

Tutorial for JADBio Classification analysis – Potatoes' quality Case study

INTRODUCTION	2
ACCESSING JADBio	3
Try for free (Sign up).....	3
To Log In (Sign in)	3
VIEWING DATA AND ANALYSES IN JADBio	4
JADBio header	4
Menu	4
Dashboard.....	5
Projects.....	8
Datasets	10
Analyses.....	13
Applied Models	14
WORKING IN JADBio	15
Creating a Project	15
Uploading a Dataset.....	16
Transformations.....	20
Background	23
Analysis.....	24
Analysis Results	28
Best Performing Model	30
Performance Overview	32
Feature Selection	36
Analysis Visualization	39
Apply Model.....	42

Introduction

JADBio is a platform designed specifically to extract value and insight from multi-omics data sets, typically thousands of measurements in a small number of samples. JADBio's uniquely tuned Automated Machine Learning (AutoML) is guided by Artificial Intelligence (AI) to provide accurate and efficient predictive models for Classification, Regression and Survival (Time to Event) analysis.

JADBio is very straightforward to use:

1. Upload your data
2. Transform your data (optional)
3. Analyze your data
4. View your results (optional)
5. Share your results
6. Test your data (optional)

Accessing JADBio

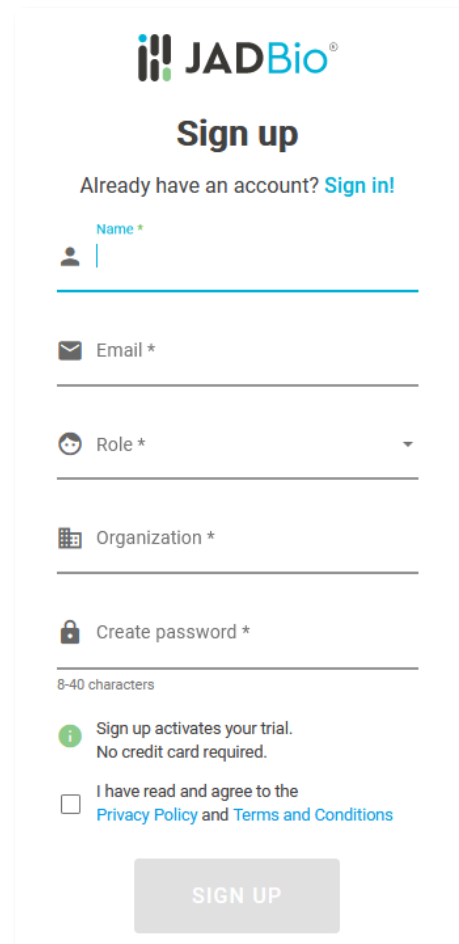
- In your browser of choice, navigate to JADBio.com.
- Click on the **LOG IN** button in top right corner if you already have an account or the **TRY FOR FREE** button if you don't.

Try for free (Sign up)

- Complete the Register Dialog.
- Click on **SIGN UP**.

After you register, you will receive an e-mail from no-reply@gnosisd.gr with the subject, JADBio Email Verification.

- To verify your email and activate your account, please open the e-mail, and click on the provided link.



The image shows a 'Sign up' dialog box for JADBio. At the top is the JADBio logo. Below it is the title 'Sign up' and a link 'Already have an account? Sign in!'. The form contains several fields: 'Name *' with a person icon, 'Email *' with an envelope icon, 'Role *' with a person icon and a dropdown arrow, 'Organization *' with a building icon, and 'Create password *' with a lock icon. Below the password field is a note '8-40 characters'. There is an information icon (i) next to the text 'Sign up activates your trial. No credit card required.' At the bottom, there is a checkbox for 'I have read and agree to the Privacy Policy and Terms and Conditions', where 'Privacy Policy' and 'Terms and Conditions' are links. A large 'SIGN UP' button is at the very bottom.

Figure 1 Registration dialog

To Log In (Sign in)

- Type in your username and password and click on **SIGN IN**.

Viewing data and analyses in JADBio

JADBio header

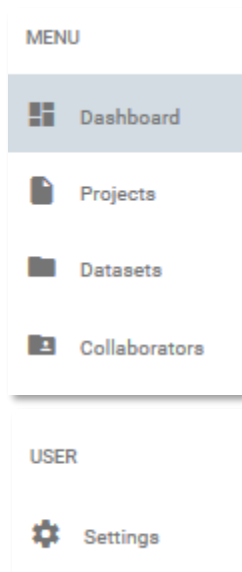
The **JADBio header** contains:

- Notifications under the bell icon on the right.
- Sign on and off function under your username.



Figure 2 JADBio header

Menu



MENU sidebar provides navigation to the **Dashboard**, to your **Projects**, to your **Datasets** and to your **Collaborators**, as well as to the **USER Settings**.

Figure 3 MENU and USER sidebars

Dashboard

The **Dashboard** provides an overview of your Account. This includes:

1. **Projects:** A graphic representation of shared and exclusive projects. To start, you will have one shared project.

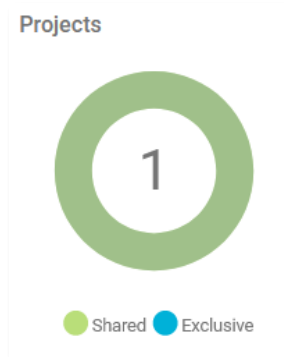


Figure 4 Projects

2. **Infrastructure Statistics:** A summary of your current Storage and Compute. Standard subscriptions include 2 or 6 cores.

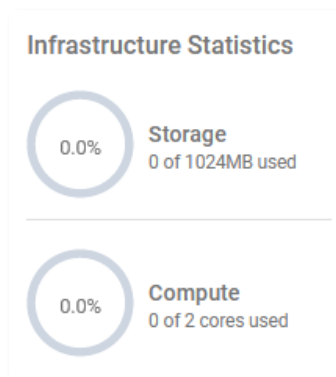


Figure 5 Infrastructure Statistics – the number of available cores is dependent on your subscription

3. **Datasets:** The total number of your Datasets.

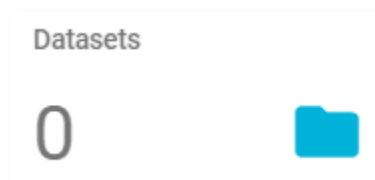


Figure 6 Datasets

4. **Analyses:** The total number of your Analyses.



Figure 7 Analyses

5. **Recently Shared:** A list of your current Collaborators. When you share a project with another subscriber, they become a collaborator, and your project is defined as shared.

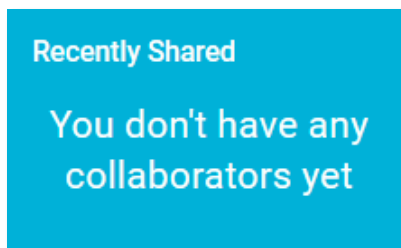


Figure 8 Recently Shared

6. **Active Subscription:** Your Subscription Plan and its expiration date.

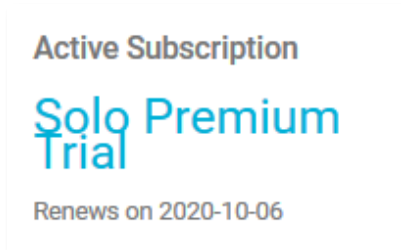


Figure 9 Subscription Plan

7. **Collaborators:** The total number of your Collaborators, other subscribers with whom you have shared projects.

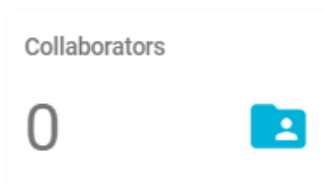


Figure 10 Collaborators

8. **Recent Analyses:** A list of all currently running Analyses.

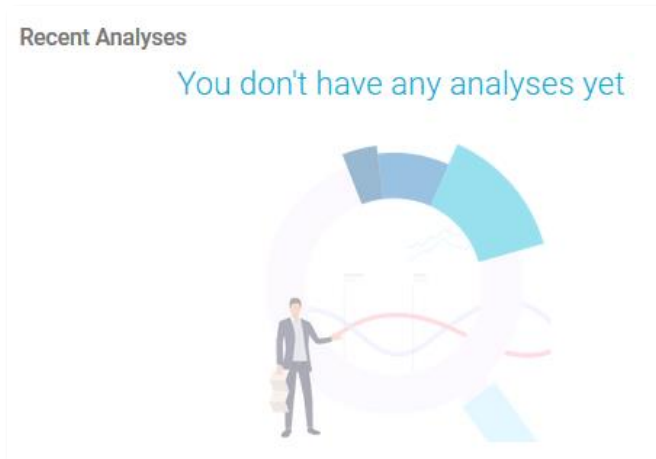


Figure 11 Recent Analyses

9. **Recent Projects:** A list with links to your current Project.





Recent Projects				
Name	Updated	Owner	Dataset	Actions
 JAD Use Cases - Demo Use Cases f	2020-09-...	vlagani		...

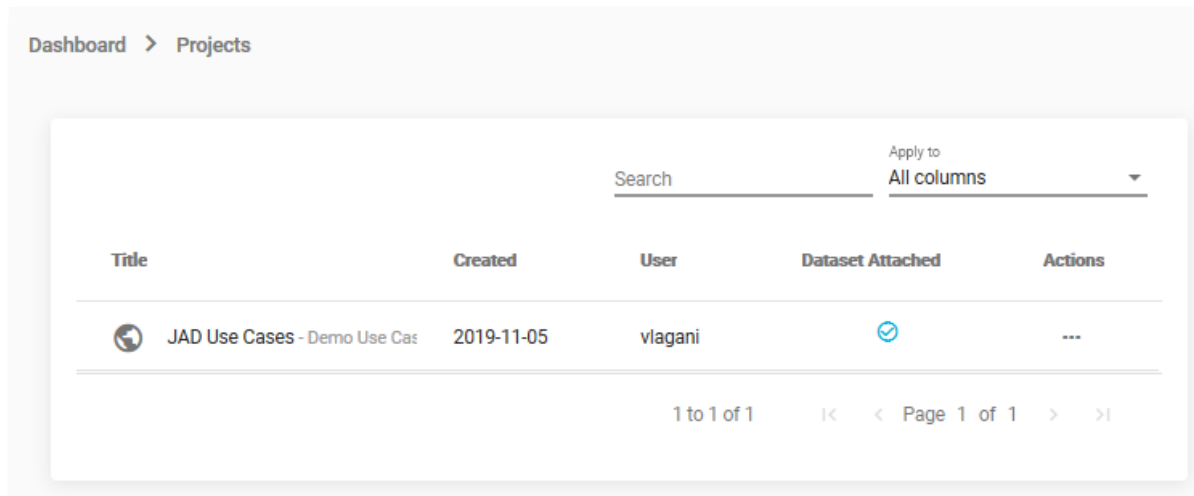
Figure 12 Recent Projects

Also, there is a  button in the top right of your dashboard that allows you to create a project and a  button that that open your **Projects** window.

