Presentación del Proyecto Final

Comision: 29725

Docente: Zurisadai Velazquez

Tutor: Juan Manuel Romero

CODER HOUSE

Felipe Aguirre - Francisco Mansilla

Introducción

En el presente, se expondrá el modelo de análisis en relación a los infartos miocardio arrojados por el dataset Heart Disease.

Las enfermedades cardiacas (EC) representan la principal causa de mortalidad en todo el mundo. Las EC son un grupo de desórdenes que afectan al corazón y a los vasos sanguíneos. La detección temprana es un factor importante para el pronóstico.

CODER HOUSE

A partir del problema y de la exposición a los conceptos vistos en clase, nos surgieron las siguientes preguntas:

- ¿Es posible generar un modelo que clasifique sujetos de alto riesgo vs de bajo riesgo de EC?
- ¿Disponemos de los datos para ello?
- ¿Qué nivel de rendimiento puede alcanzar el modelo?
- ¿Puede resultar útil en la práctica? Es decir, será lo suficientemente preciso, rápido y conveniente para que tenga utilidad clínica?

Objetivo

Hacer uso de las herramientas aprendidas en el curso de DS para diseñar, entrenar, validar y deployar un modelo de machine learning que clasifique pacientes de alto riesgo vs de bajo riesgo de Enfermedades Cardiacas con una precisión mayor a un modelo aleatorio.

CODER HOUSE

Agenda

1. Realizar un proceso apropiado de Data Acquisition
2. Realizar un EDA y una limpieza de los mismos: explorar las dimensiones del dataset, tratamiento de datos nulos (se eliminaron), análisis univariado, bivariado y multivariado.
3. Describir las variables de interés
4. Identificar las variables más correlacionadas con el target.
5. Generar insights respecto a la relevancia clínica de las distintas variables.
6. Entrenar un modelo de clasificación con los datos.
7. Evaluar su rendimiento.
8. Evaluar su potencial utilidad en un contexto clínico real.
9. Optimizar recursos en la prevención de las Enfermedades Cardiacas.

CODER HOUSE

Dataset

El proceso de búsqueda del dataset se realizó en google.
Los criterios de inclusión fueron: tener mínimo 10 columnas y 1000 filas, y poseer una columna target binaria que implique la presentación o no de la EC para cada observación.
Luego de comparar más de un dataset que cumplieron los criterios de inclusión, escogimos el dataset "Heart Disease Dataset".
Sus datos son de 1988 y se compone de cuatro subdatasets provenientes de Cleveland, Hungría, Suiza y Long Beach V.
El resto de sus características serán exploradas y descritas a continuacion

Diccionario de Variables

- **age** Edad de los individuos encuestados. Variable de tipo Numérica. Dominio: de 29 a 77.
- **sex** Genero de los individuos encuestados. Variable de tipo Categórica. Dominio: 1 significa “Male/Masculino”; 0 significa “Female/Femenino”.
- **cp (Chest Pain)** Dolor pectoral causado por la “angina” (dolor pectoral causado por el bajo flujo de sangre al corazón). Variable de tipo Categórica. Dominio: 0 significa “Stable Angina”; 1 significa “Unstable Angina”, 2 significa “Variant Angina”; 3 significa “Refractory Angina”.
- **trestbps (Resting Blood Pressure)** Presion arterial normal, medidas en mm Hg. Variable de tipo Numérica. Dominio: de 94 a 200.
- **chol (Serum Cholesterol)** Nivel de Colesterol total en sangre del individuo, en mg. Variable de tipo Numérica. Dominio: de 126 a 564.

• **fbs (Fasting Blood Sugar)** Nivel de azúcar en sangre mayor a 120 mg. Variable de tipo Categórica. Dominio: 1 significa “True”; 0 significa “False”.

• **restecg (Resting Electrocardiographic Results)** Resultado de Electrocardiograma en Reposo. Variable de tipo Categórica. Dominio: 1 significa “Normal”; 2 significa “Desviación anormal”; 3 significa “Probable o definitiva Hipertrofia Ventricular Izquierda”.

