

Capstone Project

Predecir problemas cardiacos



March 30, 2020

Capston

Buenas Tardes Mentor, primero quiero agradecerle la oportunidad de participar en el proceso de convertirme en un científico de datos, y mencionarle que a continuación va a encontrar el análisis que se me solicito para ser considerado en dicho puesto.

Para empezar, me gustaría primero mencionarle el set de datos seleccionado, además detallar las columnas y el significado de cada una de ellas. Este conjunto de datos que se utilizo para el proyecto contiene la información sobre los factores de riesgo de enfermedad cardíaca.

La base de datos original contiene 76 atributos, pero todos los experimentos publicados se refieren al uso de un subconjunto de 14 del tema y se hace referencia como conjunto de datos de la Universidad de Cleveland.

Los experimentos con la base de datos de Cleveland se han concentrado en intentar distinguir la presencia (valor 1) o la ausencia (valor 0) de enfermedad cardíaca en el paciente.

Los datos del set de datos son:

* Age =Age
* Sex- Sex
* Chest pain (typical, asymptotic, nonanginal, nontypical)
* Resting blood pressure
* Serum - cholestoral in mg/dl
* Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
* Resting electrocardiographic results
* Maximum heart rate achieved
* Exercise induced angina (1 = yes; 0 = no)
* depression induced by exercise relative to rest
* Slope of the peak exercise ST segment
* Number of major vessels colored by flourosopy (0 - 3)
* thal3 = normal; 6 = fixed defect; 7 = reversable defect)
* target = AHD - Diagnosis of heart disease (1 = yes; 0 = no)

Con lo mencionado el objetivo general será: Predecir si un paciente tiene una enfermedad cardíaca en función de un conjunto de síntomas.

Como objetivo secundario quiero destacar las características que tienen el mayor impacto con lo que se refiere a los problemas cardiacos.

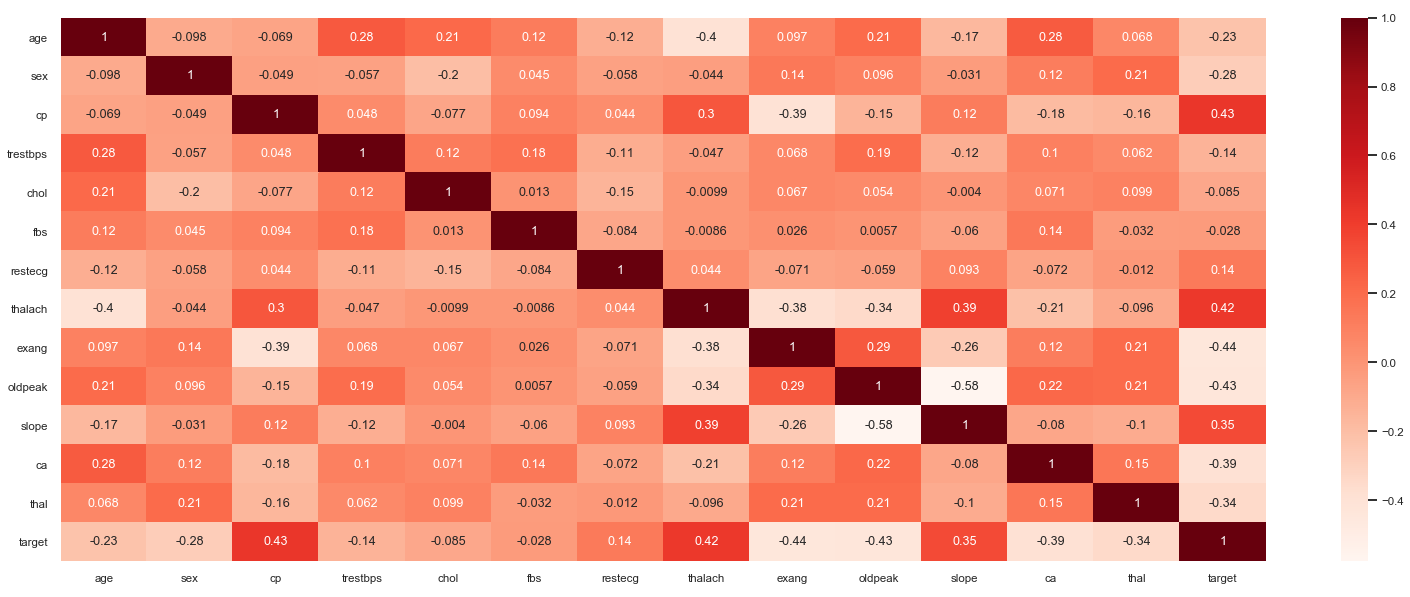
Vamos a seguir los lineamientos que se van a mencionar para estructurar el análisis y predicción:

