Machine Learning con el Corazón

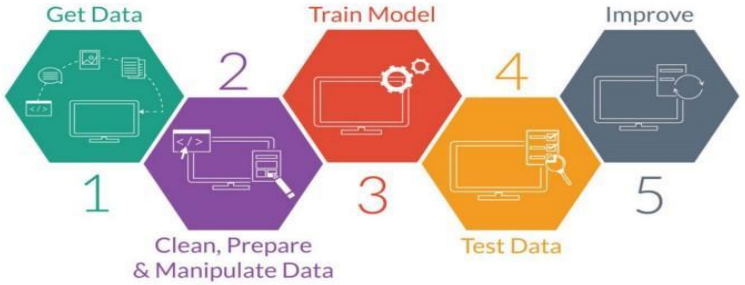
Predicción de enfermedades cardiacas

Información disponible por paciente

Partimos de un conjunto de **datos estructurados** con siguiente información por paciente:

- **slope_of_peak_exercise_st_segment**: Lectura de electrocardiografía que indica la calidad y coordinación del funcionamiento del corazón
- **thal**: Resultados de prueba de esfuerzo con talio, que mide la calidad del flujo de sangre al corazón y ser indicativo de daños.
- **resting_blood_pressure**: Presión arterial en reposo, puede ser indicativo de cardiopatías y potenciales infartos si es anormal.
- **chest_pain_type**: Tipo de dolor en el pecho.
- **Num_major_vessels**: Número de vasos principales coloreados por flouroscopia, angiograma coronario. Puede ser indicativo de anginas y otras.
- **fasting_blood_sugar_gt_120_mg_per_dl**: Azúcar en la sangre en ayunas mayor de 120 mg/dl. Mide la potencialidad de diabetes e irregularidades en la producción de insulina y gestión de la glucosa por el cuerpo.
- **resting_ekg_results**: Resultados electrocardiográficos en reposo. Si los resultados son extremos puede indicar una situación cardiovascular fatal.
- **serum_cholesterol_mg_per_dl**: Análisis de colesterol serum. El indice de colesterol no da síntomas, por lo que este análisis puede indicar una deficiencia en la situación coronaria del individuo e incluso proximidad a obstrucciones.
- **oldpeak_eq_st_depression**: Irregularidad en el nivel del ST-segment del ECG que puede indicar isquemia, condición que provoca deficiencia de riego sanguineo a tejidos del cuerpo.
- **sexo**
- **edad**
- **max_heart_rate_achieved**: Frecuencia cardíaca máxima alcanzada (latidos por minuto). Si es anormal puede indicar irregularidades coronarias.
- **exercise_induced_angina**: Dolor de pecho inducido por el ejercicio, puede ser indicativo de situaciones graves como daños coronarios y obstrucción coronaria que puede llevar a infarto.

Introducción



Las enfermedades cardiacas son la primera causa de mortalidad a nivel mundial, para aprender cómo prevenirlas primero debemos aprender a **detectarlas** de forma fiable.

La iniciativa del presente estudio viene de la plataforma **drivendata**, y los datos facilitados vienen de un estudio sobre enfermedades cardiacas que ha sido abierto al público hace años. El estudio con datos anónimos de pacientes, recopila mediciones sobre la salud y estadísticas cardiovasculares de los pacientes.

Nos enfrentamos a un problema de **Clasificación Binaria**, cuyo **objetivo** es minimizar el número de **falsos negativos** que pudieran darse, para ello se han tomado métricas que minimicen estos casos asumiendo que esto implica que se darán más falsos positivos.