• **thalach (Maximum Heart Rate Achieved)** Maximo ritmo cardiaco alcanzado. Variable de tipo Categorica. Dominio: de 71 a 202.

• **exang (Exercise Induced Angina)** Angina producida al realizar ejercicio. Variable de tipo Categórica. Dominio: 1 significa “True”; 0 significa “False”.

• **oldpeak** Relación entre la demanda de oxígeno y la disminución del flujo sanguíneo, refleja la privación de oxígeno debido al ejercicio. Variable de tipo Numérica. Dominio: de 0.0 a 6.2, significando el valor más alejado a 0.0 una privación mayor de oxígeno.

• **slope** Desvío del ritmo cardíaco al realizar ejercicio. Variable de tipo Categorica. Dominio: 0 significa “Upsloping”; 1 significa “Flat”; 2 significa “Downsloping”.

• **ca** Numero de vasos sanguineos. Tipo de variable Numérica. Dominio: de 0 a 3.

• **thal (Thalassemia)** Enfermedad hereditaria en la que la sangre tiene un número menor de glóbulos rojos a los normales. Variable de tipo Categórica. Dominio: 3 significa “Normal”; 6 significa “Fixed Defect”; 7 significa “Reversable defect”.

• **target** Infarto registrado. Variable de tipo Categorica. Dominio: 1 significa “True”; 0 significa “False”.

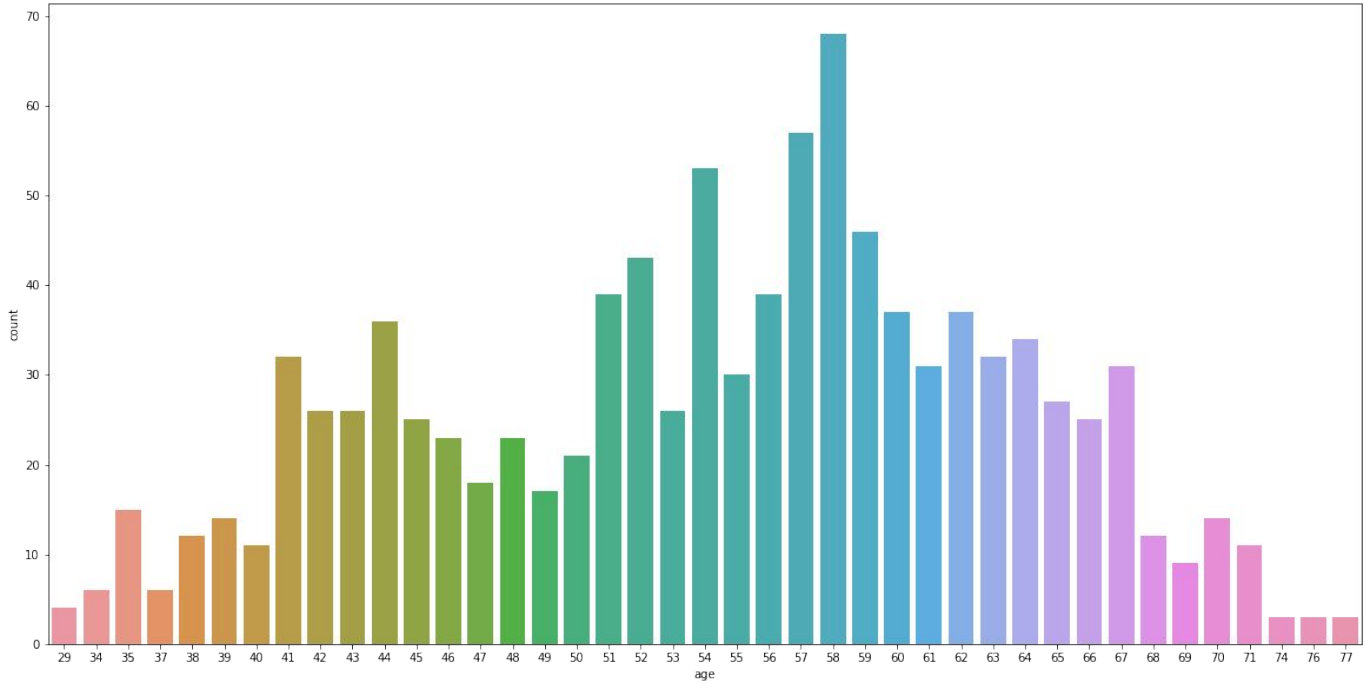
Versionado

En términos de versionado, la única modificación a destacar fue el dropeo del dataset original luego de la Primera Entrega.
A partir de la Segunda entrega, todas las modificaciones pertinentes fueron realizándose sobre el mismo Notebook.