* Visualización de datos cortos. Esta tarea detalla el aspecto de visualización del análisis de datos.
* Limpieza de datos. Esta tarea se relaciona con el aspecto de limpieza de datos del análisis.
* Generación de algoritmos. Esta tarea se relaciona con el aspecto de creación del algoritmo de predicción.

Sin más preámbulo vamos a comenzar con el análisis de los datos seleccionados.

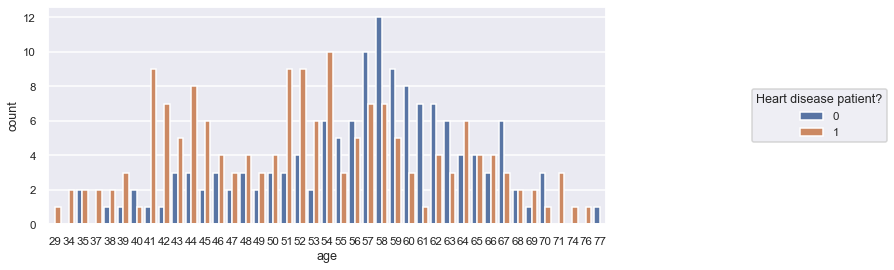
Lo primero es ver la correlación de los datos con la variable que queremos predecir, esto no ayudara a entender las características altamente correlacionadas con esta variable. Ya que una relación muy grande puede aumentar el error en los nuevos (datos no vistos), por lo cual debemos tratar de reducir este problema.

A continuación, vemos el grafico para visualizar de mejor manera las correlaciones



Pero en este caso podemos ver que no tenemos el problema antes mencionado.

Continuando vamos a distribucion de personas por edades donde veremos una vision generalizada de como se ve por edad si una parseona tiene o no problemas cardicacos:

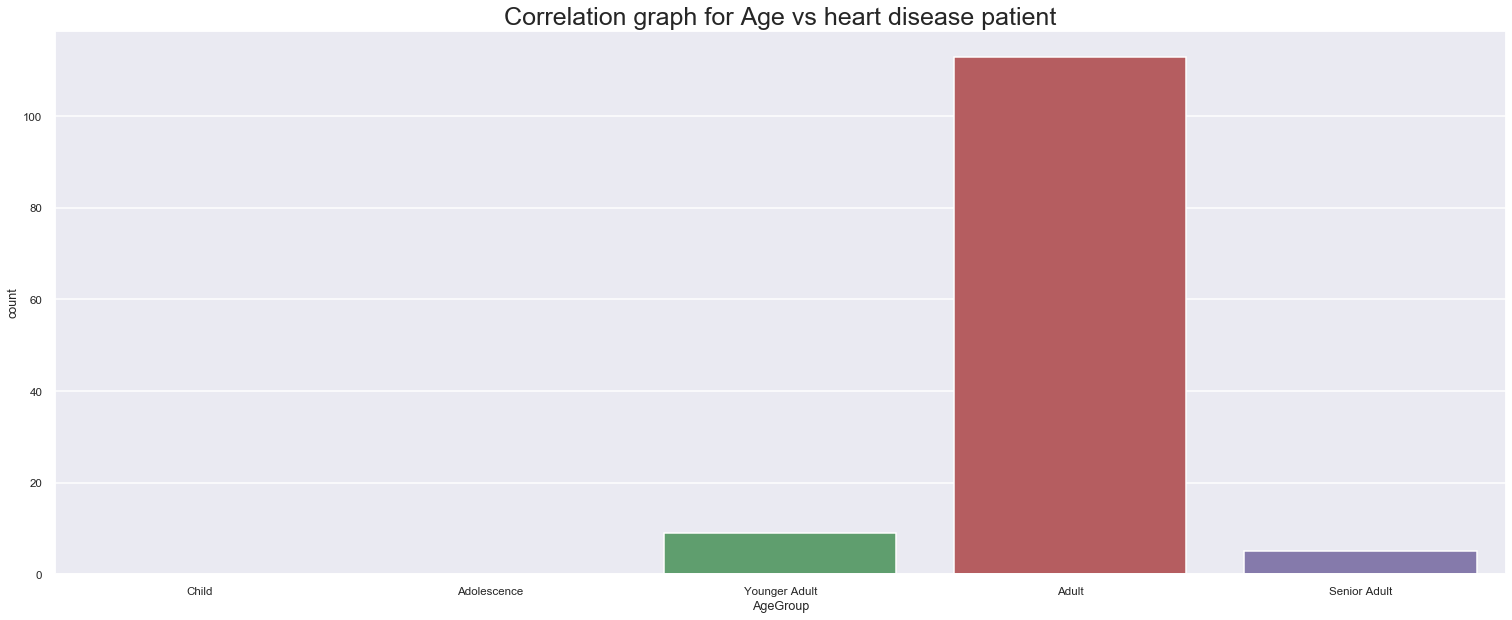


Con este grafico mostrado anteriormente se ve de una manera más general de como la enfermedad ataca por edad.

Pero para movernos a un análisis más holístico vamos agrupar esas edad en los siguientes rangos:

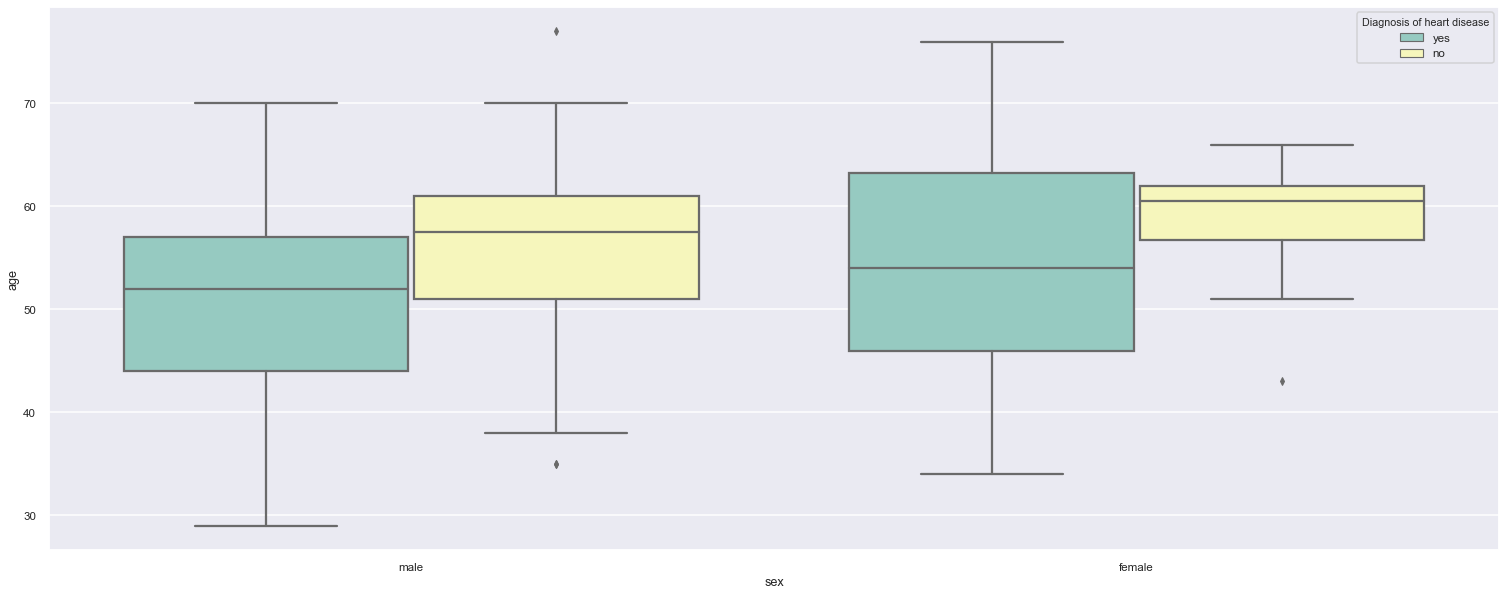
* Child = 0-12
* Adolescence = 13-18
* Younger Adult = 19-39
* Adult = 40-59
* Senior Adult = 60+

