

Algoritmo de Predicción de Enfermedades Cardíacas

Data Project 3

DataKadabra Team





indice

01

- Dataset Description y
Analisis de variables

03

- Data Preparation
 - Categoricas a dummies
 - Undersampling

02

- Data Cleaning
 - Imputación de valores faltantes por variable
 - KNN imputer

04

- Entrenamiento del modelo
 - Benchmark de modelos
 - Backward selection
 - Bucle de 1k iteraciones
 - Métricas de performance y evaluación



Objetivo

“Desarrollar un modelo de aprendizaje automático de inteligencia artificial capaz de predecir la presencia de enfermedades cardíacas”



Dataset description

VARIABLE	TIPO	TIPO DE DATO	DESCRIPCIÓN
Age	cuantitativa continua	float	edad del paciente
Sex	categórica nominal	float	género del paciente (1 = hombre; 0 = mujer)
CP	categórica nominal	float	tipo de dolor de pecho (1 = angina típica; 2 = angina atípica; 3 = dolor no-anginoso; 4 = asintomático)
Trestbps	cuantitativa continua	float	presión arterial en reposo (en mm Hg al ingreso en el hospital)
Chol	cuantitativa continua	float	colesterol sérico en mg/dl
Fbs	categórica nominal	float	dolor provocado por el esfuerzo (1 = sí; 0 = no)
Restecg	categórica nominal	float	resultados electrocardiográficos en reposo (0 = normal; 1 = presenta anormalidad de la onda ST-T; 2 = presenta probable o definida hipertrofia ventricular izquierda)
Thalach	cuantitativa continua	float	frecuencia cardiaca en reposo
Exang	categórica nominativa	float	angina inducida por el ejercicio (1 = sí; 0 = no)
Oldpeak	cuantitativa continua	float	depresión del ST inducida por el ejercicio en relación con el reposo
Slope	categórica nominativa	float	la pendiente del segmento ST en ejercicio máximo (1 = pendiente ascendente; 2 = plano; 3 = pendiente descendente)
CA	categórica ordinal	float	número de vasos mayores (0-3) coloreados por flouroscoopia
Thal	categórica nominal	float	(3 = normal; 6 = defecto fijo; 7 = defecto reversible)

-El dataset tiene **732 registros y 14 variables**.

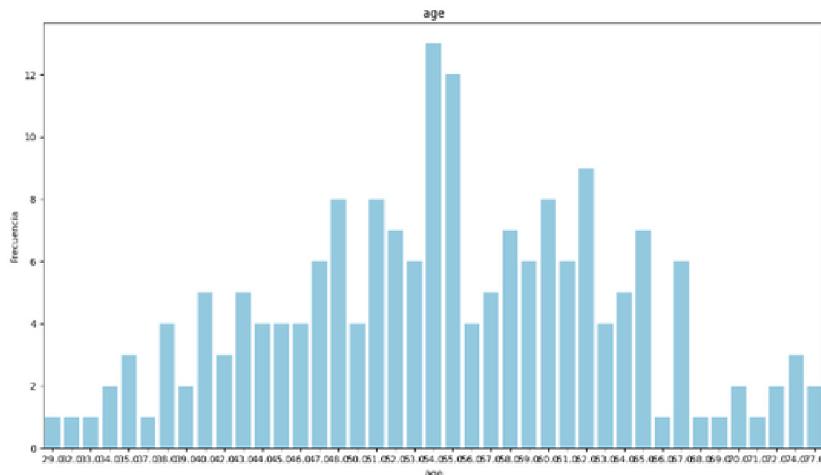
-Las variables representan diferentes **parámetros clínicos** de cada paciente.

-La columna '**label**' representa la variable **objetivo**, que indica la presencia de patología cardíaca en diferentes grados de 0 a 4.

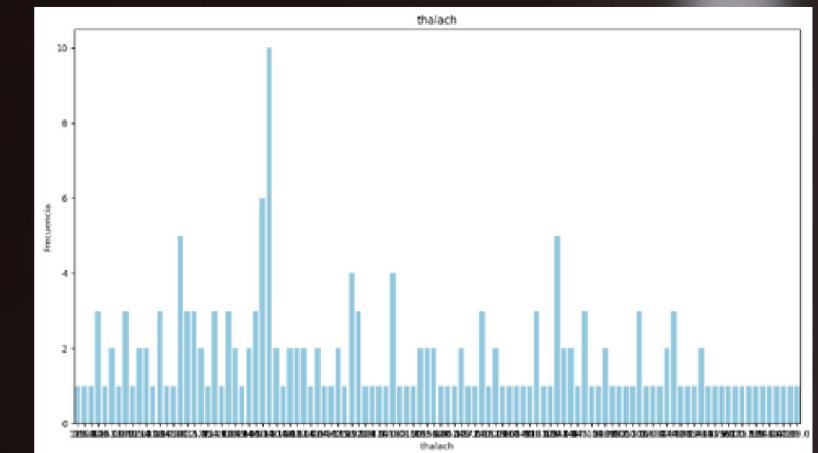
-Se observa que hay **gran cantidad de valores missing**.