- Primera Entrega: Desarrollo del Data Acquisition, Data Wrangling y Exploratory Data Analysis; Descripción de variables y distribución y relación entre las mismas, con Dataset inicial.
- Segunda Entrega: Modificación de lo realizado en la Primera entrega, en función del nuevo Dataset. Además, preparación de los datos para Data Modelling; elección de los algoritmos candidatos; y evaluación de los indicadores de desempeño.
- Tercer entrega: Revisión de todo el proyecto; e implementación de optimizaciones.

Análisis Univariado

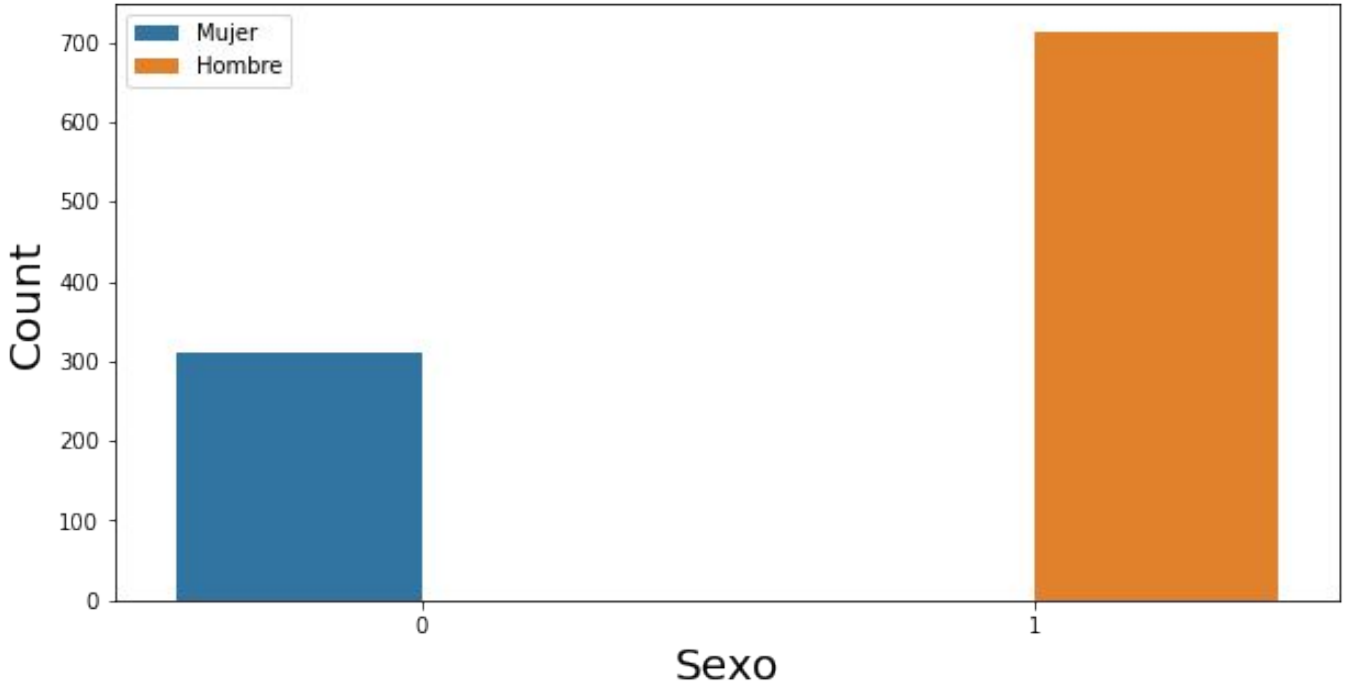
- **Edad:** Distribución aproximadamente normal, con media de 55 años.



