**Índice:**

[**Problema y objetivo**](#_heading=h.30j0zll) **2**

[**Hipótesis iniciales**](#_heading=h.41zerpirjmaa) **2**

[**Análisis Exploratorio**](#_heading=h.q4f3m4mkiz10) **4**

[**Limpieza**](#_heading=h.jh6i46ivv5pw) **8**

[**Ingeniería de Variables**](#_heading=h.s0hxiofk92es) **9**

[**Métricas**](#_heading=h.en9ne82lp1n6) **10**

[**Modelos y Selección de Variables**](#_heading=h.gr48jxbri8p4) **10**

[**Modelo 1: k-NN**](#_heading=h.1ge8b94ab1lq) **10**

[**Modelo 2: Regresión Logística**](#_heading=h.kkasu1st3y) **12**

[**Modelo 3: Árbol de Decisión**](#_heading=h.rasxxoy7wpdf) **13**

[**Modelo Seleccionado y Análisis de Resultados**](#_heading=h.qmwx2antpdvh) **13**

[**Bibliografía**](#_heading=h.g5f1wko4p1cp) **15**

# 

# Problema y objetivo

El objetivo de este trabajo es proponer una solución con un modelo de Machine Learning a una problemática de clasificación de enfermedades del corazón. Las personas con enfermedades cardiovasculares o con alto riesgo cardiovascular (por la presencia de uno o más factores de riesgo como hipertensión, diabetes, hiperlipidemia u otras enfermedades ya establecidas) necesitan una detección y manejo temprano, por lo que se propone hallar un modelo de Machine Learning que pueda ser de ayuda en esta detección temprana. Con dicho modelo se puede tener mayor control y previsibilidad en determinados estudios cardíacos y sanguíneos. Es decir, podría permitir saber en qué análisis hacer foco y saber si el paciente se encamina a un riesgo cardiovascular o no.

# 

# Hipótesis iniciales

Para comenzar con el estudio, se presentan las variables que han sido relevadas para cada uno de los pacientes

|  |  |
| --- | --- |
| PatientId | Id del paciente |
| Age | Edad del paciente |
| ChestPainType | Dolor de pecho tipo [TA: angina típica, ATA: angina atípica, NAP: dolor no anginal, ASY: asintomático] |
| RestingBP | Presión en sangre en estado de reposo [mm Hg] |
| Cholesterol | Colesterol [mm/dl] |
| RestingECG | Resultados del electrocardiograma en reposo [Normal: normal, ST: con anomalías en la onda ST-T (inversión de la onda T y/o elevación o depresión del ST > 0,05 mV), LVH: hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes] |
| MaxHR | Frecuencia cardíaca máxima alcanzada [Valor numérico entre 60 y 202] |
| ExerciseAngina | Angina inducida por el ejercicio [Y: Sí, N: No] |
| Oldpeak | pico antiguo = ST [Valor numérico medido en depresión] |
| ST\_Slope | La pendiente del segmento ST del ejercicio máximo [Up: pendiente ascendente, Flat: plana, Down: pendiente descendente] |
| HeartDisease | Output [1: enfermedad cardíaca, 0: normal] |

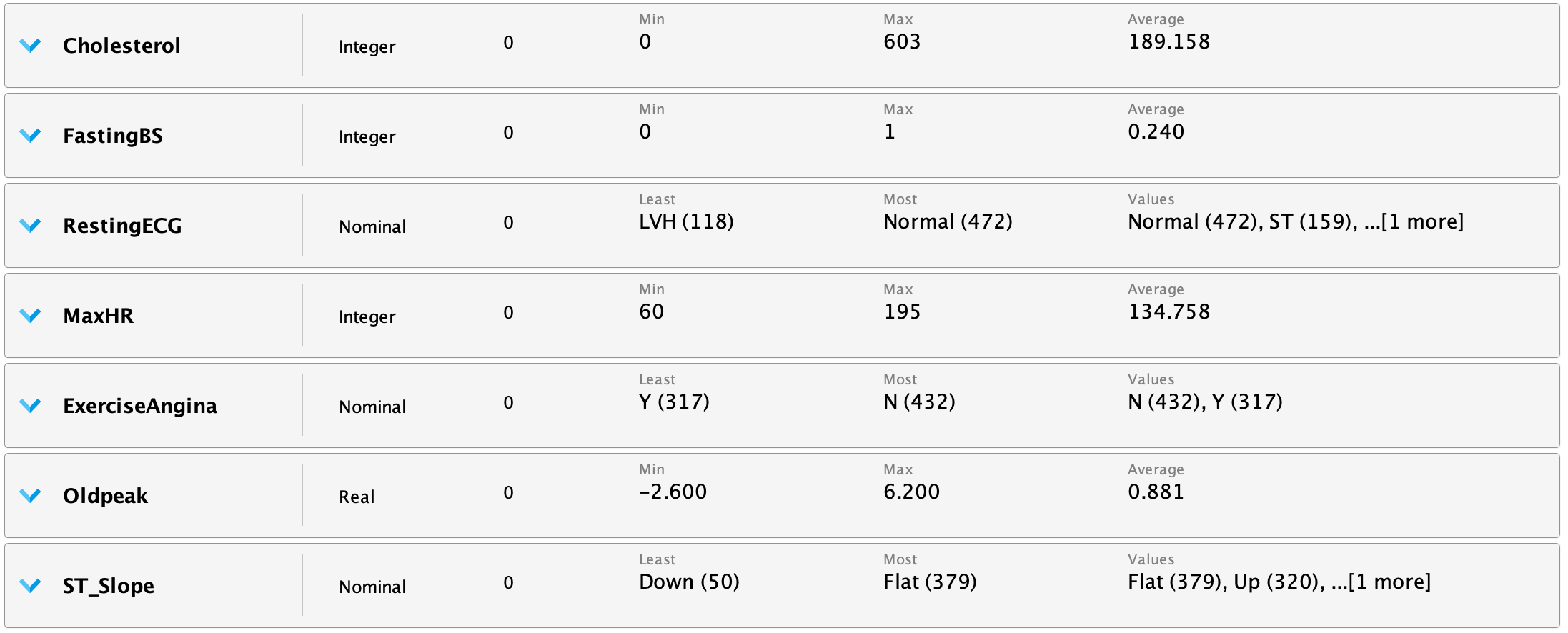
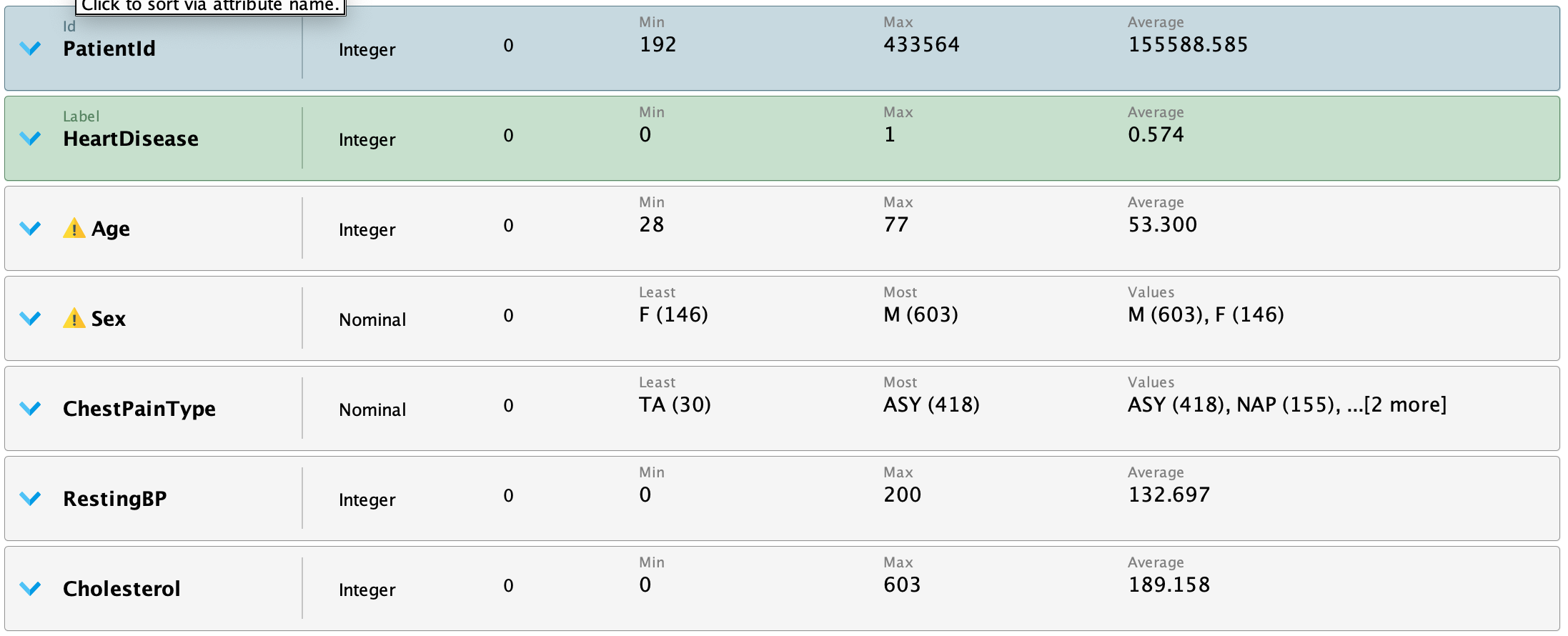
*Tabla 1: Variables registradas*

