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 **true** 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 **02:58:49**.

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
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
		splits to perform	1.0
		alphas	0.05

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, 11 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 Random Forests training 100 trees with Deviance splitting criterion, minimum leaf size = 3, and variables to split = 1.0 sqrt (nvars)

The **Area Under The Curve** is **0.818** with 95% confidence interval being [**0.792,0.844**].
The **Mean Average Precision (a.k.a. Average Area Under the Precision-Recall curve)** is **0.816** with 95% confidence interval being [**0.794,0.838**].

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

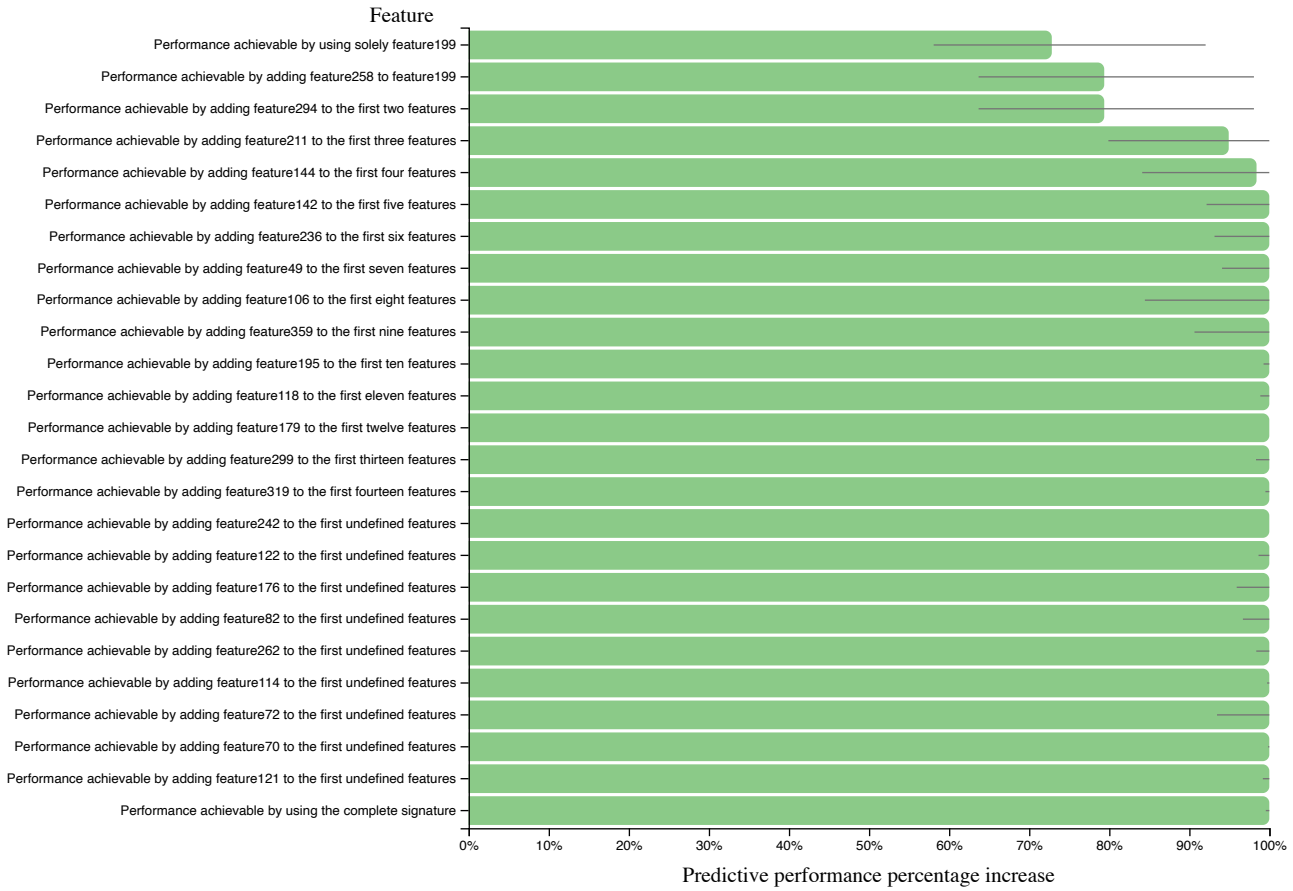
Feature Selection

There were **25** features selected out of the **400** available.