CODER HOUSE

Análisis Univariado

- **Sexo:** distribucion desbalanceada, es una limitación del estudio.



CODER HOUSE

Análisis Bivariado

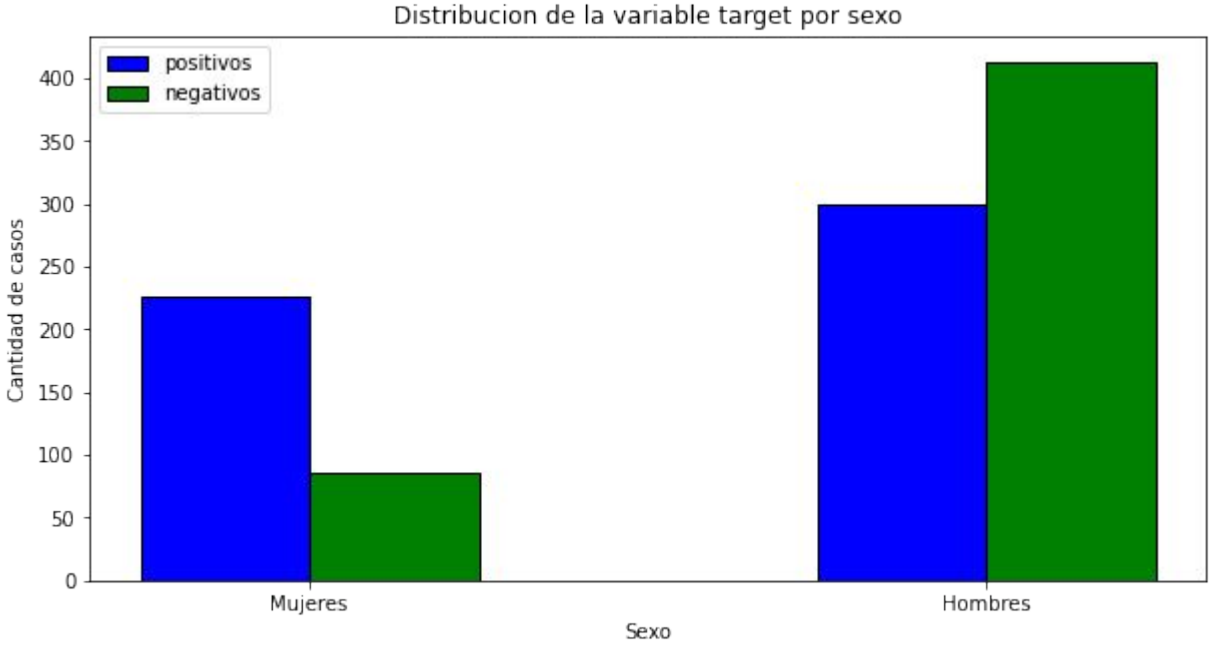
Correlaciones con el target.

- **Maximum Heart Rate Achieved:** r = 0.42.
- **Privación de oxígeno debido al ejercicio:** r = -0.42.
- **Número de vasos sanguíneos:** r = -0.38.

CODER HOUSE

Análisis Bivariado

• **Sexo:** al ser mujer, hay mas chances de tener una EC que de no tenerla. En el caso de los hombres ocurre lo contrario.

