

# Stroke Risk Classifier

## Machine Learning Classification Algorithm Selection

*Eduardo Solano Jaime*

0213663

*ECID*

*Universidad Panamericana campus Guadalajara*

### Objetivo

El objetivo de este reporte es utilizar el dataset de [Heart Attack Analysis & Prediction Dataset](#) para entrenar un modelo de algoritmo de clasificación de machine learning que prediga si una persona está en riesgo de tener un ataque cardíaco o no.

El dataset contiene información sobre 303 pacientes, incluyendo su edad, sexo, tipo de dolor de pecho, presión arterial en reposo, colesterol, frecuencia cardíaca, índice de masa corporal, glucosa en sangre, tabaquismo, consumo de alcohol, actividad física, entre otros. La variable objetivo es si el paciente testó en riesgo de un paro cardíaco.

Para entrenar el modelo, se utilizó regresión logística, no sin su previo análisis y procesamiento de las variables. Se sigue el proceso de visualización y explicación de las variables, usando métodos de transformación y métodos estadísticos para modificar las distribuciones. Finalmente se califica el modelo con métodos de validación.

Los resultados del modelo se analizarán para determinar su utilidad para la predicción de ataques cardíacos.

El cuaderno de Jupyter completo dónde se realizaron todos los cálculos/modificaciones se encuentra en mi [repositorio de GitHub](#) para mayor comprensión.

### Metodología

Los pasos a seguir para completar este reporte son los siguientes:

- Exploración del dataset para entender el significado de información de las bases de datos.
- Visualización de los datos mediante varias representaciones
- Preparación de los datos para el entrenamiento del modelo, incluyendo la transformación de las variables.
- Entrenamiento del modelo.
- Evaluación del modelo usando métodos de validación (técnicas de remuestreo).
- Análisis de los resultados para determinar su utilidad para la predicción de ataques cardíacos.

### Análisis de variables

#### Exploración del dataset

Al importar el dataset se utilizó la función `DataFrame.head()`, la cual se utiliza para devolver las primeras n filas de un objeto DataFrame. El valor predeterminado de n es 5.

El dataset se importó previamente y se declaró con el nombre '`df`', de tipo `DataFrame`.

```
df.head(5)
```

	age	sex	cp	trtbps	chol	fbp	restecg	thalachh	exng	oldpeak	sip	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Antes de continuar con la actividad, se checa la existencia de valores nulos o no-numéricos en el dataset. Por lo que se usó la función `Dataframe.info()` para poder comparar la cantidad total de entradas por variable que no sean nulas o no-numéricas contra la cantidad total de entradas. En este caso los número coinciden, lo que comprueba la no-existencia de estas variables. De igual manera se usa la función para evaluar el tipo de dato que se presenta en el dataset. Asimismo, la función provee el dato del tamaño del dataset, que en este caso es de **(303, 14)**.

Con dicha información y la documentación se puede realizar una explicación profunda de las características de los atributos para su futuro procesamiento y comprensión.

Atributo	Descripción	Tipo
age	Edad en años	Contínua / int64

Atributo	Descripción	Tipo
sex	Sexo (1 = masculino; 0 = femenino)	Categórica / int64
cp	Tipo de dolor torácico (1 = angina típica; 2 = angina atípica; 3 = dolor no anginoso; 0 = asintomático)	Categórica / int64
trestbps	Presión arterial en reposo (en mm Hg al ingresar al hospital)	Continua / int64
chol	Colesterol sérico en mg/dl	Continua / int64
fbs	Azúcar en sangre en ayunas > 120 mg/dl (1 = verdadero; 0 = falso)	Categórica / int64
restecg	Resultados del electrocardiograma en reposo (1 = normal; 2 = con anormalidad de la onda ST-T; 0 = hipertrofia)	Categórica / int64
thalach	Frecuencia cardíaca máxima alcanzada	Continua / int64
exang	Angina inducida por el ejercicio (1 = sí; 0 = no)	Categórica / int64
oldpeak	Depresión del ST inducida por el ejercicio en relación con el reposo	Continua / float
slope	Pendiente del segmento ST del ejercicio máximo (2 = ascendente; 1 = plana; 0 = descendente)	Categórica / int64
ca	Número de vasos principales (0-3) coloreados por flourosopia	Categórica / int64
thal	Talio (2 = normal; 1 = defecto fijo; 3 = defecto reversible)	Categórica / int64
num	Atributo predicho	Categórica / int64