Análisis y tratamiento de datos

Parte fundamental del Machine Learning es entender los datos y procesarlos para que los modelos puedan tratarlos correctamente , puntos clave:

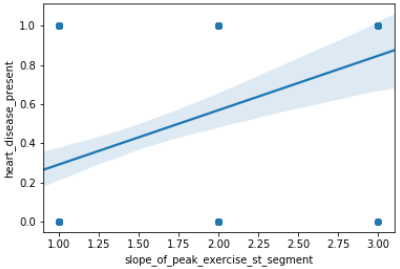
1. **Recopilar** y preprocesar datos, limpieza, tratamiento de nulos, validación... en nuestro caso no aplica.
2. **Codificación** de campos con texto a valores numéricos.
3. Verificar si los datos están **balanceados** y en su caso aplicar técnicas para minimizar impacto, en nuestro caso no aplica pues tenemos un 44% de casos con problemas cardíacos y un 56% sin enfermedad. Ejemplo con un 1% de positivos el modelo entendería que todos son negativos y tendría una precisión del 99%.
4. Análisis exploratorio de datos en busca de:
    a. Inconsistencias.
    b. Relaciones entre los datos.
    c. Posibles transformaciones que mejoren resultados.

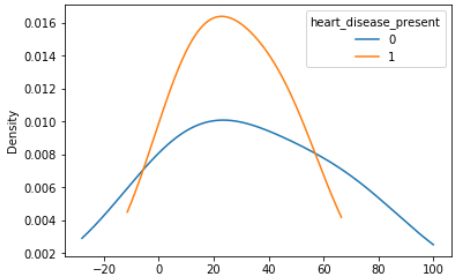
Análisis exploratorio de datos

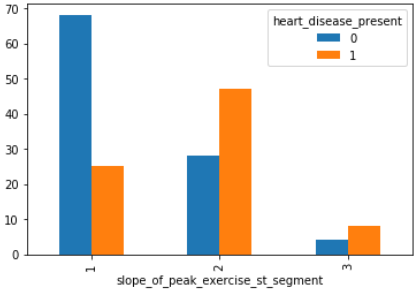
Diagramas que nos ayuda a tener imagen visual que relaciona cada variable con el target.

1. Nube de puntos con regresión lineal.
2. Diagrama de densidad, que nos muestra cómo se distribuyen los valores de la variable.
3. Número de elementos que hay en la muestra para cada valor.

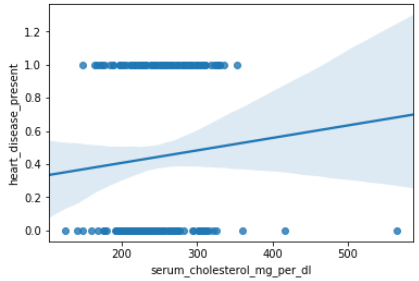
Lectura de **electrocardiografía** que indica la calidad del flujo sanguíneo al corazón