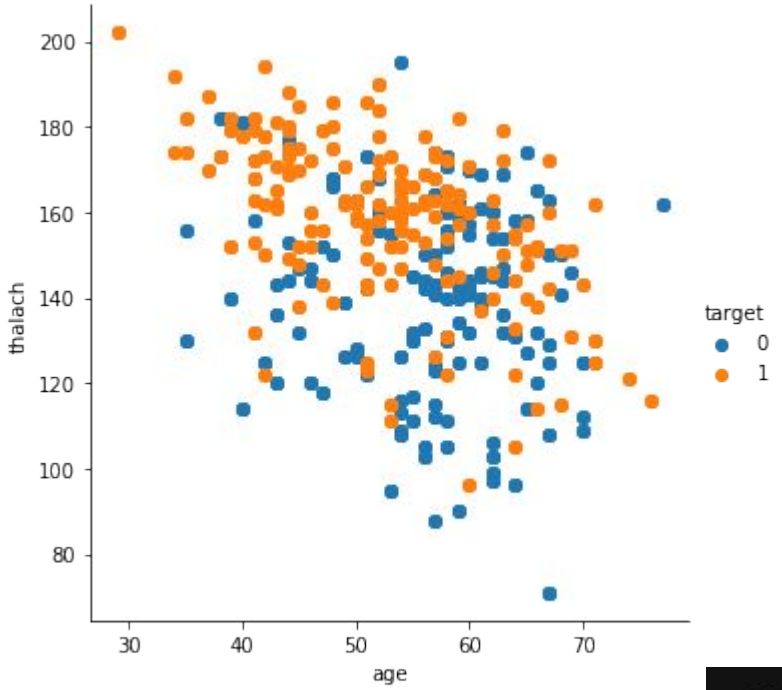


CODER HOUSE

Análisis Multivariado

Clusterización.

- Edad (eje x).
- Máximo ritmo cardíaco alcanzado (eje y).



CODER HOUSE

Modelos

Evaluamos una amplia gama de modelos de Clasificación, tales como Logistic Regression, KNN, y de Bagging y Boosting.

Para los modelos con múltiples hiperparametros, utilizamos grid search y cross validation.

A continuación comparamos sus métricas .

– Random Forest

Accuracy

```
100% de aciertos sobre el set de entrenamiento
100% de aciertos sobre el set de evaluación
```