Así vamos a tener una visión mucho más enfocada, dicho esto se muestra el siguiente gráfico:



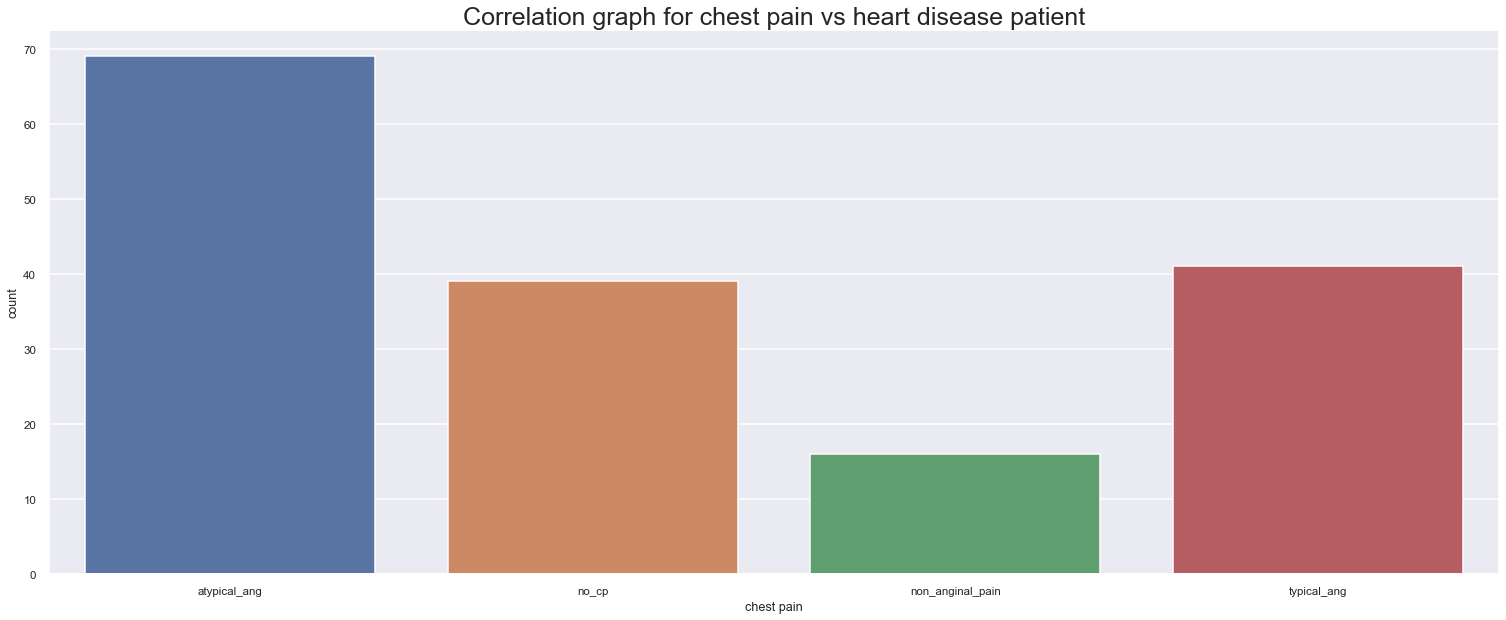
Como podemos ver la relación entre los pacientes adultos y las personas con problemas cardiacos es la más relevante, por lo que podemos decir que las personas mayores son más susceptibles a las enfermedades del corazón ya que no vemos una distribución uniforme de pacientes con enfermedades cardíacas en todas las edades.