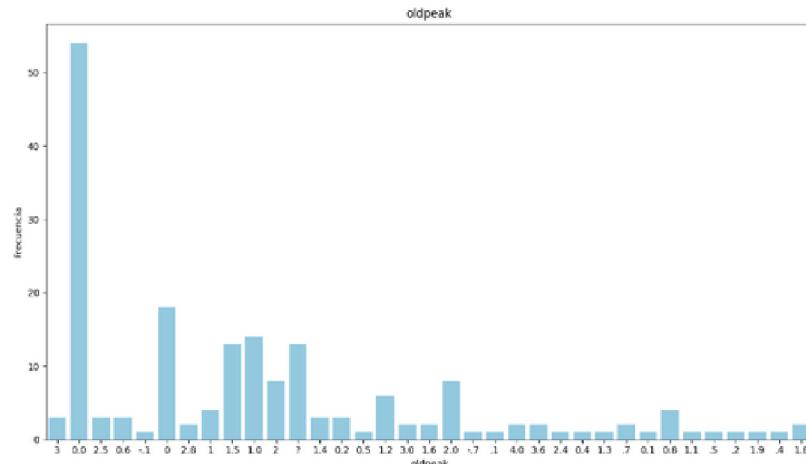
Dataset description



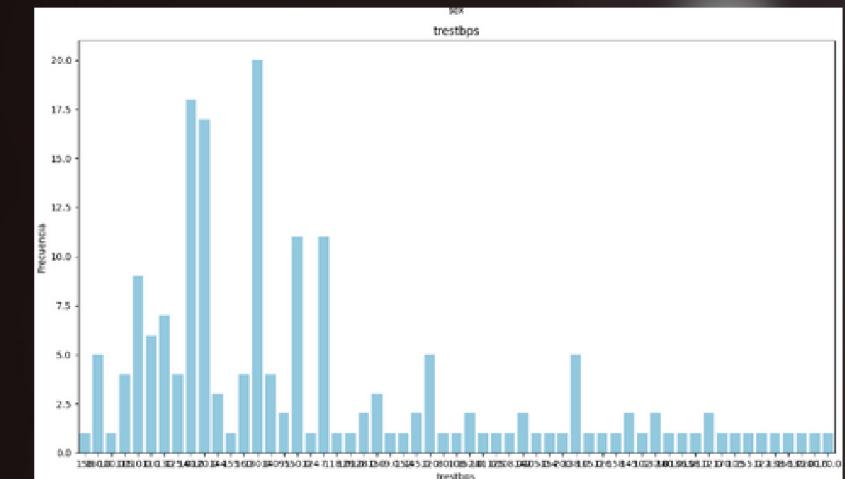
Las patologías cardíacas son más comunes en **personas mayores de 55 años**.



La frecuencia cardíaca elevada puede ser un **factor de riesgo para las enfermedades cardíacas**.



Hay una **mayor proporción de pacientes con depresión del segmento ST inducida por el ejercicio** que de pacientes con depresión del segmento ST inducida por el ejercicio baja.



La mayoría de los pacientes no presenta **dolor provocado por el esfuerzo**.



Data Cleaning

Dataproject 3 - DataKadabra Team

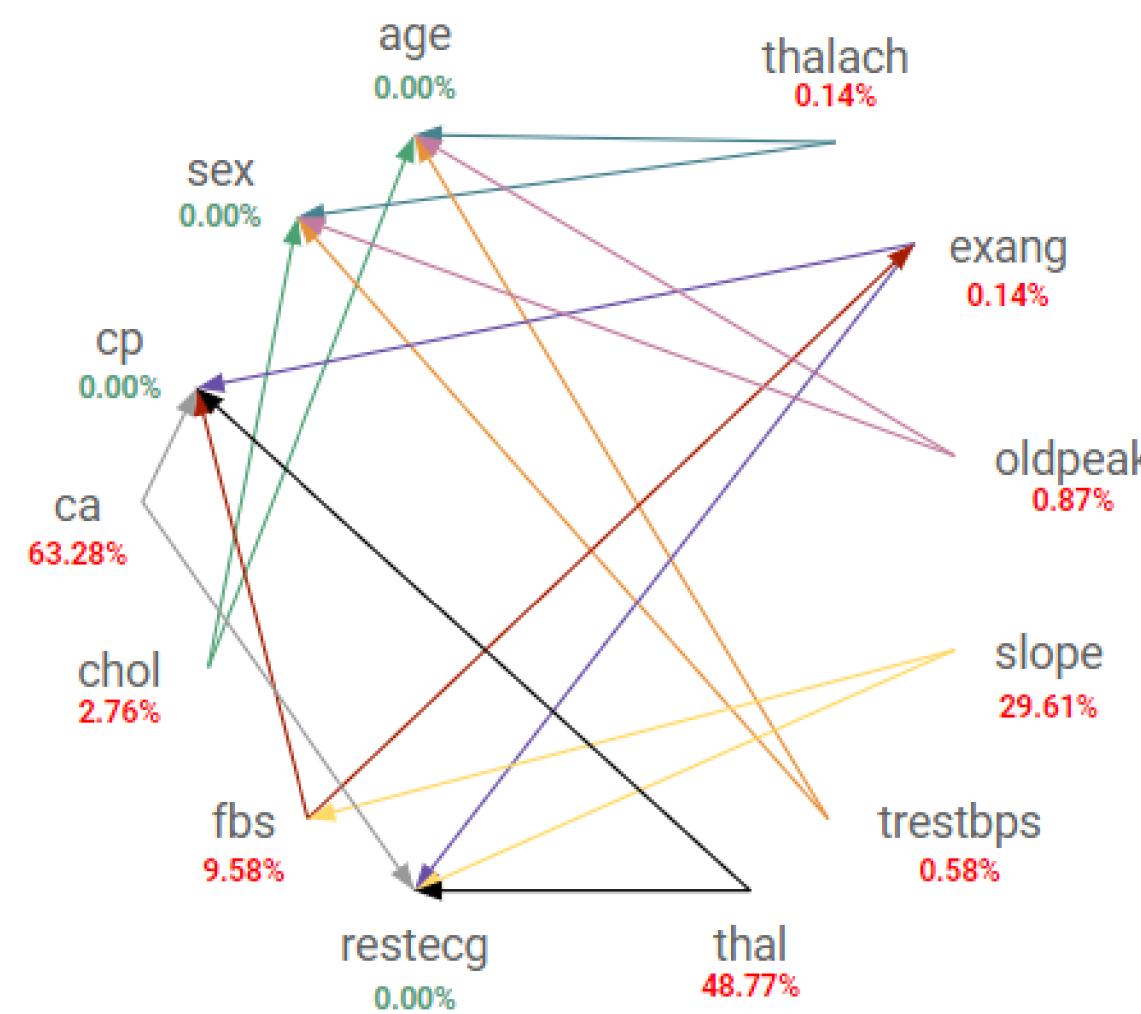
- 01**
 - Reemplazo de **caracteres especiales** y **atípicos** por NaN.
- 02**
 - Definición de los **DataTypes**.
- 03**
 - Eliminació **53 casos** con 7 o más variables sin registros.
- 04**
 - **Imputación de outliers en variables continuas** que estén por encima de 1,5 desviaciones estándar.