Finalmente usamos la función `DataFrame.describe()` para poder ver los descriptores estadísticos de los datos.

	age	sex	cp	trestbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	ou
count	303	303	303	303	303	303	303	303	303	303	303	303	303	303
mean	54.3663	0.683168	0.966997	131.624	246.264	0.148515	0.528053	149.647	0.326733	1.0396	1.39934	0.729373	2.31353	0.544
std	9.0821	0.466011	1.03205	17.5381	51.8308	0.356198	0.52586	22.9052	0.469794	1.16108	0.616226	1.02261	0.612277	0.498
min	29	0	0	94	126	0	0	71	0	0	0	0	0	0
25%	47.5	0	0	120	211	0	0	133.5	0	0	1	0	0	2
50%	55	1	1	130	240	0	1	153	0	0.8	1	0	0	2
75%	61	1	2	140	274.5	0	1	166	1	1.6	2	1	1	3
max	77	1	3	200	564	1	2	202	1	6.2	2	4	4	3

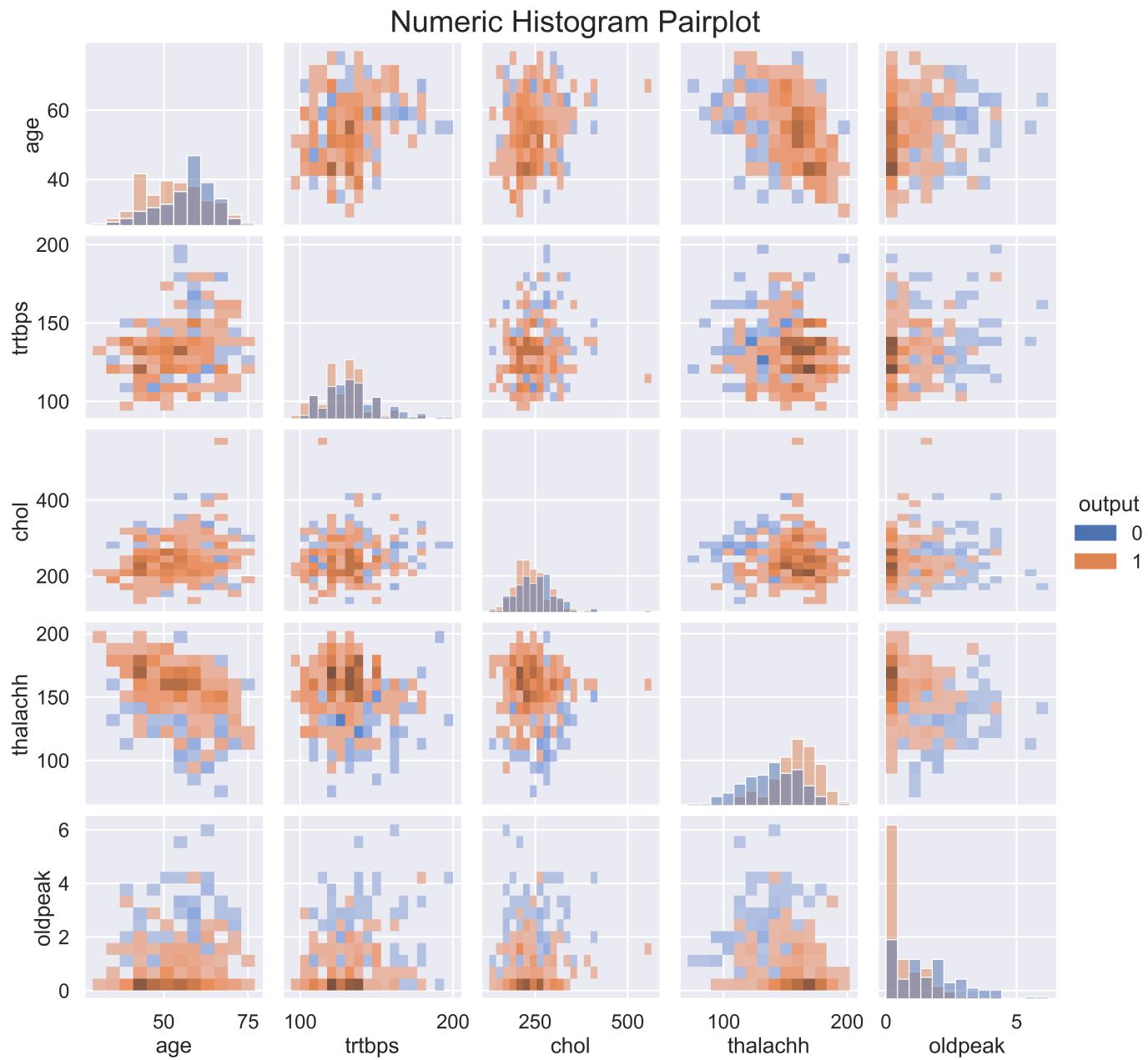
La tabla proporciona los valores de **conteo, media, desviación estándar, mínimo, percentil 25, percentil 50, percentil 75 y máximo** para cada columna. A continuación se presenta una interpretación estadística de cada columna:

- **Edad:** La edad promedio de los pacientes en el conjunto de datos es de 54.4 años. La edad más joven registrada es de 29 años y la edad más avanzada es de 77 años.
- **Sexo:** El 68.3% de los pacientes son hombres y el 31.7% son mujeres.
- **Cp:** El tipo de dolor en el pecho que experimentan los pacientes se divide en 47.5% de los pacientes con tipo 0, el 40.3% tipo 1, el 9.6% tipo 2 y el 2.6% tipo 3.
- **Trestbps:** La presión arterial en reposo (en mm Hg) de los pacientes tiene una media de 131.6 y una desviación estándar de 17.5. La más baja registrada es de 94 y la más alta es de 200.
- **Chol:** El colesterol sérico (en mg/dl) tiene una media de 246.3 y una desviación estándar de 51.8. El nivel de colesterol más bajo registrado es de 126 y el más alto es de 564.
- **Fbs:** El nivel de azúcar en sangre en ayunas (en mg/dl) tiene una media de 120.9 y una desviación estándar de 30.1. El nivel más bajo registrado es de 78 y el más alto es de 275.
- **Restecg:** El resultado electrocardiográfico en reposo de los pacientes se divide en 50.8% de los pacientes normal, el 47.5% tiene una anormalidad de ST-T y el 1.7% tiene una hipertrofia ventricular izquierda probable o definitiva.
- **Thalach:** La frecuencia cardíaca máxima alcanzada por los pacientes durante el ejercicio tiene una media de 149.6 y una desviación estándar de 22.9. La frecuencia cardíaca más baja registrada es de 71 y la más alta es de 202.
- **Exang:** El 32.7% de los pacientes experimenta angina inducida por ejercicio.
- **Oldpeak:** La depresión del segmento ST inducida por el ejercicio tiene una media de 1.0 y una desviación estándar de 1.2. El valor más bajo es de 0 y el más alto es de 6.2.
- **Slope:** Para pendiente del segmento pico del ejercicio el 46.8% tiene una pendiente ascendente, el 46.8% una pendiente plana y el 6.4% una pendiente descendente.
- **Ca:** El número de vasos principales coloreados por flourosopia tiene una media de 0.7 y una desviación estándar de 1.0. El valor más bajo registrado es de 0 y el más alto es de 4.
- **Thal:** El 54.5% de los pacientes tiene un resultado normal, el 38.1% tiene un resultado defectuoso y el 7.4%