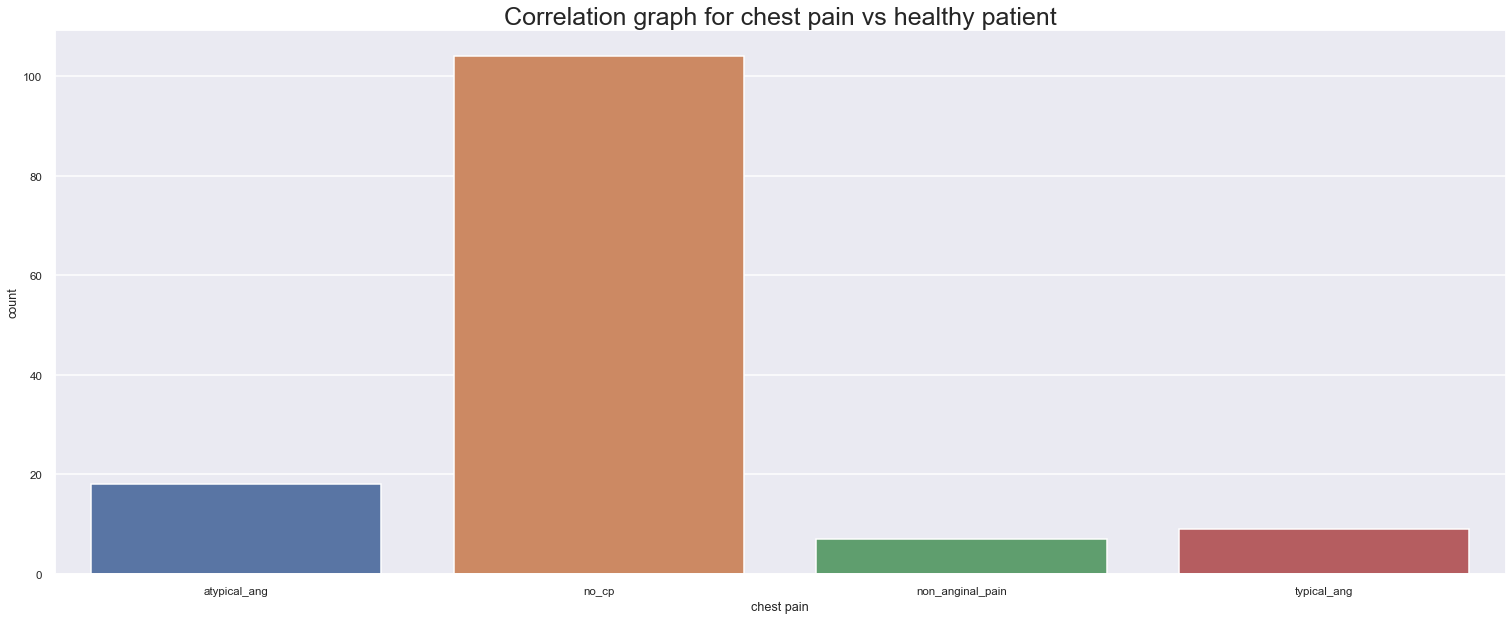
Seguimos ahora con la relación sexo vrs enfermedad cardiaca:



Podemos ver que el echo de tener que la proporción de mujeres con y sin enfermedades cardiacas es bastante similar, mas el caso de hombres con enfermedades cardiacas si es más elevado que los hombres que están sanos, por lo cual con la muestra que tenemos vemos que los hombres son más propensos a estas enfermedades, para confirmar la hipotesis se podría volver hacer el estudio, pero con mayor cantidad de personas.