Preprocesamiento de datos

- **Imputación de valores faltantes por variable:** Agrupación de variables clínicamente relacionadas para encontrar el valor más frecuente.



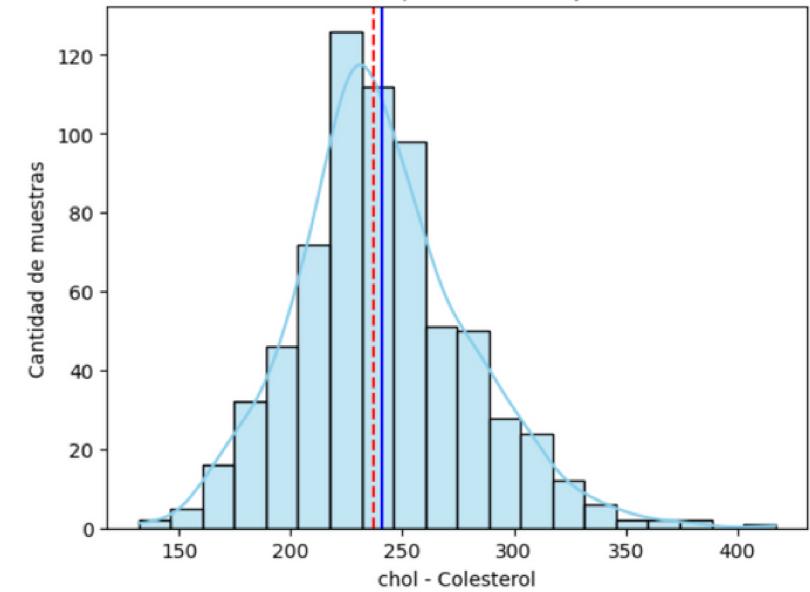
- **Imputación de Valores Faltantes KNN (K-Nearest Neighbors):** Aplicación del algoritmo de vecinos más cercanos para la imputación de valores missing remanentes que no pudieron ser imputados mediante la moda del groupby de variables clínicamente relacionadas. El KNN imputa los valores faltantes utilizando los valores de las observaciones más similares en función de otras variables. Calcula los vecinos más cercanos y usar sus valores para imputar los valores faltantes.



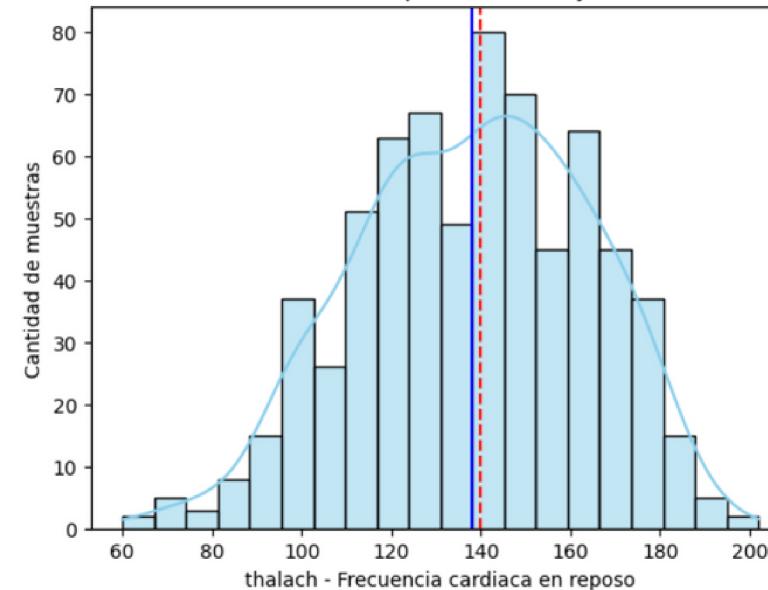
Preprocesamiento de datos

*Promedio de valores para cada **combinación de edad y sexo..**

Distribución de la columna "chol" después del manejo de valores faltantes y outliers

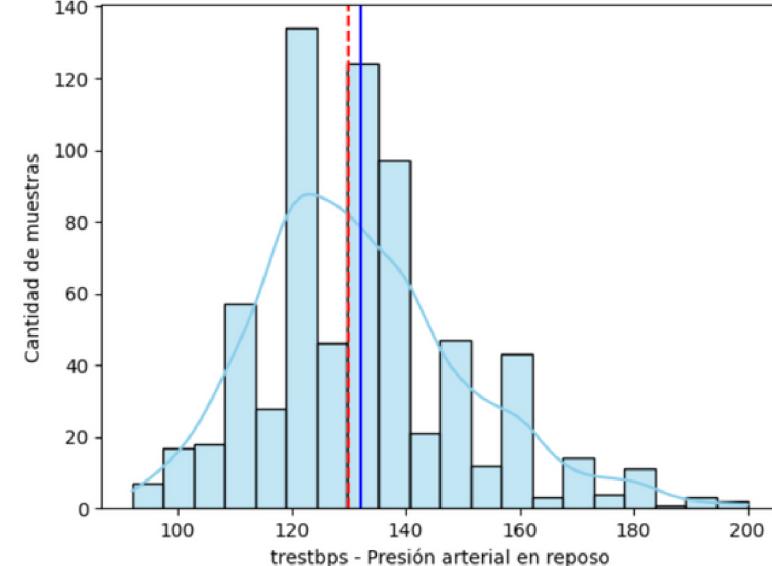


Distribución de la columna "thalach" después del manejo de valores faltantes y outliers