Projects

- Click on the **LIST PROJECTS** button.

Note: The **Projects** includes a shared project, **JAD Use Cases**, which includes several datasets in order for you to have some examples to work with in JADBio. You cannot change the data in this public project, but, as you can see later, you will be able to import this data into another project.





Title	Created	User	Dataset Attached	Actions
 JAD Use Cases - Demo Use Cas	2019-11-05	vlagani		...

Figure 13 Projects list

Each project is described by the following characteristics:

1. An icon that describes the number of collaborators and sharing status. The **JAD Use Cases**, has a global icon, because this data is shared publicly.
2. The name of the project, and a short description, both of which can be used as values for a project search in the **Filter** tool at the top of the **Projects** window.
3. The date the project was **Created**.
4. The initial creator of the Project.
5. A flag to indicate the presence of a dataset.
6. **Actions** that will allow you to view or remove the project.

- For **JAD Use Cases**, under **Actions**, Click on the **View project** icon.

Within the **JAD Use Cases** project you will have several layers of information:

1. Below the JADBio header, you will now see that you are in **Dashboard > Projects > JAD Use Cases**.

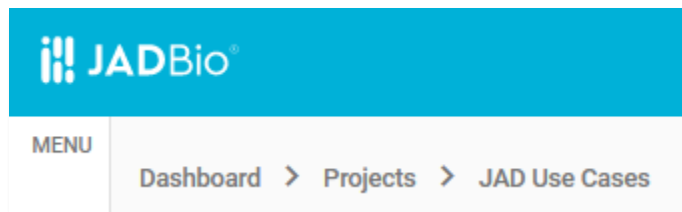


Figure 14 Breadcrumbs, JAD Use cases

2. In the top left, in the **ACTIONS** sidebar, you will have five buttons that will allow you to perform a variety of functions with the project:
 - a. **Add Data**
 - b. **Transform Data**
 - c. **Analyze Data**
 - d. **Apply Model**
 - e. **Delete Project**
3. In the same sidebar, JADBio provides the option to view/add **collaborators** to your Project.
4. In the **PROJECT DETAILS** sidebar, JADBio provides information about your Project: **Name**, **Owner**, and a short **Description** of your Project, if one was created.
5. JADBio provides navigation to three different views of your Project window: **Datasets**, **Analyses**, and **Applied Models**.

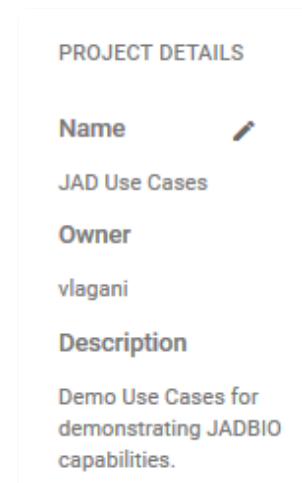


Figure 15 Project info

Datasets

- Click on the **Datasets** label, to view the available datasets in the **JAD Use Cases** project.

Datasets includes a description of all of the datasets in a tabular format:

- Name
- Created
- Features
- Samples
- Size
- Under the **Actions** column, JADBio provides buttons to:
 - Preview Dataset
 - Perform Analysis
 - Detach Dataset

Datasets Analyses Applied Models					
Search					Apply to All columns
Name	Created	Features	Samples	Size	Actions
COVID-19_GSE152075 - RNA	2020-08-13	35787	484	38.11MB	---
BrCa_paired_methylation - T	2020-07-24	485580	116	871.70MB	---
BrCa_paired_transcriptomic	2020-07-01	48703	114	45.21MB	---
MultiClassExpressionCance	2020-06-04	16064	190	11.60MB	---
Psoriasis_SNP_test - The ori	2020-05-13	1082	2250	12.20MB	👁️ ▶️ ✕
Psoriasis_SNP_train - The ori	2020-05-13	1082	5103	27.65MB	---
P_falciparum_CNV - This dat	2020-04-27	11374	183	15.46MB	---
Oyster_transcriptomics - Thi	2020-04-26	31921	177	40.56MB	---
Multiple_sclerosis_proteomi	2020-04-24	9304	80	5.80MB	---

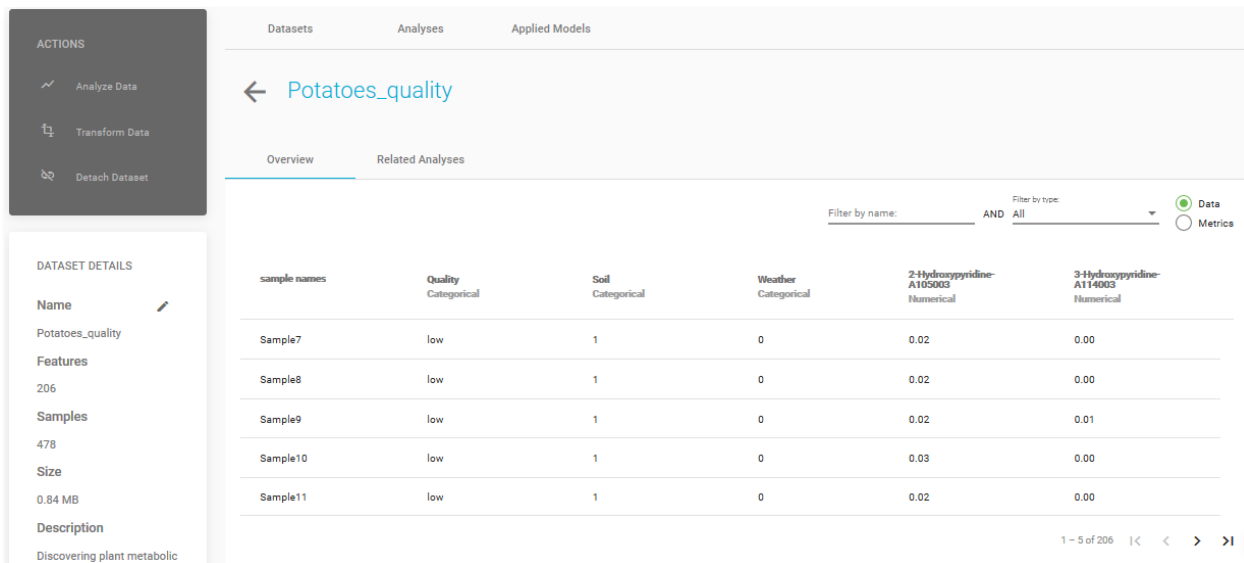
Figure 16 JAD Use Cases, Datasets

- Find the **Potatoes_quality** dataset.
- Click on the three dots, to open the **Actions** menu associated with the **Potatoes_quality** dataset.

- Click on the **Preview dataset**  icon.

In Preview dataset:

- ACTIONS** sidebar, provides 3 possible actions to perform with the dataset: to **Analyze Data**, to **Transform Data** or to **Detach Dataset**.
- DATASET DETAILS** sidebar, includes: **Name**, number of **Features**, number of **Samples**, file **Size** and a short **Description** of the dataset, if one was created.
- Overview** displays the dataset's column and row labels for the first five samples and first five features with their assigned data types. Tools to navigate or Filter, by name or by type, are embedded in the preview.
- Related Analyses** displays the previous run analyses of the dataset, when the data have been analyzed previously, as in this Project.



The screenshot shows the JADBio interface for the 'Potatoes_quality' dataset. On the left, the 'ACTIONS' sidebar offers 'Analyze Data', 'Transform Data', and 'Detach Dataset'. Below it, the 'DATASET DETAILS' sidebar lists the dataset name, 206 features, 478 samples, 0.84 MB size, and the description 'Discovering plant metabolic'. The main panel shows the 'Overview' tab for 'Potatoes_quality'. It includes a table with 6 columns: 'sample names', 'Quality' (Categorical), 'Soil' (Categorical), 'Weather' (Categorical), '2-Hydroxypyridine-A105003' (Numerical), and '3-Hydroxypyridine-A114003' (Numerical). The table displays data for samples Sample7 through Sample11. At the top right of the table, there are filter options for 'Filter by name' and 'Filter by type' (Data/Metrics).

sample names	Quality Categorical	Soil Categorical	Weather Categorical	2-Hydroxypyridine- A105003 Numerical	3-Hydroxypyridine- A114003 Numerical
Sample7	low	1	0	0.02	0.00
Sample8	low	1	0	0.02	0.00
Sample9	low	1	0	0.02	0.01
Sample10	low	1	0	0.03	0.00
Sample11	low	1	0	0.02	0.00

Figure 17 Preview dataset, Overview

- Select the **Metrics** radio button to view graphical and numerical summaries of each feature in the dataset (i.e. histograms).

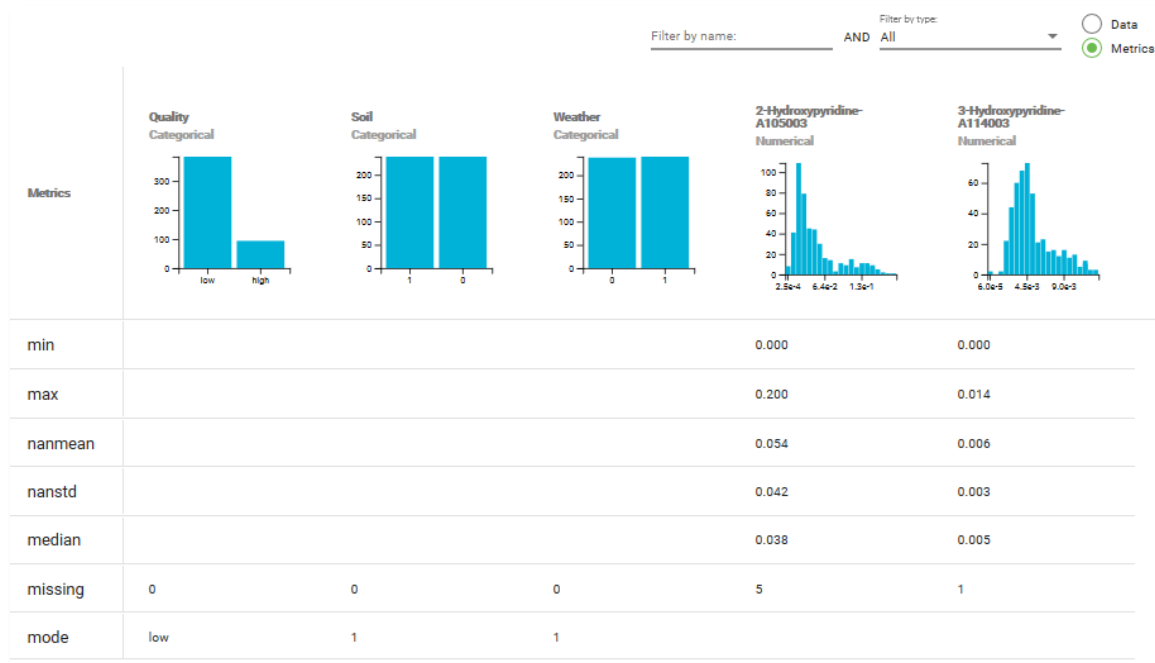


Figure 18 Metrics view

In the **Metrics** view, For the quality feature, one can see that of the 478 samples, 384 are in the low-quality group and 94 are in the high quality group.

