JADBio Description of Performed Analysis

Setup

JADBio version **1.4.174** ran on dataset **diabetes_prediction_dataset** with **100000** samples and **8** features to create a predictive model for outcome named **diabetes**. 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 **BACC** 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 **2**.
The execution time was **01:25:50**.

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
		equivalenceThreshold	0.01
		stoppingThreshold	0.01
	Test-Budgeted Statistically Equivalent Signature (SES)	maxK	2.0
		alpha	0.05
	Univariate	alpha	0.01
		maxVars	100
	LASSO	penalty	1.0
Modeling	Classification Decision Tree with Deviance splitting criterion	alpha	0.05
		minLeafSize	3
	Classification Random Forest with Deviance splitting criterion	nTrees	100
		minLeafSize	3.0

Leading to **17** 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, 21 models were set out to train.

JADBio Results Summary

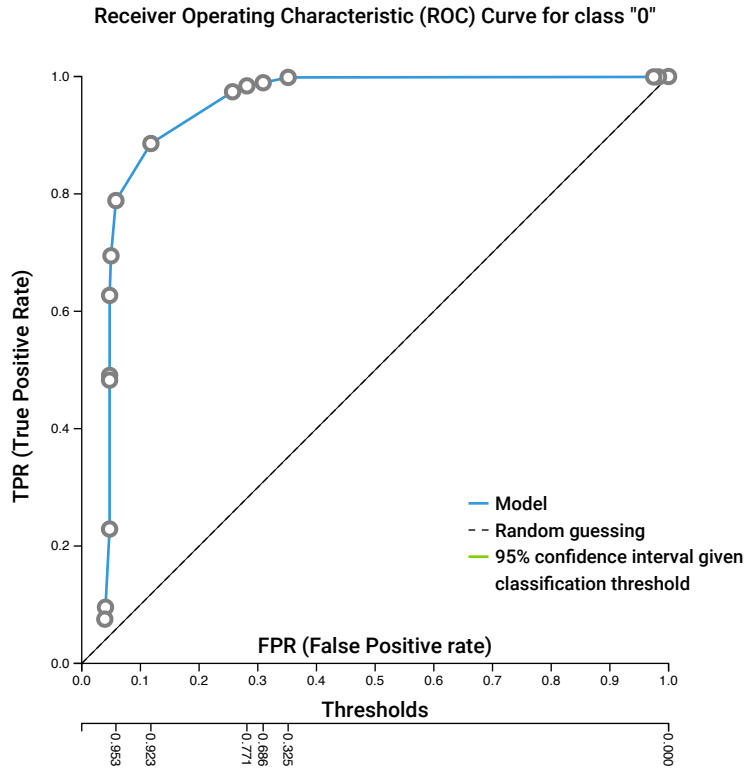
Overview

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 Decision Tree with Deviance splitting criterion and hyper-parameters: minimum leaf size = 3, and pruning parameter alpha = 0.05

The **Area Under The Curve** is **0.933** with 95% confidence interval being [**0.914,0.962**].
The **Mean Average Precision (a.k.a. Average Area Under the Precision-Recall curve)** is **0.896** with 95% confidence interval being [**0.876,0.922**].

The Balanced Accuracy is shown in the figure below:



Selecting to classify as class: **0** any sample with predicted probability to be in this class above **NaN**, the model achieves:

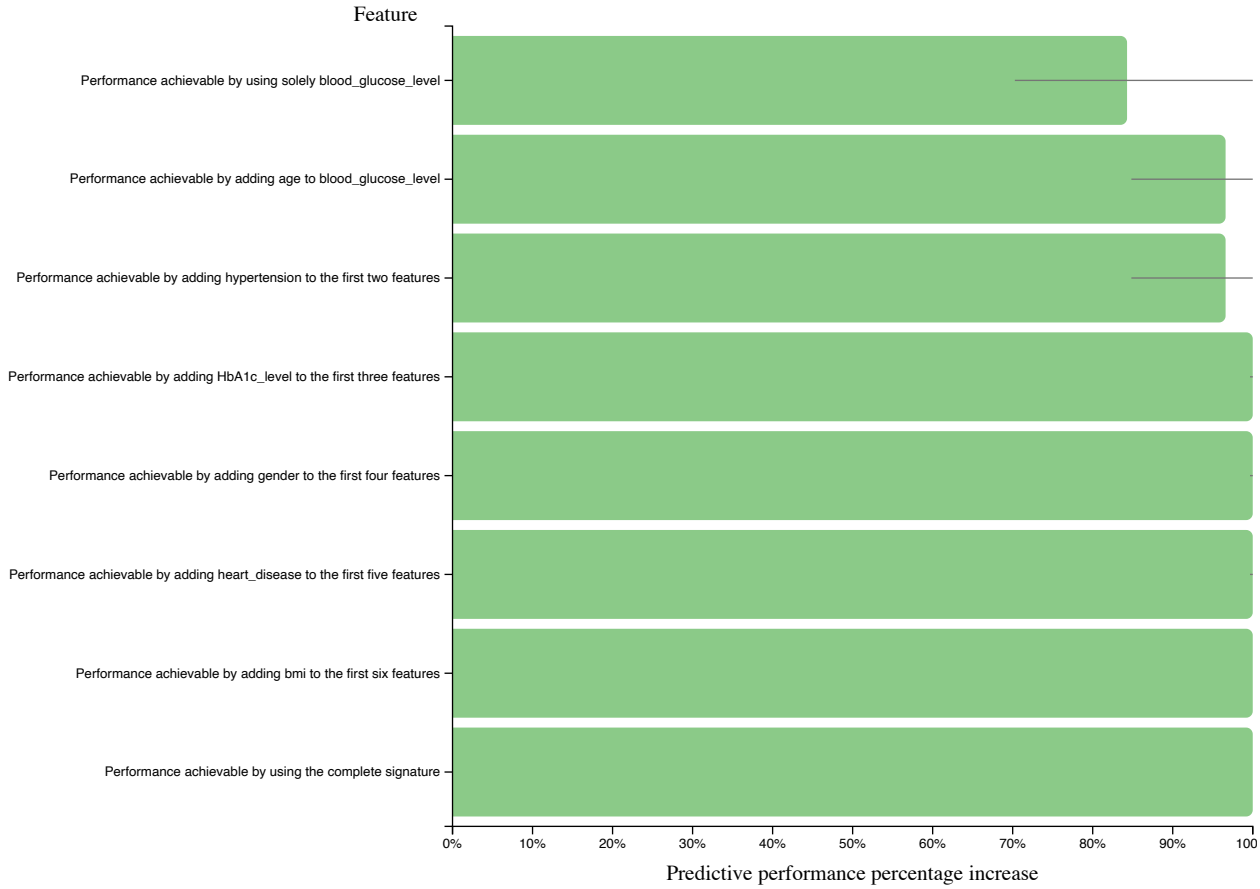
Metric	Mean estimate	CI

Feature Selection

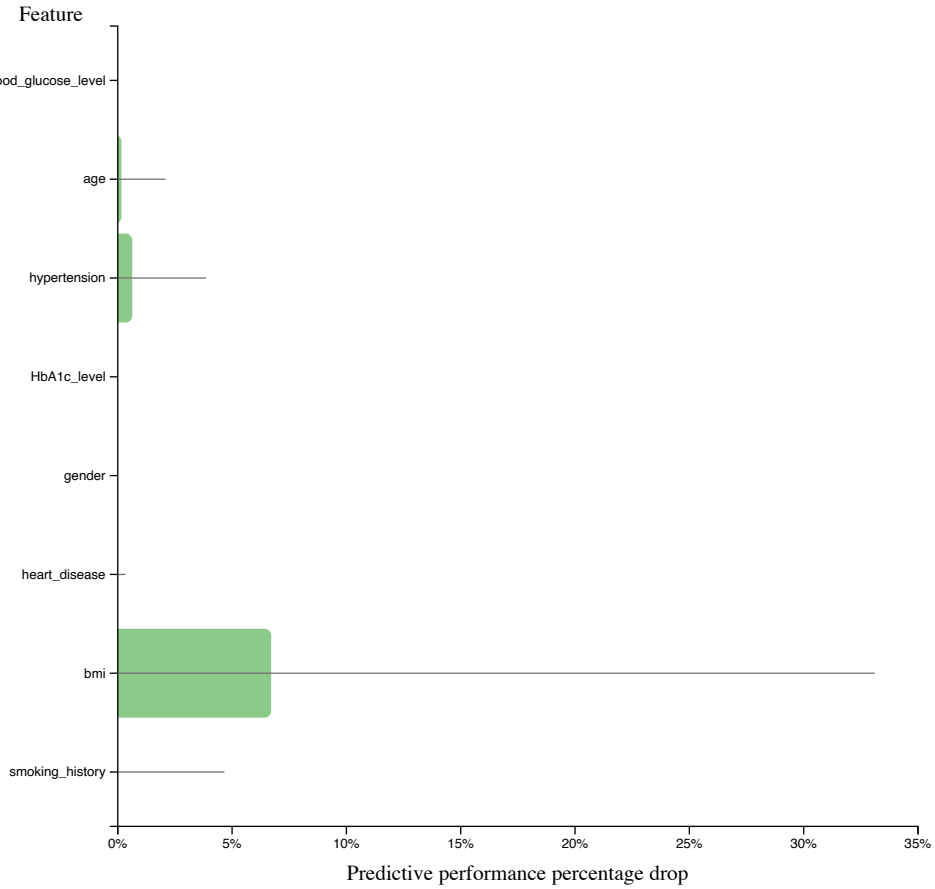
There were **8** features selected out of the **8** available.