* **Edad**: se considera que a mayor edad el paciente es más propenso a sufrir una enfermedad cardiaca. Esto puede deberse al deterioro del corazón en el tiempo, que pierde fuerza de bombeo de sangre al ser un músculo que se va debilitando.
* **Presión Sanguínea (en reposo)**: generalmente se relaciona con problemas de hipertensión. La presión en sangre elevada produce que las paredes de las arterias sean presionadas. Generalmente, la hipertensión se define como la presión arterial por encima de 140/90 y se considera grave cuando está por encima de 180/120.
* **Colesterol**: si se encuentra en niveles elevados, puede producir una peligrosa acumulación en las paredes de las arterias (aterosclerosis). Estos depósitos pueden reducir el flujo sanguíneo a través de las arterias, lo que puede provocar complicaciones cardiacas.
* **Fasting Blood Sugar**: los niveles altos de azúcar en la sangre pueden dañar los vasos sanguíneos y los nervios que controla el corazón, esto provoca que el corazón no bombee sangre de forma eficiente y sea susceptible a un infarto.
* **Frecuencia Cardiaca Máxima:** La frecuencia cardiaca máxima es el número máximo de latidos que puede alcanzar el corazón durante un minuto sometido a esfuerzo. Estos valores pueden ser elevados si tienen una explicación obvia como por ejemplo fiebre, embarazo, ansiedad o efectos de la cafeína, sino podrían ser riesgosos. También puede ocurrir que los valores se encuentren bajos y presenten un peligro. Los valores de referencia se encuentran entre 60 y 202 latidos por minuto.
* **Angina a causa de ejercicio:** la angina puede provocar un dolor de pecho debido a la disminución de la irrigación sanguínea del corazón**.**
* **Picos del electrocardiograma (Oldpeak):** representa la depresión del segmento ST. Si la elevación del segmento ST del electrocardiograma supera 1 milímetro y más 80 milisegundos puede ser indicio de un infarto miocardio. Se halló que este indicador suele arrojar una tasa de falsos negativos del 20-30%, por lo que se entiende que esta variable puede ser crítica a la hora de determinar la presencia de enfermedades cardiacas y debe poder utilizarse correctamente.[[1]](#footnote-1) Esta variable se puede categorizar entre crítico o no crítico, siendo que supera los parámetros límites mencionados anteriormente.
* **Pendiente del electrocardiograma (ST Slope):** este indicador representa las pendientes de las ondas del electrocardiograma. Si las mismas se encuentran con una orientación plana o descendente pronunciada puede indicar una isquemia coronaria. Esto es cuando el flujo sanguíneo al corazón se reduce, lo que impide que el músculo cardíaco reciba suficiente oxígeno.
* **Electrocardiograma (en reposo):** con el estudio del electrocardiograma se puede saber si se encuentra atrofiado el ventrículo izquierdo del corazón (LVH). Esto ocurre frecuentemente cuando se encuentra un descenso del ST asimétrico, con un descenso inicial suave y ondas T negativas en las derivaciones laterales. También si se pueden obtener valores anómalos al hallar una depresión de la onda ST de más de 0,05 mV. Un caso fuera de peligro sería cuando el segmento ST oscila entre -0,05 mV y 0,05 mV.

# 

# Análisis Exploratorio

Luego, se cargaron ambas tablas (Entrenamiento y Testeo) al software RapidMiner para poder comenzar con el análisis. A este fin, se debe entender el tipo de variables y estudiar las distribuciones, a modo de análisis preliminar para asegurar que el software pueda interpretar la información correctamente y en base a eso generar las predicciones correctas.



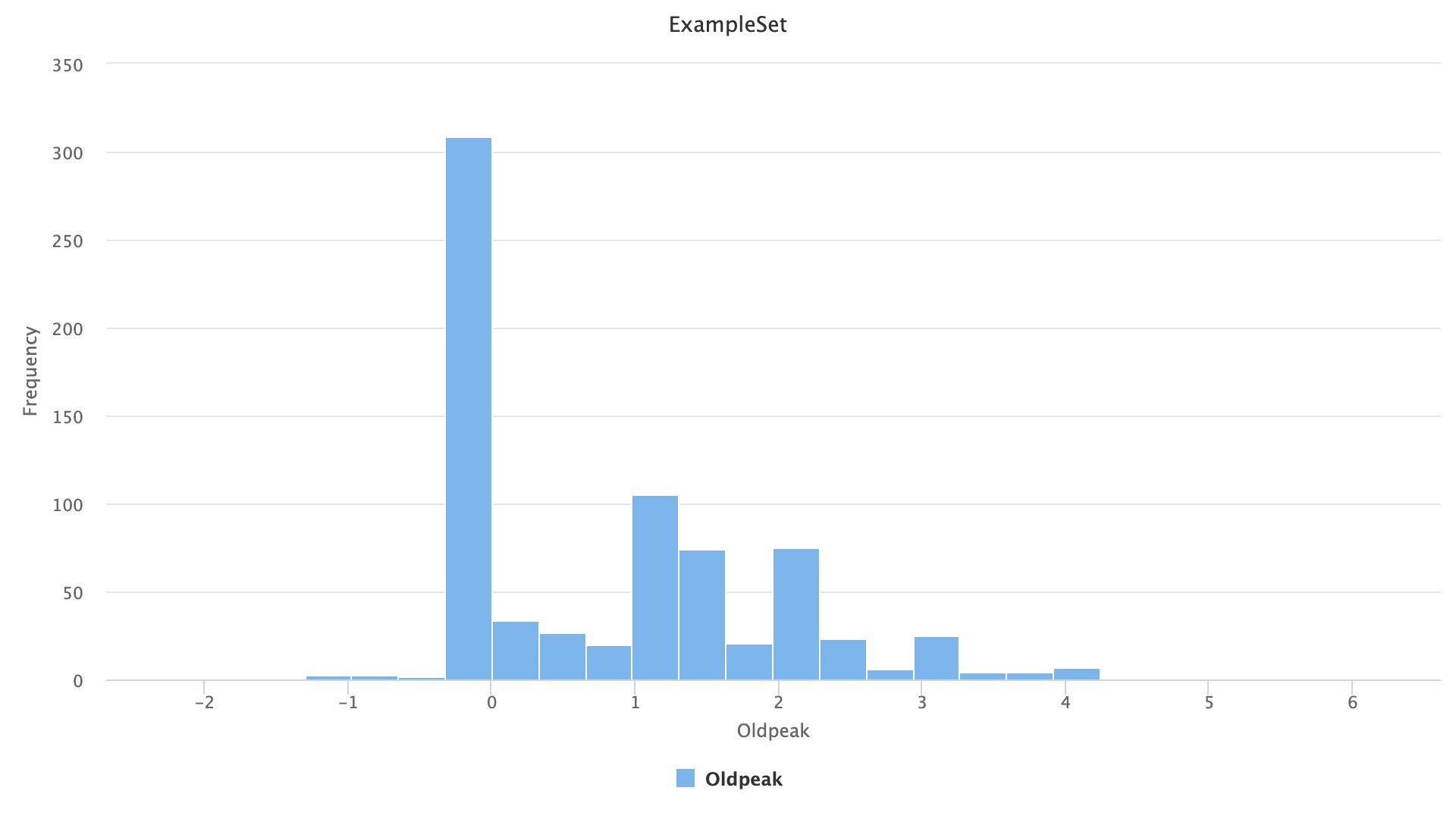
*Figura 1: Análisis exploratorio de las variables*

En primer lugar se tiene “HeartDisease” (Resaltada en verde), que es la variable que se desea predecir. Esta solo tiene dos valores posibles, 0 y 1, donde 1 implica propensión a desarrollar una enfermedad cardiovascular, y 0 implica tener bajo riesgo. En el modelo, esta variable debe configurarse como ROL=LABEL para que este pueda identificar a esta variable como la variable respuesta.

Luego se registran variables para identificar a los pacientes como Age, Sex y PatientId. Estas variables permitirán segmentar en grupos, pero a su vez pueden generar distorsión en los números, dado que estas variables pueden tener correlación con otras.

