**Universidad de los Andes**

**Facultad de ingeniería**

**Departamento de ingeniería industrial**

**Analítica computacional**

**Proyecto 1**

**14/03/2022**

**Alumno: Pablo José Cortés Sanabria, Isaac Gregory**

**Código: 201111837**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Reporte análisis exploratorio de datos:**

1. **Reporte de entre 5 y 15 páginas con los resultados principales del análisis exploratorio de datos y la modelización.**

El conjunto de datos usado en este análisis fue tomado de [1]; el dataset corresponde al conjunto de datos processed\_cleveland.data. El archivo fua cargado en un notebook de jupyter usando la sintaxis de la ilustración 1. Es importante resaltar que se añadió como argumento del comando pd.read\_csv, na\_values=’?’, con el objetivo de convertir los datos que fueran iguales a ‘?’ a nans, para de esta forma facilitar su manipulación más adelante.

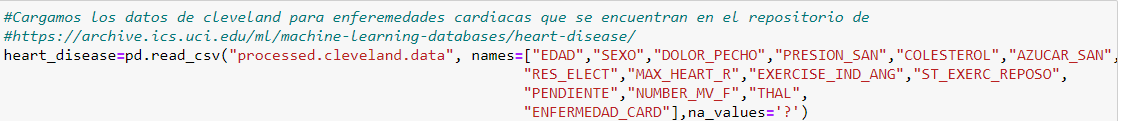


Ilustración . Código usado para leer los datos en el entorno jupyter

El datase en total consta de 303 datos, cada uno con 14 características o columnas. En la ilustración 2 se observa la descripción de cada una de las 14 variables, tal como se especifica en [1]. Usando la función de la librería pandas, .info(), es posible obtener un resumen de las características de las diferentes variables. En la ilustración 3 se aprecia el resultado de aplicar dicha función al conjunto de datos, tal como se cargaron inicialmente.

Como se puede observar en la ilustración 3, todas las variables excepto ENFERMEDAD\_CARD son tipo float64; ENFERMEDAD\_CARD es tipo int64. Adicionalmente se observa que hay unos pocos valores nulos en las variables NUMBER\_MV\_F y THAL. De acuerdo con la información provista por [1], las 14 variables tienen las características que se observa en la ilustración 4 y 5. Se procede entonces a convertir los tipos de datos de cada variable según lo especificado por los autores del dataset. En la ilustración 6 se aprecia el código usado, y en la ilustración 7 el resultado de aplicar la función info() al conjunto modificado de datos.

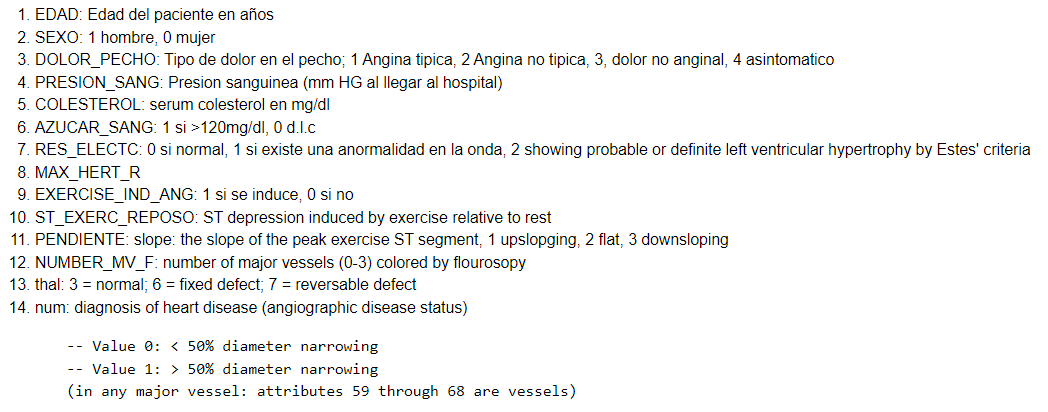


Ilustración . Breve descripción del significad de cada variable

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

Ilustración . Información de las variables proporcionada por la función pandas.info()

Con el objetivo de eliminar los objetos nulos presentes en el dataset, se procede a eliminar las correspondientes filas. Para ello se usó el código de la ilustración 8; en la ilustración 8 puede verse también el total de datos resultante luego de la eliminación de los objetos nulos.