Precision

```
Precision: 1.0
```

Recall

```
Recall: 1.0
```

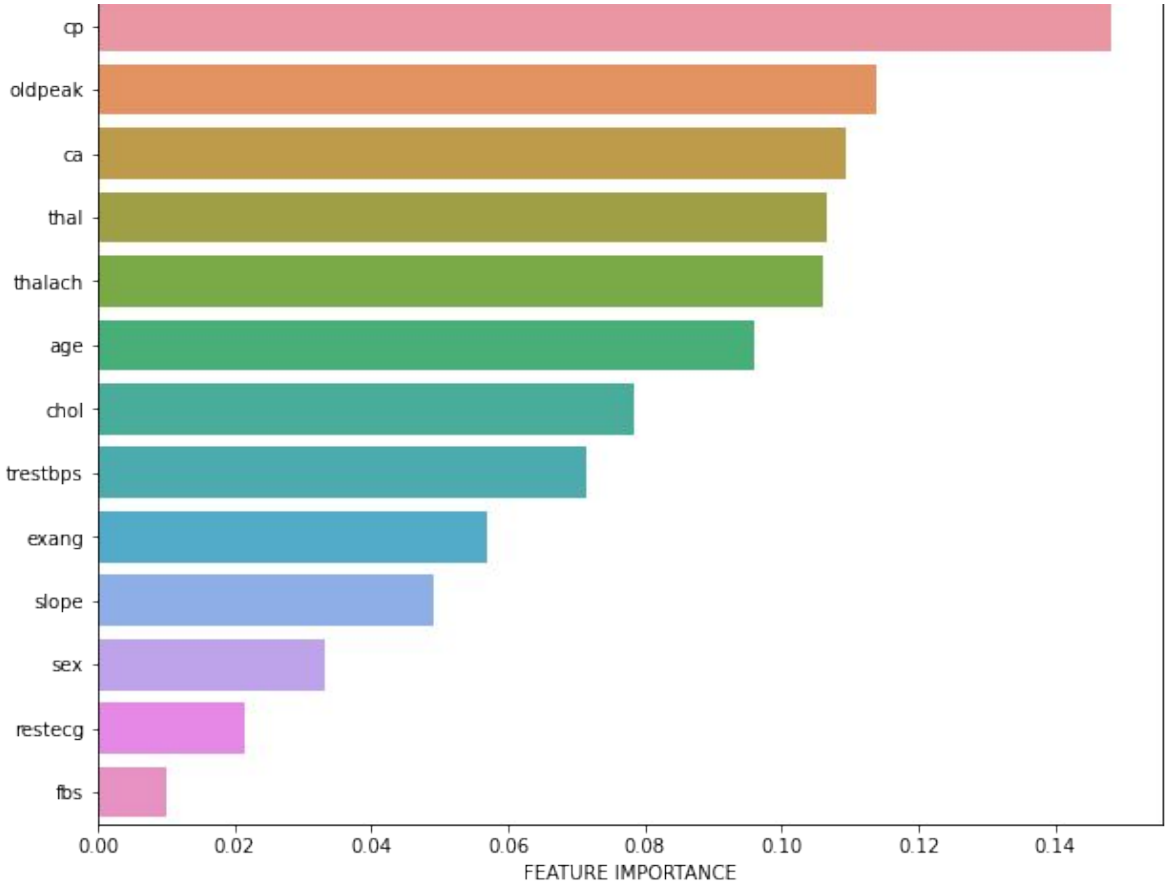
F1 Score

```
F1 score: 1.0
```

Coeficiente AUC

```
AUC score: 1.0
```

Importancia de cada variable para la predicción final del modelo:



CODER HOUSE

– K-Nearest Neighbor

Accuracy

98% de aciertos sobre el set de entrenamiento
87% de aciertos sobre el set de evaluación

Precision

Precision: 0.869281

Recall

Recall: 0.869281

F1 Score

F1 score: 0.869281

Coeficiente AUC

AUC score: 0.870124

– Logistic Regression

Accuracy

```
85% de aciertos sobre el set de entrenamiento
85% de aciertos sobre el set de evaluación
```

Precision

```
Precision: 0.820359
```

Recall

```
Recall: 0.895425
```

F1 Score

```
F1 score: 0.856250
```

Coeficiente AUC

```
AUC score: 0.850938
```

CODER HOUSE

De los modelos simples, el mejor modelo (Random Forest) clasifica correctamente la totalidad de datos de prueba. Podemos esperar que en producción el modelo detecte correctamente casi la totalidad de personas con enfermedades cardiacas así como las personas sin enfermedades cardiacas.

A continuación, probamos modelos más complejos de bagging y boosting.

– Xgboost

Accuracy

100% de aciertos sobre el set de evaluación

Precision

Precision: 1.0

Recall

Recall: 0.986

F1 Score

F1 score: 0.993

Coeficiente AUC

AUC score: 0.993

– Lightgbm

Accuracy

100% de aciertos sobre el set de evaluación

Precision

Precision: 1.0

Recall

Recall: 1.0

F1 Score

F1 score: 1.0

Coeficiente AUC

AUC score: 1.0

– BaggingClassifier

Accuracy

```
100% de aciertos sobre el set de evaluación
```

Precision

```
Precision: 0.986
```

Recall

```
Recall: 0.986
```

F1 Score

```
F1 score: 0.986
```

Coeficiente AUC

```
AUC score: 0.987
```

Múltiples modelos obtuvieron una performance perfecta en todas las métricas de evaluación. De estos modelos, sugerimos utilizar Random Forest, dado que es más simple y tarda menos tiempo en entrenarse.

Conclusión

Asumiendo como premisa que los datos utilizados son representativos de la realidad, es posible generar un modelo predictivo de riesgo de enfermedad cardíaca que sea eficaz y eficiente. A su vez, este análisis provee información relevante sobre las variables asociadas a la posibilidad de sufrir estas enfermedades. La utilización de este modelo en la clínica es en principio posible, pero depende de la recolección de los datos (algunas variables pueden tomar demasiado tiempo/recursos en calcularse).

Limitaciones

Como limitaciones, la población femenina está subrepresentada en el dataset, la fecha del mismo es antigua, y contiene pocas observaciones.

Futuras líneas

Futuras líneas de investigación deberían abordar las limitaciones de la presente: utilizar (o crear) un dataset con más observaciones y en el cual no existan diferencias de género. También, se sugiere el trabajo interdisciplinario con personal de salud para la recolección de variables especialmente relevantes.
Estas investigaciones podrían encontrar mejores resultados al utilizar modelos más complejos, por lo que recomendamos recorrer nuevamente el presente pipeline.