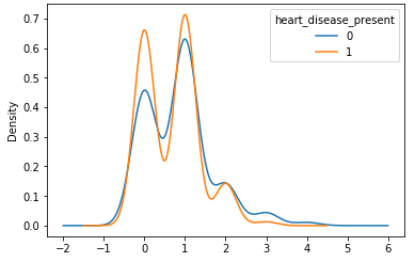


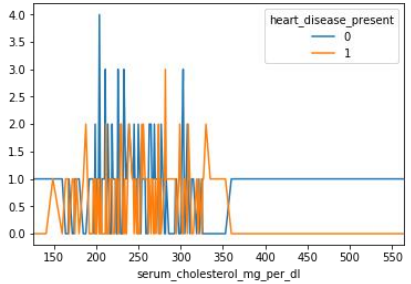




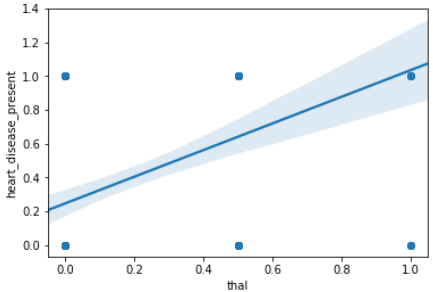
Colesterol serum

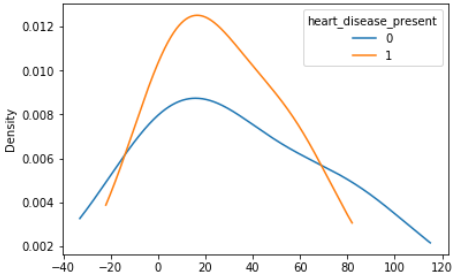


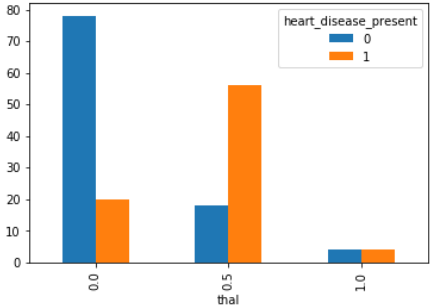




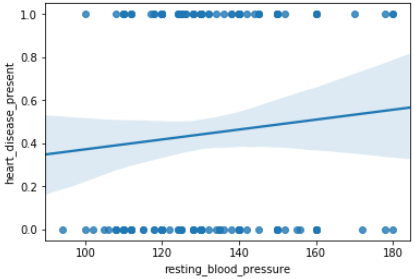
Resultados de **prueba de esfuerzo con talio**, que mide el flujo de sangre al corazón.

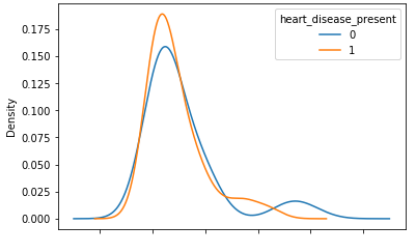


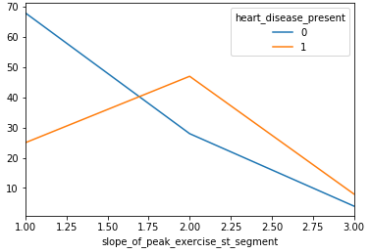




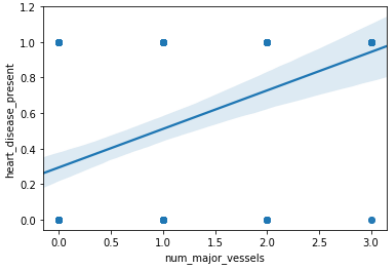
Presión arterial en reposo

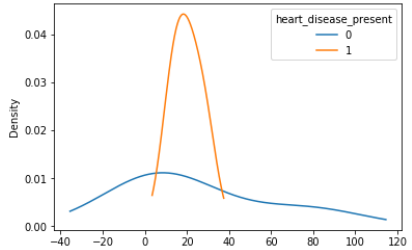


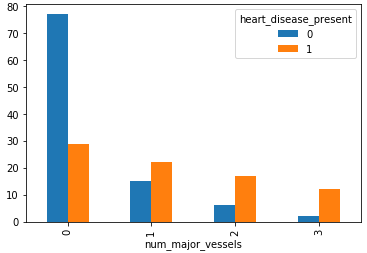




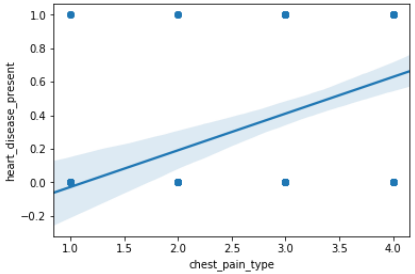
Número de vasos principales coloreados por flourosopy.

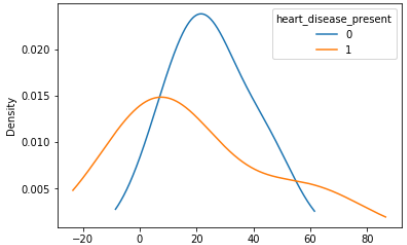


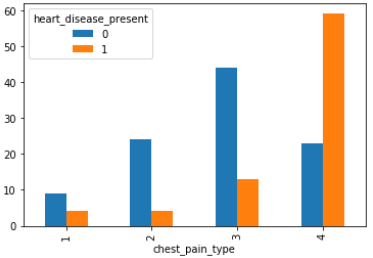




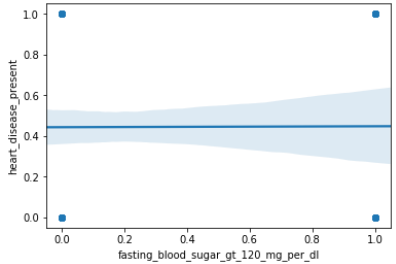
Tipo de dolor en el pecho.