Tabla

Descripción generada automáticamente

Ilustración . Información de variables. Tomado de [1]

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ilustración . Información de variables. Tomado de [1]

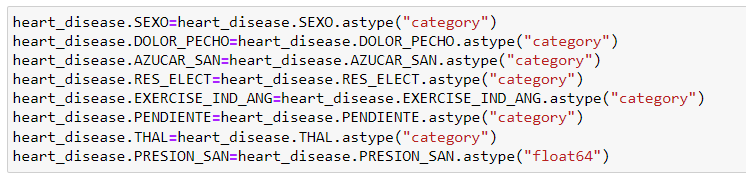


Ilustración . Código usado para convertir el tipo de dato de cada variable.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente con confianza media

Ilustración . Resultado de la función info() luego de modificar el tipo de datos.

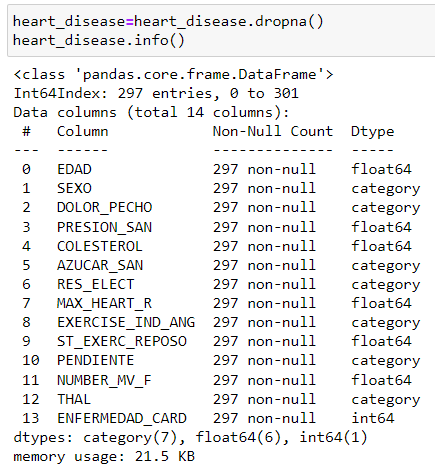


Ilustración . Código usado para eliminar los datos nulos, y resultado de la función info() luego de la eliminación.

Usando la función.describe(), podemos obtener las estadísticas descriptivas del dataset. En la ilustración 9 se observa el resultado al aplicarlo sobre el conjunto de datos estudiado.

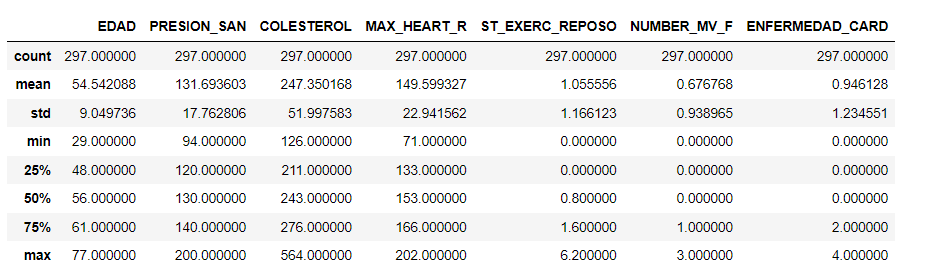


Ilustración . Estadísticas descriptivas de las variables numéricas del dataset estudiado

De la ilustración 9 se puede ver que la variable ENFERMEDAD\_CARD toma valores enteros de 0 a 4, representando 0 la ausencia de enfermedad, y 1, 2, 3 y 4 diversos grados de la misma, según la valoración del especialista. En este estudio procedemos a convertir esta variable a binaria, con valor de 0 para no enfermedad, y valor de 1 para enfermedad. En la ilustración 10 se observa el código utilizado para ello.



Ilustración . Codigo para convertir la variable ENFERMEDAD\_CARD a una variable binaria.

Usando la librería seaborn, es fácil dibujar un “pairplot”, que puede ser útil para estudiar las relaciones entre las diferentes variables númericas. En la ilustración 11 se observa el código usado para generar dicha gráfica, y en la ilustración 12 se aprecia el resultado.

