JADBio Description of Performed Analysis

Setup

JADBio version **1.4.118** ran on dataset **Crop_recommendation** with **2200** samples and **7** features to create a predictive model for outcome named **label**. The outcome was discrete leading to a **classification** modeling.

The preferences of the analysis were set to **true** for feature selection and **false** for full feature models tried.
The **AUC** metric was used to optimize for the best model.
The maximum number of features to select was set to **25**.
The effort to spend on tuning the algorithms were set to **Quick**.
The number of CPU cores to use for the analysis was set to **5**.
The execution time was **00:12:05**.

Configuration Space

JADBio's AI decide to try the following algorithms and tuning hyper-parameter values:

Algorithm Type	Algorithm	Hyper-parameter	Set of Values
Preprocessing	Mean Imputation		
	Mode Imputation		
	Constant Removal		
	Variable Normalization		
Feature Selection	Epilogi	stoppingCriterion	Independence Test
		stoppingThreshold	0.001
		equivalenceThreshold	0.01
	Test-Budgeted Statistically Equivalent Signature (SES)	alpha	0.05
		maxK	2.0
	LASSO	penalty	1.0
Modeling	Classification Decision Tree with Deviance splitting criterion	minLeafSize	3
		alpha	0.05
	Ridge Logistic Regression	lambda	1.0
	Classification Random Forest with Deviance splitting criterion	minLeafSize	3.0
		nTrees	100
	Support Vector Machines (SVM) of type C-SVC with Linear Kernel	cost	1.0
	Support Vector Machines (SVM) of type C-SVC with Polynomial Kernel	cost	1.0
		degree	3
		gamma	1.0
	Support Vector Machines (SVM) of type C-SVC with Gaussian Kernel	cost	1.0
		gamma	1.0

Leading to **25** combinations and corresponding configurations (machine learning pipelines) to try. For the full configurations tested see the Appendix.

Configuration Estimation Protocol

JADBio's AI system decided to estimate the out-of-sample performance of the models produced by each configuration using **Repeated 10-fold CV without dropping (max. repeats = 20).** Overall, 25 configurations × 20 repeats × 10 folds = 250 models were set out to train.

