

Machine Learning: Heart Imputations

Problem Story

Lemon AI secured a contract with the prestigious Riga Cardiology Center. Ivan was tasked with organizing the hospital's historical patient data to train a model that can predict the presence of heart disease.

However, the hospital's database server was extremely old and glitchy. During the data export, about 10% of the values were corrupted and lost (replaced with `NaN`).

Furthermore, the Chief Cardiologist was very strict: she believed that failing to detect a patient with heart disease is **3 times worse** than falsely diagnosing a healthy person.

"If we miss one sick patient, that's a life at risk," she explained to Ivan. "A false positive just means more tests. I consider a false negative 3 times worse than a false positive."

Ivan needed to clean up the corrupted data and build a model that satisfied the Chief Cardiologist's strict safety requirements. This was his toughest challenge yet — a problem with real consequences.

Problem

You are given a dataset of patient health metrics:

- `X_train.csv`: The features for the training set. Contains missing values.
- `y_train.csv`: The target labels for the training set (`1` = heart disease, `0` = no heart disease).
- `X_test.csv`: The features for the test set. Contains missing values.

Your task is to predict the target for the test set.

Hints

- Preprocessing: Look into `sklearn.impute` to handle the `NaN` values in the CSVs.
- Optimization: Pay close attention to the Grading formula.
- You may find this guide by scikit-learn helpful: <https://scikit-learn.org/stable/modules/impute.html>

Grading

Submissions are scored based on a weighted accuracy metric. The formula is:

$$Score = \frac{3 \cdot TP + TN}{3 \cdot (TP + FN) + (TN + FP)}$$

Where:

- **TP (True Positive)**: You predicted Disease, and they actually have it. (Weighted x3)
- **TN (True Negative)**: You predicted Healthy, and they are actually healthy.
- **FP (False Positive)**: You predicted Disease, but they are healthy.
- **FN (False Negative)**: You predicted Healthy, but they actually have disease. (This hurts the score the most because it reduces the numerator's high-value term).

Essentially, the score represents accuracy on a dataset where every positive patient is duplicated 3 times.

Final score for the task is $(\text{score} - \text{baseline}) / (\text{max_score} - \text{baseline})$.