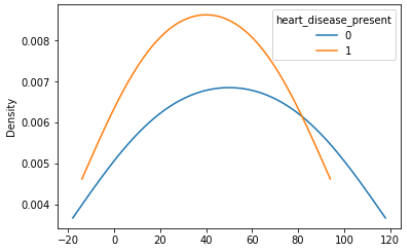


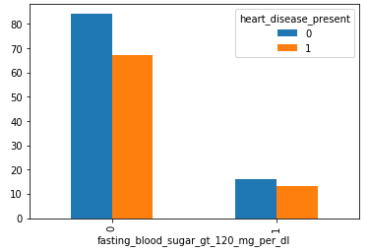




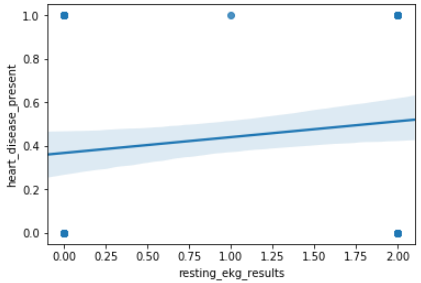
Azúcar en la sangre en ayunas mayor de 120 mg/dl.

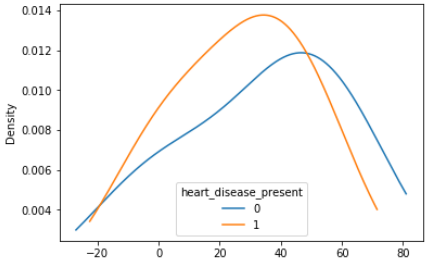


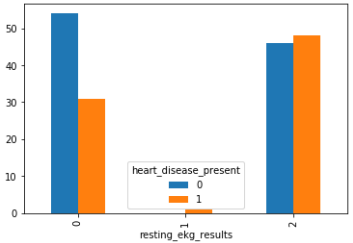




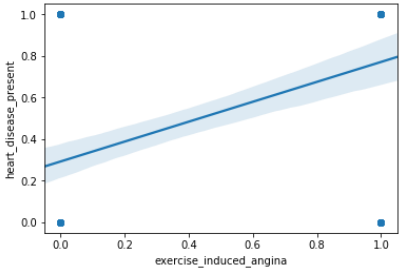
Electrocardiográficos en reposo.

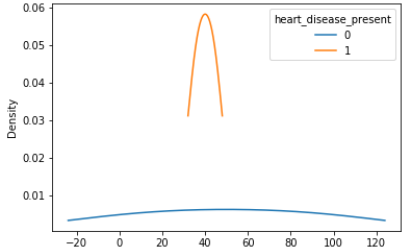


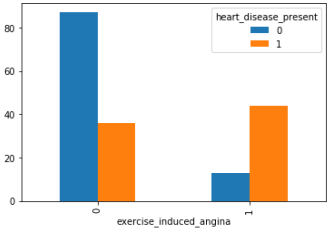




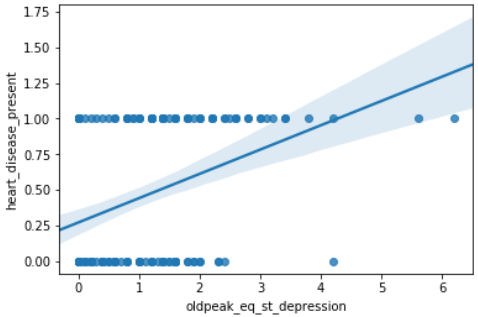
Dolor de pecho inducido por el ejercicio.

