JADBio Description of Performed Analysis

Setup

JADBio version **1.4.69** ran on dataset **classification_peptides_binary** with **63400** samples and **400** features to create a predictive model for outcome named **feature0**. The outcome was discrete leading to a **classification** modeling.

The preferences of the analysis were set to **false** 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 **Preliminary**.
The number of CPU cores to use for the analysis was set to **1**.
The execution time was **03:10:05**.

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	Mode imputation		
	Mean imputation		
	Contant Removal		
	Standardization		
Feature Selection	Test-Budgeted Statistically Equivalent Signature (SES)	alpha	0.05
		maxk	2
	LASSO	penalties	1.0
	FullSelector		
Modeling	Polynomial Support Vector Machines	gammas	], costs=[
		costs	], degrees=[
		degrees	
	RBF Support Vector Machines	gammas	], costs=[
		costs	
	Logistic Regression	lambdas	1.0
	Random Forests	min leaf sizes	3
		vars to split	1.154 sqrt (nvars), 1.0 sqrt (nvars), 0.816 sqrt (nvars)
		splits to perform	1.0
		ntrees	100
	Decision Tree	min leaf sizes	3
		vars to split	nvars // 1.0

Algorithm Type	Algorithm	Hyper-parameter	Set of Values
		splits to perform	1.0
		alphas	0.05

Leading to **16** 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, 16 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	FullSelector	Classification Random Forests training 100 trees with Deviance splitting criterion, minimum leaf size = 3, and variables to split = 0.816 sqrt (nvars)

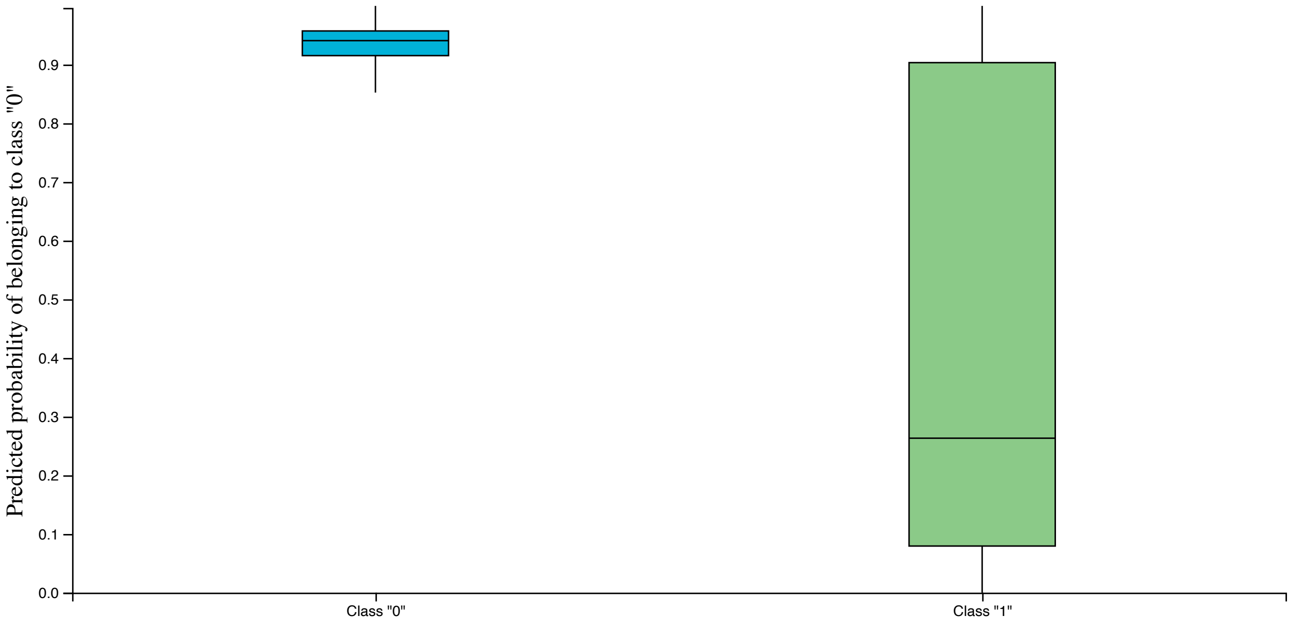
The **Area Under The Curve** is **0.878** with 95% confidence interval being [**0.855,0.902**].
The **Mean Average Precision (a.k.a. Average Area Under the Precision-Recall curve)** is **0.889** with 95% confidence interval being [**0.870,0.909**].

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

Feature Selection

Jadbio selected **all** features in the original dataset for the reference signature. Note that **375** features that were found constant are excluded.

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	Test-Budgeted Statistically Equivalent Signature (SES) algorithm	maxK = 2, alpha = 0.05, budget = 3 * nvars	Classification Random Forests with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.8036284115770936	00:03:09.189804	true
2	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES) algorithm	maxK = 2, alpha = 0.05, budget = 3 * nvars	Ridge Logistic Regression	lambda = 1.0	0.7750902023241212	00:03:08.188789	true
3	Mean Imputation, Mode Imputation, Constant Removal, Standardization	FullSelector	-	Ridge Logistic Regression	lambda = 1.0	0.8364060274444931	00:00:19.19835	false
4	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO Feature Selection	penalty = 1.0	Ridge Logistic Regression	lambda = 1.0	0.7736133060229292	00:07:56.476420	true

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
5	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES) algorithm	maxK = 2, alpha = 0.05, budget = 3 * nvars	Classification Random Forests with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.801961259179293	00:03:09.189574	true
6	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO Feature Selection	penalty = 1.0	Classification Random Forests with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.8158630894963035	00:07:57.477278	true
7	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES) algorithm	maxK = 2, alpha = 0.05, budget = 3 * nvars	Classification Decision Tree with Deviance splitting criterion	minimum leaf size = 3, alpha = 0.05	0.371297189463876	00:03:10.190060	true
8	Mean Imputation, Mode Imputation, Constant Removal, Standardization	FullSelector	-	Classification Random Forests with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.8794206322166996	00:00:11.11542	false
9	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO Feature Selection	penalty = 1.0	Classification Random Forests with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.8186801932525788	00:07:57.477559	true
10	Mean Imputation, Mode Imputation, Constant Removal, Standardization	Test-Budgeted Statistically Equivalent Signature (SES) algorithm	maxK = 2, alpha = 0.05, budget = 3 * nvars	Classification Random Forests with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.8036284115770936	00:03:10.190021	false
11	IdentityFactory	NoSelector	-	Trivial model	-	0.5000000000000001	00:00:00.000	false
12	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO Feature Selection	penalty = 1.0	Classification Decision Tree with Deviance splitting criterion	minimum leaf size = 3, alpha = 0.05	0.4819473234846214	00:07:57.477843	true

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
13	Mean Imputation, Mode Imputation, Constant Removal, Standardization	FullSelector	-	Classification Decision Tree with Deviance splitting criterion	minimum leaf size = 3, alpha = 0.05	0.4153376789166692	00:00:05.5773	true
14	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO Feature Selection	penalty = 1.0	Classification Random Forests with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.8186801932525788	00:07:57.477569	false
15	Mean Imputation, Mode Imputation, Constant Removal, Standardization	FullSelector	-	Classification Random Forests with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.8784697622106655	00:00:14.14636	false
16	Mean Imputation, Mode Imputation, Constant Removal, Standardization	FullSelector	-	Classification Random Forests with Deviance splitting criterion	ntrees = 100, minimum leaf size = 3	0.8794064707596714	00:00:15.15782	false