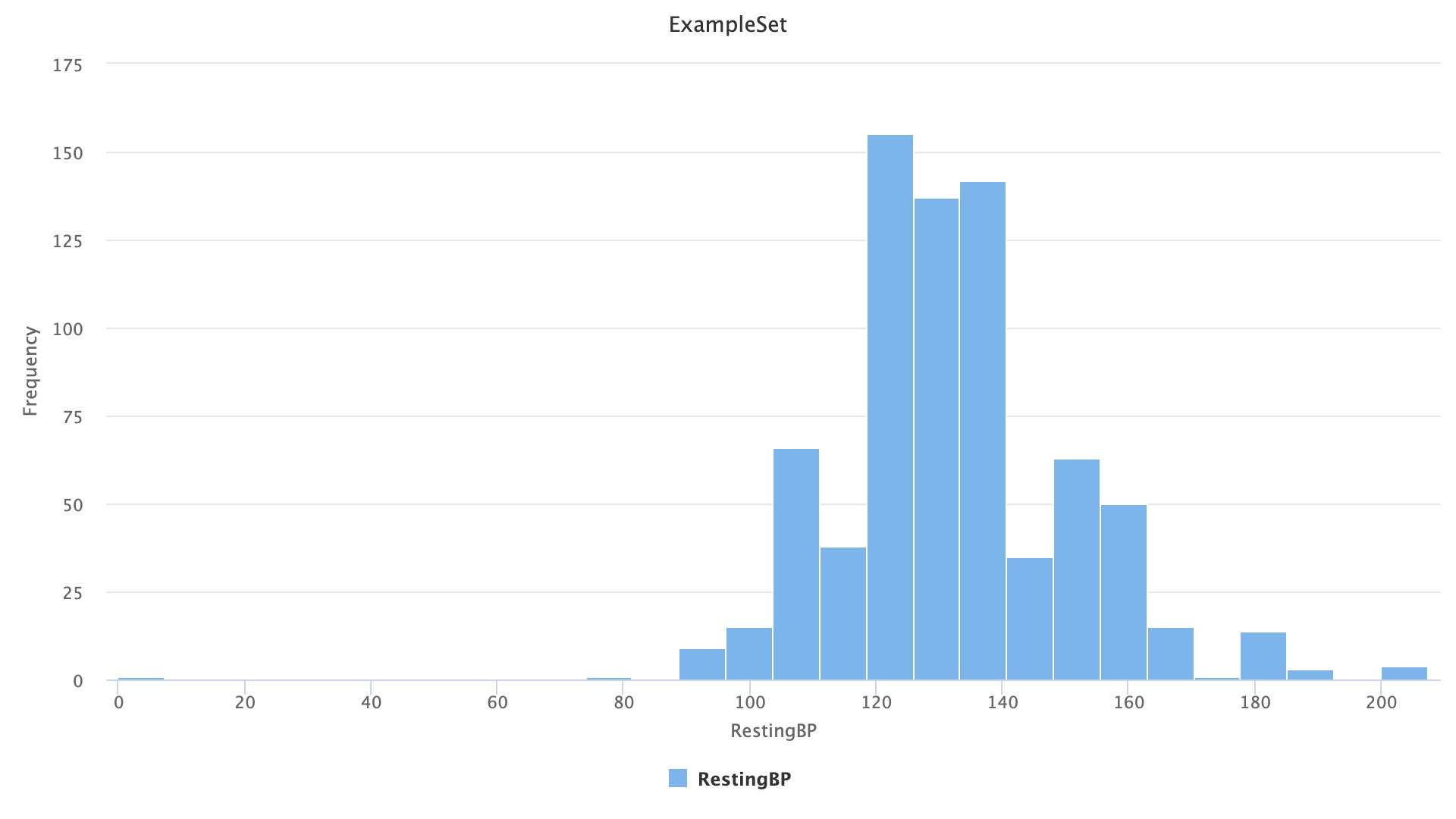
Por último, se registran los valores de distintas pruebas realizadas que describen el estado del sistema circulatorio de cada paciente. Se espera que luego permitan predecir la potencialidad de desarrollar enfermedades cardiacas.

Habiendo definido los grupos de variables, se puede entonces comenzar a estudiar las distribuciones de algunas de estas, para poder entender la validez de los datos.



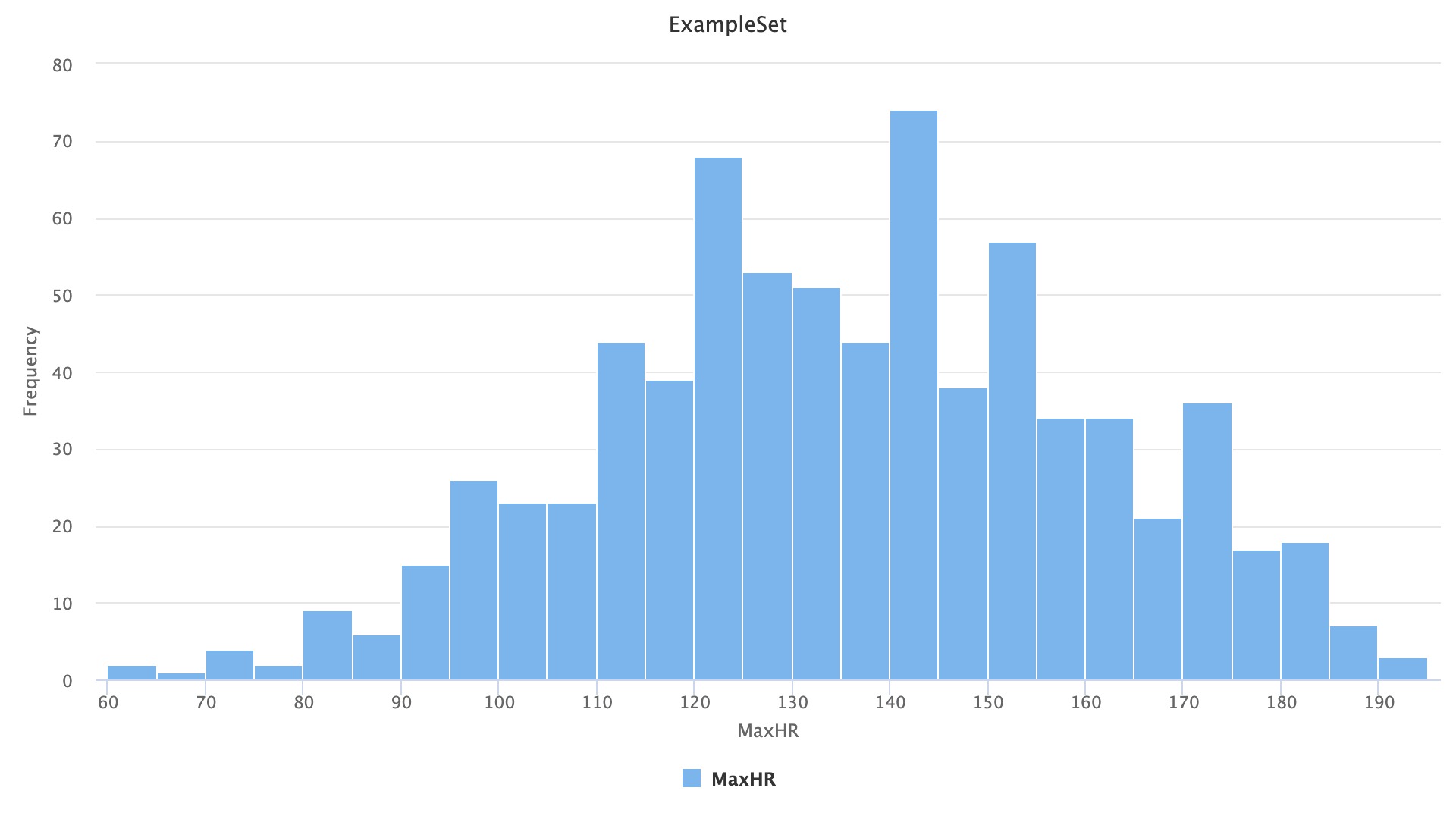
*Figura 2: Distribución de la variable “Oldpeak”*

Graficando las variables, se encuentra la variable OldPeak, que representa anomalías en la onda registrada por el electrocardiograma realizado al paciente. Como puede verse en la figura 2, el 41% de los registros (307 de 750) toma un valor de 0, lo cual representaría un ECG que no presenta valores anómalos. El 60% restante se distribuye de manera símil Normal, centrada entre 1 y 2,33, lo cual indicaría valores que se alejan del valor normal, y podrían indicar alguna condición en el corazón.