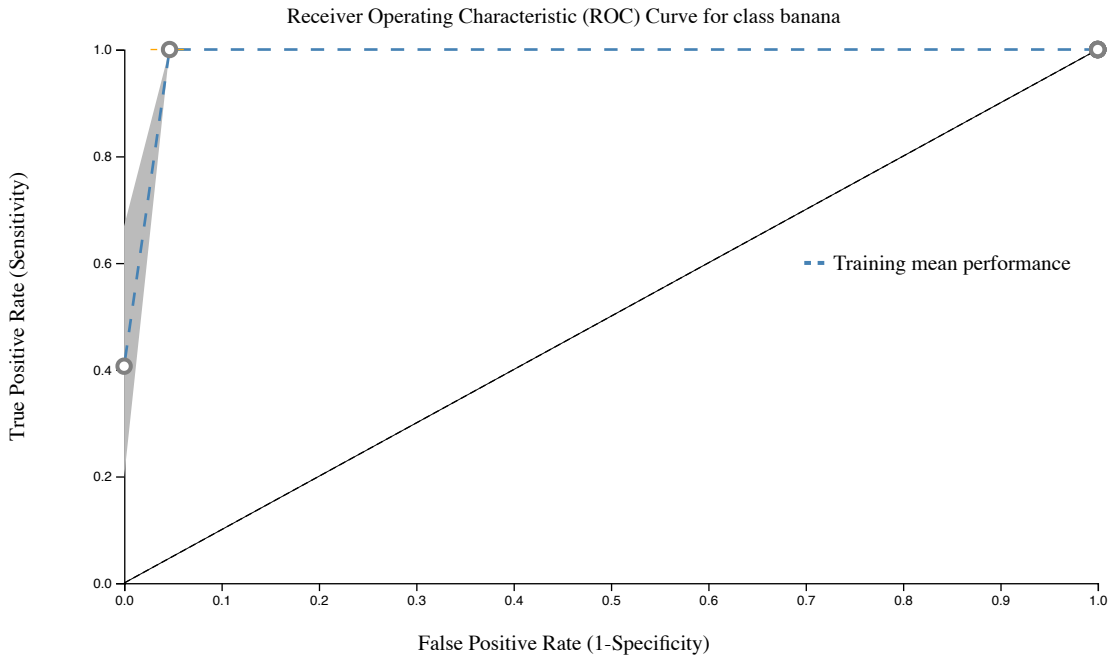
JADBio Results Summary

Overview

A result summary is presented for analysis optimized for Performance. The model is produced by applying the algorithms in sequence (configuration) on the training data:

Preprocessing	Feature Selection	Predictive algorithm
Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO Feature Selection (penalty=1.0)	Classification Random Forest training 100 trees with Deviance splitting criterion, minimum leaf size = 3, splits = 1, alpha = 1, and variables to split = 0.816 sqrt (nvars)

The Area Under the ROC Curve is shown in the figure below:



Metric	Mean estimate	CI
Area Under the ROC Curve	1.000	[1.000, 1.000]
Mean Average Precision (a.k.a. Average Area Under the Precision-Recall curve)	1.000	[1.000, 1.000]
Accuracy	0.995	[0.990, 0.999]
Balanced Accuracy	0.995	[0.989, 0.999]
Average F1 score	0.994	[0.989, 0.999]
Average Matthews correlation	0.994	[0.988, 0.998]
Precision for class apple	1.000	[1.000, 1.000]
Precision for class banana	1.000	[1.000, 1.000]
Precision for class blackgram	0.997	[0.971, 1.000]
Precision for class chickpea	1.000	[1.000, 1.000]
Precision for class coconut	1.000	[1.000, 1.000]
Precision for class coffee	1.000	[1.000, 1.000]
Precision for class cotton	0.991	[0.957, 1.000]
Precision for class grapes	1.000	[1.000, 1.000]
Precision for class jute	0.938	[0.857, 1.000]
Precision for class kidneybeans	1.000	[1.000, 1.000]
Precision for class lentil	0.990	[0.955, 1.000]

Metric	Mean estimate	CI
Precision for class maize	0.997	[0.967, 1.000]
Precision for class mango	1.000	[1.000, 1.000]
Precision for class mothbeans	0.991	[0.956, 1.000]
Precision for class mungbean	1.000	[1.000, 1.000]
Precision for class muskmelon	1.000	[1.000, 1.000]
Precision for class orange	1.000	[1.000, 1.000]
Precision for class papaya	1.000	[1.000, 1.000]
Precision for class pigeonpeas	1.000	[1.000, 1.000]
Precision for class pomegranate	1.000	[1.000, 1.000]
Precision for class rice	0.989	[0.950, 1.000]
Precision for class watermelon	1.000	[1.000, 1.000]
MCC for class apple	1.000	[1.000, 1.000]
MCC for class banana	1.000	[1.000, 1.000]
MCC for class blackgram	0.996	[0.971, 1.000]
MCC for class chickpea	1.000	[1.000, 1.000]
MCC for class coconut	1.000	[1.000, 1.000]
MCC for class coffee	1.000	[1.000, 1.000]
MCC for class cotton	0.995	[0.979, 1.000]
MCC for class grapes	1.000	[1.000, 1.000]
MCC for class jute	0.959	[0.909, 0.992]
MCC for class kidneybeans	1.000	[1.000, 1.000]
MCC for class lentil	0.988	[0.961, 1.000]
MCC for class maize	0.993	[0.968, 1.000]
MCC for class mango	1.000	[1.000, 1.000]
MCC for class mothbeans	0.987	[0.959, 1.000]
MCC for class mungbean	1.000	[1.000, 1.000]
MCC for class muskmelon	1.000	[1.000, 1.000]
MCC for class orange	1.000	[1.000, 1.000]
MCC for class papaya	1.000	[1.000, 1.000]
MCC for class pigeonpeas	1.000	[1.000, 1.000]
MCC for class pomegranate	1.000	[1.000, 1.000]
MCC for class rice	0.954	[0.881, 0.992]
MCC for class watermelon	1.000	[1.000, 1.000]
True Positive Rate for class apple	1.000	[1.000, 1.000]
True Positive Rate for class banana	1.000	[1.000, 1.000]
True Positive Rate for class blackgram	0.997	[0.970, 1.000]
True Positive Rate for class chickpea	1.000	[1.000, 1.000]