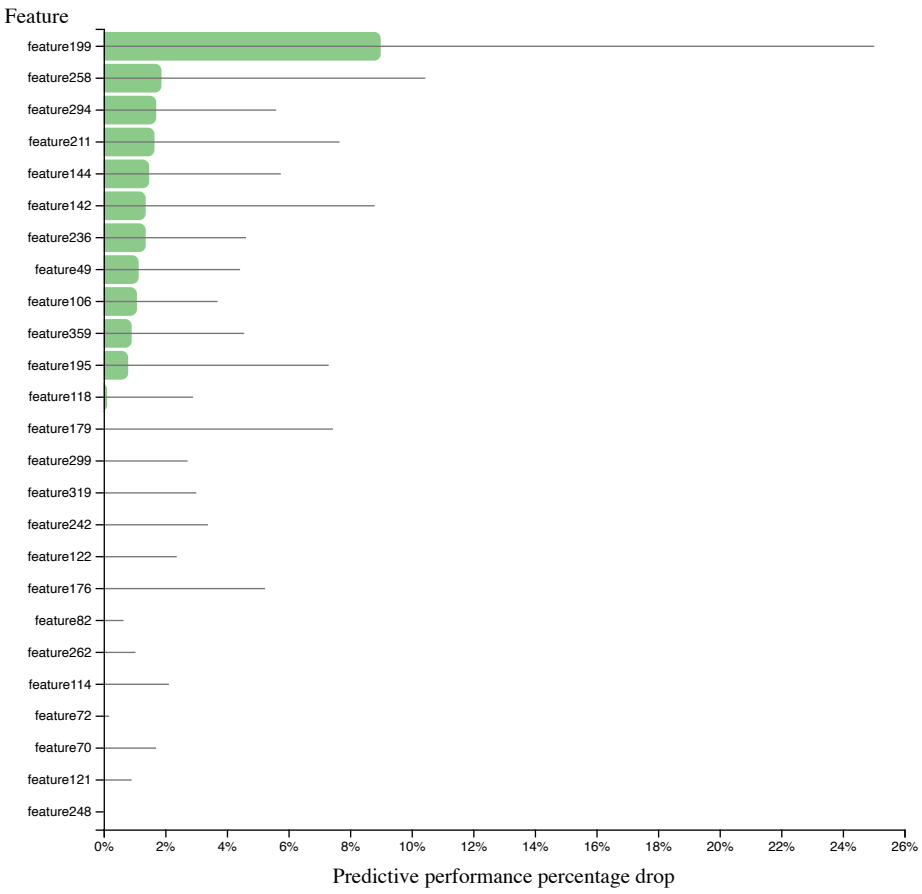
The selected features consist of the following subset called a signature. **There was a single signature identified.** The first signature identified by the system is the set: **feature49, feature70, feature72, feature82, feature106, feature114, feature118, feature121, feature122, feature142, feature144, feature176, feature179, feature195, feature199, feature211, feature236, feature242, feature248, feature258, feature262, feature294, feature299, feature319, feature359** in order of importance. The following features cannot be substituted with others and still obtain an equal predictive performance: **feature49, feature70, feature72, feature82, feature106, feature114, feature118, feature121, feature122, feature142, feature144, feature176, feature179, feature195, feature199, feature211, feature236, feature242, feature248, feature258, feature262, feature294, feature299, feature319, feature359**.

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:

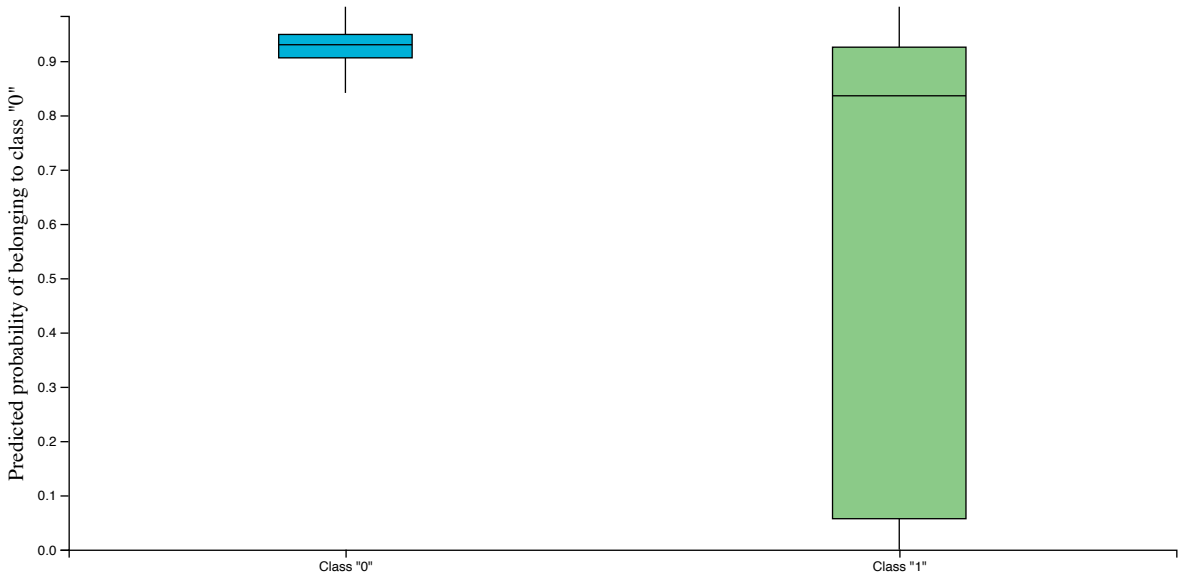


Some features may not seem to add predictive performance to the model; however, the feature selection algorithms include them as an effort to make the final model more robust to noise. The performances achieved by a model that contains all features except one, relative to the performance achieved when 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	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	00:03:06.186634	true
2	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	00:03:06.186375	true
3	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	00:08:02.482703	false
4	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	00:03:06.186387	false
5	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	00:03:05.185278	false
6	IdentityFactory	NoSelector	-	Trivial model	-	0.5000000000000001	00:00:00.000	false
7	Mean Imputation, Mode Imputation, Constant Removal, Standardization	LASSO Feature Selection	penalty = 1.0	Ridge Logistic Regression	lambda = 1.0	0.7736133060229292	00:08:01.481152	true
8	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	00:08:02.482562	true
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	00:08:02.482662	false

Configuration	Preprocessing	Name	Hyperparams	Name	Hyperparams	Performance (unadjusted)	Time (miliseconds)	Dropped
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	0.801961259179293	00:03:06.186207	true
11	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	00:08:02.482302	false