JADBio Description of Performed Analysis

Setup

JADBio version **1.4.118** ran on dataset **Sleep_Efficiency** with **452** samples and **11** features to create a predictive model for outcome named **Sleep efficiency**. 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 **1**.
The execution time was **00:00:18**.

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	LASSO	penalty	1.0
	Test-Budgeted Statistically Equivalent Signature (SES)	alpha	0.05
		maxK	2.0
Modeling	Support Vector Regression Machines (SVR) of type epsilon-SVR with Linear Kernel	epsilon	0.1
		cost	1.0
	Ridge Linear Regression	lambda	1.0
	Regression Random Forest with Mean Squared Error splitting criterion	minLeafSize	5.0
		nTrees	100
	Support Vector Regression Machines (SVR) of type epsilon-SVR with Gaussian Kernel	epsilon	0.1
		cost	1.0
		gamma	1.0
	Regression Decision Tree with Mean Squared Error splitting criterion	minLeafSize	5
		alpha	0.05

Leading to **15** 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 without dropping.** Overall, 90 models were set out to train.

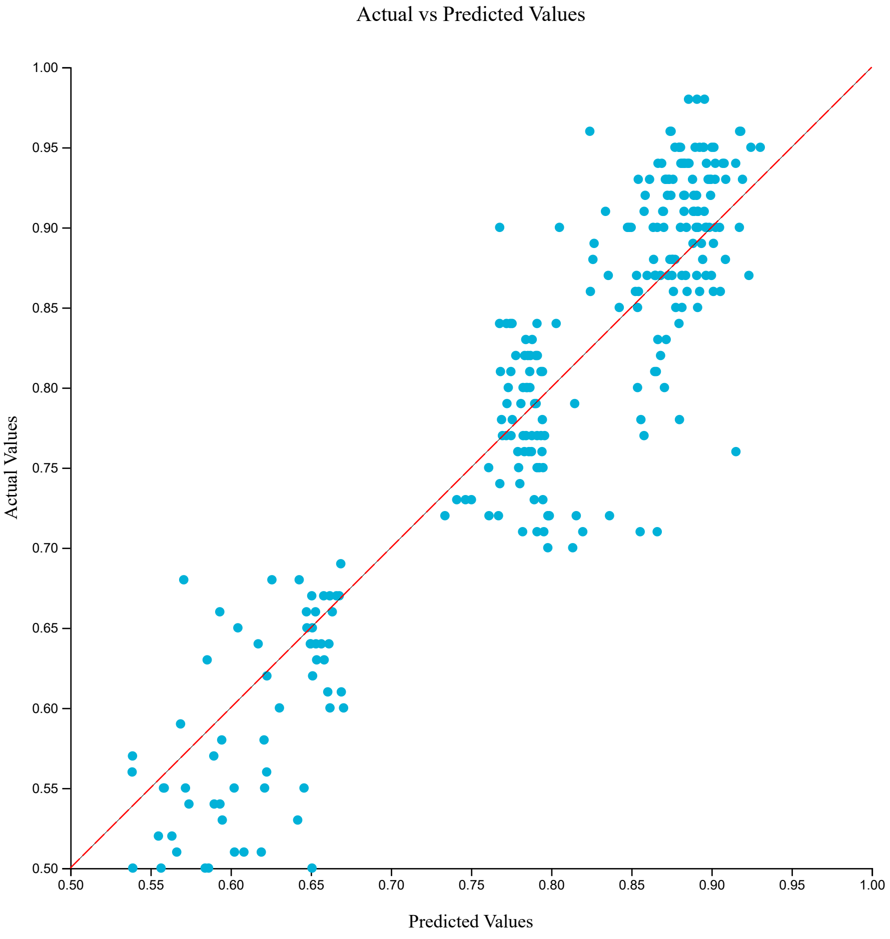
JADBio Results Summary

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)	Regression Random Forest training 100 trees with Mean Squared Error splitting criterion, minimum leaf size = 5, splits = 1, alpha = 1, and variables to split = nvars // 3.0

The R-squared is shown in the figure below:



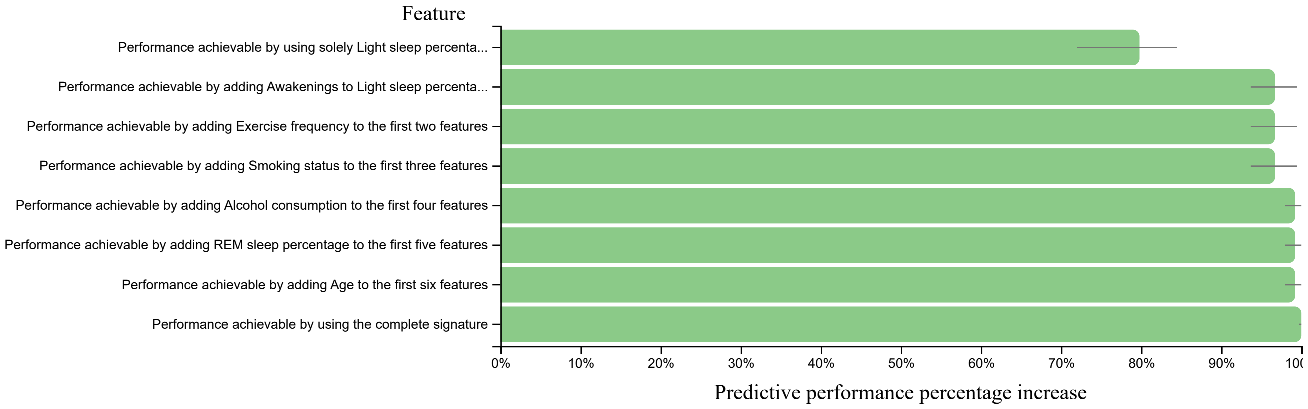
Metric	Mean estimate	CI
R-squared	0.834	[0.752, 0.881]
Mean Absolute Error	0.041	[0.036, 0.047]
Mean Squared Error	0.003	[0.002, 0.004]
Relative Absolute Error	0.395	[0.337, 0.472]
Relative Squared Error	0.179	[0.125, 0.276]
Correlation Coefficient	0.920	[0.871, 0.951]