Metric	Mean estimate	CI
True Positive Rate for class coconut	1.000	[1.000, 1.000]
True Positive Rate for class coffee	1.000	[1.000, 1.000]
True Positive Rate for class cotton	1.000	[1.000, 1.000]
True Positive Rate for class grapes	1.000	[1.000, 1.000]
True Positive Rate for class jute	0.988	[0.944, 1.000]
True Positive Rate for class kidneybeans	1.000	[1.000, 1.000]
True Positive Rate for class lentil	0.989	[0.950, 1.000]
True Positive Rate for class maize	0.991	[0.955, 1.000]
True Positive Rate for class mango	1.000	[1.000, 1.000]
True Positive Rate for class mothbeans	0.987	[0.947, 1.000]
True Positive Rate for class mungbean	1.000	[1.000, 1.000]
True Positive Rate for class muskmelon	1.000	[1.000, 1.000]
True Positive Rate for class orange	1.000	[1.000, 1.000]
True Positive Rate for class papaya	1.000	[1.000, 1.000]
True Positive Rate for class pigeonpeas	1.000	[1.000, 1.000]
True Positive Rate for class pomegranate	1.000	[1.000, 1.000]
True Positive Rate for class rice	0.932	[0.844, 1.000]
True Positive Rate for class watermelon	1.000	[1.000, 1.000]
Sensitivity for class apple	1.000	[1.000, 1.000]
Sensitivity for class banana	1.000	[1.000, 1.000]
Sensitivity for class blackgram	0.997	[0.970, 1.000]
Sensitivity for class chickpea	1.000	[1.000, 1.000]
Sensitivity for class coconut	1.000	[1.000, 1.000]
Sensitivity for class coffee	1.000	[1.000, 1.000]
Sensitivity for class cotton	1.000	[1.000, 1.000]
Sensitivity for class grapes	1.000	[1.000, 1.000]
Sensitivity for class jute	0.988	[0.944, 1.000]
Sensitivity for class kidneybeans	1.000	[1.000, 1.000]
Sensitivity for class lentil	0.989	[0.950, 1.000]
Sensitivity for class maize	0.991	[0.955, 1.000]
Sensitivity for class mango	1.000	[1.000, 1.000]
Sensitivity for class mothbeans	0.987	[0.947, 1.000]
Sensitivity for class mungbean	1.000	[1.000, 1.000]
Sensitivity for class muskmelon	1.000	[1.000, 1.000]
Sensitivity for class orange	1.000	[1.000, 1.000]
Sensitivity for class papaya	1.000	[1.000, 1.000]
Sensitivity for class pigeonpeas	1.000	[1.000, 1.000]

