HEART DISEASE MACHINE LEARNING PREDICTION DOCUMENTATION

Overview and Objectives

This notebook is a machine learning solution aimed at addressing the problem of predicting heart disease based on medical data. The primary objective is to develop an accurate model that can determine the probability of an individual having heart disease using patient data. The notebook includes steps for data extraction, transformation, modeling, and inference.

Objectives:

- Leverage machine learning techniques to create a model that can analyze relevant features and provide reliable predictions for early heart disease detection.
- Optimize model performance using hyper-parameter tuning and feature engineering.
- Provide performance metrics to evaluate model accuracy and effectiveness.

ETL Process

Extract:

- **Data Source**: Heart disease dataset available from Zindi platform.
- **Data Format**: CSV file containing patient records.
- **Extraction Method**: The data is extracted using standard Python libraries (pandas).

Transform:

- **Transformation Logic**:
  - Feature selection is performed based on correlation and domain knowledge.

Load:

- **Data Loading**:
  - Transformed data is loaded into the machine learning models for training.
  - Data is split into training and testing sets using an 80-20 split.

Data Modeling

Model Description:

- **Model Type**: Stacked ensemble model consisting of Random Forest and XGBoost, using Logistic Regression as the meta-model.

- **Feature Selection**: Feature selection is done based on correlation analysis and importance ranking from a Random Forest.
- **Model Training**:
    - **Random Forest**: Hyper-parameter tuning includes setting the number of trees and max depth.
    - **XGBoost**: Hyper-parameter tuning is applied for learning rate, max depth, and boosting rounds.
- **Evaluation Metrics**:
    - Accuracy

Validation:

- The model's performance is tested on a separate test set via the Zindi platform, and evaluation metrics are logged.

Inference

Deployment:

- The trained model is used for inference on new patient data.
- Inference is performed by loading the model, passing in new data, and interpreting the output as either positive (presence of heart disease) or negative (no heart disease).

Model Updates:

- Retraining strategy includes updating the hyper-parameters and re-validating the model.

Run Time

- **Data Preprocessing**: approx. 1 minute
- **Model Training**:
    - Random Forest: approx. 7 minutes
    - XGBoost: approx. 5 minutes
    - Stacked Model: approx. 3 minutes
- **Inference**: < 1 second per new data point

Performance Metrics

- **Public Score (Zindi)**: 0.920987654
- **Private Score (Zindi)**: 0.928457869

- **Metric Used**:
    - Accuracy: 0.8208744710860366

Additional Metrics:

- Other metrics such as Precision, Recall and F1 score were used during model selection and optimization to balance false positives and false negatives.

Error Handling

- **Error Handling**:
    - No error to handle, training data appeared mostly clean.

Maintenance and Monitoring

- **Monitoring**:
    - The model's performance is monitored using validation metrics. Regular checks are done to ensure that the model does not drift from its original performance.

Environment Setup

- Use the provided requirements.txt file to install the necessary libraries.
- Notebook was ran and training was done on Google Colab.