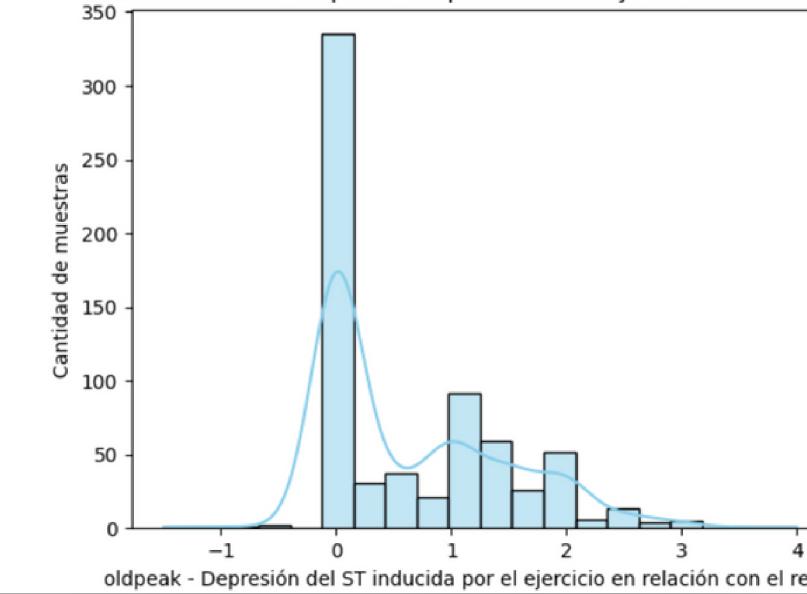


*Imputación y **manejo de outliers** en las variables chol, thalach, trestbps, oldpeak

Distribución de la columna "trestbps" después del manejo de valores faltantes y outliers



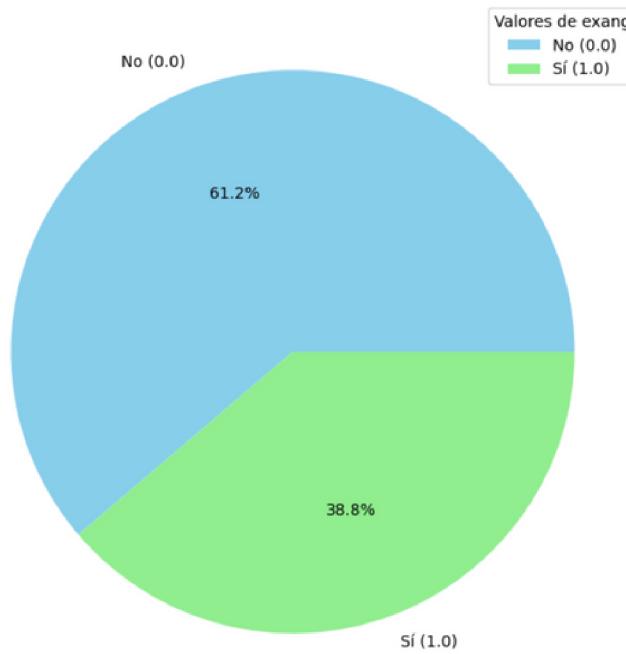
Distribución de la columna "oldpeak" después del manejo de valores faltantes y outliers



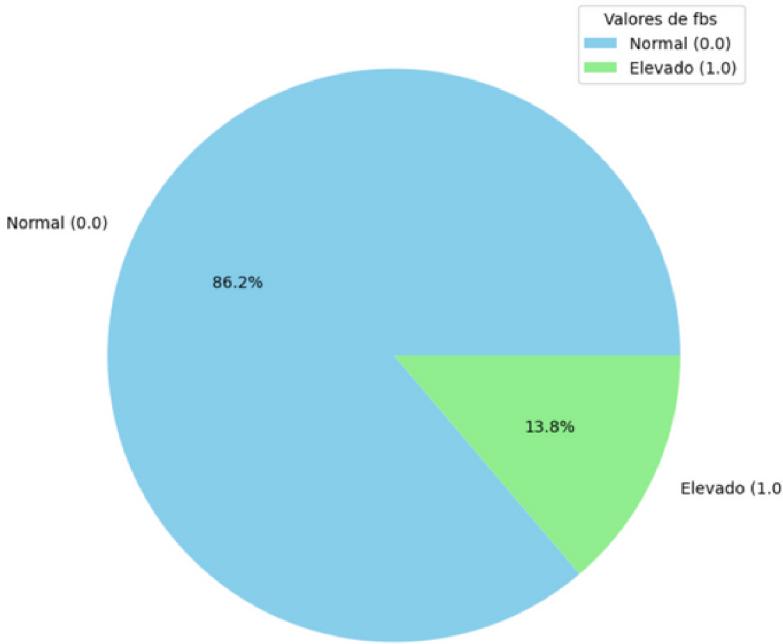


Preprocesamiento de datos

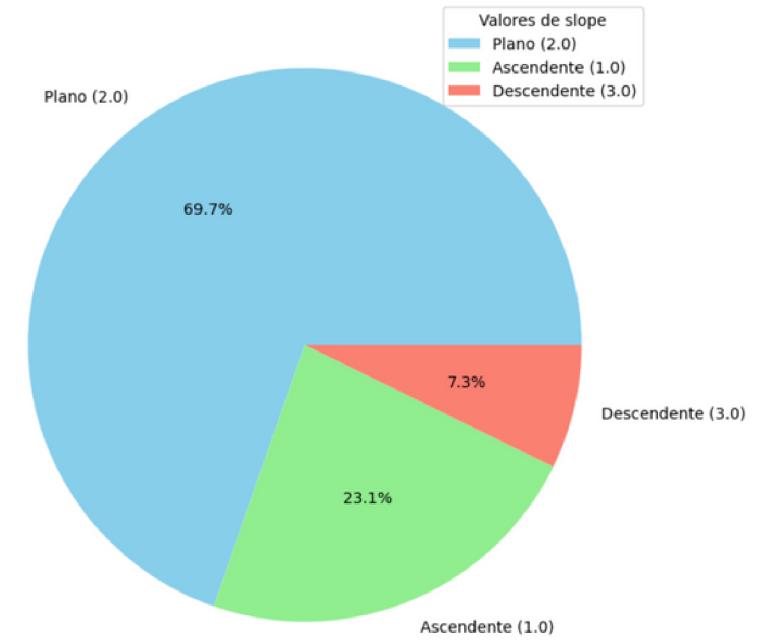
Distribución de valores en la columna "exang"