Metric	Mean estimate	CI
Sensitivity for class pomegranate	1.000	[1.000, 1.000]
Sensitivity for class rice	0.932	[0.844, 1.000]
Sensitivity for class watermelon	1.000	[1.000, 1.000]
Specificity for class apple	1.000	[1.000, 1.000]
Specificity for class banana	1.000	[1.000, 1.000]
Specificity for class blackgram	1.000	[0.999, 1.000]
Specificity for class chickpea	1.000	[1.000, 1.000]
Specificity for class coconut	1.000	[1.000, 1.000]
Specificity for class coffee	1.000	[1.000, 1.000]
Specificity for class cotton	1.000	[0.999, 1.000]
Specificity for class grapes	1.000	[1.000, 1.000]
Specificity for class jute	0.997	[0.993, 1.000]
Specificity for class kidneybeans	1.000	[1.000, 1.000]
Specificity for class lentil	1.000	[0.999, 1.000]
Specificity for class maize	1.000	[0.999, 1.000]
Specificity for class mango	1.000	[1.000, 1.000]
Specificity for class mothbeans	1.000	[0.999, 1.000]
Specificity for class mungbean	1.000	[1.000, 1.000]
Specificity for class muskmelon	1.000	[1.000, 1.000]
Specificity for class orange	1.000	[1.000, 1.000]
Specificity for class papaya	1.000	[1.000, 1.000]
Specificity for class pigeonpeas	1.000	[1.000, 1.000]
Specificity for class pomegranate	1.000	[1.000, 1.000]
Specificity for class rice	0.999	[0.997, 1.000]
Specificity for class watermelon	1.000	[1.000, 1.000]
Average Precision for class apple	1.000	[1.000, 1.000]
Average Precision for class banana	1.000	[1.000, 1.000]
Average Precision for class blackgram	1.000	[1.000, 1.000]
Average Precision for class chickpea	1.000	[1.000, 1.000]
Average Precision for class coconut	1.000	[1.000, 1.000]
Average Precision for class coffee	1.000	[1.000, 1.000]
Average Precision for class cotton	1.000	[1.000, 1.000]
Average Precision for class grapes	1.000	[1.000, 1.000]
Average Precision for class jute	0.996	[0.980, 1.000]
Average Precision for class kidneybeans	1.000	[1.000, 1.000]
Average Precision for class lentil	1.000	[1.000, 1.000]
Average Precision for class maize	1.000	[1.000, 1.000]

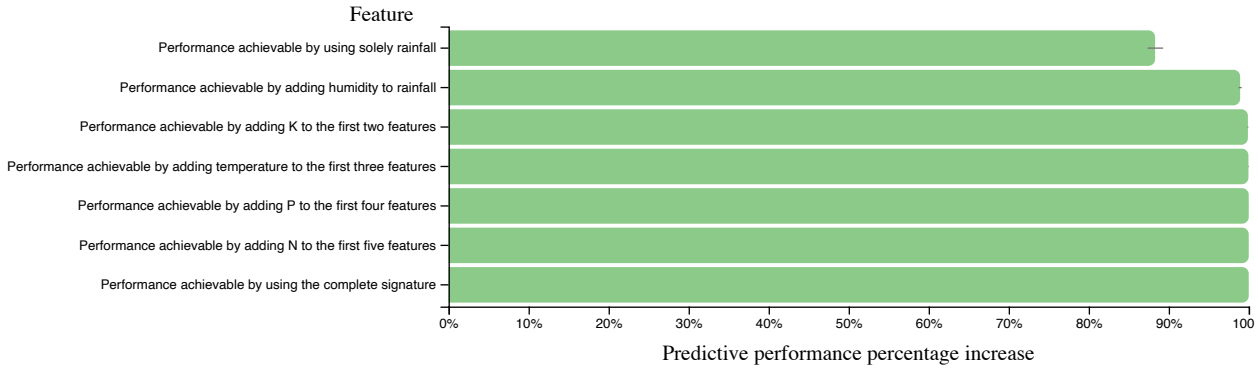
Metric	Mean estimate	CI
Average Precision for class mango	1.000	[1.000, 1.000]
Average Precision for class mothbeans	1.000	[1.000, 1.000]
Average Precision for class mungbean	1.000	[1.000, 1.000]
Average Precision for class muskmelon	1.000	[1.000, 1.000]
Average Precision for class orange	1.000	[1.000, 1.000]
Average Precision for class papaya	1.000	[1.000, 1.000]
Average Precision for class pigeonpeas	1.000	[1.000, 1.000]
Average Precision for class pomegranate	1.000	[1.000, 1.000]
Average Precision for class rice	0.998	[0.984, 1.000]
Average Precision for class watermelon	1.000	[1.000, 1.000]