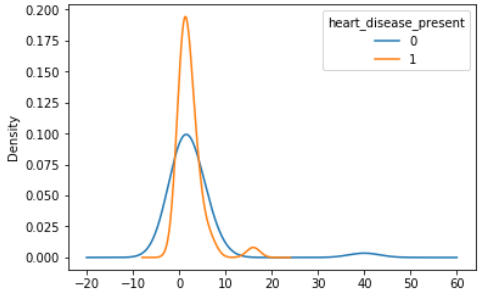


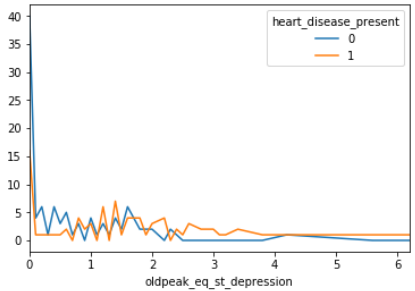




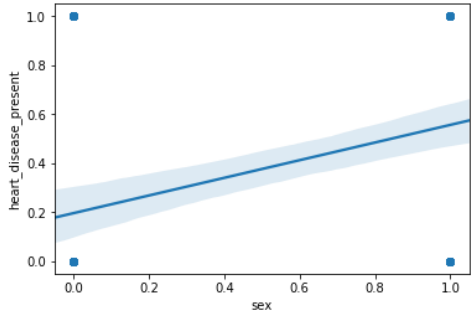
Depresión inducida por el ejercicio en relación con el reposo, medida de anomalía en los electrocardiogramas.