## Visualización de variables

### Variables numéricas

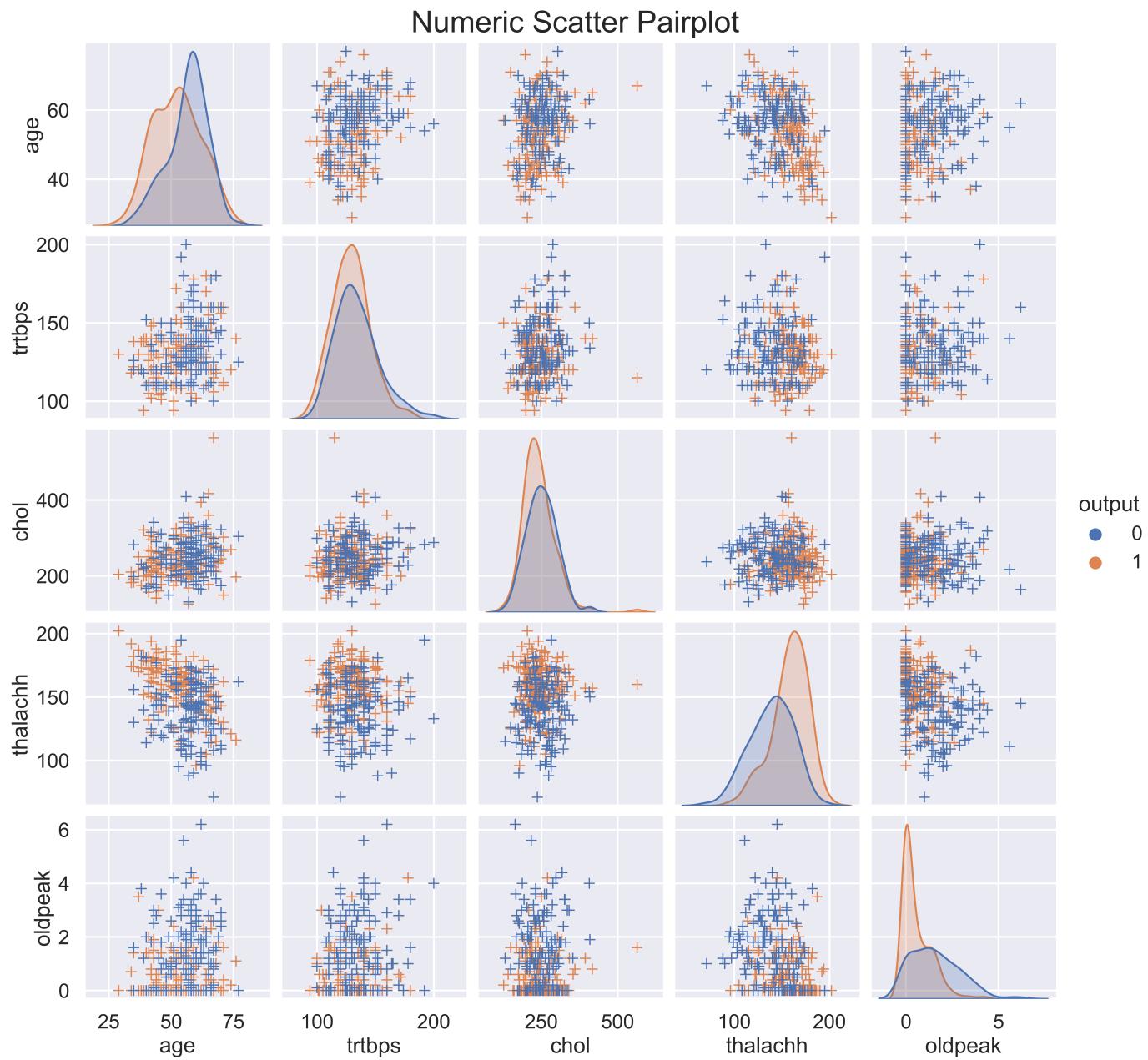
Haciendo uso de la función `seaborn.pairplot(df[numéricos], hue='stroke', kind='hist')` se puede representar las variables numéricas comparadas entre sí con respecto a su frecuencia de la siguiente manera (aquellos con riesgo cardíaco representados en naranja y aquellos sin riesgo cardíaco representados en azul):