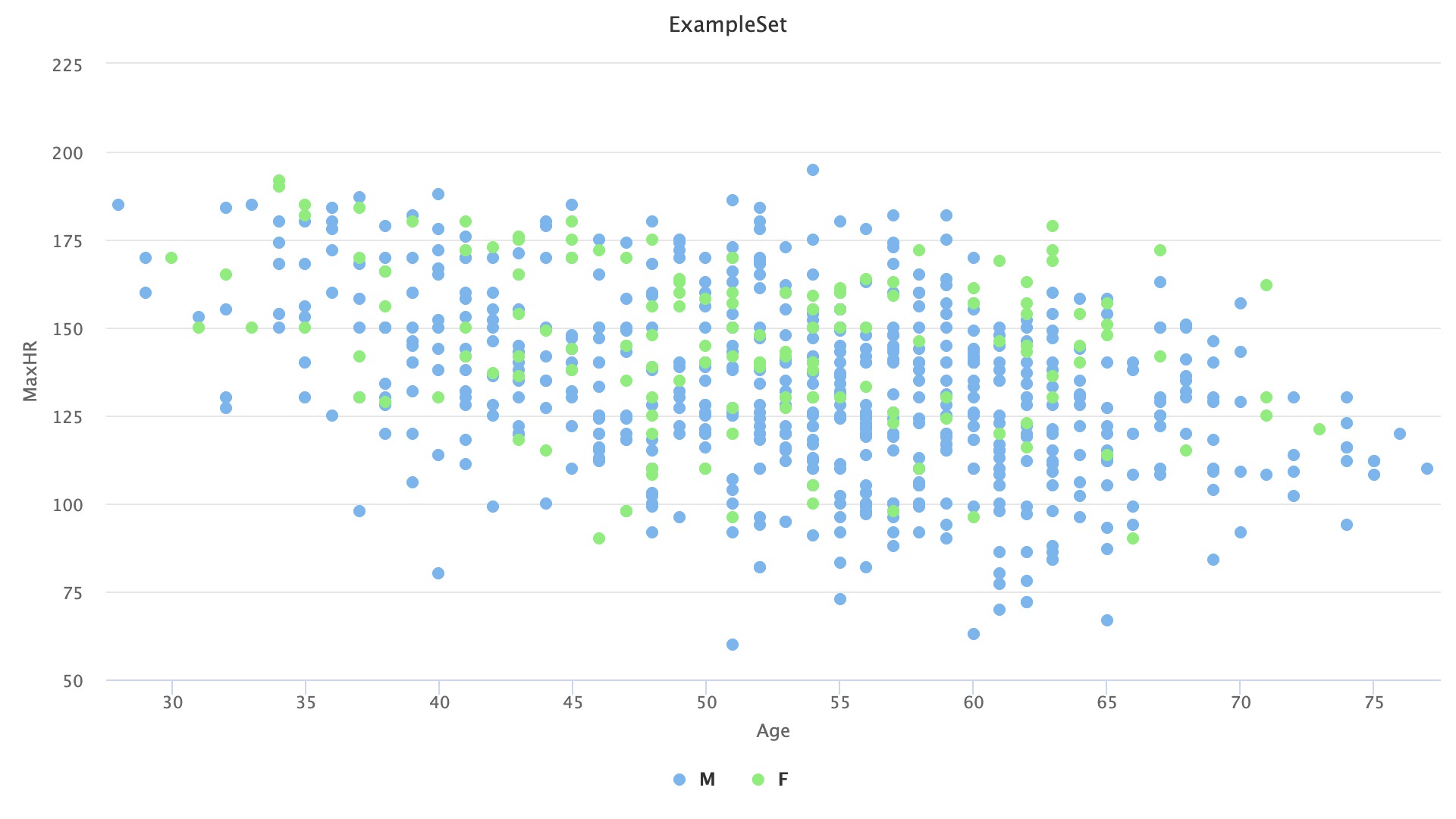


*Figura 3: Distribución de la variable “RestingBP”*

Luego, se estudia la variable “RestingBP”, medida en mm de Hg. Ésta tiene una distribución normal, pero llama la atención un registro donde la presión sanguínea toma un valor nulo, que en el próximo apartado se verá cuál será el tratamiento que este recibirá.

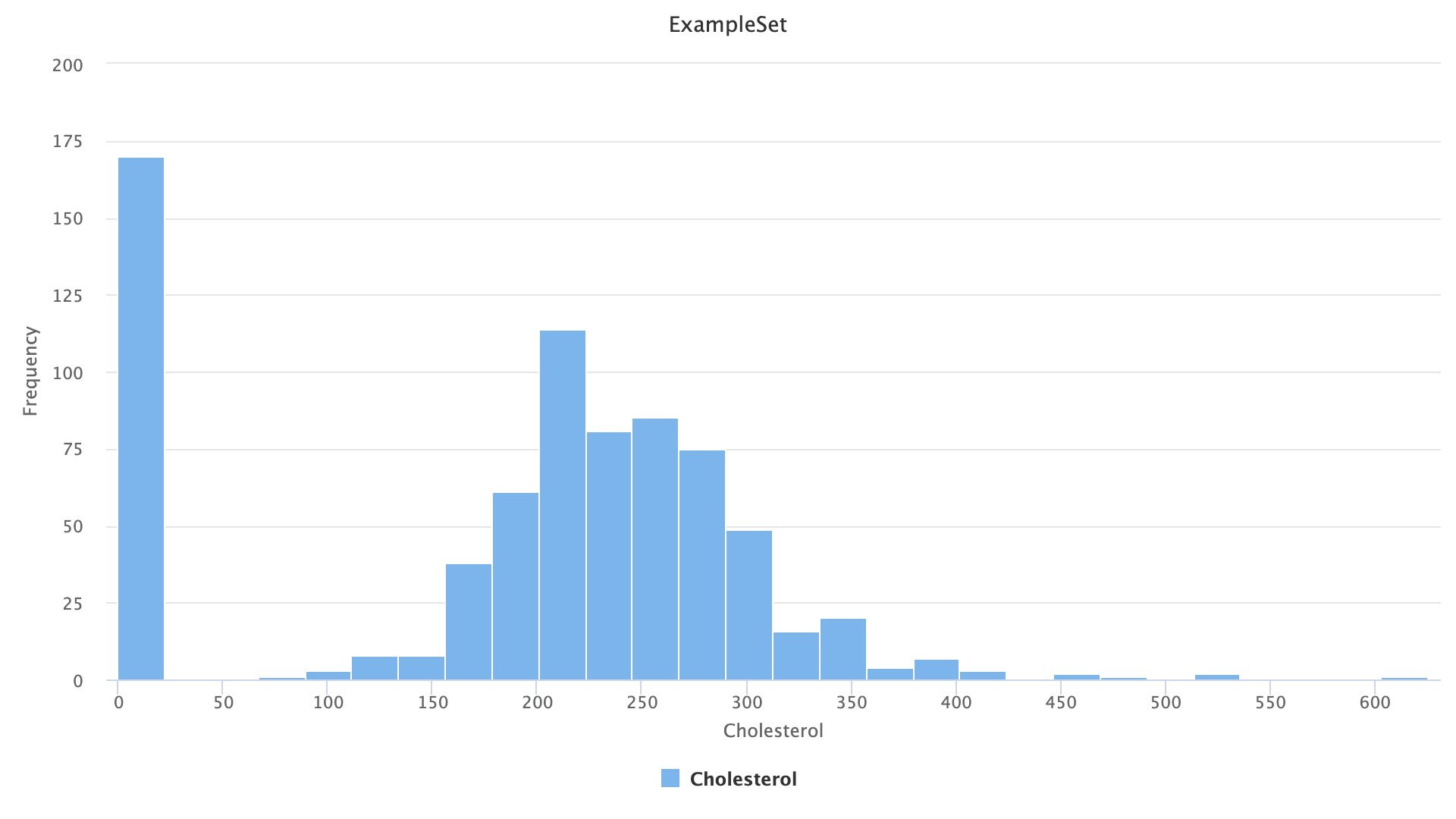
*Figura 4: Distribución de la variable “MaxHR”*

Otra de las variables estudiadas es el “MaxHR”, medido en latidos por minuto. Esta variable también tiene una distribución normal, con una media de 134. El valor máximo de latidos por minuto aceptable se calcula como la diferencia entre 220 y la edad del individuo, lo cual permite agrupar en pacientes que superan la frecuencia cardiaca máxima y los que no.