Feature Selection

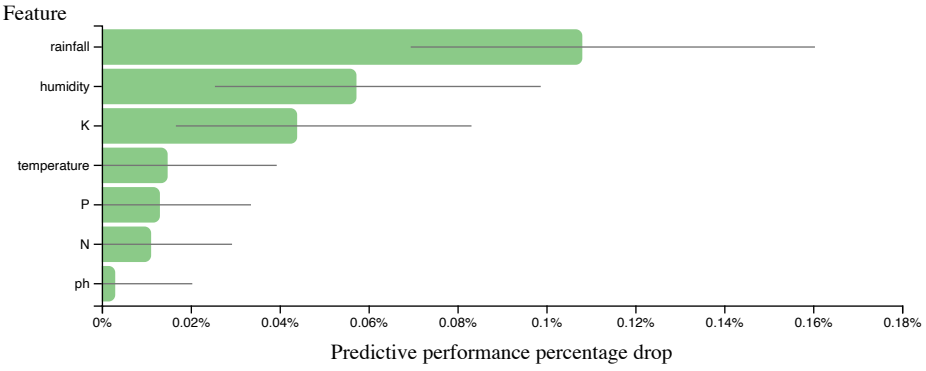
There were **7** features selected out of the **7** available.

The selected features consist of the following subset called a signature. **There was a single signature identified.** The first signature identified by the system is the set: **N, P, K, temperature, humidity, ph, rainfall** in order of importance. The following features cannot be substituted with others and still obtain an equal predictive performance: **N, P, K, temperature, humidity, ph, rainfall**.

The performance achieved by adding each feature in sequence to the model relative to the performance of the final model with all selected features is shown below. The features are added in order of importance:

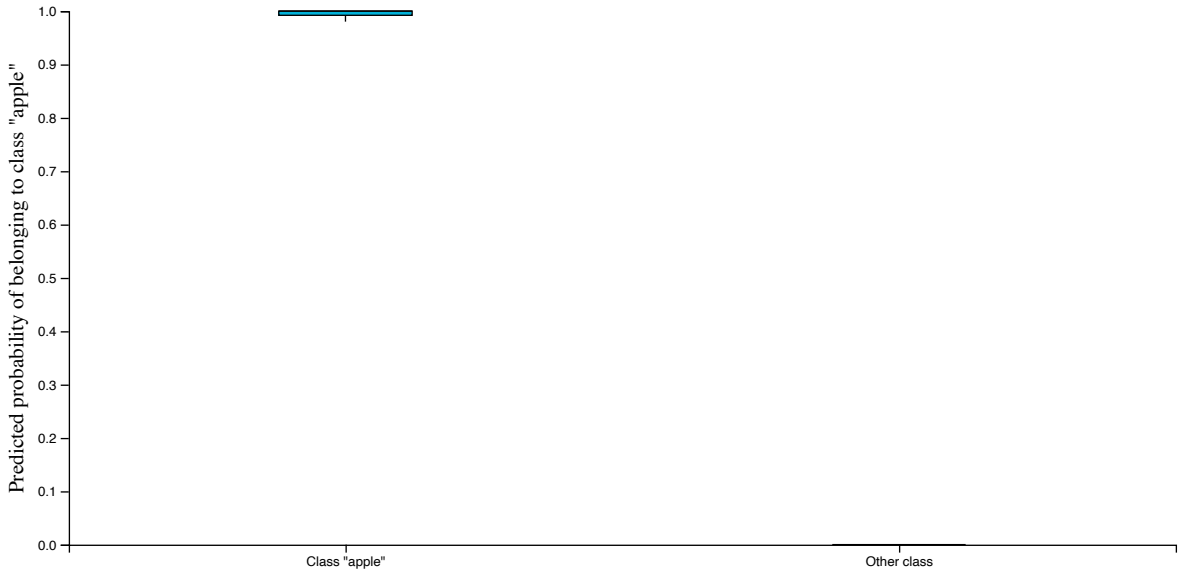


Some features may not seem to add predictive performance to the model; however, the feature selection algorithms include them as an effort to make the final model more robust to noise. The performances achieved by a model that contains all features except one, relative to the performance achieved when the feature is removed is shown below:



For some features there is no noticeable drop in performance when they are removed because they carry predictive information that is shared by other features selected.

