PROJECT LINK <https://www.google.com/url?q=https%3A%2F%2Fwww.kaggle.com%2Fcompetitions%2Fwidsdatathon2024-challenge1%2Frules>

DESCRIPTION: The challenge is to predict if a patient is diagnosed with Metastatic TNBC within a 90 day diagnostic period based on a set of features. This is a binary classification problem where the goal is to use the provided features to classify the tumors accurately. Data Description: We are using a real-world evidence dataset from Health Verity (HV), one of the largest healthcare data ecosystems in the US, as the main data source. The features include patient ID,patient race, patient age and patient gender and bmi. Validation Accuracy: 0.6427 while theirs was 0.63

SUMMARY OF WORK DONE

Data Overview

* Data Type: Tabular
* train.csv
  + 12906 rows, 83 columns
* test.csv
  + 5792 rows, 82 columns
  + Omits the target column ('DiagPeriodL90D')

DATA VISULAIZATION

* **Class Distribution**:
  + Highlighted imbalance between the two classes.

A screenshot of a graph

Description automatically generated

**Preprocessing / Cleanup**

* **Handling Missing Values**:
  + Numerical columns: Imputed using the median.
  + Categorical columns: Missing values filled proportionally based on other column distributions
* **Encoding Categorical Variables**:
  + Label Encoding as there were already many columns
* **Feature Selection**:
  + Removed highly correlated geographic features to reduce redundancy.
  + Dropped irrelevant features that had no importance.

**Problem Formulation**

* **Input/Output**:
  + Input: Preprocessed patient data.
  + Output: Binary classification (diagnosis within 90 days or not).
* **Model**: XGBoost Classifier for its robustness and best performance from 3 baseline models.

**Training**

* **Process**:
  + Training split: 60% for training, 40% for validation.

Distinct Separation Between Classes: Features where the distribution of values for Class 0 and Class 1 (or for each defined class in regression) is significantly different are likely more informative for the model. High Variability: Features that exhibit a large spread or variability in their values across the classes may be more useful for prediction.