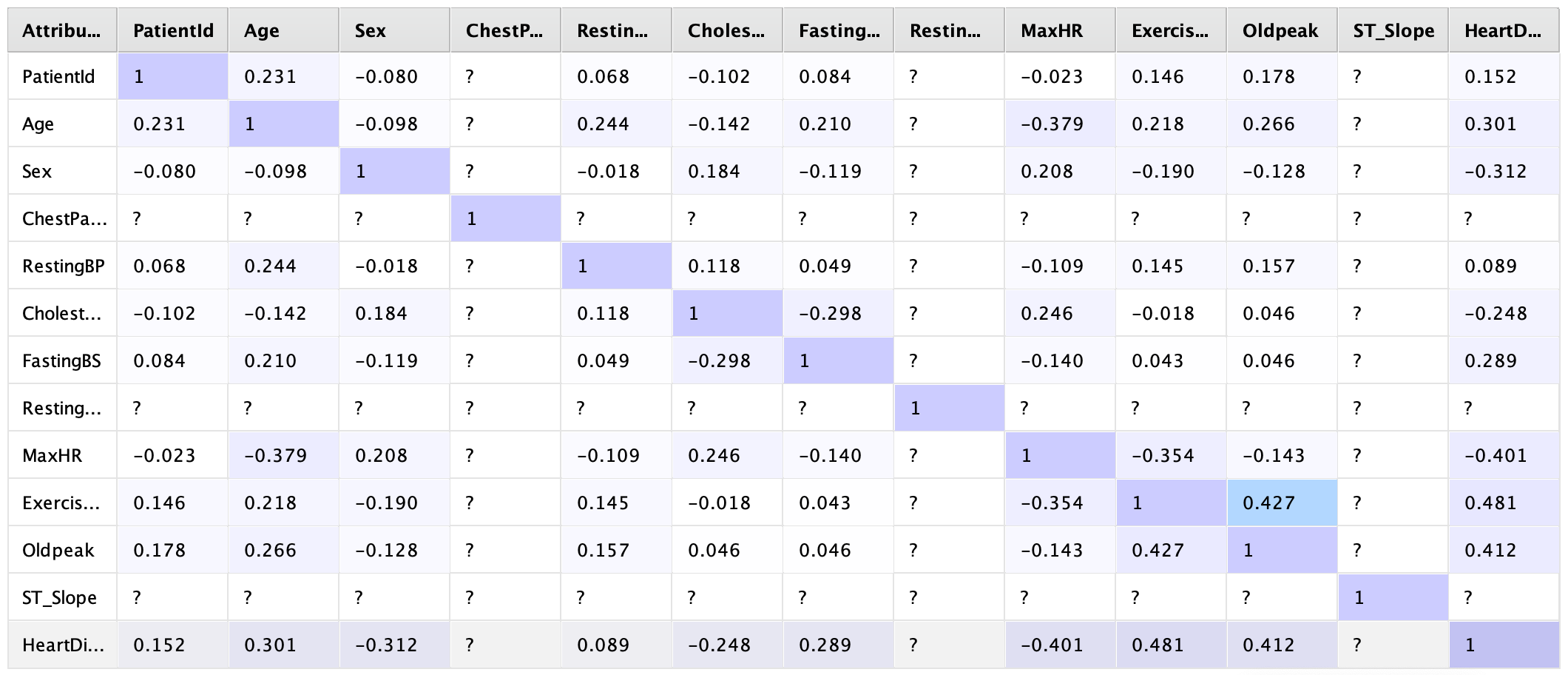
*Figura 5: Correlación entre las variables “Age” y “MaxHR”*

Profundizando el análisis de la variable “MaxHR”, parece notarse una relación entre la edad del paciente y la frecuencia cardiaca máxima registrada, lo cual a su vez despierta la necesidad de estudiar las relaciones entre el resto de las variables, que posteriormente se analizará.



*Figura 6: Distribución de la variable “Cholesterol”*

La última de las distribuciones que resaltan es la del “Cholesterol”, que tiene 170 registros nulos, lo cual en la realidad no es un valor posible por lo cual se deberá dar una interpretación especial, a desarrollar nuevamente en el próximo apartado.



*Tabla 2: Matriz de correlaciones*

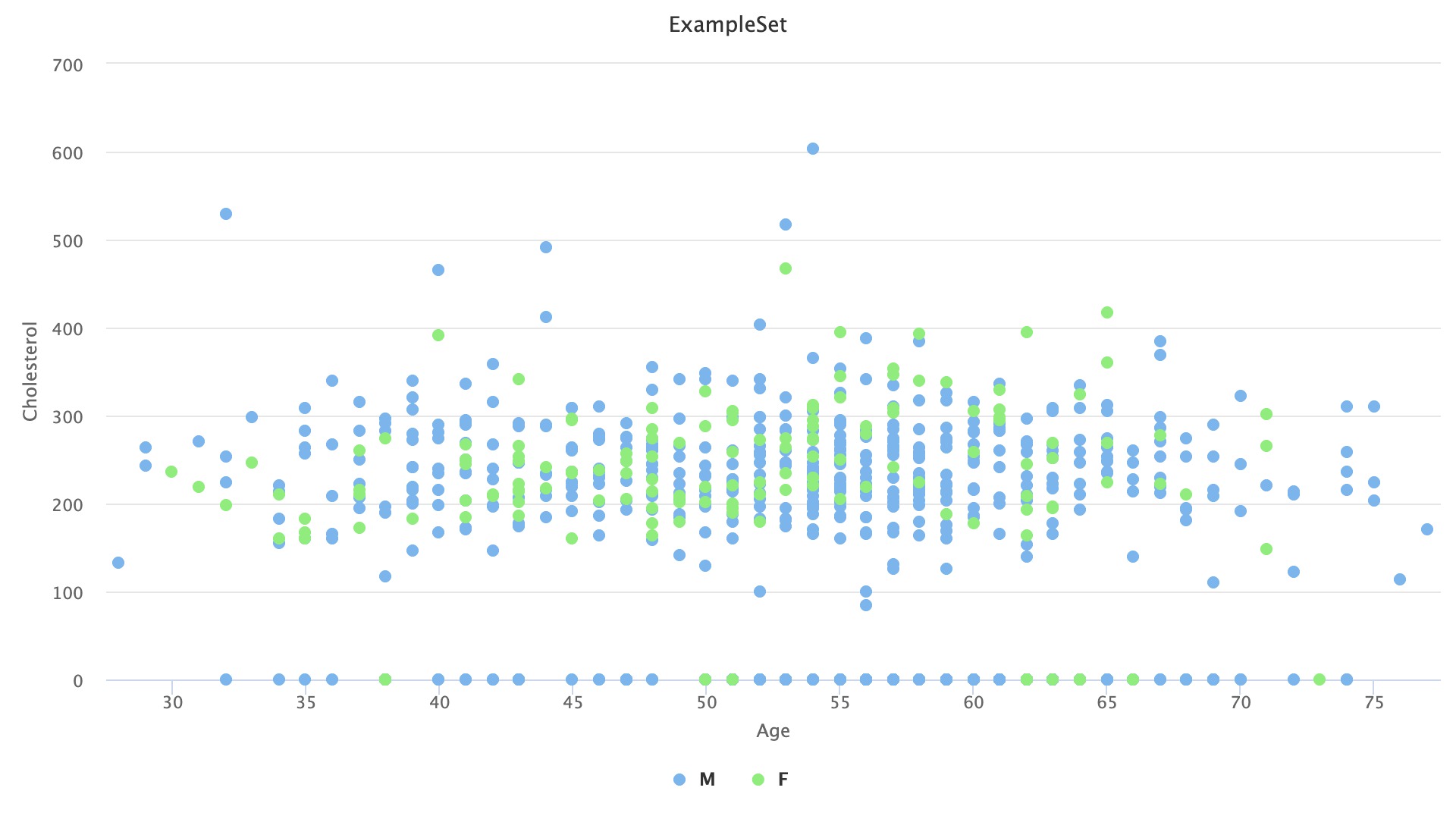
Para finalizar el análisis exploratorio se realiza un estudio de correlación entre las variables. Esto permite entender si hay dependencias entre variables, para descartar aquellas que son dependientes y evitar duplicar información en el modelo, además de sesgar la interpretación de la información por parte del software.

Realizando la matriz mencionada anteriormente, se nota que ningún par de variables supera el umbral del 95% estipulado para ser considerada como correlación directa, por lo cual no se eliminará ninguna de las variables estudiadas, y se procede a la limpieza de los datos, donde se volverá a estudiar la correlación para validar que los cambios a realizar no afecten las relaciones entre variables.

# Limpieza

Para comenzar con el análisis del conjunto de datos de entrenamiento brindado por la cátedra, primero se realizó un estudio preliminar buscando datos erróneos o celdas sin datos. No se encontraron celdas sin información.

En primer lugar, para la variable *Colesterol*, se hallaron 170 registros con valor de colesterol igual a 0, lo cual carece de sentido. Se optó por reemplazar estos valores nulos por el promedio de 244.7 mg/dL en el dataset de entrenamiento previo a utilizarlo como input en el software RapidMiner.