Distribución de valores en la columna "fbs"



Distribución de valores en la columna "slope"



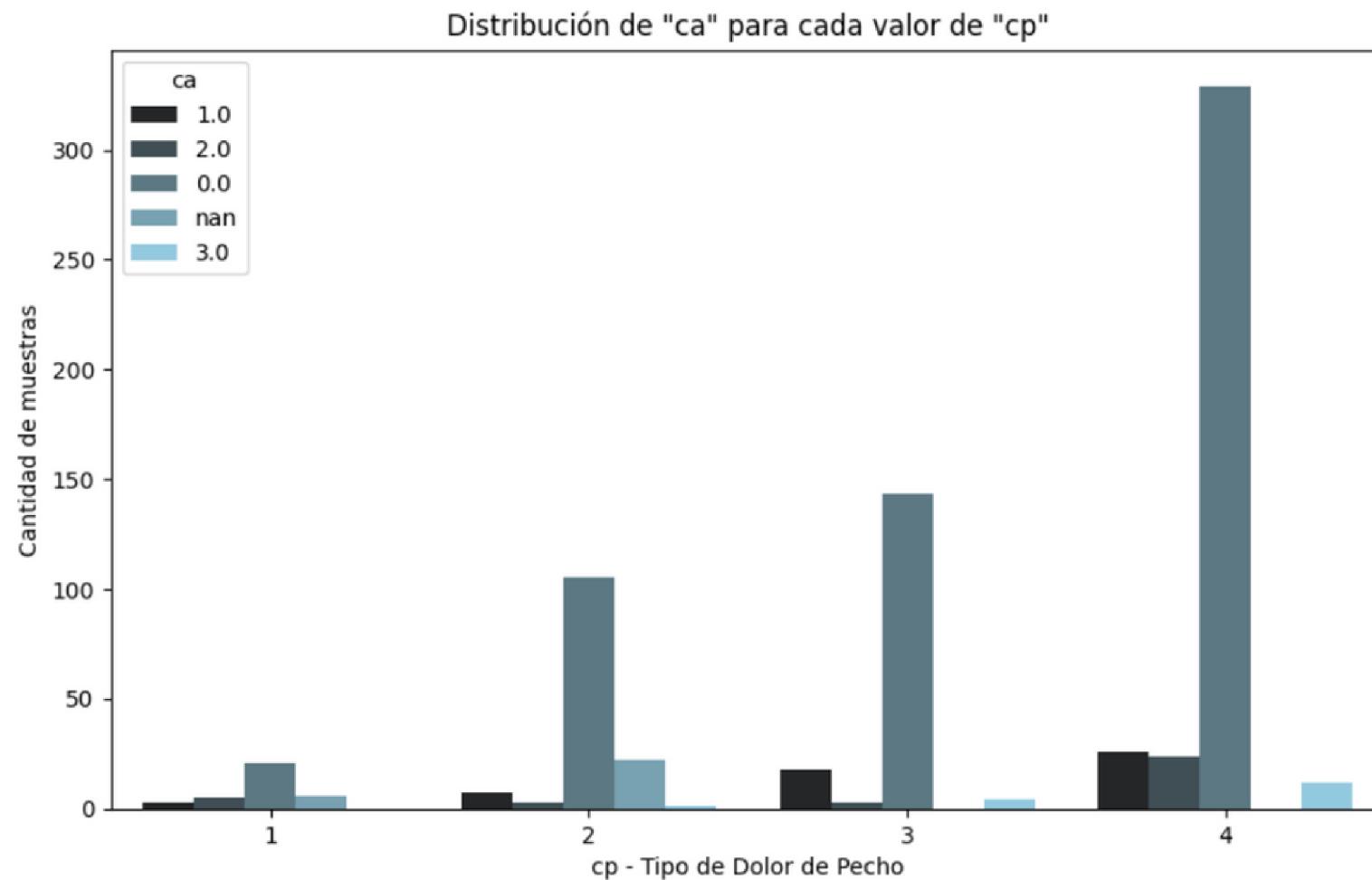
Imputación y el **manejo de outliers** en las variables exang, fbs, slope