Under **Actions**, JADBio gives you the option to **Perform Analysis** with this dataset. We will perform an analysis with this data later in the tutorial, but in a new Project.

Analyses

- Click on the **Analyses** label, to view the previously run analyses in the **JAD Use Cases** project.





















<div> Datasets Analyses Applied Models </div>								
<div> <div>Search</div> <div>Apply to</div> <div>All columns</div> </div>					<div> Columns to display Name, Started, Dataset, M... </div>			
Name	Started	Dataset	Metric	Metric Value	Outcome	Type	Progress	Actions
 Alzheimer's_transc	2020-09-07 11:34	Alzheimer's_transc	AUC	0.98	diagnosis		FINISHED (100%)	...
 COVID-19_GSE152	2020-09-02 23:53	COVID-19_GSE1520	AUC	0.96	covid		FINISHED (100%)	...
 Tuberculosis_prote	2020-07-15 12:39	Tuberculosis_prote	AUC	0.77	Status		FINISHED (100%)	...
 Parkinson's_diseas	2020-07-02 23:16	Parkinson's_diseas	R2	0.84	total_UPDRS		FINISHED (100%)	...
 Chemosensitivity_I	2020-07-02 16:56	Chemosensitivity_P	AUC	0.78	target		FINISHED (100%)	...
 Chemosensitivity_I	2020-07-02 16:46	Chemosensitivity_P	AUC	0.79	target		FINISHED (100%)	...
 Parkinson's_diseas	2020-07-02 08:37	Parkinson's_diseas	R2	0.84	total_UPDRS		FINISHED (100%)	...
 MultiClassExpressi	2020-06-26 17:39	MultiClassExpressi	AUC	0.90	class		FINISHED (100%)	...
 Tuberculosis_prote	2020-06-25 18:53	Tuberculosis_prote	AUC	0.77	Status		FINISHED (100%)	...
 Oyster_transcriptoi	2020-06-25 18:17	Oyster_transcriptoi	AUC	0.92	Population		FINISHED (100%)	...

Figure 19 JAD Use Cases, Analyses

Within the **Analyses** page, JADBio describes all analyses, including those currently running. Furthermore, the analysis type is described by an icon. Here, we have **Regression**, **Classification**, and **Survival (Time to Event)** analyses.

Analyses also includes a description in tabular format:

- Name
- Started (date)
- Dataset
- Metric
- Metric Value
- Outcome
- Type
- Progress
- Actions, includes: View results, Apply model and Remove results buttons

Applied Models

- Click on the **Applied Models** label, to view the results from the reapplication of a model on a novel or test dataset.

Applied Models description includes:

- Dataset
- Analysis used
- Progress
- Actions, includes:
 - View Results
 - Remove Applied Model

Datasets Analyses <u>Applied Models</u>			
<div> <div>Search</div> <div>Apply to All columns ▼</div> </div>			
Dataset	Analysis used	Progress	Actions
Microbiome_test	Microbiome_training_typical	FINISHED (100%)	...
Psoriasis_SNP_test	Psoriasis_SNP_train_preliminary	FINISHED (100%)	...

1 to 2 of 2 |< < Page 1 of 1 > >|

Figure 20 JAD Use Cases, Applied Models

Working in JADBio

Creating a Project

All of the datasets and analyses within the **JAD Use Cases** project are read only. In order to start working with data, you must create a new project. For this example, you can create a new project and upload data from the **JAD Use Cases** project.

- On the **MENU** sidebar at the left of the JADBio window, click on **Projects**.

This will bring you to the **Projects** window, which includes the **Create a new Project** function.

- Click on **CREATE** button.
- Within the **Name your Project** dialog, enter the following **Project title** and **Project Description**.

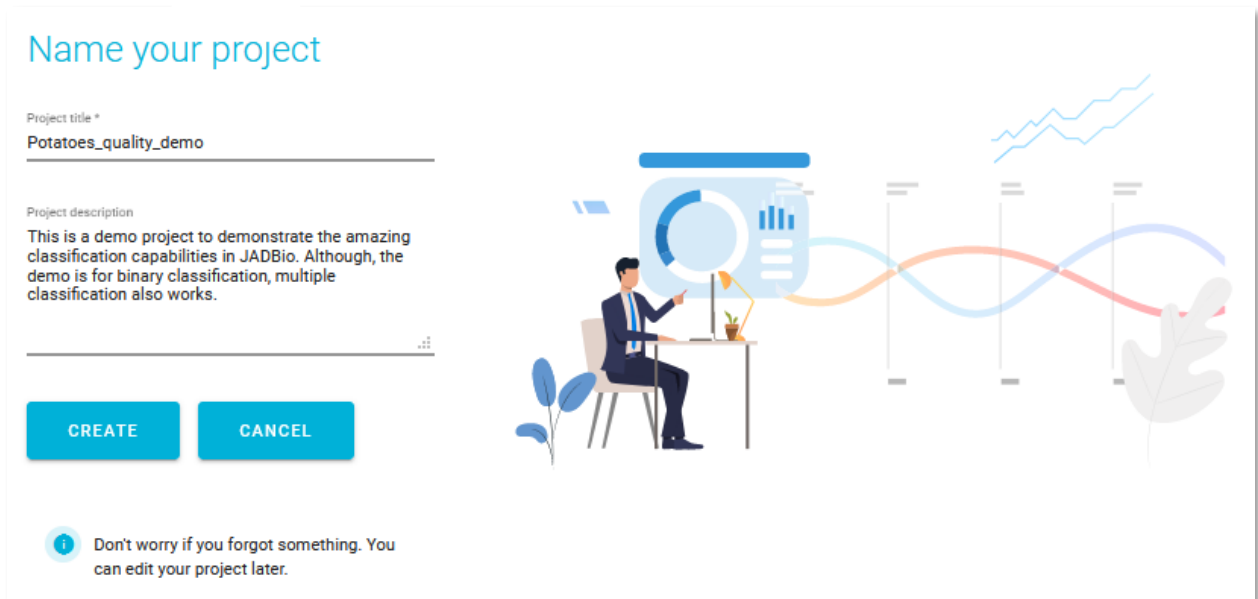


Figure 21 Creating a new project

- Click on the **CREATE** button.

Your new project will now appear in the **Projects** window.

Uploading a Dataset

- Under **ACTIONS**, in the top left corner, click **Add Data** icon to add a dataset to this project.

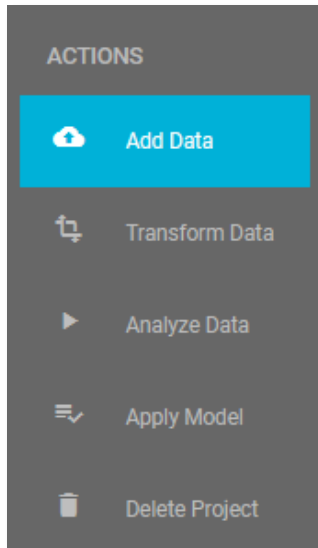


Figure 22 ACTIONS options in project window, prior to the upload of any datasets.

You will have two options. If you were working with an external file, you would select the **Upload a file** option, but for this demonstration, you will use the existing **Potatoes_quality** dataset from the **JAD Use cases** project.

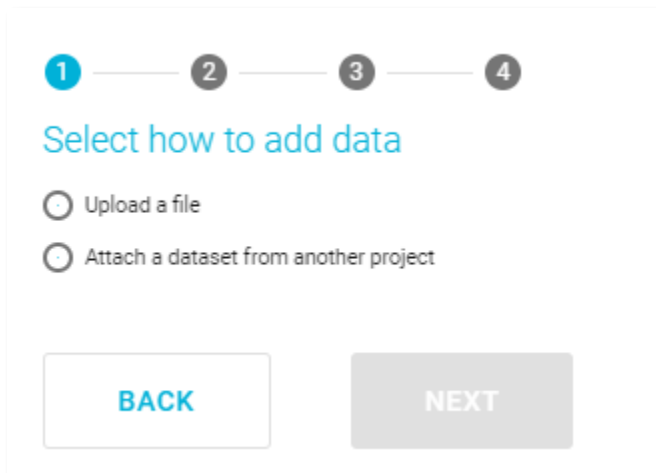


Figure 23 Add Data, Select how to add data

- Click on the radio button, **Attach a dataset from another project**.

JADBio will now present you will a list of datasets in a tabular format.

- Select the **Potatoes_quality** dataset. (Note: You may have to search for it or scroll down to see it.)



✓ — 2 — 3 — 4

Select a dataset to attach to this project

Search: Apply to:

Name	Created	Features	Samples	Projects
<input checked="" type="checkbox"/> Potatoes_quality	2019-11-29	206	478	Potatoes_Quality-Demo, Steven Test, al...

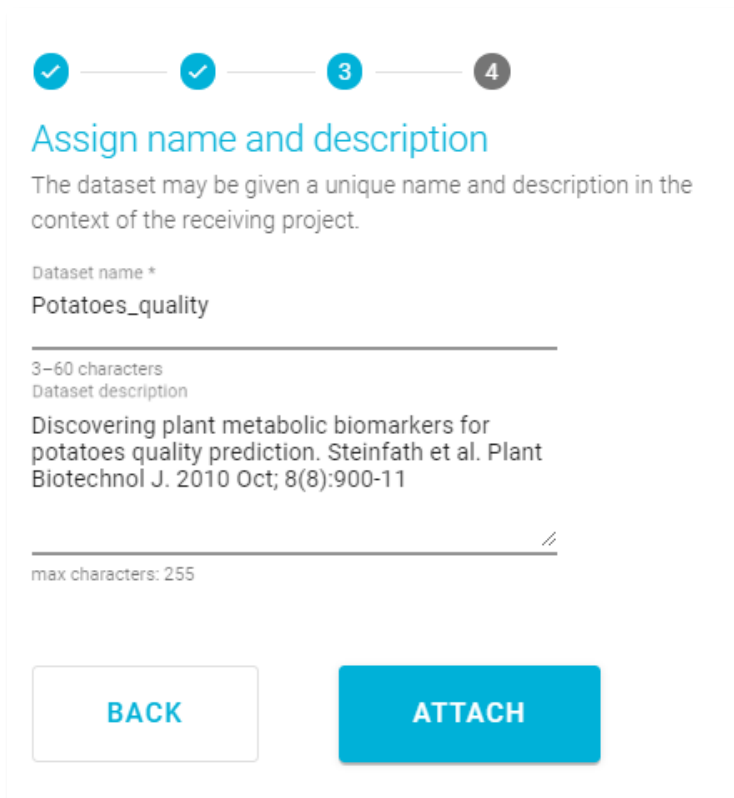
1 to 1 of 1 |< < Page 1 of 1 > >|

Figure 24 Select a dataset to attach

- Click the **NEXT** button.