*Figura 7: Correlación entre las variables “Cholesterol” y “Age”*

Por otro lado, también se halló que la variable “RestingBP*”* contenía un valor único con valor igual a 0 que no tenía coherencia alguna ya que esto representa una presión sanguínea nula. Al ser poco representativo un único valor en todo el dataset de 750 registros, se lo eliminó para no perjudicar el resultado del modelo.

Por último, se eliminó la variable “FastingBS*”* dado que al investigar este tipo de estudio en sangre, se halló que este representa el nivel de glucosa en sangre (en ayuno), lo cual es un valor continuo que puede encontrarse dentro de un rango de [70 ; 100] mg/dL. Como dicha variable únicamente presentaba valores binarios de 0 y 1, se decidió eliminarla ya que se entiende que estos valores ensuciaron el aprendizaje del programa.

# Ingeniería de Variables

A continuación, se realizaron las siguientes transformaciones:

Se cambió el rol de la columna “PatientId*”* por “id” y el rol de la variable predictora *HeartDisease* por “label” ya que esta variable es nuestra variable objetivo a categorizar a futuro con el conjunto de datos de testeo.

También, la variable “HeartDisease***”*** que se la transformó a una variable dicotómica en vez de numérica ya que el programa no puede estimar resultados numéricos, sumado a que se desea obtener una respuesta categórica como el de “tiene una enfermedad cardiaca” o “no tiene una enfermedad cardiaca”.

Por otro lado, se transformaron las variables polinómicas (texto) a datos numéricos en pos de agilizar el aprendizaje del programa. Entre ellas:

* “ChestPainType”: se le asignaron números enteros de 0, 1, 2 y 3 en lugar de los valores polinómicos de TA, ATA, NAP, ASY. No necesariamente el modelo le asignó esos valores a las variables respectivamente.
* “RestingECG”: se reemplazaron los valores enteros de 0, 1 y 2 en lugar de Normal, ST o LVH.
* “ExcerciseAngina”: se reemplazaron los valores binarios de 0 y 1 en lugar de N y Y, respectivamente.
* “ST\_Slope”: se reemplazaron los valores down, flat y up por 0, 1 y 2. Nuevamente, el modelo no tiene porque haber asignado tal valores a las variables en forma respectiva.

# 

# 

# 

# 

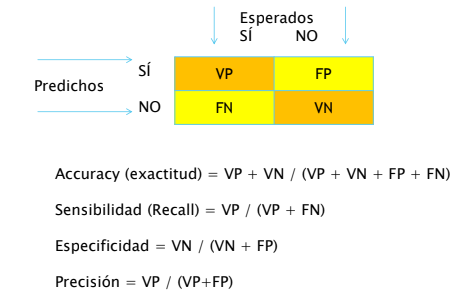
# 

# 

# 

# Métricas

Teniendo en cuenta la naturaleza del problema, la métrica más adecuada para evaluar la performance de los distintos modelos que se tuvieron en cuenta es la sensibilidad.



*Tabla 3: Matriz de confusión*

Como se puede apreciar en la tabla 3, la sensibilidad define el porcentaje de casos verdaderos (efectivamente tienen una enfermedad cardiaca) correctamente identificados sobre el total de casos enfermos (Correctamente predichos e incorrectamente descartados por el modelo).

Esta métrica fue elegida ya que es muy importante conocer cuántos de los pacientes que verdaderamente tienen enfermedades se identifican, y se hace el foco en poder alertar a la totalidad de las personas enfermas. No identificar un positivo podría ser un grave problema, empeorando la situación del paciente al no tratarse a tiempo.

# Modelos y Selección de Variables

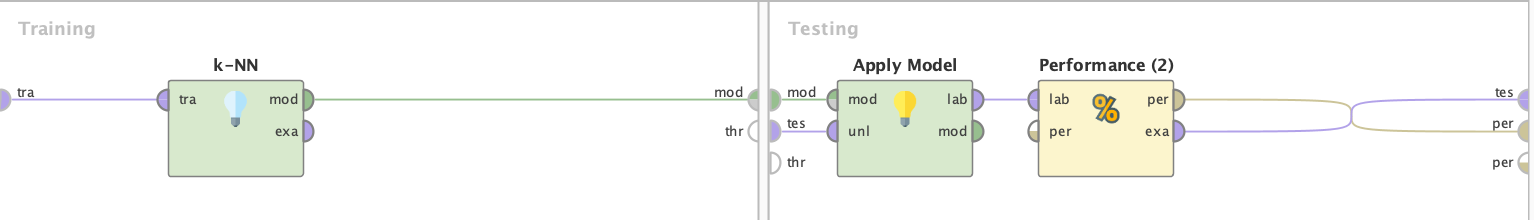
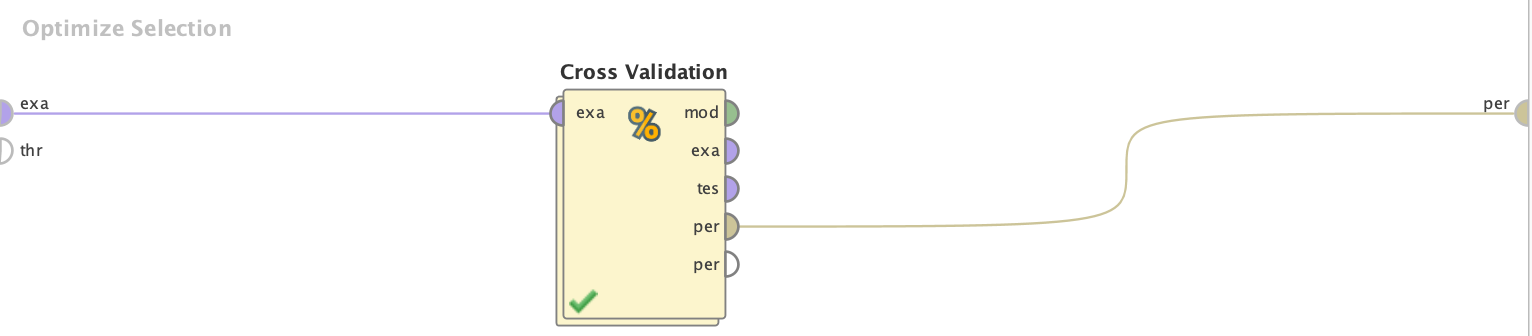
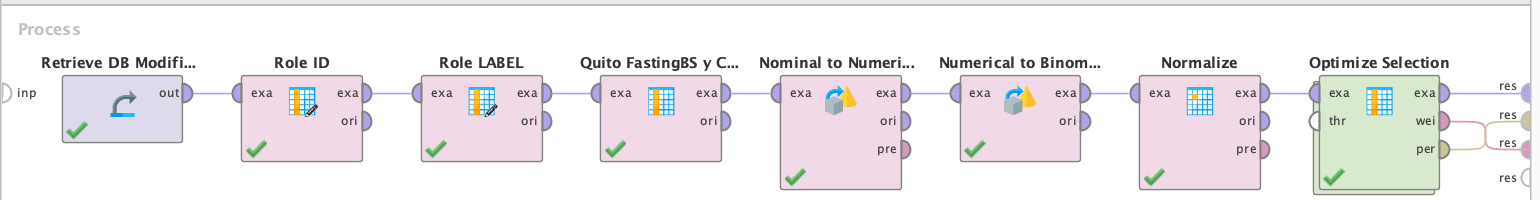
Los modelos que se evaluaron fueron:

* k-NN: K-Nearest Neighbours (k vecinos más cercanos)
* Regresión Logística
* Árbol de Decisión

Se evaluaron los 3 modelos y se seleccionó el que tenía mejor ajuste y presentaba la mayor adecuación al caso de estudio.

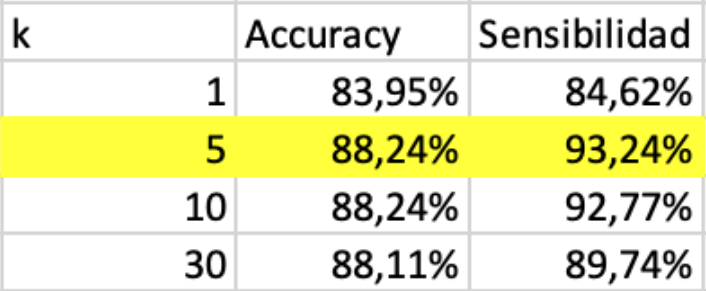
## Modelo 1: k-NN

Se corrió el modelo con un Cross Validation dentro del Optimize Selection para que seleccione las variables adecuadas que maximizan la performance del método predictor del modelo.