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: **gender, age, hypertension, heart_disease, smoking_history, bmi, HbA1c_level, blood_glucose_level** in order of importance. The following features cannot be substituted with others and still obtain an equal predictive performance: **gender, age, hypertension, heart_disease, smoking_history, bmi, HbA1c_level, blood_glucose_level**.

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:

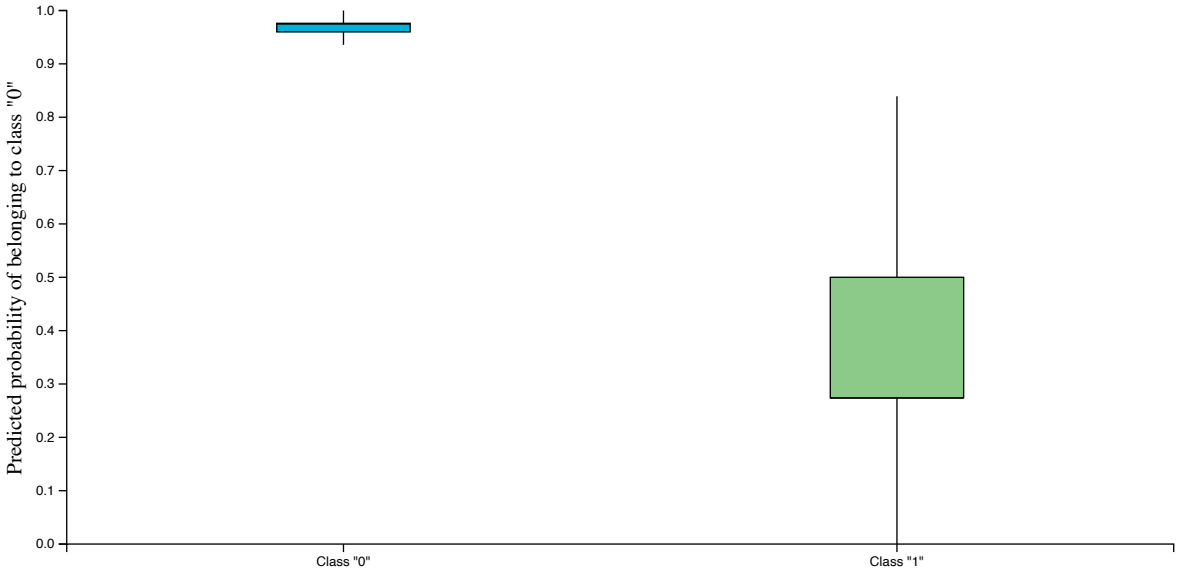


Some features may not seem to add predictive performance to the model; however, the feature selection algorithms include them as an effort to make the final model more robust to noise. The performances achieved by a model that contains all features except one, relative to the performance achieved when 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 = 0.01	Classification Decision Tree with Deviance splitting criterion	minimum leaf size = 3, alpha = 0.05	0.7185053037608486	00:00:11.11243	false
2	IdentityFactory	FullSelector	-	Trivial model	-	0.5	00:00:00.003	false
3	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	0.7006396657023465	00:00:12.12327	false
4	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Univariate feature selection with Benjamini-Hochberg correction	alpha = 0.01, max vars = 100	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.5852041144326583	00:00:09.9329	false
5	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.01	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.5866248794599807	00:00:11.11098	false
6	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.7094085503053681	00:00:18.18344	false
7	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.6029636772741883	00:00:16.16718	false
8	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	0.5852041144326583	00:00:09.9318	false
9	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.01	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.5866248794599807	00:00:11.11083	false
10	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Epilogi	equivThresh = 0.01, stopping criterion = Independence Test, stopping threshold = 0.01	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.7025233044037287	00:00:12.12754	false

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
11	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Univariate feature selection with Benjamini-Hochberg correction	alpha = 0.01, max vars = 100	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.5852041144326583	00:00:09.9309	false
12	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO	penalty = 1.0	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.6029636772741883	00:00:16.16715	false
13	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	0.719038894246223	00:00:16.16734	false
14	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	0.7185053037608486	00:00:08.8345	false
15	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Univariate feature selection with Benjamini-Hochberg correction	alpha = 0.01, max vars = 100	Classification Random Forest with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.7006396657023465	00:00:10.10890	false
16	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	0.5852041144326583	00:00:08.8937	false
17	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Univariate feature selection with Benjamini-Hochberg correction	alpha = 0.01, max vars = 100	Classification Decision Tree with Deviance splitting criterion	minimum leaf size = 3, alpha = 0.05	0.7185053037608486	00:00:09.9306	false