JADBio Description of Performed Analysis

Setup

JADBio version **1.4.118** ran on dataset **healthcare-dataset-stroke-data** with **5110** samples and **10** features to create a predictive model for outcome named **age**. 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 **00:00:34**.

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
	Epilogi	stoppingCriterion	Independence Test
		stoppingThreshold	0.001
		equivalenceThreshold	0.01
Modeling	Regression Random Forest with Mean Squared Error splitting criterion	nTrees	100
		minLeafSize	5.0
	Support Vector Regression Machines (SVR) of type epsilon-SVR with Linear Kernel	epsilon	0.1
		cost	1.0
	Ridge Linear Regression	lambda	1.0

Leading to **11** 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 **90.00 % - % 10.00 hold-out.** Overall, 22 models were set out to train.

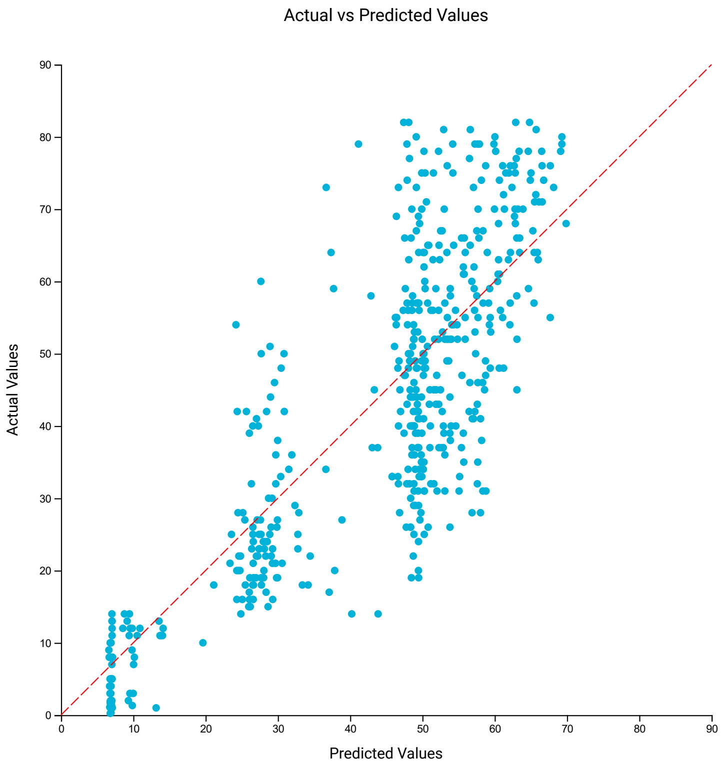
JADBio Results Summary

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 // 5.0

The R-squared is shown in the figure below:



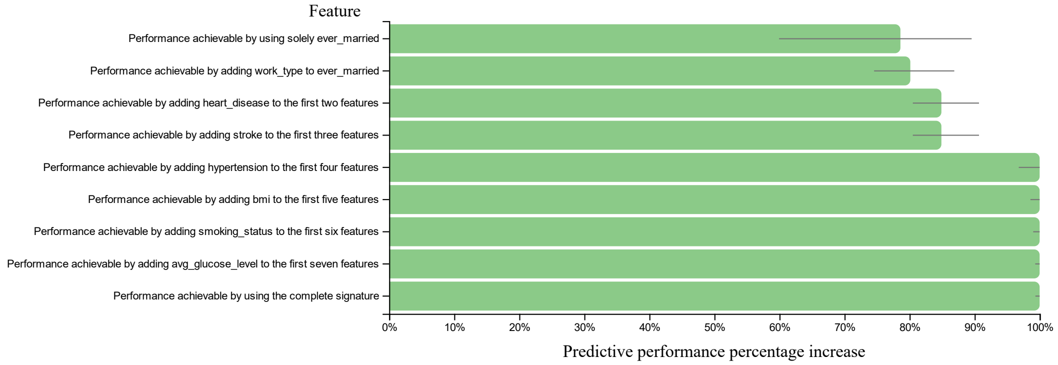
Metric	Mean estimate	CI
R-squared	0.655	[0.578, 0.718]
Mean Absolute Error	10.042	[9.173, 10.986]
Mean Squared Error	163.922	[137.492, 195.696]
Relative Absolute Error	0.549	[0.490, 0.616]
Relative Squared Error	0.347	[0.284, 0.425]
Correlation Coefficient	0.811	[0.763, 0.850]

Feature Selection

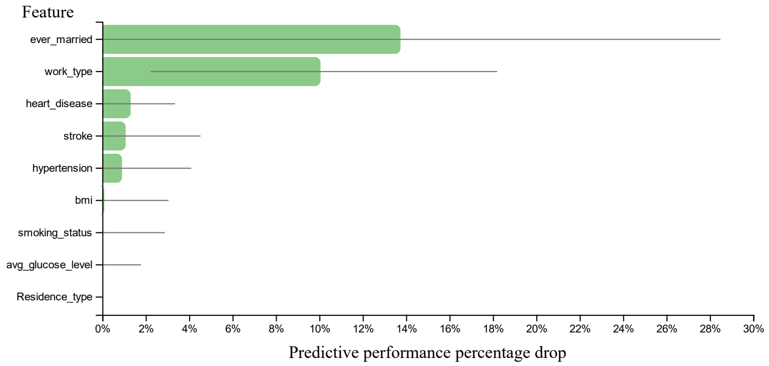
There were **9** features selected out of the **10** available.

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: **hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, stroke** in order of importance. The following features cannot be substituted with others and still obtain an equal predictive performance: **hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, stroke**.

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 Random Forest with Mean Squared Error splitting criterion	ntrees = 100, minimum leaf size = 5	0.6569819578742223	00:00:00.275	false
2	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	0.6467894968363956	00:00:02.2576	false
3	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Ridge Linear Regression	lambda = 1.0	0.6224556971522954	00:00:01.1619	false
4	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.001	Support Vector Regression Machines (SVR) of type epsilon-SVR	kernel = 'Linear Kernel', cost = 1.0, epsilon = 0.1	0.647533135581357	00:00:04.4797	false
5	IdentityFactory	FullSelector	-	Trivial model	-	2.220446049250313e-16	00:00:00.000	false
6	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	0.6619208606579539	00:00:01.1763	false

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
7	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	0.647533135581357	00:00:02.2168	false
8	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	0.6553688897991209	00:00:00.569	false
9	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.001	Regression Random Forest with Mean Squared Error splitting criterion	ntrees = 100, minimum leaf size = 5	0.652889984014042	00:00:04.4196	false
10	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	0.655760229933386	00:00:01.1930	false
11	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.001	Regression Random Forest with Mean Squared Error splitting criterion	ntrees = 100, minimum leaf size = 5	0.6551167368037157	00:00:04.4161	false