The separation of the predictions of the classes achieved by the model is shown in the box-plots below. These are the out-of-sample predictions made by model produced by the same configuration as the final model when the sample was used for testing (e.g.., during cross-validation) and was not used to train the model.



Appendix

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
1	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.001	Classification Decision Tree with Deviance splitting criterion	minimum leaf size = 3, alpha = 0.05	0.9896082251082242	00:00:06.6972	false
2	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Ridge Logistic Regression	lambda = 1.0	0.522517316017316	00:00:16.16897	false
3	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.001	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.9973647186147188	00:00:06.6893	false
4	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.9999870129870132	00:00:21.21015	false
5	IdentityFactory	FullSelector	-	Trivial model	-	0.5	00:00:00.000	false
6	Mean Imputation, Mode Imputation, Constant	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Support Vector Machines (SVM) of type C-SVC	kernel = 'Linear Kernel', cost = 1.0	0.9681493506493509	00:00:16.16100	false

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
	Removal, Standardization							
7	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.001	Support Vector Machines (SVM) of type C-SVC	kernel = 'Polynomial Kernel', cost = 1.0, gamma = 1.0, degree = 3	0.9877164502164502	00:00:06.6892	false
8	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Support Vector Machines (SVM) of type C-SVC	kernel = 'Linear Kernel', cost = 1.0	0.9855909090909092	00:00:20.20979	false
9	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.001	Support Vector Machines (SVM) of type C-SVC	kernel = 'Linear Kernel', cost = 1.0	0.9875779220779222	00:00:06.6891	false
10	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Support Vector Machines (SVM) of type C-SVC	kernel = 'Polynomial Kernel', cost = 1.0, gamma = 1.0, degree = 3	0.9915887445887448	00:00:20.20979	false
11	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Support Vector Machines (SVM) of type C-SVC	kernel = 'Polynomial Kernel', cost = 1.0, gamma = 1.0, degree = 3	0.9833809523809528	00:00:16.16063	false
12	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.9996060606060605	00:00:16.16112	false
13	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Ridge Logistic Regression	lambda = 1.0	0.5714285714285715	00:00:21.21774	false
14	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.9995757575757578	00:00:15.15953	false
15	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Classification Decision Tree with Deviance splitting criterion	minimum leaf size = 3, alpha = 0.05	0.9875519480519469	00:00:16.16013	false

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
16	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Support Vector Machines (SVM) of type C-SVC	kernel = 'Gaussian Kernel', cost = 1.0, gamma = 1.0	0.9996623376623376	00:00:21.21163	false
17	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.9999870129870132	00:00:21.21019	false
18	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.9996060606060605	00:00:16.16012	false
19	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Support Vector Machines (SVM) of type C-SVC	kernel = 'Gaussian Kernel', cost = 1.0, gamma = 1.0	0.9986277056277058	00:00:16.16275	false
20	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.001	Ridge Logistic Regression	lambda = 1.0	0.5801428571428572	00:00:07.7727	false
21	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Classification Decision Tree with Deviance splitting criterion	minimum leaf size = 3, alpha = 0.05	0.9930887445887439	00:00:21.21032	false
22	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.9999848484848485	00:00:21.21073	false
23	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.001	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.9974469696969699	00:00:06.6924	false
24	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.001	Support Vector Machines (SVM) of type C-SVC	kernel = 'Gaussian Kernel', cost = 1.0, gamma = 1.0	0.9948917748917748	00:00:06.6999	false
25	Mean Imputation,	Epilogi	equivThresh = 0.01, stopping	Classification Random	ntrees = 100, minimum leaf	0.9974469696969699	00:00:06.6919	false

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
	Mode Imputation, Constant Removal, Standardization		criterion = Independence Test, stopping threshold = 0.001	Forest with Deviance splitting criterion	size = 3			