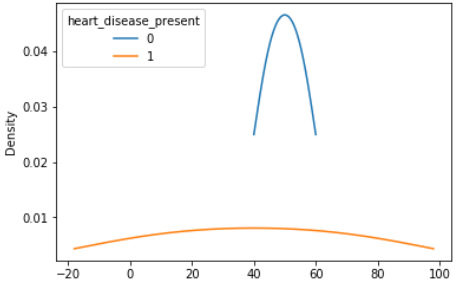


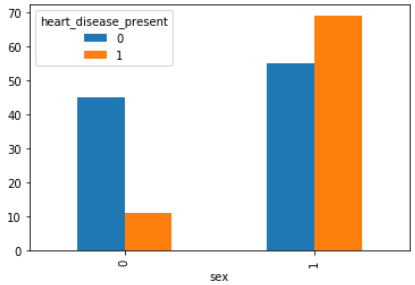




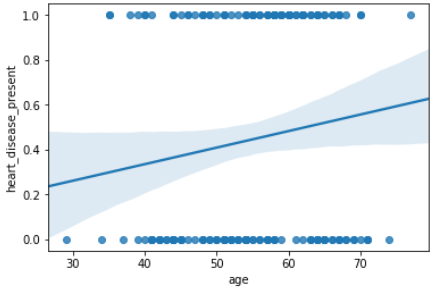
Género

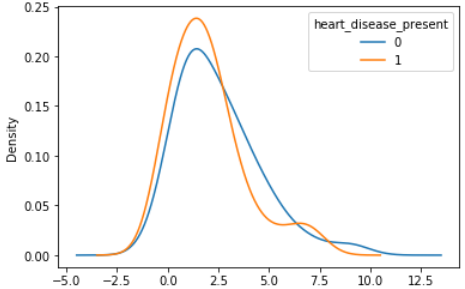


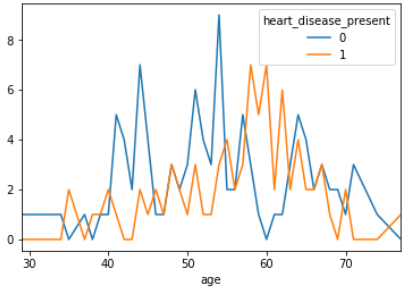




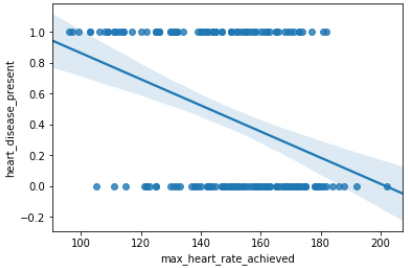
Edad

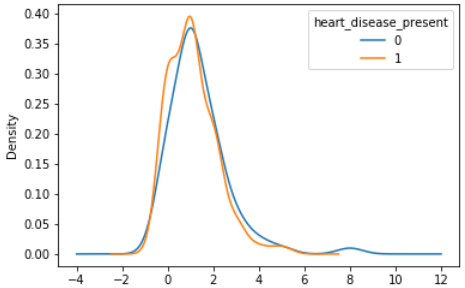


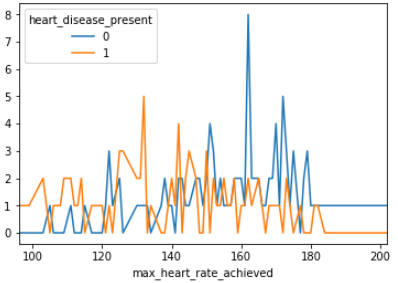




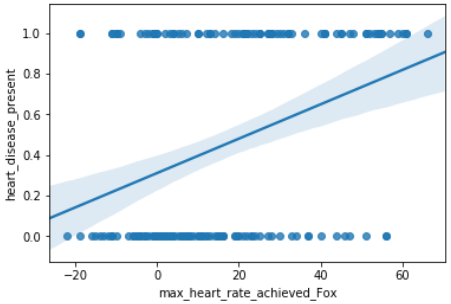
Frecuencia cardíaca máxima alcanzada (latidos por minuto)

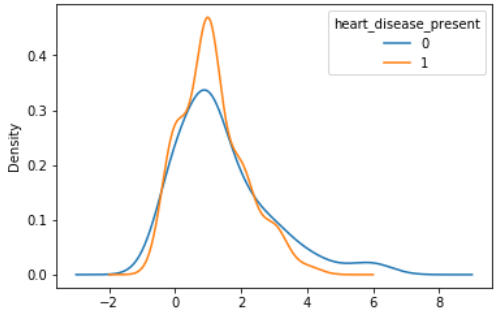


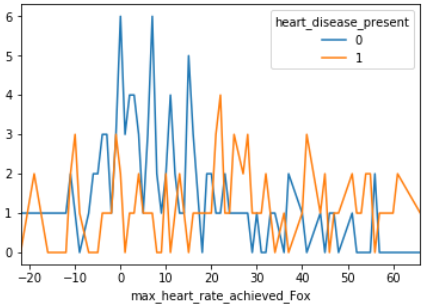




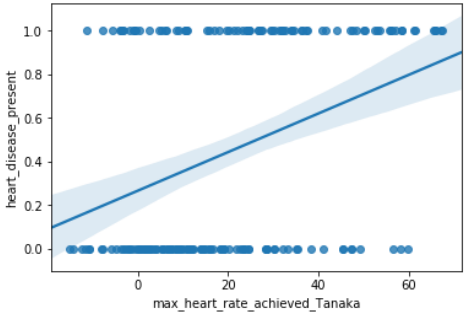
Variable compuesta, fórmula Fox que relaciona frecuencia cardiaca máxima con edad (220-(Edad+Frecuencia Cardiaca)).

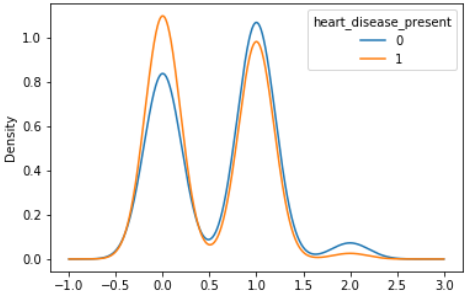


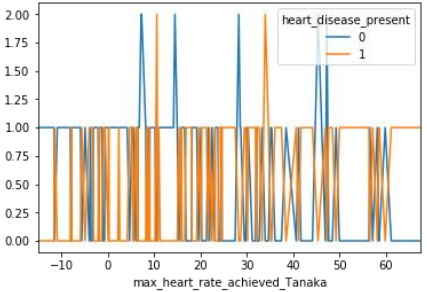




Variable compuesta, fórmula Tanaka que relaciona frecuencia cardiaca máxima con edad (208-(0.7*Edad+Frecuencia Cardiaca)).