This opens the **Assign name and a description** window. Here, you are able to change your Dataset's name and add a description for your dataset.

- Click the **ATTACH** button, to connect this dataset with your project.



✓ — ✓ — 3 — 4

Assign name and description

The dataset may be given a unique name and description in the context of the receiving project.

Dataset name *

Potatoes_quality

3–60 characters

Dataset description

Discovering plant metabolic biomarkers for potatoes quality prediction. Steinfath et al. Plant Biotechnol J. 2010 Oct; 8(8):900-11

max characters: 255

BACK ATTACH

Figure 25 Assign name and description

This opens the **Assign feature types** window. JADBio automatically assigns feature types based on the content of each column, but you can change the feature types here or in a transformation step to suit your understanding of your data. Here, you can change the type of features i.e. to assign an event or a time-to-event feature type. Also, as in the **Preview dataset** window, you are able to navigate across the dataset, search and filter the dataset, and view the metrics of each feature in the **Metrics** view.

Note: Feature types are critical to the process of using the JADBio platform. If you are not familiar with feature types, please reach out to technical support for additional information.

✓

✓

✓

4

Assign feature types

Almost done! The dataset is now part of your project. If needed, it can be given different feature types in this context.

sample names	Quality Categorical		Weather Categorical
Sample7	low		0
Sample8	low		0
Sample9	low	1	0
Sample10	low	1	0
Sample11	low	1	0

SKIP

SUBMIT

Numerical

Categorical

Event

Time to Event

Identifier

Figure 26 Assign feature types

- Click **SKIP** to move the data, without changing any feature type, into your project.

You are now in the **Datasets** window of the **Potatoes_quality_demo** Project. You can see that your new dataset has 206 Features and 478 Samples. Under the **Actions** header on the three dots, you have three options: **Preview Dataset**, **Perform Analysis** and **Detach Dataset** from project.




Datasets		Analyses		Applied Models	
				Search	Apply to All columns
Name	Created	Features	Samples	Size	Actions
Potatoes_quality - Discoverii	2019-11-29	206	478	0.84MB	  
1 to 1 of 1 < < Page 1 of 1 > >					

Figure 27 Uploaded dataset.

Transformations

JADBio provides numerous transformation tools today and we will continue to augment this functionality to support your needs. For this demonstration, you will split the data into two datasets, one for training and one for testing.

- Click on **Transform Data** in the **ACTIONS** sidebar.

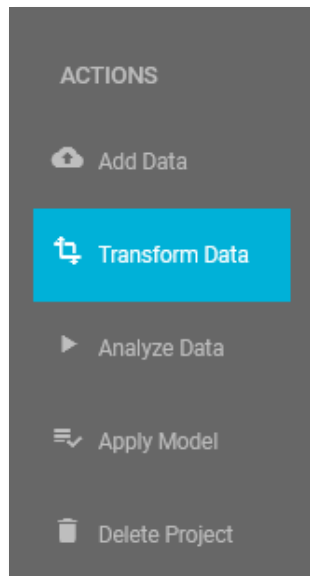
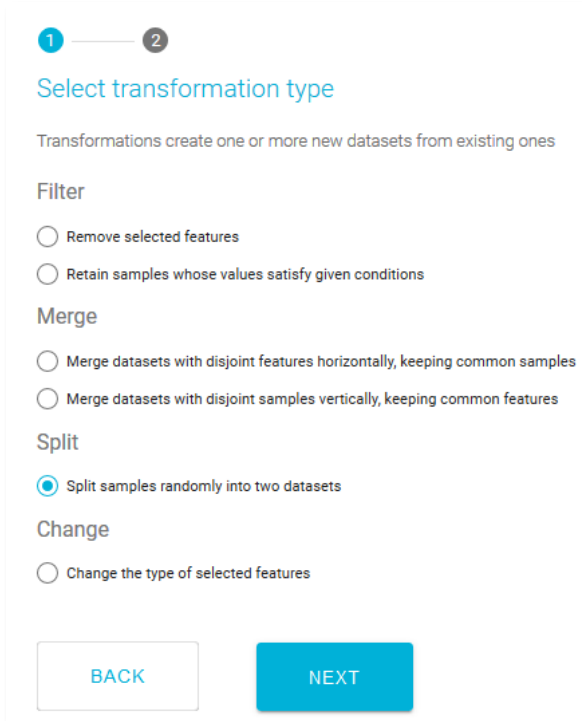


Figure 28 ACTIONS sidebar, Transform Data

Data Transformation takes place in two steps: **1. Select transformation type** and **2. Transformation configuration**, and JADBio currently provides four different transformation types: **Filter, Merge, Split and Change** dataset. You are going to create two datasets, **test** and **training**, from the imported dataset, so you will use the **Split** function.

- Select the **Split** option – **Split samples randomly into two datasets**.
- Click **NEXT**.



1 — 2

Select transformation type

Transformations create one or more new datasets from existing ones

Filter

☐ Remove selected features

☐ Retain samples whose values satisfy given conditions

Merge

☐ Merge datasets with disjoint features horizontally, keeping common samples

☐ Merge datasets with disjoint samples vertically, keeping common features

Split

☒ Split samples randomly into two datasets

Change

☐ Change the type of selected features

BACK NEXT

Figure 29 Transform Data

This opens the wizard, **Configure sample splitting**.

- From the **Select input dataset** pulldown, select the **Potatoes_quality** dataset.
- Name the first output dataset **Potatoes_training** and name the second, **Potatoes_test**.
- Set split percentage at **60%**.

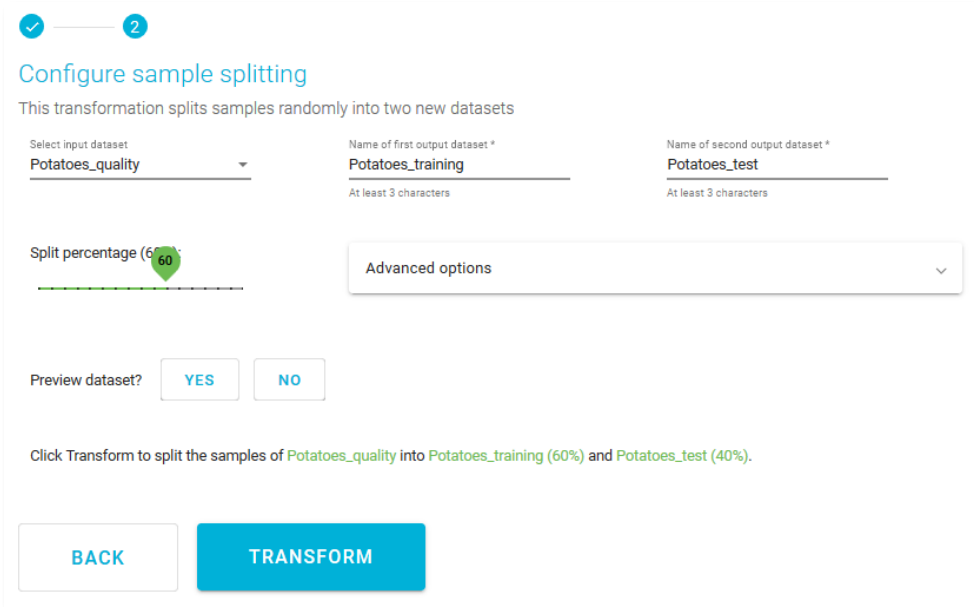


Figure 30 Transformation configuration

Note: Advanced options will allow you to select a feature for stratified splitting and/or a feature that defines how samples are grouped. This is not necessary for this demonstration.

- Click on the **TRANSFORM** button.

You will now have three datasets in your **Potatoes_quality_demo** project.

- Note the sample split in the two new datasets.

Datasets					
Analyses					
Applied Models					
Search					
Apply to All columns					
Name	Created	Features	Samples	Size	Actions
Potatoes_test	2020-09-26	206	191	0.34MB	---
Potatoes_training	2020-09-26	206	287	0.51MB	---
Potatoes_quality - Discovering	2019-11-29	206	478	0.84MB	---

Figure 31 Project Datasets

Background

Up to now, the prediction of complex phenotypes in plants like potatoes was based on growing plants and assaying the organs of interest in a time intensive process. Conventional genetic biomarkers are commonly applied in quality assessment and progeny selection but their application is still problematic for complex traits such as yield, disease resistances or stress tolerance.




A simple and accurate predictive test for potatoes yield quality could be derived from the analysis of potato metabolic biomarkers. In this demonstration, you will create, train and test a model based on the data from on 20 different potato cultivars grown in two Northern German potato farms, as described in the Steinfath et al. publication referenced below. These data include 203 metabolites measured on the tubers, Soil (loamy versus sandy) and climate (coastal versus inland) information and two traits 'susceptibility to black spot bruising' and 'chip quality". You will see that with JADBio you will produce accurate predictive models that are on par with the ones presented in the original publication, regardless of your expertise in advanced machine learning techniques.

Original publication: Steinfath et al, Plant Biotech Journal 2010

Analysis

With both a training dataset and a test dataset, within JADBio you are able to train and test a predictive signature based on the potato metabolic profiles and climate/soil information.

- From the table of **Datasets**, in the **Potatoes_training** dataset click on **Perform Analysis**.

Name	Created	Features	Samples	Size	Actions
Potatoes_test	2020-09-26	206	191	0.34MB	---
Potatoes_training	2020-09-26	206	287	0.51MB	  
Potatoes_quality - Discovering ;	2019-11-29	206	478	0.84MB	Perform Analysis

1 to 3 of 3 |< < Page 1 of 1 > >|

Figure 32 Potatoes_training dataset, Perform Analysis

- Select outcome: Quality**, by placing a checking the box next to the name of the feature and click **Next**.

1

2

Select outcome

For regression analysis, select a Numerical feature.
For classification analysis, select a Categorical feature.
For survival analysis, select both an Event and a Time to Event.

sample names	Quality <input checked="" type="checkbox"/> Categorical	Soil <input type="checkbox"/> Categorical	Weather <input type="checkbox"/> Categorical
Sample8	low	1	0
Sample9	low	1	0
Sample10	low	1	0
Sample11	low	1	0
Sample12	low	1	0

Everything looks good.
Click Next to perform a **classification** analysis, finding a model that predicts **Quality**.