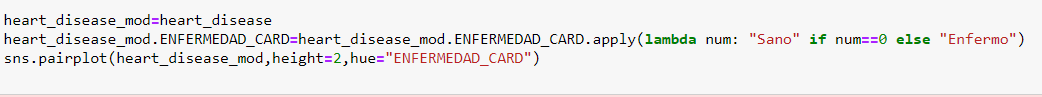


Ilustración . Código para generar el pair plot usando la librería seaborn de Python

Del pair plot de la ilustración es posible extraer varias conclusiones; en primer lugar, la distribución de la edad para los pacientes enfermos parece estar sesgada a la derecha, con una mayor proporción de personas enfermas con edad avanzada (primera grafica de la diagonal del pair plot). Esta información se puede corroborar al graficar un diagrama de cajas como el de la ilustración 13. Esta hipótesis además confirmada por la literatura, donde varios estudios han demostrado que el incremento en la edad del paciente aumenta la incidencia de las enfermedades cardiacas [2].

Si observamos las distribuciones de la presión sanguínea, visualmente son muy similares (segunda grafica de la diagonal del pair plot). Al construir un diagrama de cajas no obstante, se aprecia que existe una diferencia entre los percentiles para los casos de pacientes enfermos y sanos, como ser observa en la ilustración 14. Si bien no es del todo claro que exista una relación entre la hipertensión y la enfermedad cardiaca, este mismo fenómeno ha sido observado en repetidas ocasiones. En particular, en los últimos años, debido a un mejor control de la hipertensión mediante tratamientos médicos se ha reducido la incidencia de esta, pero aun así se reconoce una conexión entre ambas [3]. La relación entre presión sanguínea y enfermedad cardiaca es mas evidente para el conjunto de datos analizado, si definimos un limite a los niveles de presión normal, y graficamos una variable binaria con dos posibles valores: normal e hipertenso. Dicha grafica se observa en la ilustración 15, con un límite de presión normal de 130 mm de Hg en la presión sistólica, según lo establecido por la asociación americana de enfermedades cardiacas [4]. Puede verse (ilustración 15), que la proporción de pacientes hipertensos es mayor en el caso de la población que presentó enfermedad cardiaca que en aquella que no.

Diagrama

Descripción generada automáticamente

Ilustración . Pair plot generado usando la librería seaborn

En el caso del colesterol (tercera grafica de la diagonal del pair plot), a partir de la distribución se observa que hay un pico a niveles más bajos de colesterol para los pacientes sanos. Esta observación es más evidente en el diagrama de caja de la ilustración 16. Según la literatura consultada, la relación entre el colesterol y la enfermedad cardiaca no es directa. Existe una clara relación entre una alta masa corporal, y un sinnúmero de enfermedades como hipertensión, enfermedades cardiacas, diabetes, etc. Así mismo, una alta masa corporal esta relacionada con cambios adversos en el metabolismo lipídico, que ocasionan altas concentraciones de colesterol total [5]. La relación anteriormente descrita entre diabetes y colesterol se manifiesta para el conjunto de datos analizado en la ilustración 17.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Ilustración . Diagrama de cajas para la edad en función del estado de salud del paciente. Puede verse que la edad de los pacientes efectivamente enfermos es superior a la de los sanos.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Ilustración . Diagrama de cajas para la presión sanguínea en función del estado de salud del paciente. Puede verse que la edad de los pacientes efectivamente enfermos es superior a la de los sanos.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Ilustración . Grafico de barras para la frecuencia de pacientes según si su presión sanguínea se encuentra por debajo de 130 mm Hg y presentan o no enfermedad cardiaca. Se aprecia una mayor proporción de pacientes hipertensos entre los que sufrieron enfermedad cardiaca que entre aquellos que no.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Ilustración . Diagrama de caja para los niveles de colesterol, en función de la enfermedad cardiaca.

