JADBio Description of Performed Analysis

Setup

JADBio version **1.4.118** ran on dataset **Obesity_among_children_and_adolescents_aged_2_19_years__by_selected_characteristics__United_States** with **633** samples and **14** features to create a predictive model for outcome named **STUB_LABEL_NUM**. The outcome was continuous leading to a **regression** modeling.

The preferences of the analysis were set to **true** for feature selection and **false** for full feature models tried.
The **R2** 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 **6**.
The execution time was **00:00:05**.

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	Test-Budgeted Statistically Equivalent Signature (SES)	maxK	2.0
		alpha	0.05
	LASSO	penalty	1.0
Modeling	Regression Decision Tree with Mean Squared Error splitting criterion	alpha	0.05
		minLeafSize	5
	Ridge Linear Regression	lambda	1.0
	Regression Random Forest with Mean Squared Error splitting criterion	nTrees	100
		minLeafSize	5.0

Leading to **7** 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 **Incomplete 10-fold CV with dropping.** Overall, 36 models were set out to train.

Eventually, 36 had their estimation protocol completed.

JADBio Results Summary

Overview

A result summary is presented for analysis optimized for Aggressive Feature Selection. 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	Test-Budgeted Statistically Equivalent Signature (SES) algorithm with hyper-parameters: maxK = 2, alpha = 0.05 and budget = 3 * nvars	Regression Decision Tree with Mean Squared Error splitting criterion and hyper-parameters: minimum leaf size = 5, and pruning parameter alpha = 0.05

The R-squared is shown in the figure below:

Metric	Mean estimate	CI
R-squared	0.997	[0.996, 0.998]
Mean Absolute Error	0.062	[0.056, 0.068]
Mean Squared Error	0.005	[0.004, 0.006]
Relative Absolute Error	0.064	[0.053, 0.077]
Relative Squared Error	0.003	[0.002, 0.005]
Correlation Coefficient	0.998	[0.998, 0.999]
Mean Squared Logarithmic Error	0.000	[0.000, 0.000]

Feature Selection

There were **2** features selected out of the **14** available.

The selected features consist of the following subset called a signature. **There were multiple signatures identified.** The first signature identified by the system is the set: **STUB_NAME, SE** in order of importance. The following features cannot be substituted with others and still obtain an equal predictive performance: **STUB_NAME, SE**.

Alternatively, in the following table, the features that could substitute for some selected feature are listed and still obtain a statistically indistinguishable predictive performance:

Feature	Could be substituted with
STUB_NAME	STUB_NAME_NUM,STUB_LABEL

Feature	Could be substituted with
SE	

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.

Appendix

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
1	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Regression Decision Tree with Mean Squared Error splitting criterion	minimum leaf size = 5, alpha = 0.05	0.996971321222108	00:00:00.027	false
2	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Ridge Linear Regression	lambda = 1.0	0.9963280881576398	00:00:00.310	false
3	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Regression Random Forest with Mean Squared Error splitting criterion	ntrees = 100, minimum leaf size = 5	0.9970040268789256	00:00:00.323	false
4	IdentityFactory	FullSelector	-	Trivial model	-	1.132164191119725e-16	00:00:00.000	false
5	Mean Imputation,	LASSO	penalty = 1.0	Regression Random	ntrees = 100, minimum	0.9970040268789256	00:00:00.322	false

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
	Mode Imputation, Constant Removal, Standardization			Forest with Mean Squared Error splitting criterion	leaf size = 5			
6	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Regression Decision Tree with Mean Squared Error splitting criterion	minimum leaf size = 5, alpha = 0.05	0.9970459783721677	00:00:00.312	false
7	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Regression Random Forest with Mean Squared Error splitting criterion	ntrees = 100, minimum leaf size = 5	0.9970040268789256	00:00:00.322	false