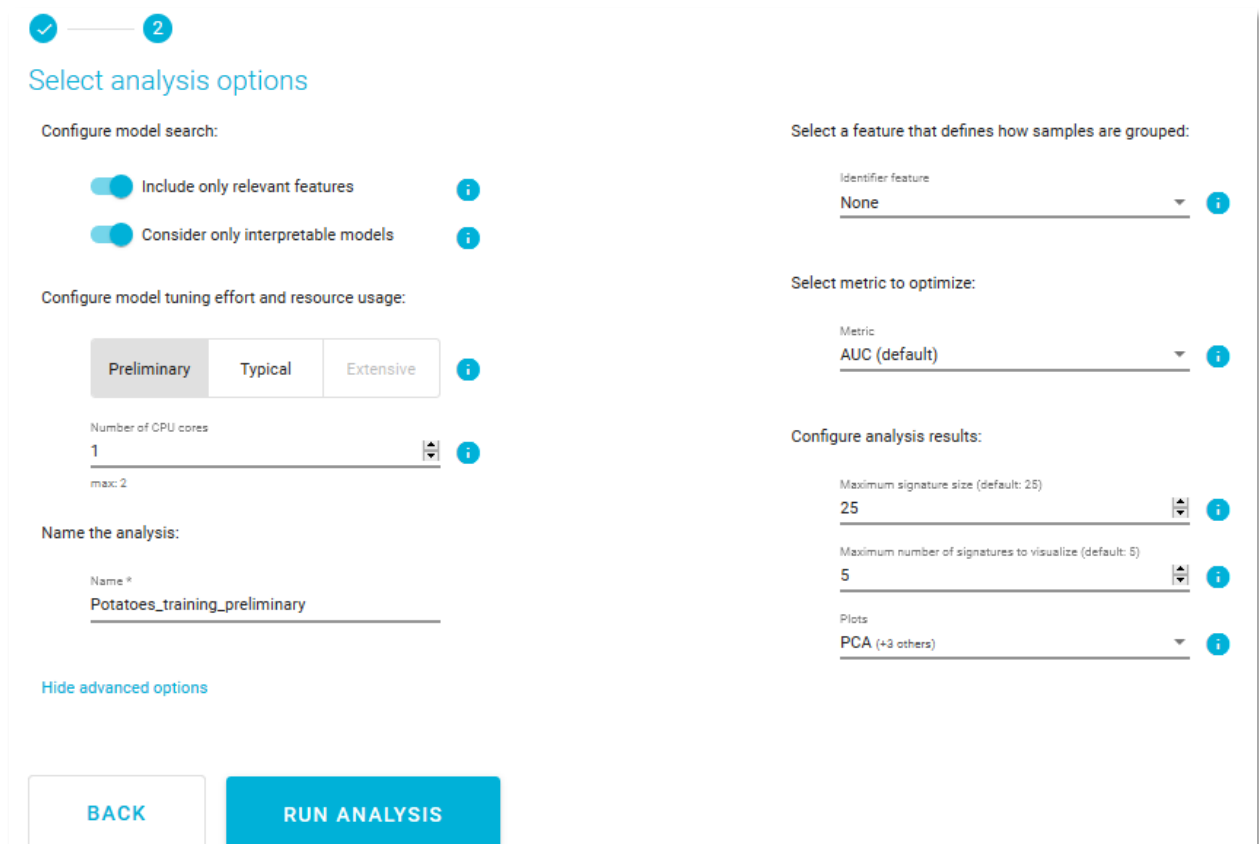
Figure 33 Select outcome - predicted feature

Note: Because you have chosen an outcome in which the values are distributed into two classes, a categorical feature, JADBio will create a model based on a binary classification analysis.

- **Select analysis options:** Set the Predictive analysis options as follows:
 - ✓ Include only relevant features - ☒ (This ensures that the models will only include the most relevant features.)
 - ✓ Consider only interpretable models - ☒ (Consider only models that are easy to interpret: Decision trees and linear regression models)
 - ✓ Tuning effort: **Preliminary**, for a quick first assessment
 - ✓ Number of CPU cores: **1**

JADBio will automatically create a descriptive name for your analysis based on your selections, which you can change.

- Name the Analysis: *Potatoes_training_preliminary*



Select analysis options

Configure model search:

- ☒ Include only relevant features
- ☒ Consider only interpretable models

Configure model tuning effort and resource usage:

Tuning effort: **Preliminary** (Selected), Typical, Extensive

Number of CPU cores: **1** (max: 2)

Name the analysis:

Name: **Potatoes_training_preliminary**

[Hide advanced options](#)

Select a feature that defines how samples are grouped:

Identifier feature: **None**

Select metric to optimize:

Metric: **AUC (default)**

Configure analysis results:

Maximum signature size (default: 25): **25**

Maximum number of signatures to visualize (default: 5): **5**

Plots: **PCA (+3 others)**

BACK **RUN ANALYSIS**

Figure 34 Select analysis options

While there are many other options, the default values for the advanced settings are the following:

- ✓ There is no requirement for an **Identifier feature**, to group the samples of the dataset.
 - ✓ The analysis will optimize performance based on the AUC (area under the ROC curve).
 - ✓ Maximum signature size is 25 features.
 - ✓ Maximum multiple signatures to visualize is 5.
 - ✓ JADBio will create four plots, PCA, UMAP, ICE, and Probabilities.
- Click **RUN ANALYSIS** to start the analysis.

As soon as you begin the analysis, JADBio reports **Progress** in the **Analyses** table.



Datasets Analyses Applied Models								
				Search	Apply to All columns	Columns to display Name, Started, Dataset, M...		
Name	Started	Dataset	Metric	Metric Value	Outcome	Type	Progress	Actions
 Potatoes	2020-09-26 21:08	Potatoes_trainin	AUC		Quality		RUNNING (67%) <div><div></div></div>	...

Figure 35 Running analysis

- Click on the blue progress bar to view learning progress, you can watch each step in the analysis progression.

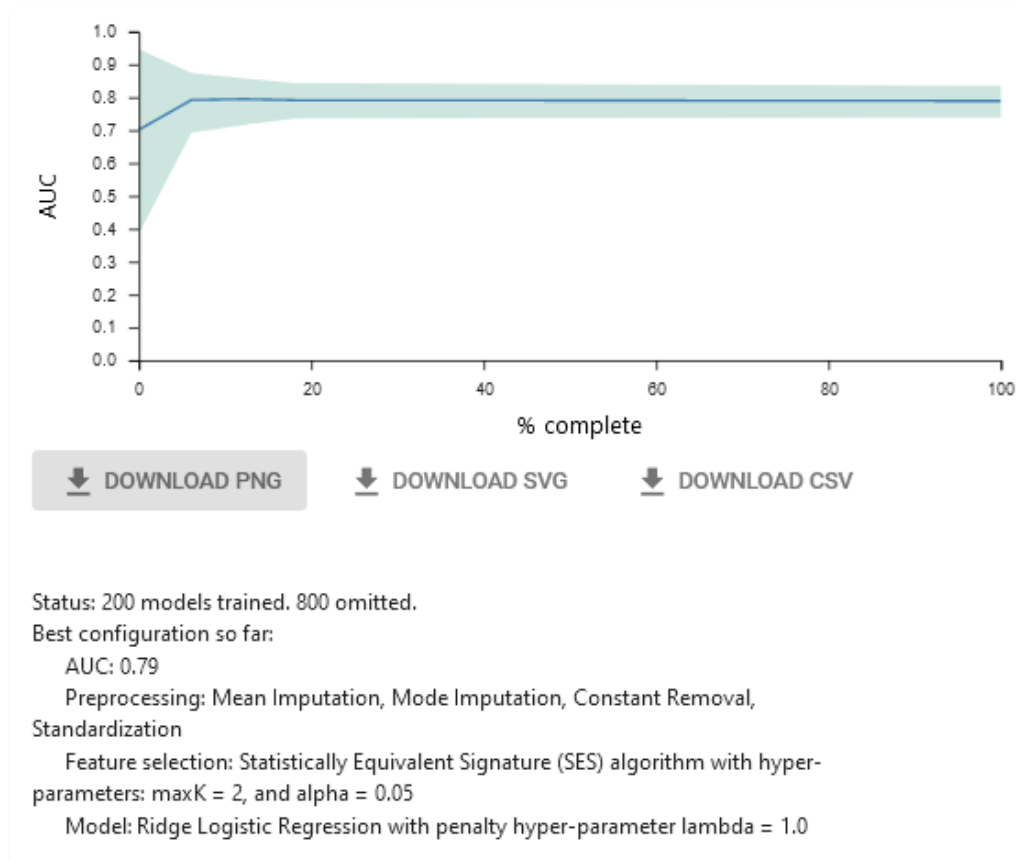


Figure 36 Learning progress

- Click on the main window to return to close the **Analyses** window.
- In the **Actions** column, click on **View results**.

Analysis Results

Note: The results you achieve may be different than the results reported here, because we are constantly updating and improving JADBio.

In the **ANALYSIS ACTIONS** sidebar, JADBio provides options to:

1. Summarize the analysis (as an **Analysis Report**),
2. **Show Model** (if it is an interpretable one)
3. **Download Predictions** of the samples
4. **Apply Model** to new samples.

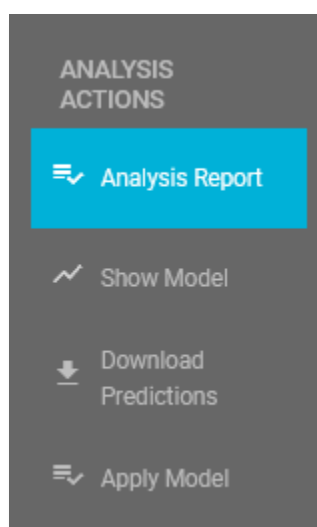
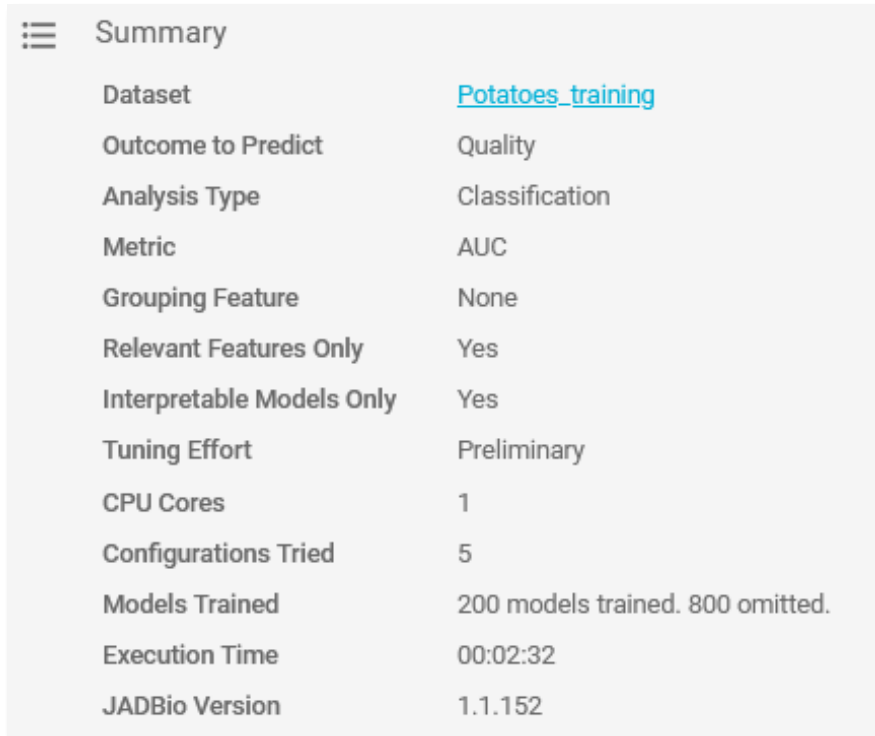


Figure 37 ANALYSIS ACTIONS sidebar

- Click on **Analysis Report**.