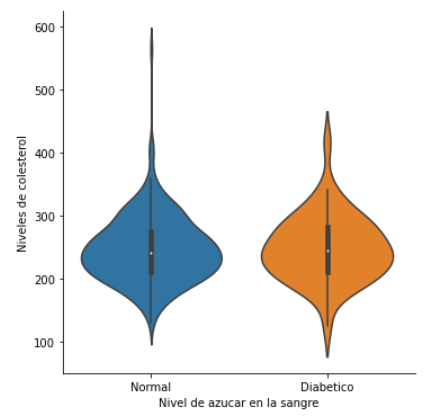


Ilustración . Diagrama de violín para los niveles de colesterol en función del nivel de azúcar en la sangre. Puede verse que, a altos niveles de colesterol, hay una mayor cantidad de pacientes diabéticos que sanos.

La frecuencia cardiaca máxima parece ser una de las variables que mejor diferencia a las poblaciones de personas sanas y con enfermedad cardiaca (cuarta grafica de la diagonal del pair plot). Puede verse que la distribución de la frecuencia cardiaca máxima esta sesgada a la derecha, hacia valores más altos, en el caso de las personas sanas. Dicha conclusión es más evidente en la ilustración 18.

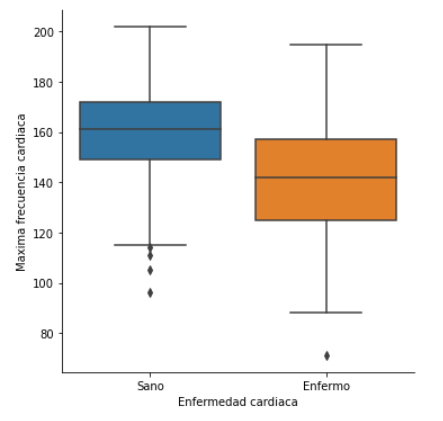


Ilustración . Diagrama de cajas para la máxima frecuencia cardiaca en función de la enfermedad cardiaca.

Otra de las variables que mejor divide las poblaciones de pacientes enfermos y sanos es ST\_EXERC\_REPOSO, que corresponde la depresión inducida por la prueba de esfuerzo. Dicha depresión es un nivel anormalmente bajo presente en un electrocardiograma. Puede verse que la distribución de la población sana esta claramente sesgada a la izquierda, hacia valores bajos de dicha depresión. En la ilustración 19 es más clara dicha diferencia.

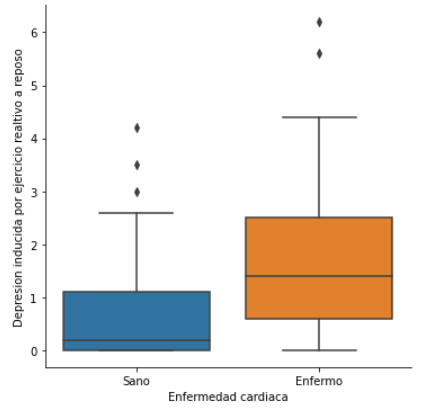


Ilustración . Diagrama de caja de la depresión inducida por prueba de estrés de ejercicio vs enfermedad cardiaca.

Finalmente, las distribuciones para pacientes enfermos y sanos de la variable NUMBER\_MV\_F, que corresponde al numero de arterias coronarias que se colorean en un examen de fluorescencia, son claramente diferentes. En el caso de los pacientes sanos se tiene un pico muy pronunciado para un valor de cero vasos sanguíneos relatados, mientras en el caso de los pacientes enfermos, la distribución es más uniforme. Esta apreciación es evidente en el diagrama de caja de la ilustración 20.