Moda de valores para cada combinación de fbs y resteng

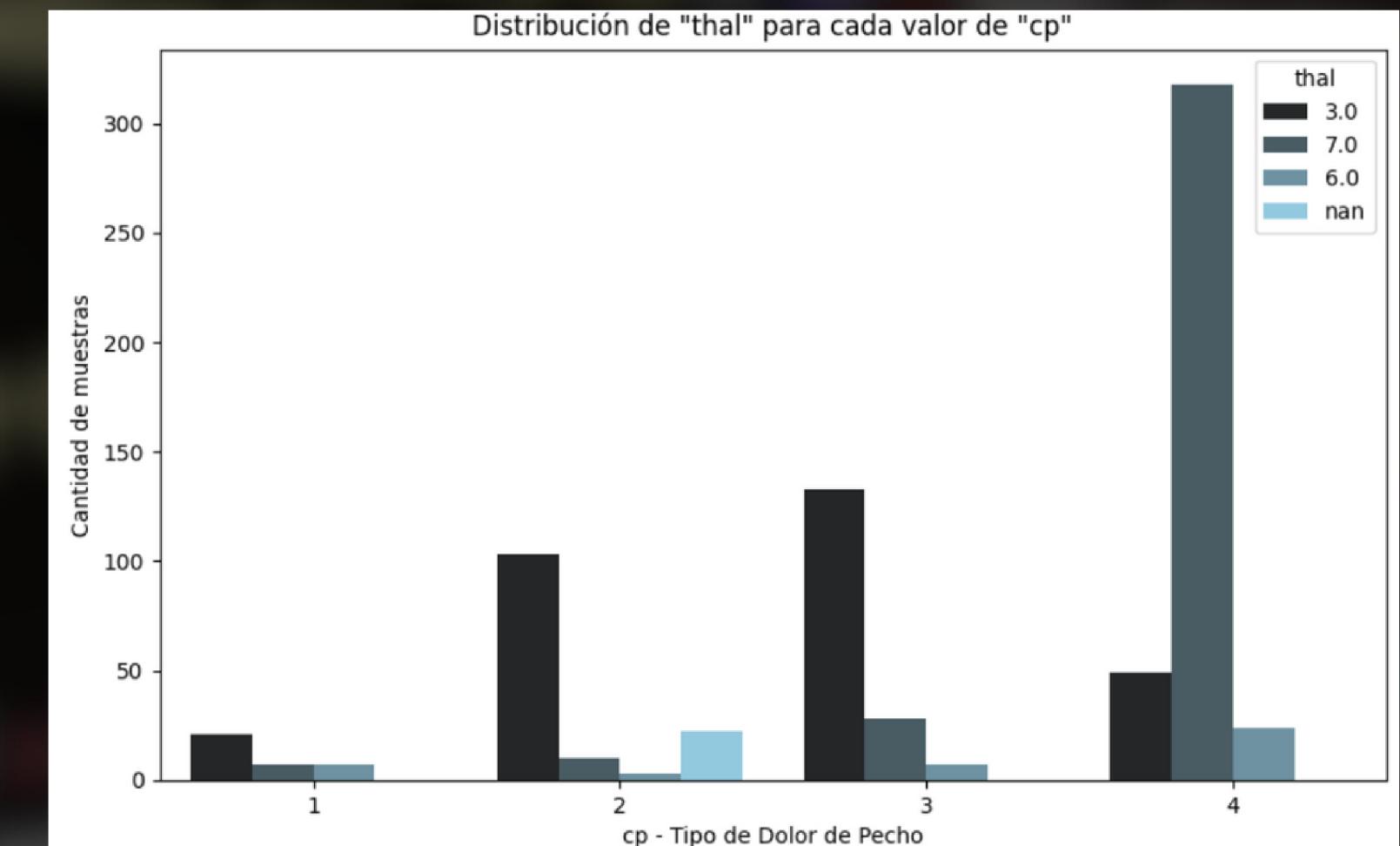
Moda de valores para cada combinación de cp y restecg



Preprocesamiento de datos



Para lograr obtener los valores faltantes de "ca" (número de vasos coloreados), hemos creído conveniente usar las variables "cp" (tipo de dolor de pecho) y "restecg" (electrocardiograma en reposo), ya que creemos que guardan relación con "ca".



A la hora de completar los valores inválidos de "thal" (valor "3" si el paciente está bien, valor "6" tiene un defecto fijo o valor "7" si es uno reversible), hemos acudido a las variables "cp" (tipo de dolor de pecho) y "restecg" (electrocardiograma en reposo) ya que creemos que guardan relación y por lo tanto nos pueden ayudar a completar mejor sus valores faltantes.



Data preparation

Transformación de variables

Para mejorar la interpretación y eficacia del modelo, se decidió transformar las variables categóricas, como 'cp', 'restecg', 'ca', 'slope' y 'thal', en **variables dummy**. Este enfoque permite tener en cuenta de manera más precisa la influencia de cada categoría en la variable objetivo, lo que garantiza un uso más adecuado de los datos durante el entrenamiento del modelo. También se elimina la ultima categoría de cada variable para **evitar el efecto de co-linealidad**.

fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
0	2	125.0	1	1.40	1	1	3.0
0	0	150.0	1	1.50	2	0	7.0
0	2	149.0	0	2.00	1	0	7.0
0	0	140.0	0	0.00	2	0	3.0