Histograma de variables numéricas

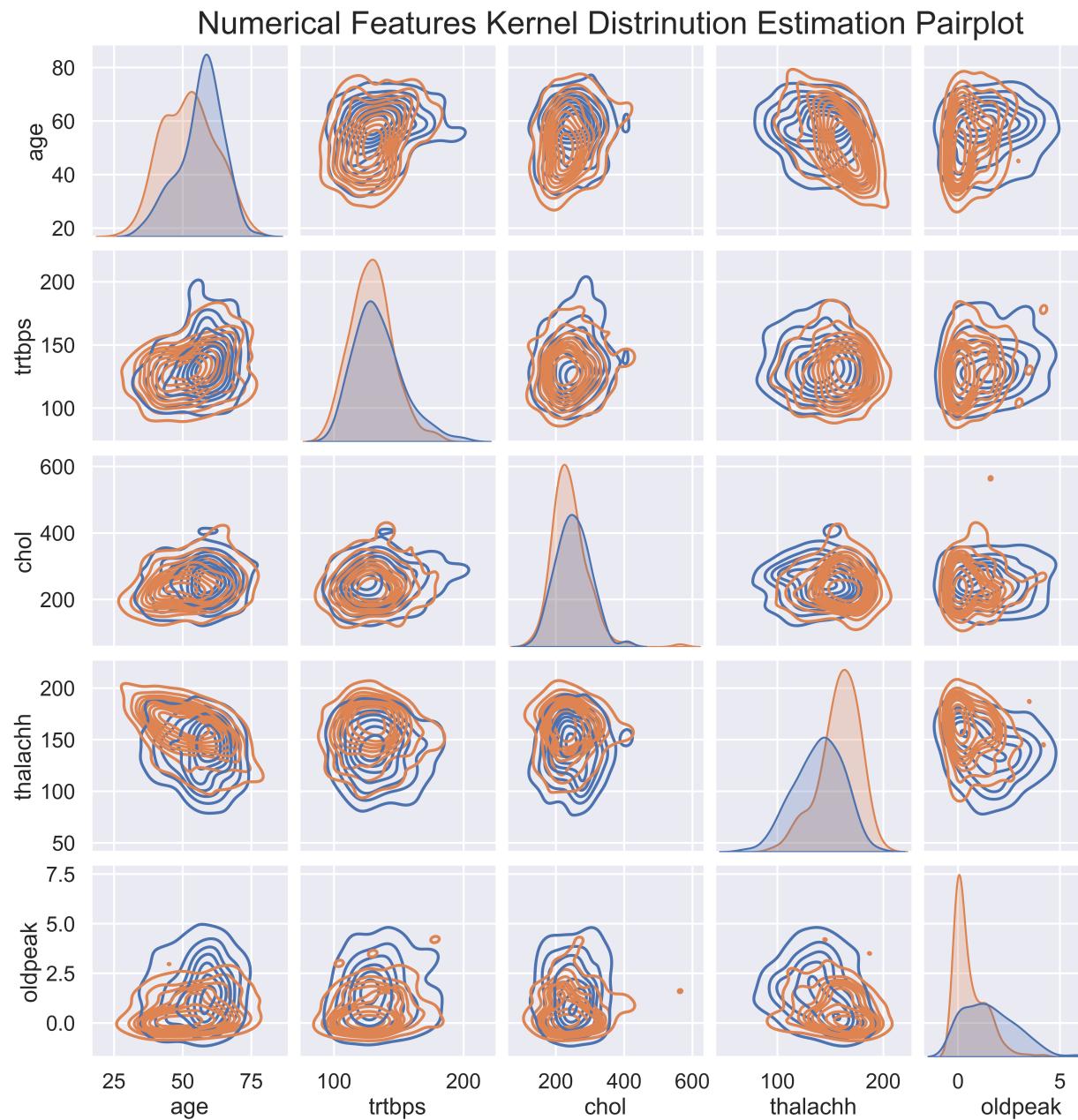
- Edad (age):** La distribución de la edad parece ser ligeramente diferente entre los dos grupos. Los individuos con riesgo cardíaco tienden a ser mayores que aquellos sin riesgo cardíaco.
- Presión arterial en reposo (trtbps):** Ambos grupos muestran una distribución similar en la presión arterial en reposo. Sin embargo, parece que hay una ligera tendencia hacia valores más altos en el grupo con riesgo cardíaco.
- Colesterol sérico en mg/dl (chol):** No parece haber una diferencia significativa en los niveles de colesterol entre los individuos con y sin riesgo cardíaco.
- Frecuencia cardíaca máxima alcanzada (thalachh):** Los individuos con riesgo cardíaco parecen alcanzar una frecuencia cardíaca máxima más alta en comparación con aquellos sin riesgo cardíaco.
- Depresión del ST inducida por el ejercicio en relación con el reposo (oldpeak):** Los individuos con riesgo cardíaco muestran valores más altos de oldpeak en comparación con aquellos sin riesgo cardíaco, lo que indica una mayor depresión del ST durante el ejercicio.

De igual manera, y con las mismas conclusiones, se analizó la distribución de las variables categóricas mediante el uso de una gráfica de dispersión `seaborn.pairplot(df[numericos], hue='stroke', kind='scatter')` donde la diagonal de la cuadrícula muestra la distribución de cada variable con un gráfico de densidad superpuesto y los gráficos fuera de la diagonal muestran la relación entre cada par de variables. Los puntos están coloreados por la variable objetivo.



Gráfica de dispersión de variables numéricas

Finalmente se usa la estimación de densidad kernel (KDE) `seaborn.pairplot(df[numericos], hue='stroke', kind='kde')` que a diferencia de los histogramas, proporcionan una estimación suave, continua, y precisa de la densidad de probabilidad. Esto se logra utilizando una función de kernel, que es una función no negativa que se coloca en cada observación de la muestra.



Gráfica de densidad kernel de variables numéricas