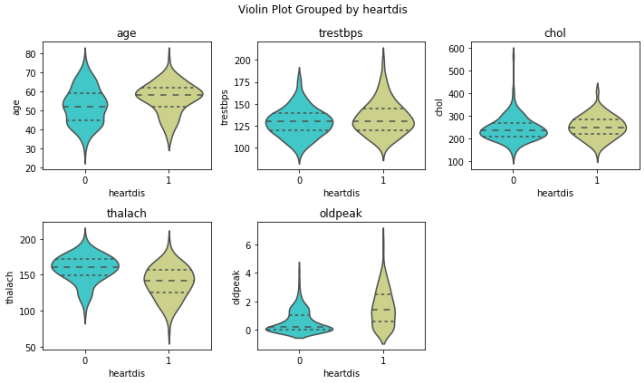

Figura 24. Multivariado heartdis

La presencia de enfermedad cardiaca si afecta o se ve afectada por las variables numéricas age, thalach y oldpeak. Esto dado que hay una diferencia en la distribución de estas variables numéricas si hay presencia de enfermedad o no.

1.3.3 Numéricas vs. Numéricas

Para el análisis de variables numéricas contra variables numéricas se tomará el coeficiente de correlación lineal (ρ) y diagramas de dispersión como base para identificar relaciones entre las variables.

Figura 25. Matriz de dispersión y correlación

Se puede observar que no hay relaciones altamente correlacionadas y que muestren un patrón entre las variables. Sin embargo, es posible destacar las relaciones thalach-age y age-trestbps, que tienen los coeficientes de correlación lineal más significativos $\rho_{thalach,age} = -0.39$; $\rho_{age,trestbps} = 0.28$; mostrando que entre mayor sea la edad, menor será la frecuencia cardíaca máxima y mayor será la presión arterial en reposo.

2 Modelo de Red Bayesiana

Con base en el análisis de las variables del modelo se construye un grafo para representar el problema por medio de una red bayesiana, este grafo se puede observar en el repositorio en la carpeta `Red Bayesiana` en el archivo `Red Bayesiana Inicial.png` (1)Sin embargo, esta red es bidireccional por lo que necesitamos volver los arcos unidireccionales para poder construirla. Se usaron diversas fuentes como revistas y papers de la National Library of Medicine (2), Frontiers in Public Health (3), Journal of the American College of Cardiology (4), International Journal of Cardiology (6), Genetics in Medicine (7), esta versión del grafo unidireccional se puede encontrar en el repositorio en la carpeta `Red Bayesiana` en el archivo `Red Bayesiana Unidireccional.png` (1). Sin embargo, se debe tener en cuenta que no puede haber ciclos, ya que de existir no sería una red bayesiana sobre la que se pueda hacer inferencia, por lo tanto, es necesario eliminar estos ciclos. Este último grafo se encuentra en el repositorio en la carpeta `Red Bayesiana` en el archivo `Red Bayesiana Final.png` (1). Se muestra a continuación:

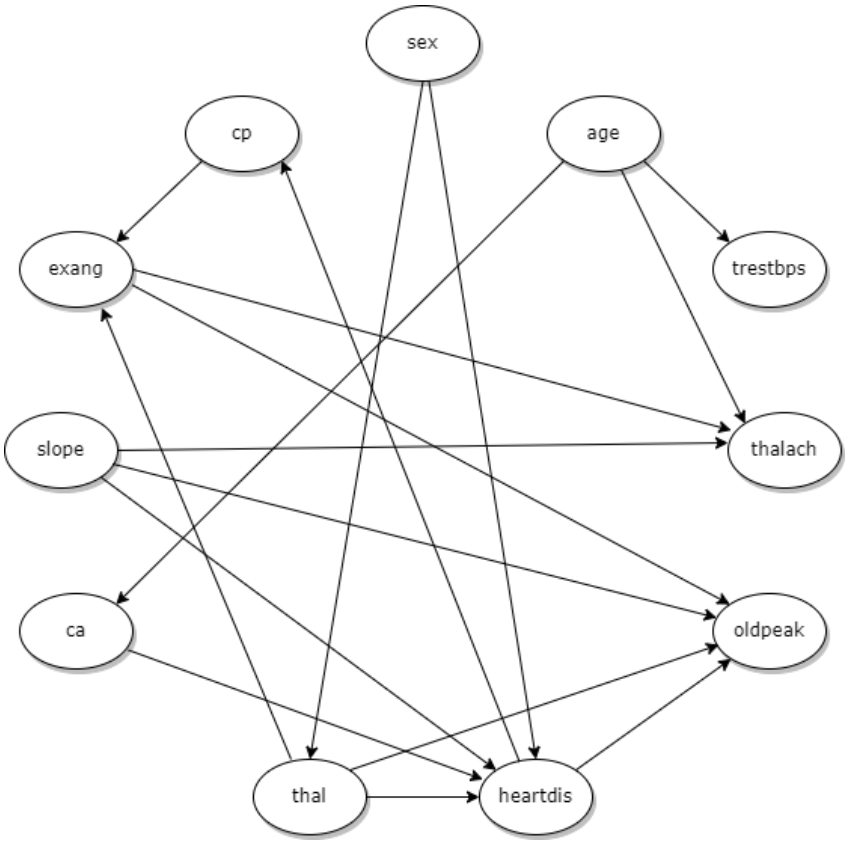

Figura 26. Red Bayesiana Final

Para poder crear el modelo por medio de redes bayesianas es necesario discretizar las variables numericas; age, treastbps, thalach y oldpeak, que se utilizaron en el modelo.

La variable edad presenta una distribución de campana, algo simétrica y las variables afectadas no muestran algo que permita una clasificación, a excepción de ca donde desde alrededor de 52 años (mediana cuando ca es 0) cambia la distribución. Utilizar este valor como separador puede ser sesgado para las inferencias deseadas por lo que se plantearan rangos a partir de los cuartiles para discretizar:

$$[-\infty\ -Q_1] = 1\ \ [Q_1\ -Q_2] = 2\ \ [Q_2\ -Q_3] = 3\ \ [Q_3\ -\infty] = 4$$

Las variables treastbps y thalach presentan una distribución de campana, casi simétrica. Por tal motivo, se usa la media para discretizar:

$$[-\infty - \overline{x}] = 1\ \ [\overline{x} - \infty] = 2$$

Por último, para la variable oldpeak devido a que depende o afecta las variables restecg, exang, slope, ca, thal y heartdis se planteó la siguiente discretización:

$$[0 - 0.5] = 1\ \ [0.5 - 1] = 2\ \ [1 - 1.5] = 3\ \ [1.5 - 2] = 4\ \ [2 - \infty] = 5$$

3 Bibliografía

1. **González Montealegre, Santiago y Cárdenas Barragán, Juliana Carolina.** GitHub. *Proyecto 1*. [En línea] https://github.com/AnaliticaComputacional-2023-10/Proyecto1.

2. **Rodgers, Jennifer, y otros.** National Library of Medicine. *Cardiovascular Risks Associated with Gender and Aging.* [En línea] 27 de Abril de 2019. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616540/.

3. **Birnbaumer, Philipp, y otros.** Frontiers in Public Health. *Heart Rate Performance Curve Is Dependent on Age, Sex, and Performance.* [En línea] 2 de Abril de 2020. https://www.frontiersin.org/articles/10.3389/fpubh.2020.00098/full#:~:text=Aging%20was%20significantly%20related%20to,ANOVA%20(p%20%3C%200.05).

4. Tofler, G. H. (2001). Resting heart rate and the risk of sudden cardiac death in apparently healthy men. Journal of the American College of Cardiology, 38(2), 478-485.

5. Hiatt, W. R. (1990). Association between changes in systolic blood pressure and changes in cardiovascular morbidity and mortality rates: the Systolic Hypertension in the Elderly Program. Circulation, 82(5), 1925-1931.

6. Eichenauer, A. C. (2017). High resting heart rate is associated with peripheral arterial disease in young and middle-aged women. International Journal of Cardiology, 236, 13-17.

7. Cao, A., & Galanello, R. (2010). Beta-thalassemia. Genetics in Medicine, 12(2), 61-76.

8. Hahalis G, Kremastinos DT, Terzis G. (2002) Global myocardial dysfunction in patients with beta-thalassemia major: a Doppler echocardiographic study. Annals of Hematology, 81(6):311-316.