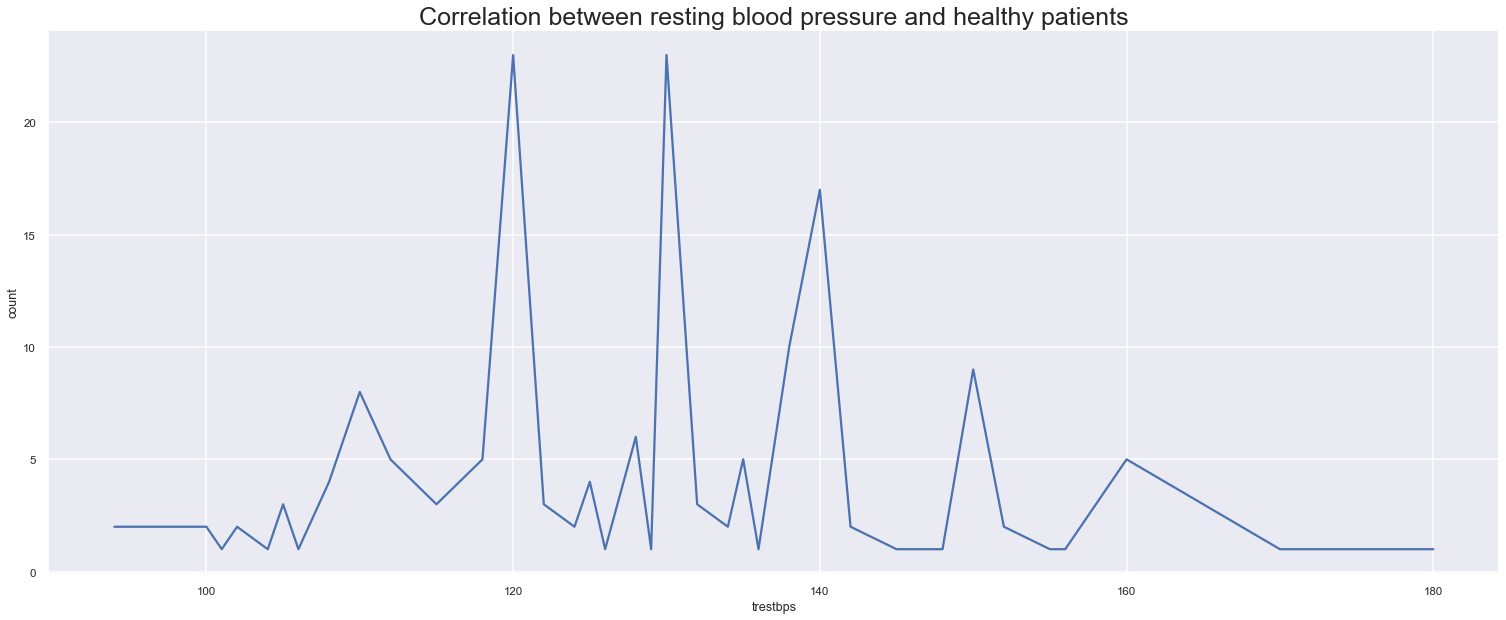
Ahora comparamos personas que han tenido algún tipo de dolor de pecho tanto en personas que se les haya diagnosticado enfermedades cardiacas como personas que están completamente sanas:





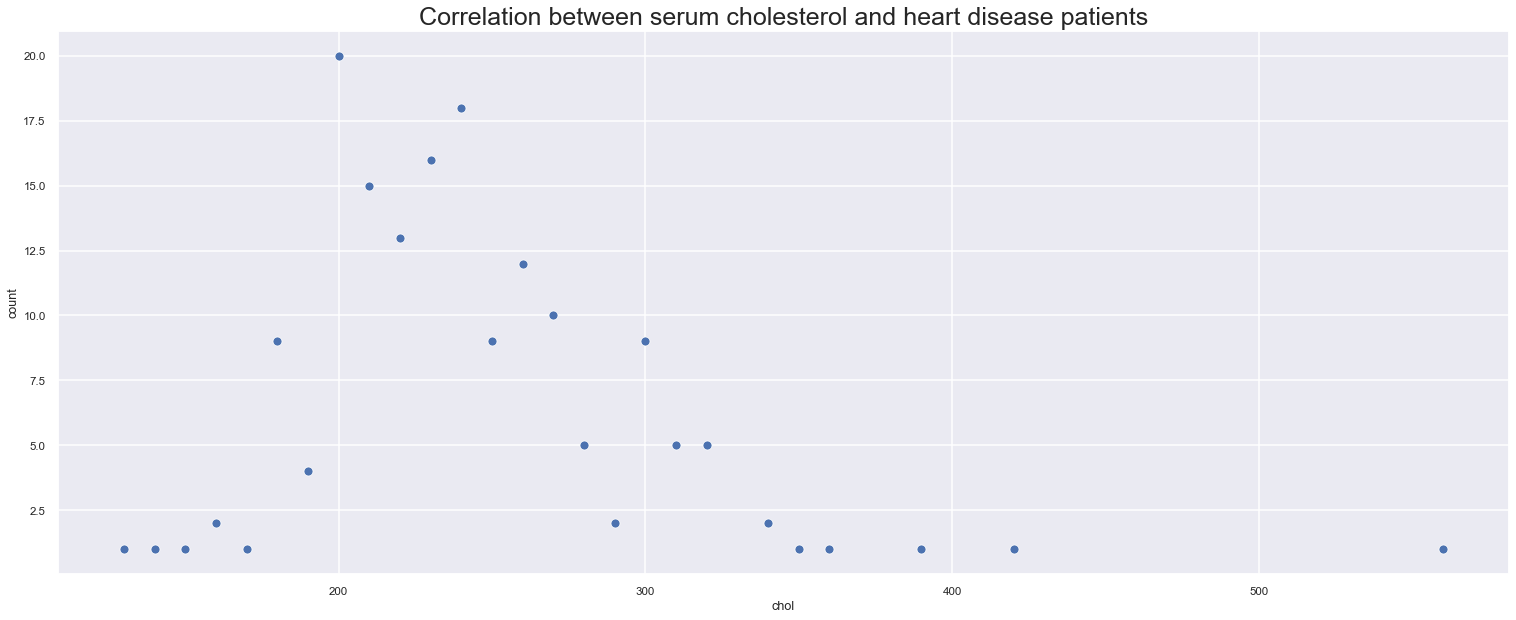
Podemos ver aquí es que los pacientes con enfermedades del corazón tienden a experimentar los 3 tipos de dolor en el pecho, mientras que los pacientes sanos generalmente no experimentan ningún dolor en el pecho. Por lo tanto, sin ninguna prueba estadística, podemos decir que definitivamente existe una correlación entre el dolor en el pecho y la enfermedad cardíaca.