Si bien intuitivamente podría pensarse que no debería existir una relación entre el sexo del paciente y la enfermedad cardiaca, al graficar un diagrama de barras como el de la ilustración 21, se hace evidente que, al menos para el conjunto de pacientes del conjunto de datos, hay una mayor proporción de hombres que presentan enfermedades cardiacas que mujeres. Esta observación es corroborada por la literatura médica, en el sentido de que la enfermedad cardiaca en mujeres se presenta en numerosos casos asintomática, favoreciendo que muchas de ellas no sufran del característico síntoma de dolor en el pecho. Adicionalmente se ha identificado que las mujeres son menos propensas a sufrir una muerte súbita por falla cardiaca (37% de las mujeres con enfermedad cardiaca) que los hombres (57%) [6].

El dolor en el pecho, como se mencionó anteriormente, es uno de los síntomas más característicos de la enfermedad cardiaca. En la ilustración 22 se observa que la proporción de pacientes que reportan algún tipo de dolor en el pecho es muy superior para la población que efectivamente padece la enfermedad cardiaca que para la que no la padece.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Ilustración . Diagrama de caja para el número de casos coloreados en un examen de fluorescencia en función de la enfermedad cardiaca.

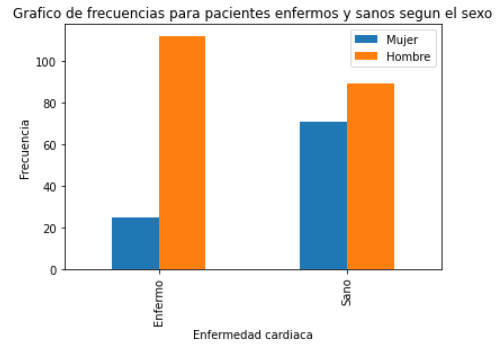


Ilustración . Grafico de barras para la frecuencia de observaciones de enfermedad cardiaca según el sexo del paciente.

En la ilustración 23 se aprecia el diagrama de frecuencias para los niveles de azúcar en la sangre, según si los pacientes presentan o no enfermedad cardiaca. Puede verse claramente que hay una mayor proporción de personas diabéticas entre aquellos pacientes que efectivamente presentaron enfermedad cardiaca. De acuerdo con [2], existe una asociación directa entre la diabetes y la enfermedad cardiaca; los pacientes que presentan diabetes tipo 1 o tipo 2 tienen un mayor riesgo de presentar enfermedades cardiacas.

La angina inducida por ejercicio también es un claro marcador de posible enfermedad cardiaca, como se deduce de la ilustración 24. Igualmente sucede con la pendiente del pico de ejercicio en la prueba de estrés, como se observa en la ilustración 25.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Ilustración . Grafico de frecuencias para pacientes con dolor en el pecho, según si presentan o no enfermedad cardiaca.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Ilustración . Grafico de frecuencias para el nivel de azúcar en la sangre de los pacientes, según sí estos presentaron enfermedad cardiaca o no.

Puede observarse que, en el caso de los pacientes enfermos, hay una mayor tendencia a presentar pendientes planas que en el caso de los pacientes sanos.

La talasemia es una enfermedad hereditaria que consiste en un desorden sanguíneo que impide al cuerpo producir suficiente hemoglobina. Como consecuencia los glóbulos rojos no funcionan bien, y su vida se reduce considerablemente. Una de las principales causas de muerte en pacientes que presentan talasemia es la falla cardiaca [7]. Esta evidencia de la bibliografía también se ve reflejada en el conjunto de datos, como se ve en la ilustración 26, donde existe una mayor proporción de pacientes con talasemia entre la muestra de pacientes con enfermedad cardiaca.