Here, JADBio provides a summary audit of the analysis including: **Dataset, Outcome to Predict, Analysis Type, etc.** and a description of the selected options as well as the JADBio version. Version 1.1.152 was used for this tutorial.



☰	Summary
Dataset	Potatoes_training
Outcome to Predict	Quality
Analysis Type	Classification
Metric	AUC
Grouping Feature	None
Relevant Features Only	Yes
Interpretable Models Only	Yes
Tuning Effort	Preliminary
CPU Cores	1
Configurations Tried	5
Models Trained	200 models trained. 800 omitted.
Execution Time	00:02:32
JADBio Version	1.1.152

Figure 38 Analysis info

JADBio also presents a list of all of the configurations that were tested in order to produce the model and selected features. The list is not long in this example, but more complex analyses may test up to 1000s of configurations. Specifically, for this analysis 5 configurations were tested

 [DOWNLOAD CSV](#)

which you can download by clicking on the button.



- Click on return button to return to the main analysis results page.

Note: In the top right corner, there is  link you can use to share your results.

Best Performing Model

The **Analysis** page provides an overview of the analysis process and a description of the **Best Interpretable Model**.

The methods include the optimal configurations for: **Preprocessing**, **Feature selection**, and **Predictive algorithm**.

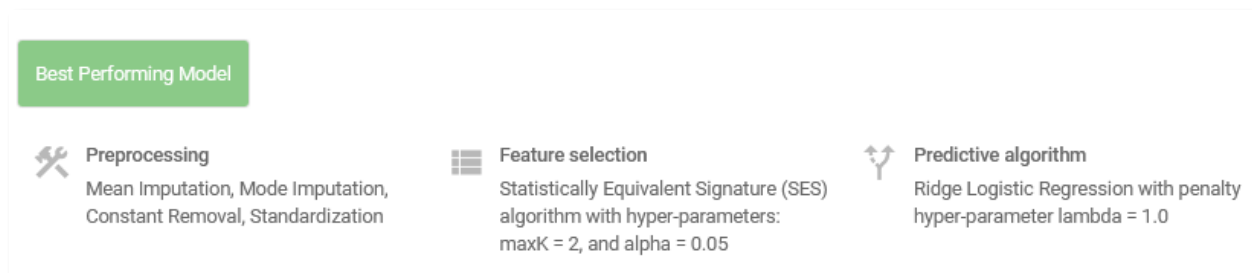


Figure 39 Best Performing Model

Note: When you selected the analysis parameters, you forced JADBio to only report interpretable models as the Best Performing Model. Otherwise, JADBio might provide two models, **Best Performing Model** and **Best Interpretable Model**.

- Click on the **Show Model** in **ANALYSIS ACTIONS** sidebar.

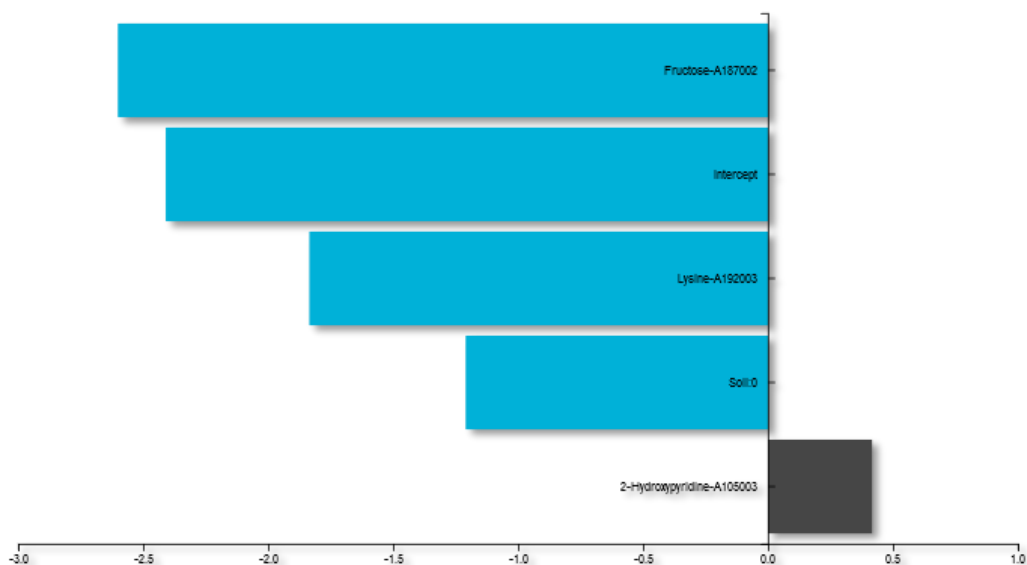


Figure 40 Show model (The image used here is from the downloaded SVG, rather than a screen shot).

JADBio displays four, out of the total 206 features, that provide the most accurate prediction of potatoes quality. The features include Fructose-A187002, Lysine-A192003 and Soil:0 whose expression are in a negative relationship to the prediction and 2-Hydroxypyridine-A105003 whose expression is in a positive relationship to the prediction.

- Hover over the features bar to visualize the values for the optimized Ridge Logistic Regression model.

The numbers provided describe the relative strength of the predictors based on the logistic model. The larger the absolute value of the feature's value in the model, the greater the impact that feature is in the analysis of any one sample's outcome. In this potatoes' quality example, Fructose-A187002, is a stronger level of evidence for the quality prediction in any sample than the other features.

Note: It is possible to download both the image and the numbers supporting the image from the



buttons.

- Click on the **Download Predictions**.

In the downloaded **analysis_predictions.txt**, you will see each of the analyzed samples and, based on the cross validation of the best configuration, their relative difficulty of predictions. For each sample you will see the **probability** the sample would be predicted low or high quality. In this dataset, 268 of 287 samples are labeled as **FALSE** (not difficult to predict). The Label column is the actual values from the dataset.

	A	B	C	D	E
1	Sample name	Prob (class = high)	Prob (class = low)	Difficult to Predict	Label
2	Sample8	0,005	0,995	FALSE	low
3	Sample9	0,058	0,942	FALSE	low
4	Sample10	0,069	0,931	FALSE	low
5	Sample11	0,265	0,735	FALSE	low
6	Sample12	0,120	0,880	FALSE	low
7	Sample13	0,027	0,973	FALSE	low
8	Sample14	0,036	0,964	FALSE	low
9	Sample15	0,089	0,911	FALSE	low
10	Sample16	0,057	0,943	FALSE	low

Figure 41 Downloaded predictions viewed in spreadsheet

Performance Overview

Define positive class

Reference class is considered the class of Positive samples and the rest are considered Negative ones. For this example, let's consider 'high' as the positive class.

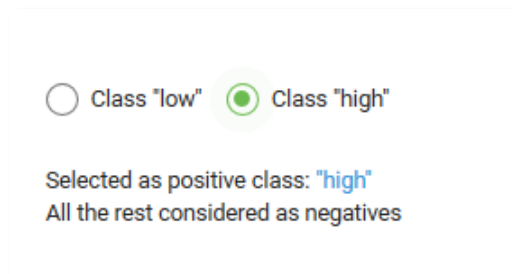


Figure 42 Define positive class

Threshold independent metrics

The performance of the binary classifier (low or high quality) can be described by the Area Under the Curve (AUC) of the ROC curve and by Average Precision of the Precision-Recall curve.

A Receiver Operating Characteristic curve (or ROC curve) summarizes the trade-off between the true positive rate (sensitivity) (y-axis) and the false positive rate (1-specificity) (x-axis) for different probability thresholds. The best ROC curves are the ones where X (false positive rate) = 0 and Y (true positive rate) = 1.

A precision-recall curve (or PR Curve) is a plot of the precision (y-axis) and the recall (x-axis) for different probability thresholds. The best PR curves are the ones where X (recall) = 1 and Y (precision) = 1.

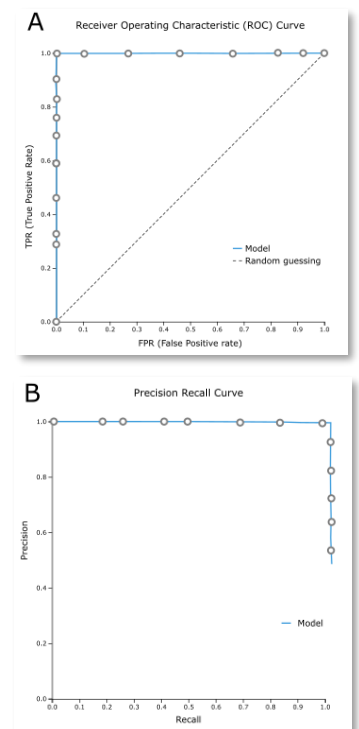


Figure 43 A. a perfect ROC curve and B. a perfect PR curve

Metric	Mean estimate	95% confidence interval	Unadjusted estimate	Base line	Statistically significant
Area Under the Curve	0.789	[0.738, 0.834]	0.789	0.500	✓
Average Precision	0.948	[0.932, 0.962]	0.951	0.822	✓

Figure 44 AUC and Average Precision metrics for the Best Interpretable model

 button, JADBio allows you to choose between three

Note: In values of significance levels.

JADBio allows you to optimize the classification threshold for a gradient of metrics for optimal specificity to optimal sensitivity.

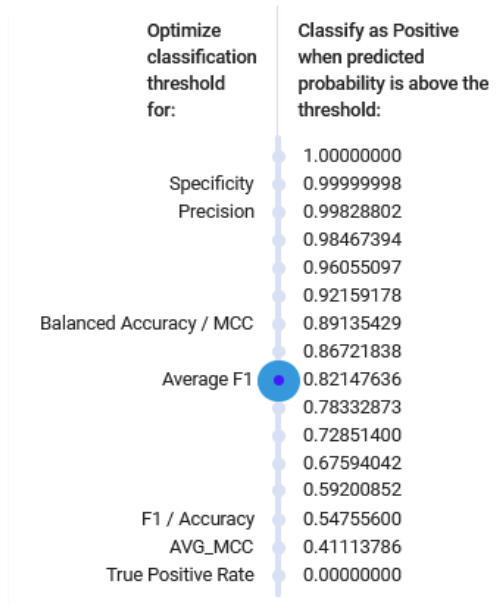


Figure 45 Optimization thresholds.

- **Optimize classification threshold as: 0.82147636**, and note the selection of the position on the ROC curve.
- Hover over the highlighted ROC curve to see the full range of metrics at this threshold.

