**How will the dataset be balanced?**

The target column, "OS event", was found to be imbalanced, meaning one class had more records than the other. To correct this, SMOTE (Synthetic Minority Over-sampling Technique) was used. This method creates synthetic examples for the minority class, making both classes more even. A balanced dataset helps the model learn from both classes equally and avoids bias.

**How will the dataset be split?**

After balancing, the dataset was split into three parts. First, 70% of the data was used for training. The remaining 30% was split again into two parts: 20% for validation, to tune the model and check for overfitting, and 10% for testing, to measure final model performance. Stratified splitting was used so that the class proportions stayed the same in all parts of the data. A random state was set to make sure the splits are consistent each time the process is repeated.

# Import necessary libraries

from sklearn.model\_selection import train\_test\_split

from imblearn.over\_sampling import SMOTE

import pandas as pd

import os

# Load the prepared dataset

prepared\_data = pd.read\_csv('/home/ec2-user/SageMaker/data/final\_prepared\_data.csv')

# Separate features (X) and target (y)

X = prepared\_data.drop("OS event", axis=1) # All columns except OS event

y = prepared\_data["OS event"] # OS event is the target (0 = alive, 1 = death event)

# Apply SMOTE to balance the dataset

smote = SMOTE(random\_state=42) # random\_state=42 for reproducibility

X\_resampled, y\_resampled = smote.fit\_resample(X, y)

# Split into train (70%) and temp (30%) sets

X\_train, X\_temp, y\_train, y\_temp = train\_test\_split(

X\_resampled, y\_resampled,

test\_size=0.3, # 30% will be split further into validation and test

stratify=y\_resampled, # preserve the proportion of classes

random\_state=42

)

# Split the temp set into validation (20%) and test (10%)

X\_val, X\_test, y\_val, y\_test = train\_test\_split(

X\_temp, y\_temp,

test\_size=0.33, # 0.33 \* 30% = ~10% test set

stratify=y\_temp, # preserve class distribution

random\_state=42

)