Ingeniería de variables

- **One hot encoding**: Sobre las variables codificadas, se han añadido tantas columnas como posibles valores codificando su contenido en SI o No.

- **Combinación** de variables, se han combinado variables de edad y frecuencia cardiaca máxima según fórmulas de Fox y Tanaka que con estas variables obtiene como se desvían de los datos normales de forma que la posibilidad de enfermedad cardiaca aumenta cuando los valores obtenidos se alejan de cero.

Estrategia de validación

Como norma general para validar un modelo debemos dividir los datos de muestra en dos subconjuntos uno de entrenamiento y otro de validación. De forma que la precisión sobre el mismo nos indica cómo de bien se comporta el modelo para los datos de entrenamiento y la precisión con el segundo con indica cómo de bien **generaliza** nuestro modelo.

En nuestro caso tenemos un problema pues solo contamos con 180 pacientes, ya que dividir el conjunto en dos haría que o bien tuviésemos muy pocos datos para entrenar o que la validación no sea representativa. Por este motivo se ha optado por entrenar con el conjunto completo y validar con las predicciones que daremos a la propia competición.

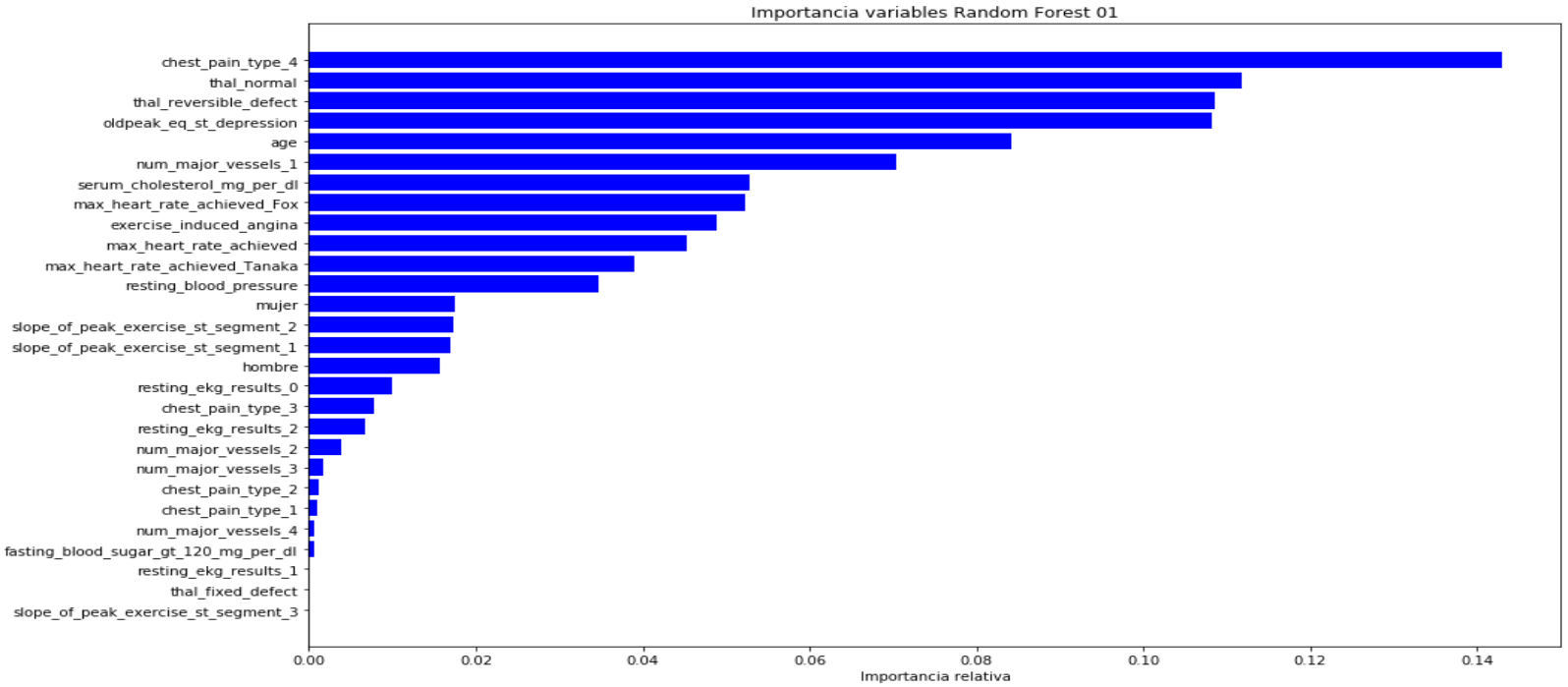
Modelos predictivos utilizados

En el presente estudio se han utilizado siguientes modelos predictivos.

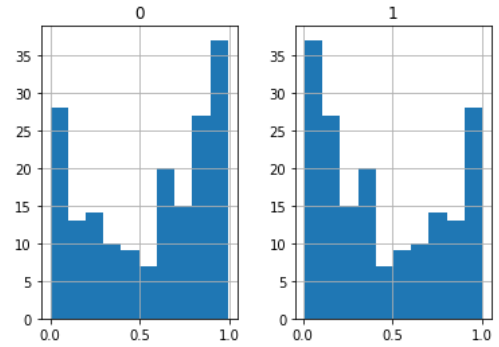
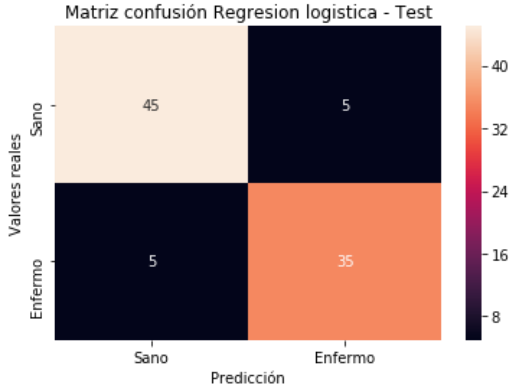
1. **Regresión logística:** Modelo enmarcado dentro del conjunto de Modelos Lineales Generalizados (GLM por sus siglas en inglés)