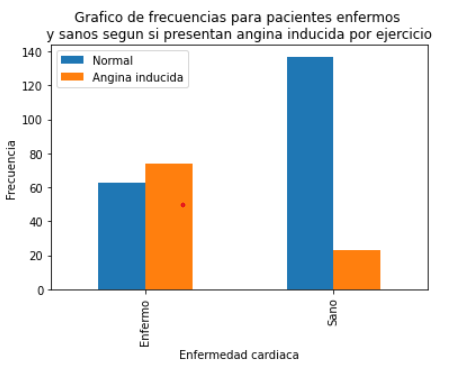


Ilustración . Grafica de frecuencias para la angina inducida, según si presentan o no enfermedad cardiaca.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Ilustración . Grafico de frecuencias para pacientes enfermos y sanos según la pendiente que presentan en el pico de ejercicio de estrés.

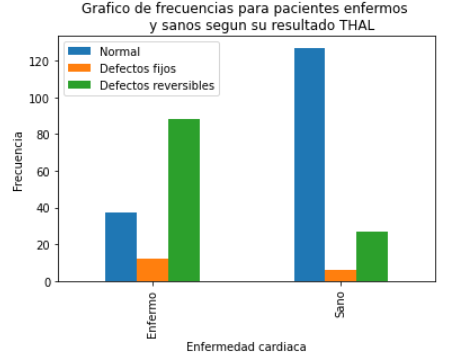


Ilustración . Grafico de frecuencia para pacientes que presentan talasemia y enfermedad cardiaca.

Otra grafica muy útil para entender las correlaciones entre variables es el mapa de calor de correlaciones que se muestra en la ilustración 27. Puede verse que existe una correlación negativa entre la edad y la máxima frecuencia cardiaca, así como también entre la máxima frecuencia cardiaca y la depresión inducida en la prueba de ejercicio. Por otro lado, existe una correlación positiva entre la edad y el número de vasos sanguíneos coloreados en el examen de fluoresencia.

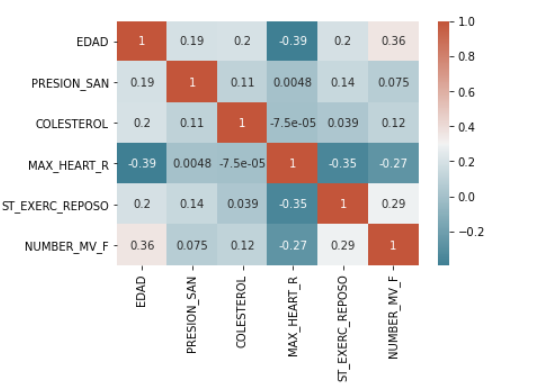


Ilustración . Mapa de calor de correlaciones entre variables numéricas del dataset

En las ilustraciones 28 y 29 se observan 2 pares de variables que separan las poblaciones de pacientes enfermos y sanos de forma casi lineal. Puede verse que estas variables serían muy útiles en el caso de construir un clasificador del tipo de regresión logística o SVM.

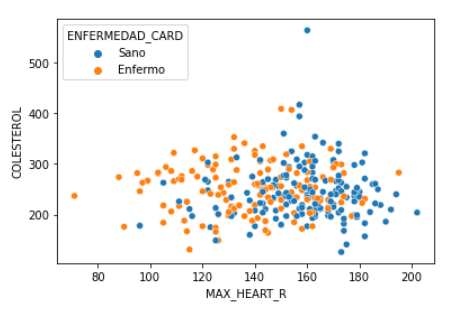


Ilustración . Grafica de dispersión de colesterol vs máxima frecuencia cardiaca, discriminados los datos por su pertenencia al grupo de pacientes sanos o enfermos

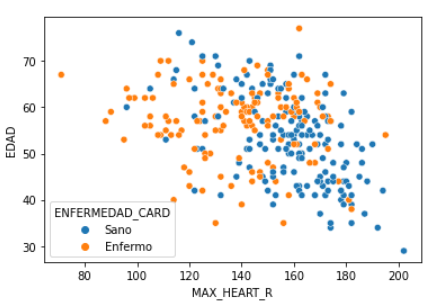


Ilustración . Grafica de dispersión de Edad vs máxima frecuencia cardiaca, discriminados los datos por su pertenencia al grupo de pacientes sanos o enfermos

Estudiemos ahora relaciones entre variables diferentes a la enfermedad cardiaca. En la ilustración 30 se observa un diagrama de caja para la edad en función de la presión sanguínea. Puede verse que una edad avanzada influirá en una mayor probabilidad de sufrir hipertensión.