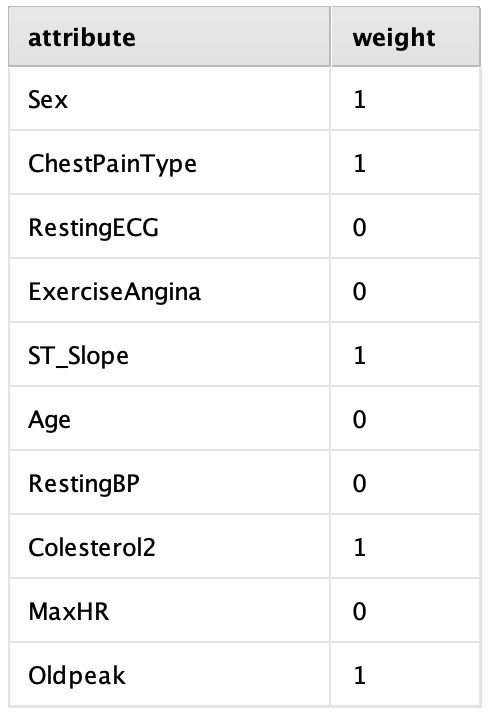
*Figura 8: Visualización del proyecto con el modelo k-NN.*

Se estudiaron diversos valores de K = [1, 5, 10, 30]. El modelo que mejor sensibilidad tuvo fue con K=5, es decir, el modelo que elige 5 “vecinos” más cercanos para predecir el resultado.



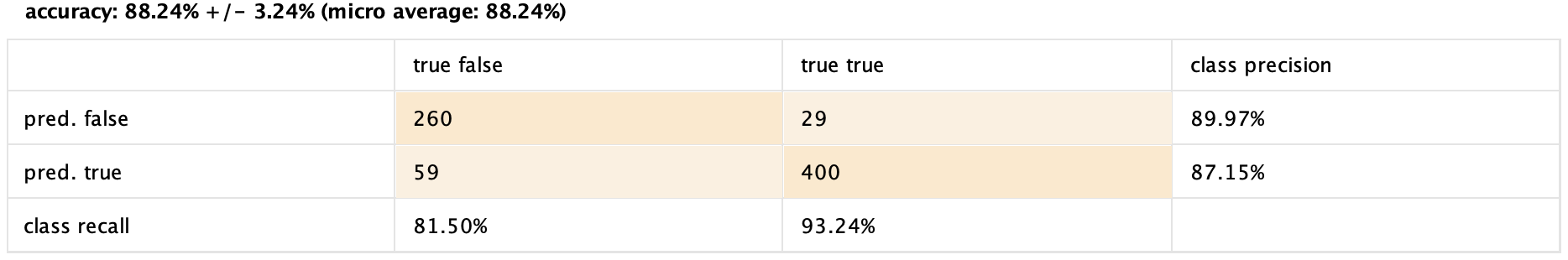
*Tabla 4: Resultados del modelo k-NN.*

En la tabla a continuación se listan las variables que fueron seleccionadas por el modelo k-NN.



*Tabla 5: Variables seleccionadas en el modelo k-NN.*

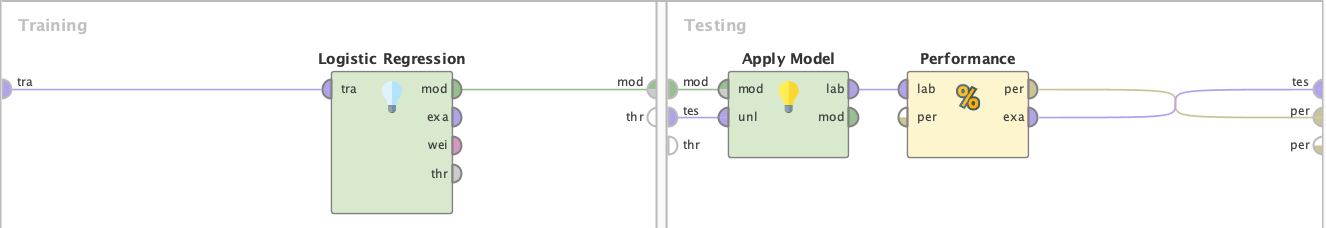
Por último, se presenta la matriz de confusión con el mejor K probado, K=5.



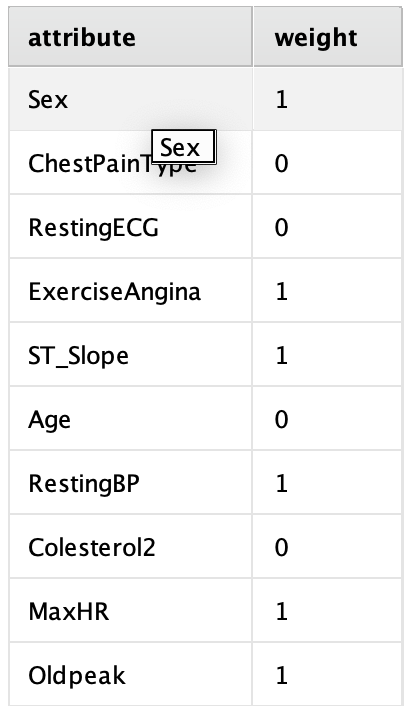
*Tabla 6: Matriz de Confusión, k-NN con K=5.*

## Modelo 2: Regresión Logística

La regresión logística es un modelo de clasificación. Devuelve valores continuos que luego deben ser llevados a clases. Se pasa por una función que transforma la salida en una variable binaria.

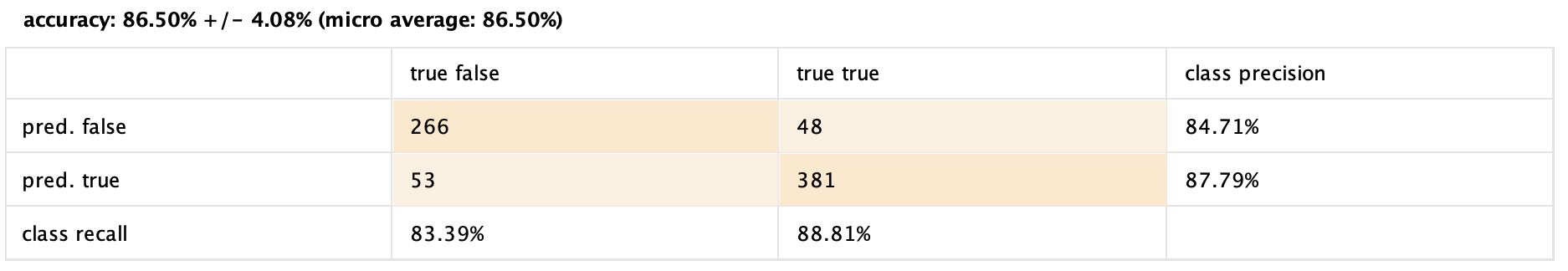


*Figura 9: Modelo de Regresión Logística.*



*Tabla 7: Variables seleccionadas por el modelo de Regresión Logística.*

Este modelo arrojó una sensibilidad del 88,81%, la cual no supera al modelo k-NN con K=5.