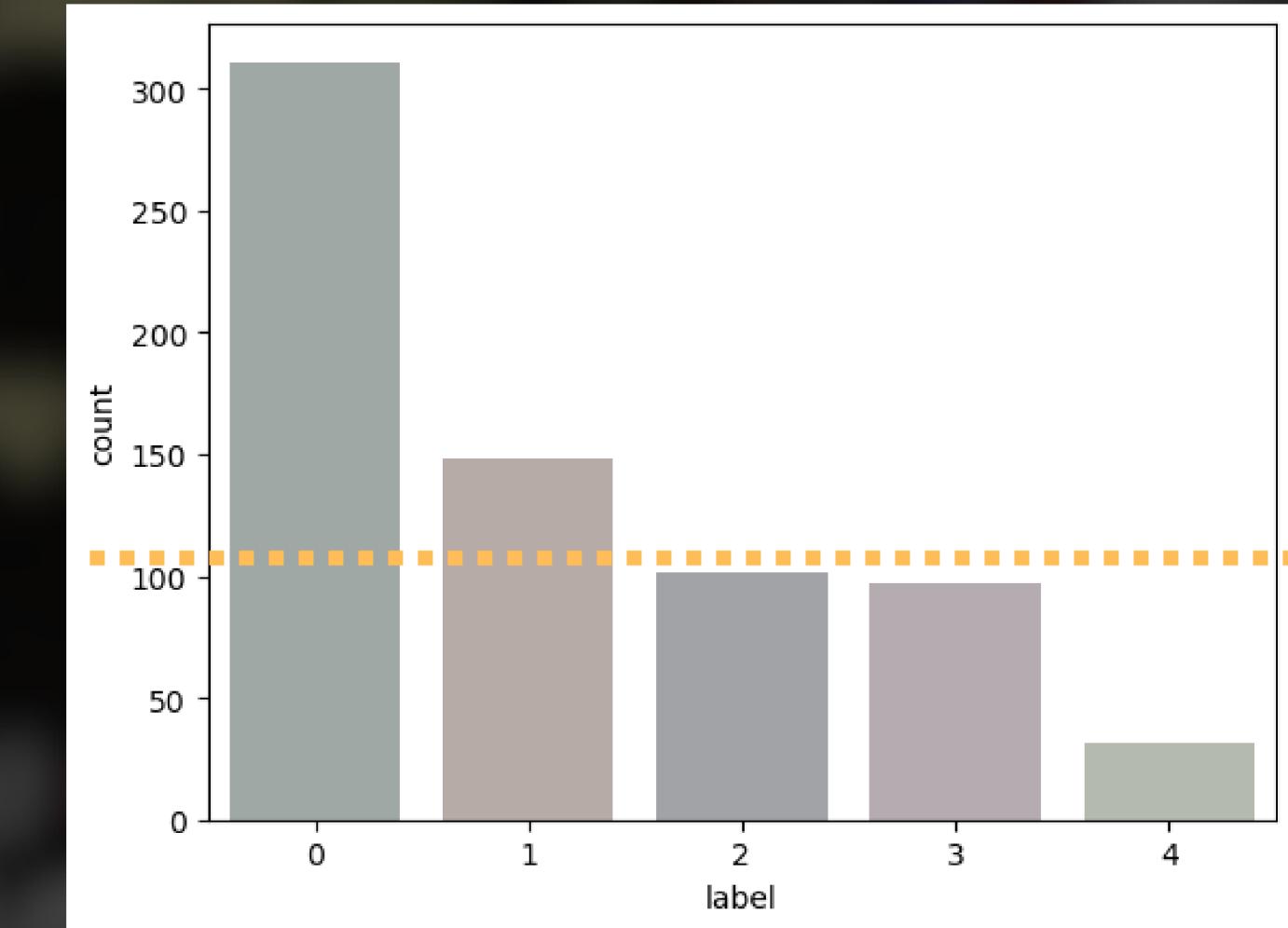
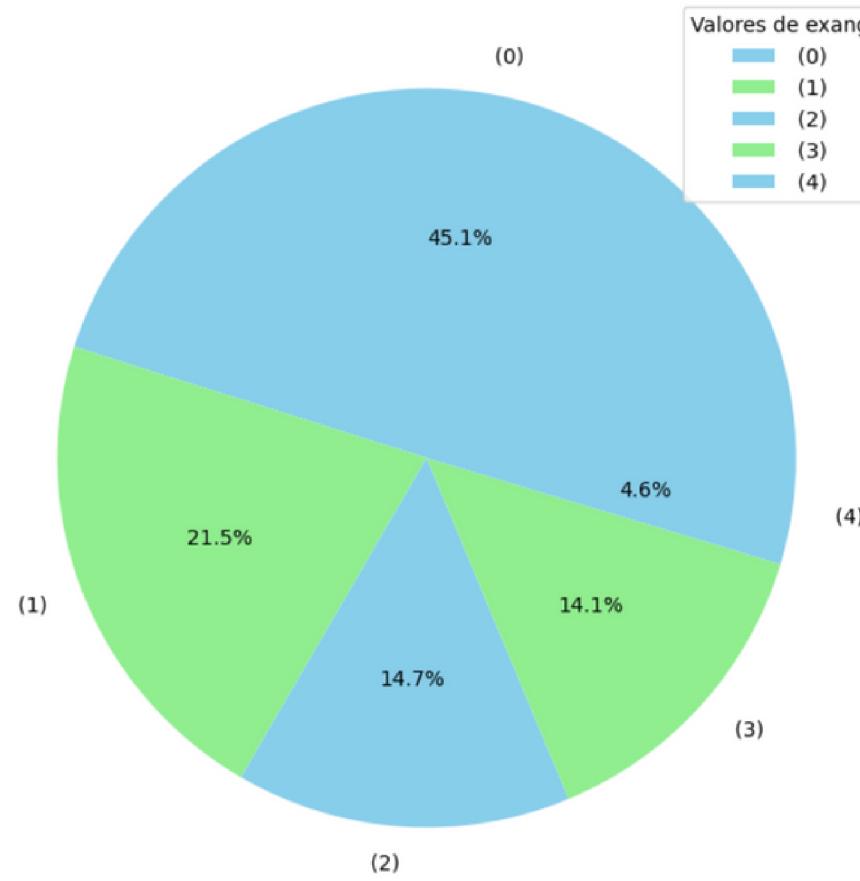
bs	thalach	exang	oldpeak	cp_1	cp_2	cp_3	restecg_0	restecg_1	ca_0	ca_1	ca_2	slope_1	slope_2	thal_3	thal_6	
1	120.0	1	0.0	0	1	0	1	0	1	0	0	0	0	1	0	0
0	135.0	0	0.0	0	0	1	0	0	0	1	0	0	0	1	1	0
0	165.0	0	1.6	0	0	1	0	0	0	1	0	0	1	0	0	0
0	162.0	0	0.0	0	1	0	1	0	0	1	0	0	1	0	1	0
0	179.0	0	0.0	0	1	0	1	0	0	0	0	1	1	0	1	0



Data preparation

Resampling de clases

Distribución de valores en la columna "label"



Utilizamos métodos de submuestreo para reducir el número de muestras en la clase mayoritaria y **equilibrar las proporciones de las clases** en el conjunto de datos. Realizamos un muestreo estratificado para garantizar que cada clase tenga un número similar de muestras.

*Reemplazamos los valores de 'label' que son 4 por 3



Modelado

Benchmark de modelos: utilizamos métricas como la precisión, la sensibilidad y el error cuadrático medio para evaluar la eficacia de cada modelo y poder seleccionar el de mejor performance.