Ahora vemos la relación entre la presión arterial y las enfermedades coronarias



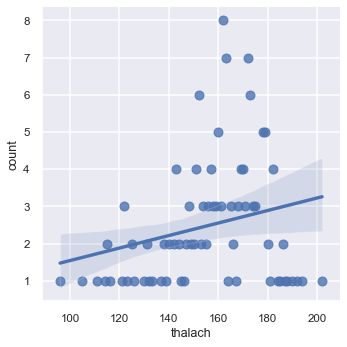
Tampoco vemos al menos con los datos analizados una relación grande entre estos 2 componentes

Seguimos con el colesterol y los pacientes con problemas cardiacos



Tampoco vemos al menos con los datos analizados una relación grande entre estos 2 componentes, se podría volver hacer el análisis con un set de datos mayor para confirmar la hipotesis.

Continuando con el analisis vemos la relacion entre las personas que alcanzaron su maximo de latidos del conrazon y su relacion con las personas que padecen de una enfermedad coronaria



Acá si vemos un poco más de relación con personas que alcanzan pulsaciones entre 150 y 180 pulsaciones por minuto.

Hemos probado la mayoría de los atributos para la correlación y, a partir de los resultados, podemos decir con confianza que tanto los resultados del ECG en reposo como los tipos de dolores en el pecho están correlacionados con la enfermedad cardíaca. Aunque vemos una correlación al realizar la prueba Chi-Sq en el atributo de género, la gran diferencia en los datos de mujeres sanas planteó una gran preocupación por su precisión.

Con esto damos terminado el análisis de los datos utilizados en el reporte ahora nos emprendemos en la parte de crear un modelo que pueda terminar con los datos que anteriormente se proporcionaron si una persona va a tener problemas coronarios.