Feature Selection

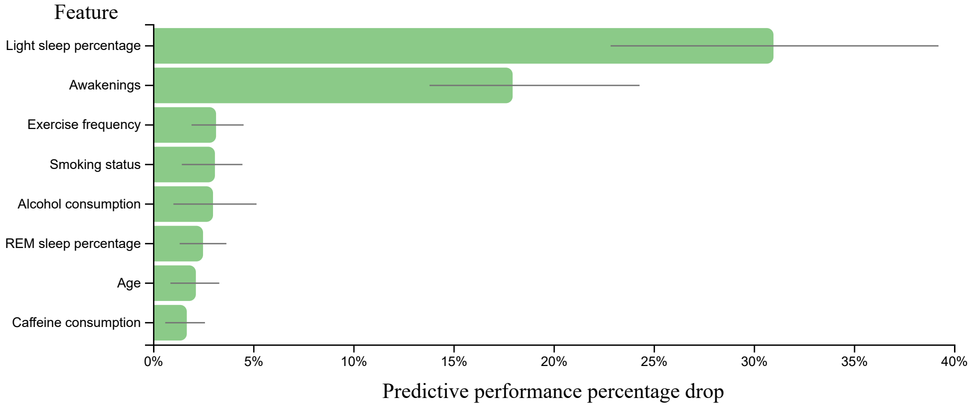
There were **8** features selected out of the **11** available.

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: **Age, REM sleep percentage, Light sleep percentage, Awakenings, Caffeine consumption, Alcohol consumption, Smoking status, Exercise frequency** in order of importance. The following features cannot be substituted with others and still obtain an equal predictive performance: **Age, REM sleep percentage, Light sleep percentage, Awakenings, Caffeine consumption, Alcohol consumption, Smoking status, Exercise frequency**.

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	LASSO	penalty = 1.0	Support Vector Regression Machines (SVR) of type epsilon-SVR	kernel = 'Linear Kernel', cost = 1.0, epsilon = 0.1	0.7714090218546393	00:00:00.583	false
2	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Ridge Linear Regression	lambda = 1.0	0.763838794041031	00:00:00.515	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	0.6985504842658058	00:00:00.527	false

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
4	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Regression Random Forest with Mean Squared Error splitting criterion	ntrees = 100, minimum leaf size = 5	0.7148695566803928	00:00:00.045	false
5	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Ridge Linear Regression	lambda = 1.0	0.7552840203134236	00:00:00.041	false
6	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Support Vector Regression Machines (SVR) of type epsilon-SVR	kernel = 'Gaussian Kernel', cost = 1.0, gamma = 1.0, epsilon = 0.1	0.8061581804087522	00:00:00.088	false
7	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	0.8305849960061965	00:00:00.526	false
8	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Support Vector Regression Machines (SVR) of type epsilon-SVR	kernel = 'Linear Kernel', cost = 1.0, epsilon = 0.1	0.7628340024196278	00:00:00.169	false
9	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Support Vector Regression Machines (SVR) of type epsilon-SVR	kernel = 'Gaussian Kernel', cost = 1.0, gamma = 1.0, epsilon = 0.1	0.686830741535398	00:00:00.531	false
10	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	0.8301132938089505	00:00:00.083	false
11	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES)	maxK = 2, alpha = 0.05, budget = 3 * nvars	Regression Random Forest with Mean Squared Error splitting criterion	ntrees = 100, minimum leaf size = 5	0.7148695566803928	00:00:00.056	false
12	IdentityFactory	FullSelector	-	Trivial model	-	-6.284435866328377e-16	00:00:00.000	false
13	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	0.6985504842658058	00:00:00.528	false
14	Mean Imputation, Mode Imputation, Constant	Test-Budgeted Statistically	maxK = 2, alpha = 0.05,	Regression Random Forest with	ntrees = 100, minimum leaf size = 5	0.8407463840060113	00:00:00.074	false

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
	Removal, Standardization	Equivalent Signature (SES)	budget = 3 * nvars	Mean Squared Error splitting criterion				
15	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	0.8471758534168807	00:00:00.552	false