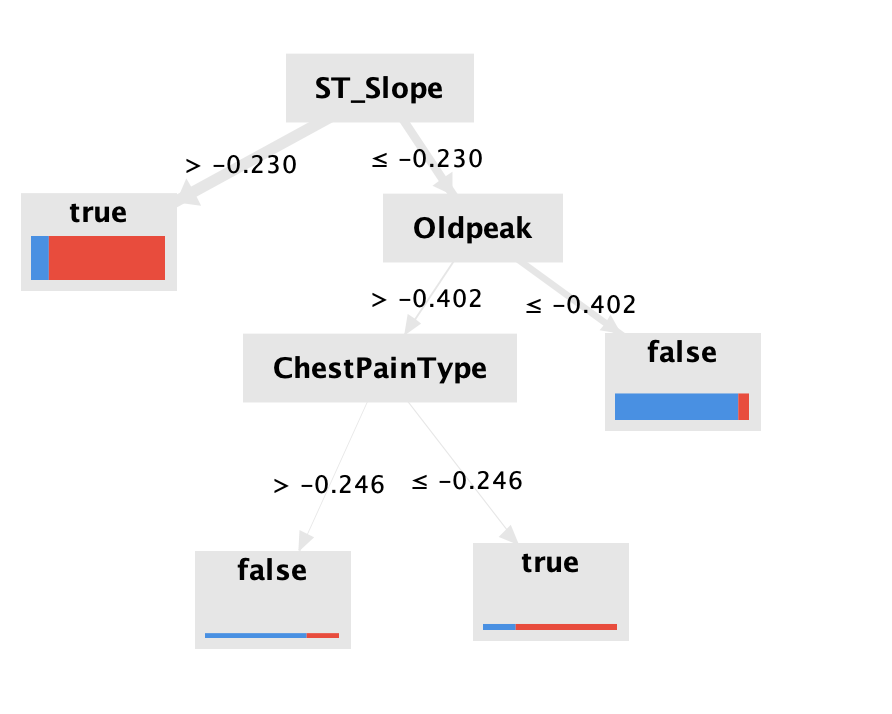


*Tabla 8: Resultado de la Regresión Logística.*

## Modelo 3: Árbol de Decisión

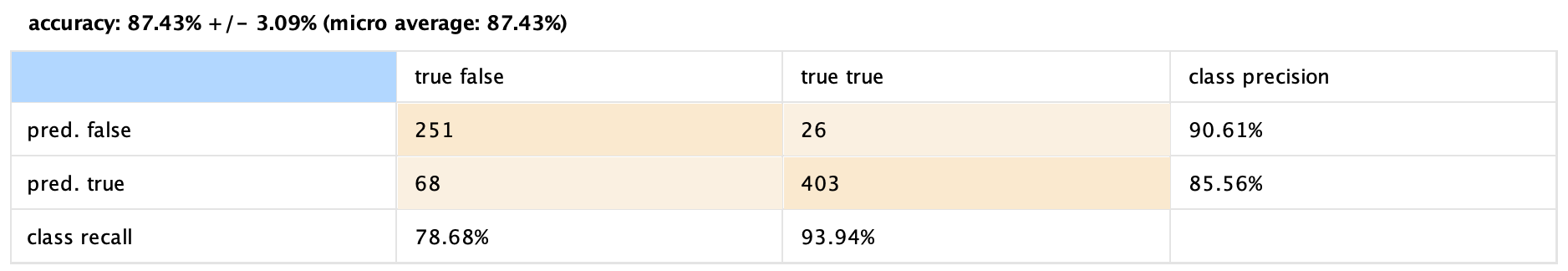
Como hiperparámetros del modelo para evitar overfitting de datos se utilizaron los valores por default y se estableció una profundidad máxima admitida por el árbol de 3 niveles.

En el siguiente gráfico se puede observar que la primera variable del árbol es la pendiente del segmento ST. A partir de esta variable se comienza a particionar los datos en subconjuntos, esta variable resulta ser el mejor atributo que genera entropía al modelo. Es decir, separar de manera más uniforme las clases. Podemos ver que la ganancia de información es máxima ya que si la pendiente supera el valor -0.23, entonces ya se concluye una respuesta de que hay riesgo de padecer una enfermedad cardiaca. De esta forma se puede ordenar las variables más informativas de manera ascendente: “ST Slope”, “Oldpeak” y “ChestPainType”.

****

*Figura 11: Decision tree desarrollado por el software*

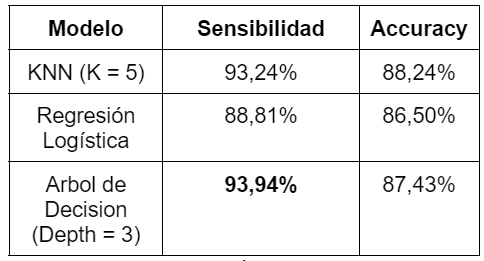
La performance obtenida con este modelo fue la siguiente:

****

*Tabla 9: Matriz de confusión del Árbol de Decisión*

# Modelo Seleccionado y Análisis de Resultados

A partir de la comparación de los 3 modelos desarrollados previamente, y considerando a la sensibilidad como el marcador más relevante para la elección, se avanza con las predicciones del modelo del Árbol de Decisión.

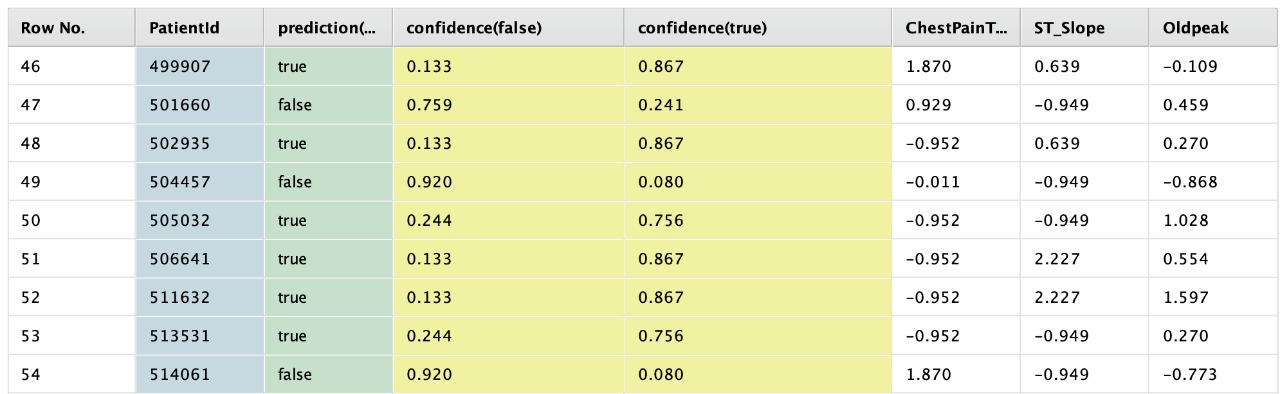


*Tabla 10: Comparación entre modelos*

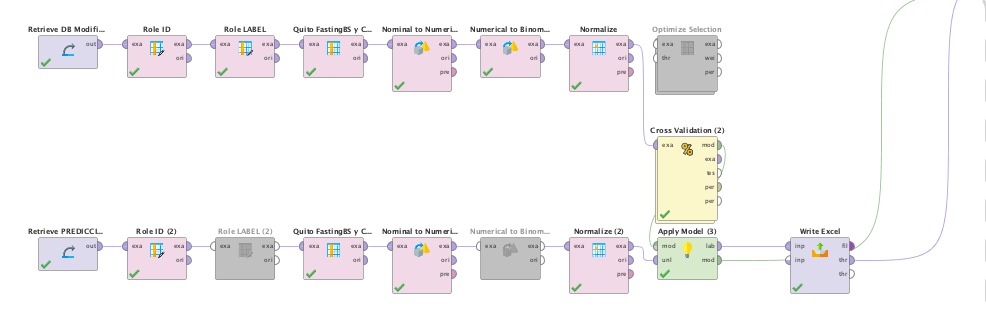
**Testeo**

Finalmente, una vez definido el modelo, se procedió a evaluarlo con el conjunto de testeo que el modelo nunca vió, con el objetivo de evaluar su performance para casos nuevos y poder determinar qué clientes tienen probabilidades de tener una enfermedad cardiaca y tomaron el préstamo.

De los 170 casos de evaluación, se determinaron que 125 tienen probabilidades de presentar una enfermedad cardiaca, mientras que 45 pacientes no padecen de peligro alguno. Esto representa un 73,53% de los pacientes muestreados para el testeo.

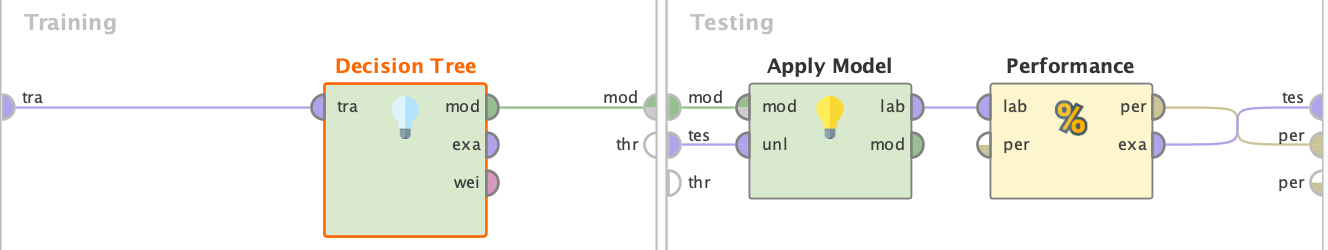
****

*Tabla 11: Muestra de resultados obtenidos del conjunto de testeo*



*Figura 12: Estructura del modelo predictivo final*

# 



*Figura 13: Modelo del Árbol de decisión*

# Bibliografía

* Tutoriales RapidMiner en Youtube de la cuenta “Dataminingincae” del Incae Business School: <https://www.youtube.com/channel/UCbdsyFBdG0Fmjroqk-ZoPcg>
* Video de Youtube de “Data Science At Incae”: <https://www.youtube.com/watch?v=jIq3rTbOaZQ>
* Informacion sobre enfermedades cardiacas, segmento ST, OldPeak y estudios cardiacos: [www.medlineplus.gov](http://www.medlineplus.gov); [www.healthline.com](http://www.healthline.com); [www.ecocardio.com](http://www.ecocardio.com) y wikipedia

1. <https://es.wikipedia.org/wiki/Segmento_ST> [↑](#footnote-ref-1)