2. **Random Forest Classifier**: Estimador que se ajusta a una serie de árboles de decisión en varias submuestras del conjunto de datos y utiliza el promedio para mejorar la precisión y el ajuste.
3. **Cat Boost Regressor**: Algoritmo de aprendizaje automático que utiliza el aumento de gradiente en los árboles de decisión.
4. **SGDClassifier**: Algoritmo de aprendizaje que usa descenso de gradiente estocástico simple.

Para **ajustar** los modelos se ha usado utilizado GridSearchCV que nos permite probar con combinaciones de configuraciones y donde hemos encontrado que su score es más representativo a la hora de generalizar.

Importancia de las variables en el modelo



Análisis de resultados

Histograma de probabilidades dadas por el modelo	Matriz de confusión
	

Conclusiones

Posibles opciones de mejora:

- Nutrir el modelo con **más datos**, como altura, peso pues son factories fáciles de medir que tienen relación con enfermedades cardiacas.
- Analizar casos donde el modelo ha fallado tratando de identificar causa.
- Ponderar probabilidades de distintos modelos, ejemplo cuando un modelo no está muy ‘seguro’ de su predicción (probabilidad cercana al 50%) cotejar con probabilidad de otros modelos...

Es proceso descrito es iterativo donde en muchas ocasiones probamos distintas técnicas para mejorar los resultados, algunas pruebas realizadas:

- Incluir como variable de un modelo las predicciones de otros.
- Estandarizar valores de variables continuas, de forma que hacer que media sea 0 y la desviación 1. Esto dio peores resultados.