Una relación similar se observa al comparar el colesterol con la presión sanguínea, como en la ilustración 31. Puede observarse que se observan niveles mayores de colesterol en las personas hipertensas que en las no hipertensas.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Ilustración . Diagrama de caja para la edad en función de la presión sanguínea.

Gráfico, Gráfico radial

Descripción generada automáticamente

Ilustración . Diagrama de violín para el nivel de colesterol en función de la presión sanguínea.

En la ilustración 32 se explora la relación entre los niveles de colesterol y el resultado del examen de fluoroscopia. Puede verse que existe una tendencia hacia valores más altos de colesterol total cuando se presentan más vasos afectados.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Ilustración . Diagrama de cajas para el nivel de colesterol vs el número de vasos sanguíneos resaltados en un examen de fluoroscopia.

En la ilustración 33 se observa la comparación de los resultados del electrocardiograma para pacientes enfermos y sanos. Puede verse claramente que la proporción de electrocardiogramas con ondas normales es mayor en el grupo de pacientes sanos, mientras que en el grupo de pacientes enfermos existe una mayor proporción de electrocardiogramas identificados con alto riesgo de tener hipertrofia ventricular.

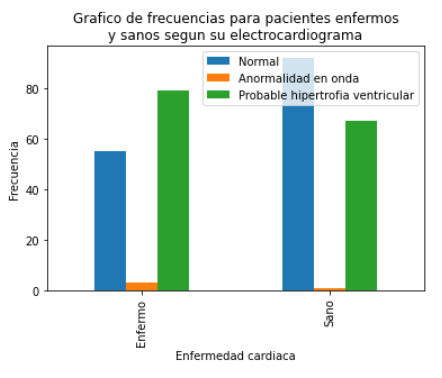


Ilustración . Resultados de electrocardiograma para pacientes enfermos y sanos, según la forma de la onda.

CONCLUSIONES:

De análisis presentado anteriormente se puede concluir:

* La edad tiene una fuerte influencia en la hipertensión y en la diabetes, y como consecuencia incide en la probabilidad de padecer una enfermedad cardiaca.
* Los altos niveles de colesterol están relacionados con afectación a las arterias coronarias, y por ende inciden en la probabilidad de padecer una enfermedad cardiaca.
* La hipertensión, si bien controlada en los últimos años por el avance en medicamentos, sigue teniendo una incidencia en las enfermedades cardiacas.
* Niveles reducidos de la máxima frecuencia cardiaca lograda son un indicativo de la presencia de una enfermedad cardiaca.
* Una depresión alta en la curva del electrocardiograma en una prueba de esfuerza es uno de los mejores indicativos de la presencia de enfermedad cardiaca
* El examen de fluorescencia es una de las pruebas más concluyentes en cuanto a la identificación de enfermedades cardiacas.
* Padecer talasemia predispone al paciente a sufrir también enfermedades cardiacas, siendo uno de los principales factores de riesgo.
* El dolor en el pecho a angina es uno de los síntomas más claros de la enfermedad cardiaca, tanto si es inducida como si no.
* El sexo de los pacientes tiene una fuerte influencia en la predisposición a padecer enfermedad cardiaca, siendo más probable que la padezcan los hombres.
* La pendiente de pico en la prueba de ejercicio es otro de los síntomas más importantes que permiten separar las poblaciones de pacientes que padecen y que no padecen enfermedades cardiacas.

**Red bayesiana**

De acuerdo con lo anteriormente expuesto, basados en la evidencia del análisis de datos y la bibliografía consultada se propone la red bayesiana de la ilustración 33.