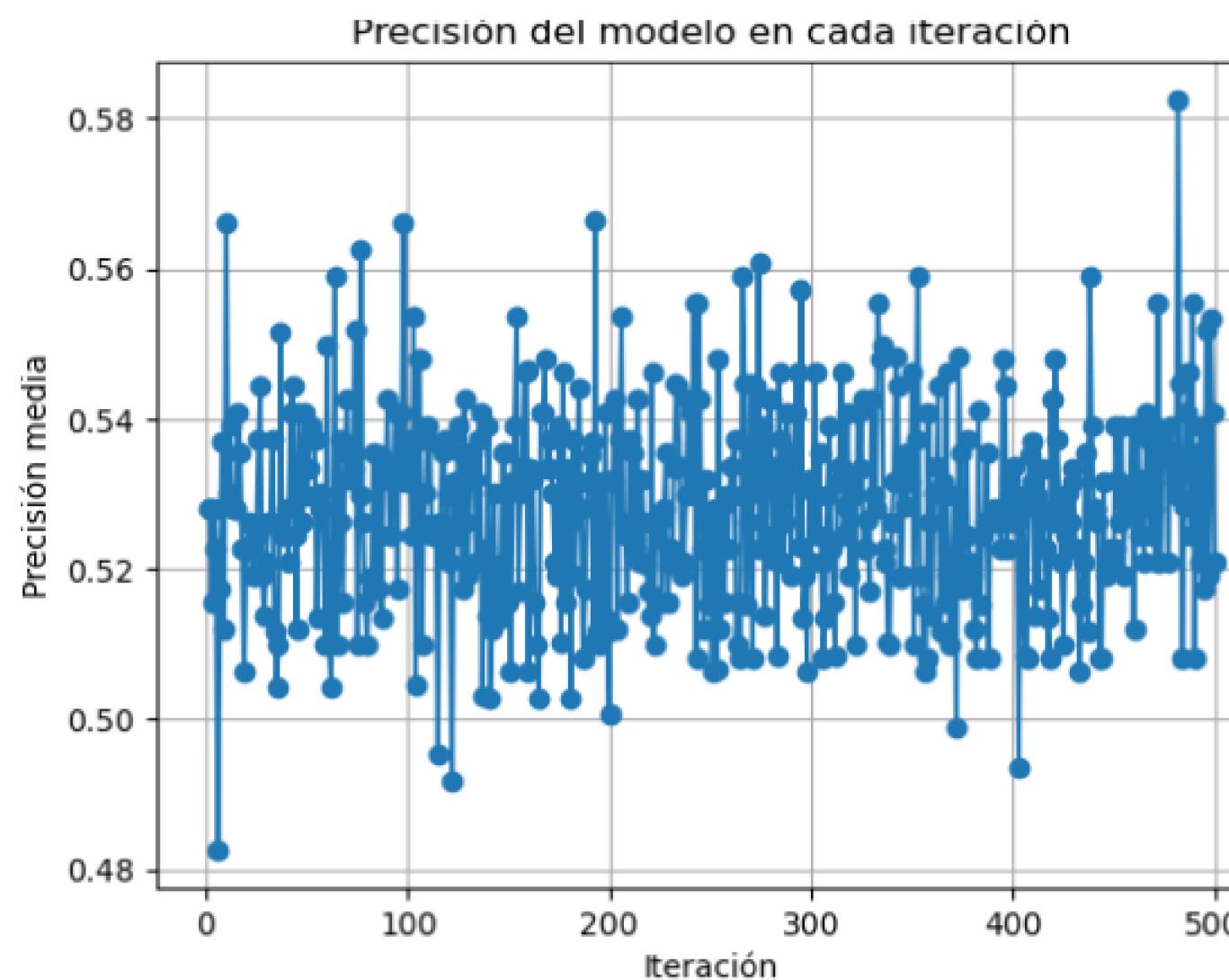
```
Model: LinearRegression, MSE: 0.8531479316398278, R2: 0.3730039281384294
Model: Ridge, MSE: 0.8519785276794326, R2: 0.3738633473111136
Model: Lasso, MSE: 1.0987843346761492, R2: 0.19248065181288998
Model: DecisionTreeRegressor, MSE: 2.4855072463768115, R2: -0.8266507158569056
Model: RandomForestRegressor, MSE: 0.9863478260869564, R2: 0.2751125689808206
Model: LogisticRegression, MSE: 1.3985507246376812, R2: -0.02782387218770488
```

Backward elimination: implementamos una técnica de selección de características que iterativamente realiza entrenamientos del modelo con todas las combinaciones posibles de variables, seleccionando la combinación con mejor performance, maximizando así su rendimiento.

```
Mejor conjunto de características: ['trestbps', 'thalach', 'exang', 'oldpeak', 'thal_filled', 'cp_1', 'cp_2', 'cp_3']
Precisión media del modelo con el mejor conjunto de características: 0.5534479934479934
```



Modelado



Bucle de 1k iteraciones: validación cruzada k-fold en el modelo de regresión logística utilizando un pipeline que incluye el escalado de características. Itera 1000 veces dividiendo los datos en conjuntos de entrenamiento y prueba, ajustando el modelo y calculando la precisión media en cada iteración dando una estimación del rendimiento.



Modelado

Métricas de performance y evaluación: Matriz de confusión y scores.

Matriz de Confusión					
Real	0	1	2	3	4
0	51	6	2	1	0
1	11	16	1	2	0
2	5	8	8	5	0
3	6	6	3	4	0
4	1	0	1	1	0
	0	1	2	3	4

F1 Score: 0.5475762825802643

Reporte de Clasificación Regresion Logistica TRAIN:

	precision	recall	f1-score	support
0	0.0	0.67	0.69	0.68
1	1.0	0.47	0.45	0.46
2	2.0	0.27	0.33	0.30
3	3.0	0.65	0.57	0.60
accuracy			0.54	81
macro avg	0.51	0.51	0.51	81
weighted avg	0.56	0.54	0.55	81

Precision TEST Clasificación:
0.5432098765432098

Gracias! ❤



Lucía Esteve



Cristian Marty



Pablo Pérez



Stas Korotchenko