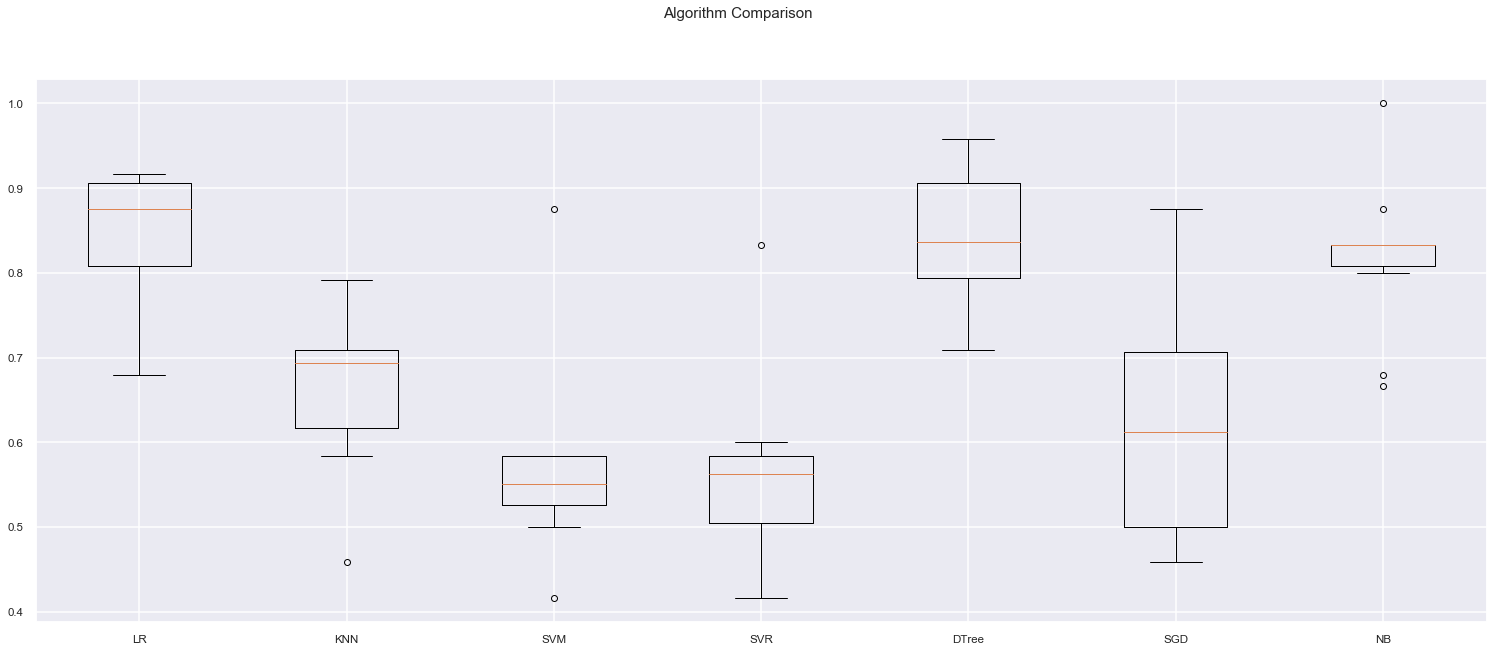
Se van a usar diferentes tipos de algoritmos y set de datos como PCA (principal component analysis) es una técnica utilizada para describir un conjunto de datos en términos de nuevas variables. Con esto lo que se quiere lograr es el mejor algoritmo para ser implementado.

Los algoritmos utilizados son los siguiente:

* **Regresión logística**: es un modelo estadístico que en su forma básica utiliza una función logística para modelar una variable dependiente binaria, aunque existen muchas extensiones más complejas.
* **KNN**: el algoritmo es una técnica para clasificar datos en un grupo conocido que esencialmente predice un valor de salida específico mediante el aprendizaje supervisado
* **Decision Tree Classifier**: es una representación simple para clasificar ejemplos
* **Gaussian NB**: algoritmo utilizado para actualizar las características y la varianza de las características
* **Support Vector Machine**: es un modelo de aprendizaje automático supervisado que utiliza algoritmos de clasificación para problemas de clasificación de dos grupos
* **Support Vector Regression**: se utiliza como método de regresión, manteniendo todas las características principales que caracterizan el algoritmo (margen máximo).
* **Clasificador SGDC**: este estimador implementa modelos lineales regularizados con aprendizaje de descenso de gradiente estocástico (SGD): el gradiente de la pérdida se estima en cada muestra a la vez y el modelo se actualiza a lo largo del camino con un programa de fuerza decreciente (también conocido como tasa de aprendizaje).

Después de los análisis con cada uno de los modelos se obtienen los siguientes resultados en términos de exactitud de los modelos:

* Linear Regression: 0.839667
* KNN: 0.665500
* SVM: 0.570500
* SVR: 0.566167
* Decision Tree: 0.834833
* Classificatory SGD: 0.619333
* Gaussian NB: 0.818833



Con los resultados anteriores decidimos seguir adelante para hacer el modelo de predicción con el algoritmo de regresión lineal, por lo cual se hace el modelo buscando obtener 2 métricas con niveles bajos las cuales son

R Squared: R-cuadrado es una medida estadística de qué tan cerca están los datos de la línea de regresión ajustada. Además, cabe recalcar que un valor con bajo valor de RMES es aceptable.

RMSE: El error cuadrático medio (RMSE) es la desviación estándar de los residuos (errores de predicción). Cuanto menor sea el RMSE es mejor para el algoritmo.

Mencionado lo anterior estos son los resultados obtenidos del modelo

* R Squared: 0.336
* RMSE: 0.405

Los cuales son prometedores y no dicen que algoritmo esta listo para entrar en su etapa de prueba y comenzar a ayudar en la detención temprana de pacientes con problema cardiovasculares.

Se le agradece de antemano poder concursar la oportunidad que se me da para convertirse en un científico de datos.

Saludes