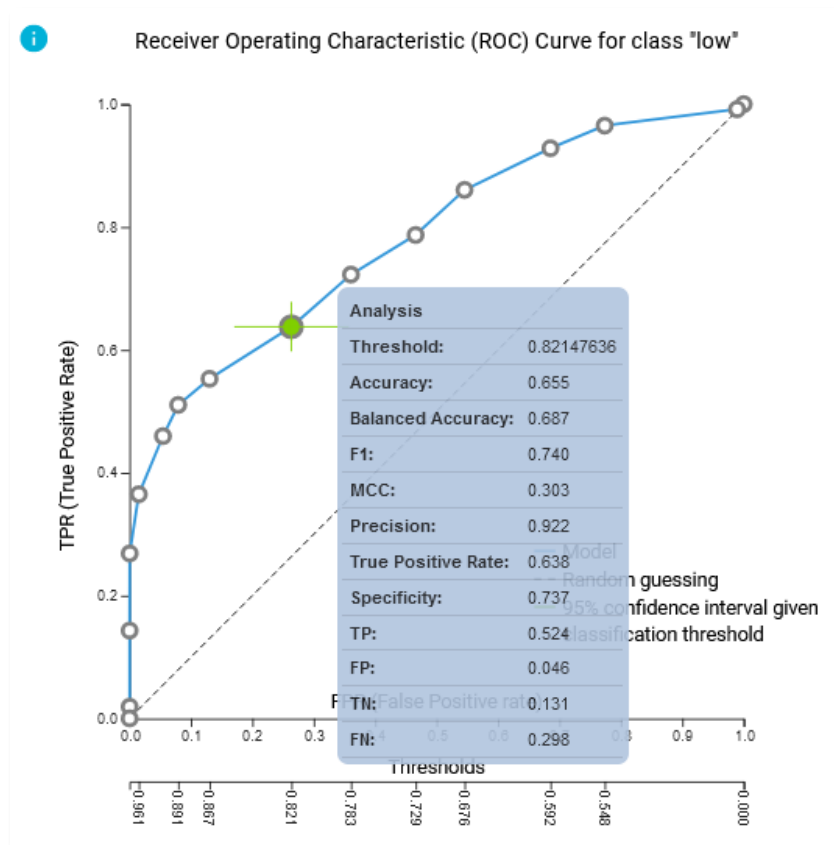


Figure 46 ROC Curve and Predictive performance for the selected threshold

ROC plot

Precision recall plot

- Click on the button to view the Precision recall plot.

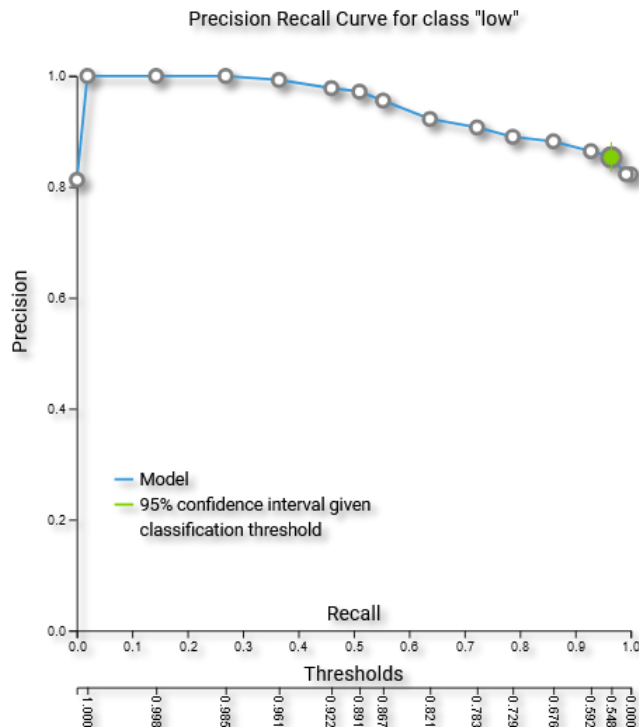


Figure 47 Precision Recall curve plot (The image used here is from the downloaded PNG, rather than a screen shot).

Note: ROC curves are appropriate when the observations are balanced between each class, whereas precision-recall curves are appropriate for imbalanced datasets.

Confusion Matrix

Confusion matrix is a table that describes the performance of a classification model (or "classifier") on predicted class (values) for which the true class (values) is known. In JADBio, confusion matrix displays either the percentages or a heatmap of the predicted values vs the real true values.

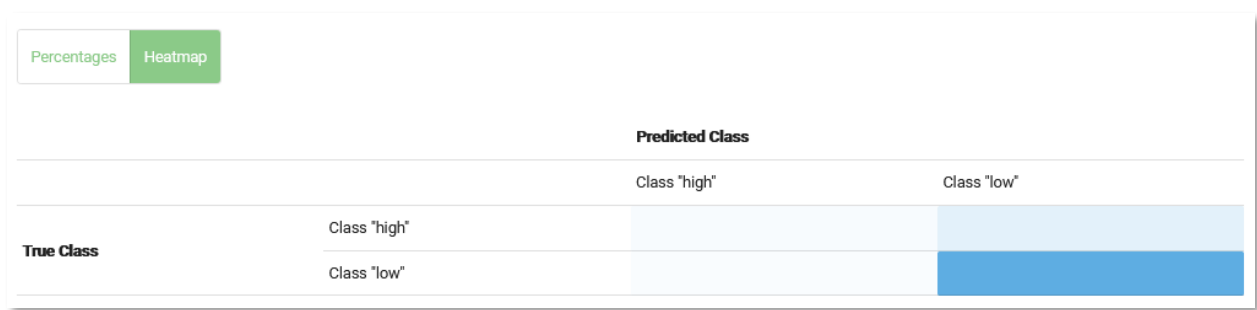


Figure 48 Confusion matrix displaying as heatmap for the selected threshold

Threshold dependent metrics

Here, JADBio reports 13 different metrics and their confidence intervals based on your Best Performing Model.

- Hover over any “i” adjacent to a metric for an explanation of the score.

How you set the thresholds will be determine the overall sensitivity and specificity of the model.

Note: As you move your cursor in the JADBio windows, JADBio will provide contextual information or links to relevant locations within the application.

Feature Selection

Feature Selection is a process that identifies a minimal-size subset of features that is maximally predictive of the outcome of interest, the selected target feature.

- Scroll back to the top of the page, and Select the **Feature Selection** tab.

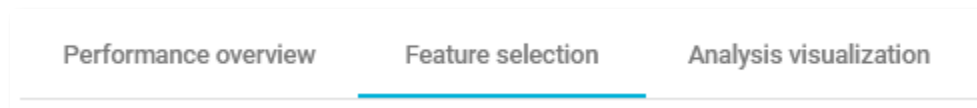


Figure 49 Feature selection

Selected Signatures

A signature is a minimal subset of predictive features that, when considered jointly, are maximally informative for an outcome of interest. As a product of each analysis, JADBio produces all signatures that perform equally well, up to the maximum limit defined in parameters. In this example, JADBio produced 4 equivalent signatures.

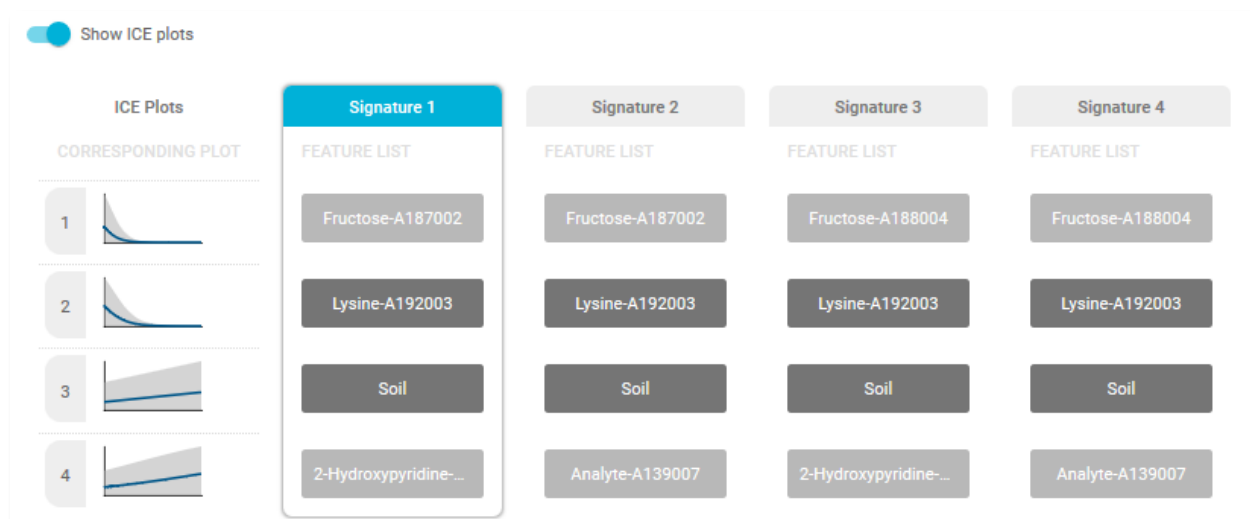


Figure 50 Selected signatures

The **Individual Conditional Expectation** (ICE) plots further reveal the nature of the contribution of each metabolite feature to the model.

- Click on the thumbprint ICE plot adjacent to Fructose-A187002 feature to enlarge the ICE plot.

This opens the ICE plot for the prediction of “high” quality classification. In this plot, you can see that as the metabolite level increases, the likelihood of a “high” quality classification decreases.

- Use the pulldown to select the **Class low**.

As one would expect for a binary classification, the level of Fructose-A187002 metabolite has the inverse correlation to a low-quality classification.

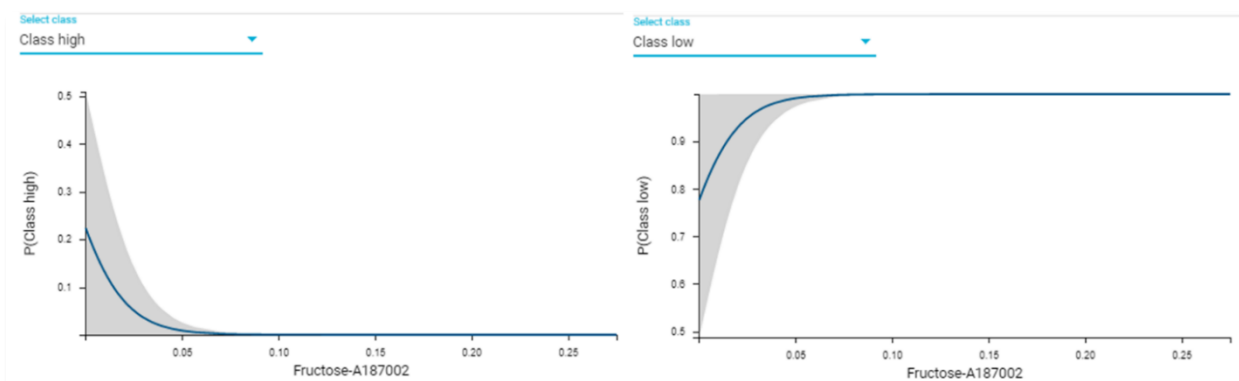


Figure 51 ICE plots for Fructose-A187002 predictor.

