|  |  |
| --- | --- |
| **Heart Disease dataset**  This dataset is collected by Hungarian Institute of Cardiology, University Hospital (Zurich, Switzerland), University Hospital (Zurich, Switzerland) and V.A. Medical Center (Long Beach and Cleveland Clinic Foundation). The dataset includes the target variable representing presence of some cardiovascular disease (0=no heart disease, 1=heart disease). The features measured in the dataset are Age, Sex, Type of Chest Pain, Resting Blood Pressure, Serum Cholesterol, Fasting Blood Sugar, Resting ECG, Max Heart Rate, Exercise Induced Angina, The slope of peak exercise ST segment, Number of Vessels colored by Fluoroscopy and presence of Fixed or Reversible Defect in Stress Echocardiography.  **Random Forest Classification**  Random Forest is an ensemble method in which many decision trees are trained by bootstrapping the data and aggregating the results at the end. For classification problems, majority decision of each tree is taken to be the overall classification prediction. In each decision tree, data is continuously divided based on values of randomly selected features and purity of resulting nodes is computed to evaluate the effectiveness of the split. Various impurity measures, such as Gini Index, is used to measure purity of the nodes. Gini Index calculates probability of randomly picked data point that is classified incorrectly. The formula to calculate the Gini Index is  The following chart depicts values of Gini Index for nodes with different splits of a target classification variable.  to be a leaf node and number of data points required to be a leaf node and whether a bootstrap should be used to build a tree instead of the full data set. | For binary classification problems, if a node contains an even split of the target variable, then the Gini Index is 0.50. Hence the probability that a randomly picked data point is classified incorrect is random (p=0.50) which is expected due to the even split of two classes. If the node contains a higher percentage of one class, Gini Index decreases.  **Cost Function**  This Random Forest classifier models presence of heart disease which is a binary variable. Standard classification metrics are used to optimize the model. This includes Area Under the Receiver Operating Curve, Recall, Precision and Accuracy.    These evaluation metrics are collected using 5-fold cross validation to avoid overfitting on the training set. This data is collected for each iteration in the hyperparameter tuning process.  **Hyperparameter Tuning**  Machine learning models use various parameter settings that have an impact on the cost function. Random Forest models involve a set of parameters that developers can optimize to evaluate the cost function. For example, one parameter setting is maximum depth of decision trees that are built. This setting allows developers to cap the depth of decision trees to avoid overfitting model on the training set. If this setting is not optimized, the tree is expanded until all leaves contain data points from the same class. This can lead to severe overfitting. The parameters tuned for modeling heart disease data optimized parameters including the number of decision trees, the number of features to consider for the best split, maximum depth of each tree, minimum number of data points required for splitting a node, minimum number of data points required |
|  |  |