Feature Importance plots

The practical use of **Feature Importance plots** is evident in the case of selecting biomarkers. For instance, the purpose of this analysis is to identify the optimal list of biomarkers that predict potato quality. However, in order to satisfy economical or technical constraints on an assay, JADBio also reports the cost to performance that occurs when one chooses to further reduce the total number of predictive biomarkers from those included in the Best Performing Model. In this way, you, can evaluate the trade-off between reducing the number of biomarkers and achieving optimal performance.

Both in the **Progressive Feature Inclusion** and in the **Feature Importance** view, JADBio displays the four features of the selected signature and their relative performance.

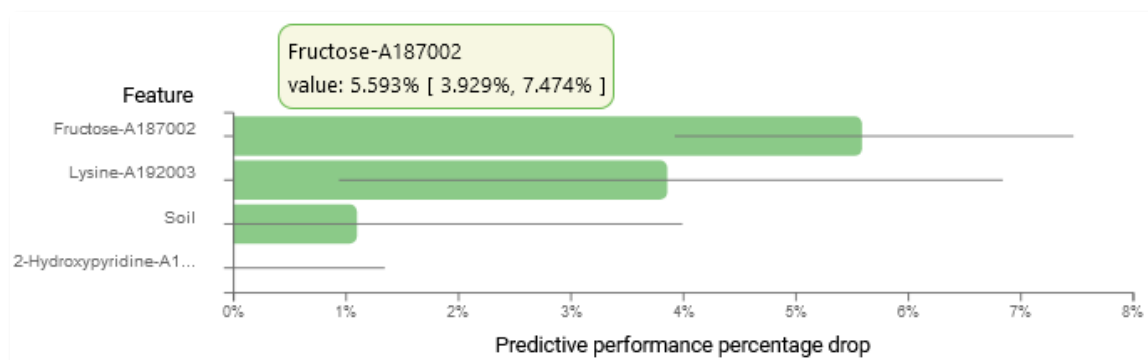


Figure 52 Feature Importance plot

Analysis Visualization

- Select the **Analysis Visualization** tab.

Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) is a dimension reduction technique that attempts to learn the high-dimensional manifold on which the original data lays, and then maps it down to two dimensions. UMAP is ideal for non-linear models.

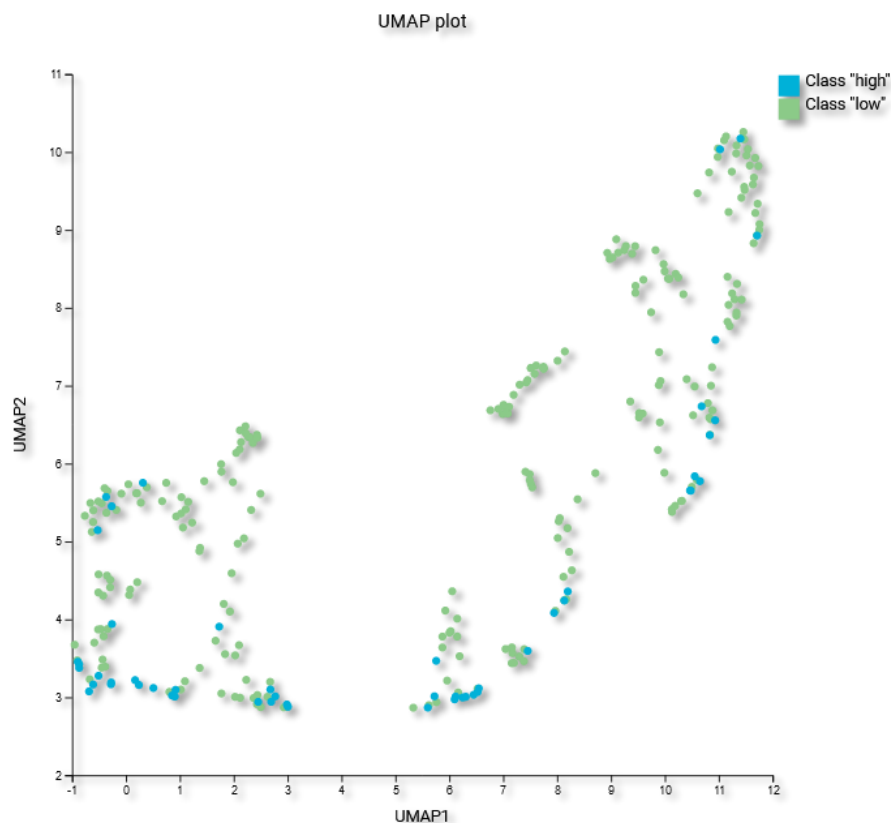


Figure 53 Uniform Manifold Approximation Projection. UMAP (The image used here is from the downloaded PNG, rather than a screen shot).

- Change the **Display as** selection from UMAP to PCA.

Principal component analysis (PCA) is another dimensionality reduction technique that seeks the linear combinations (Principal Components) of the original features such that the derived features capture maximal variance.

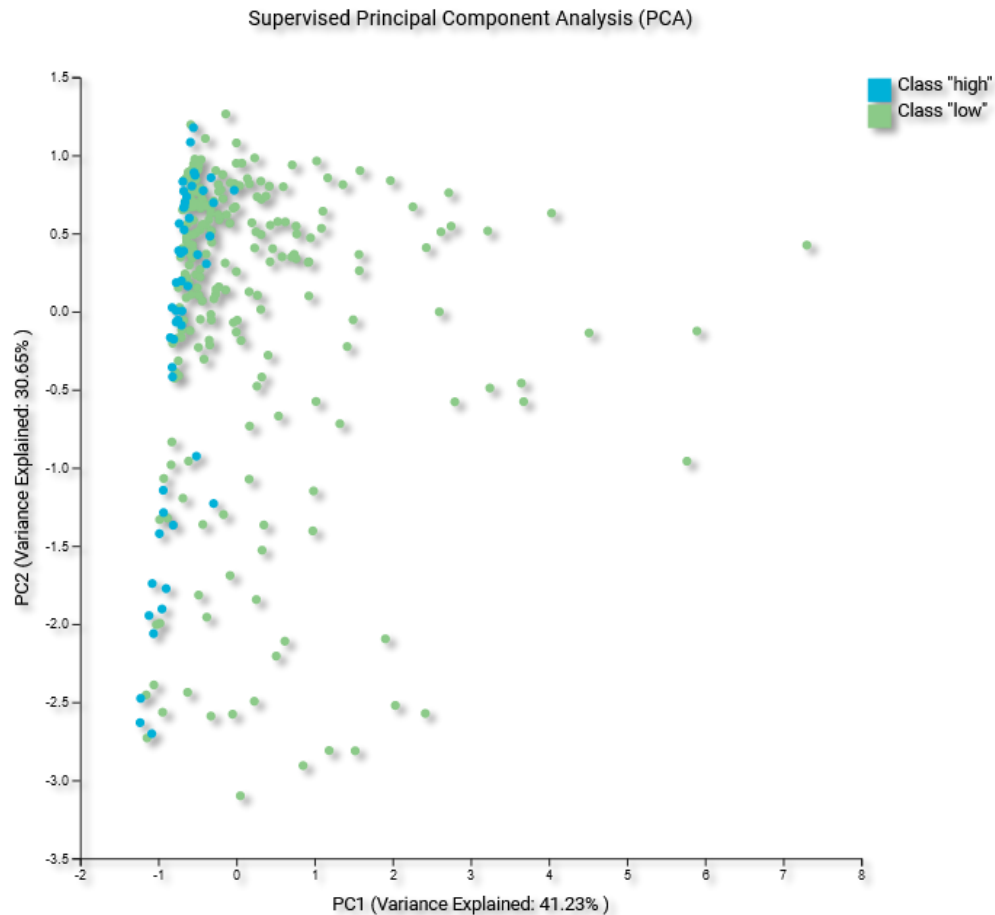


Figure 54 Principal Component Analysis plot (The image used here is from the downloaded SVG, rather than a screen shot).

The appropriate visualizations for this model are UMAP and PCA plot, in which the low and high quality predictions are visualized. Other types of analysis and other types of models would likely result in different visualizations. For instance, a survival (time to event) analysis would have resulted in a Kaplan-Meier curve.

- Scroll down to see the **Probabilities plot**.

The **Probabilities plot** shows, for both classes, the probability of the prediction resulting in the high class. An ideal plot would have complete separation between the two classes. You can also visualize the probabilities in a box plot.

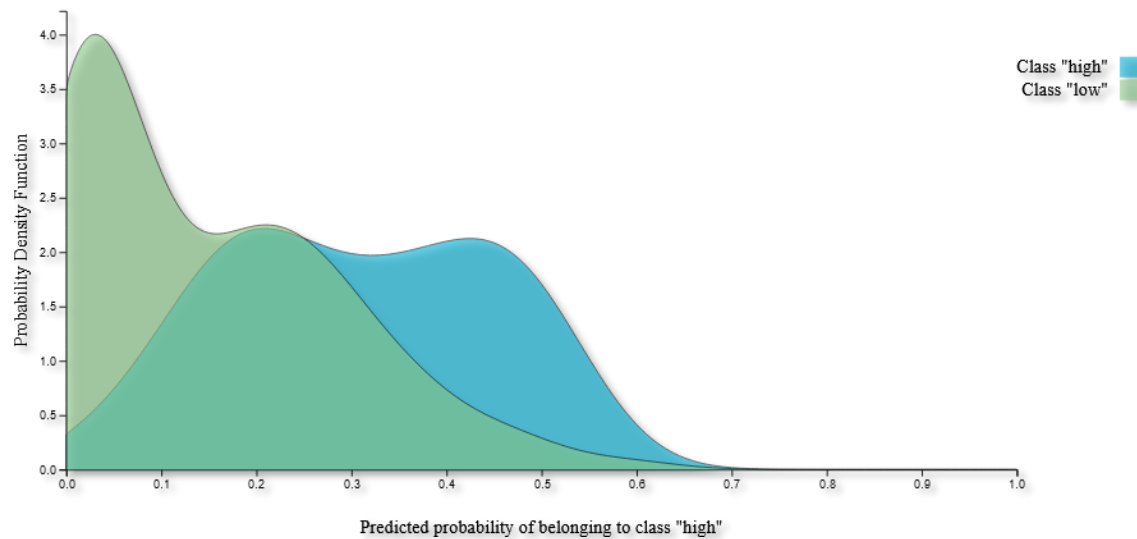


Figure 55 Probabilities plot displaying as density plot (The image used here is from the downloaded SVG, rather than a screen shot).

Apply Model

All of the above analysis was performed on the training dataset. Now that the analysis of the trained model is complete, you can run the test dataset.

- Scroll to the top of the page, and click on the **Potatoes_quality_demo** project label.
- In the **ACTIONS** sidebar, click on **Apply Model**.

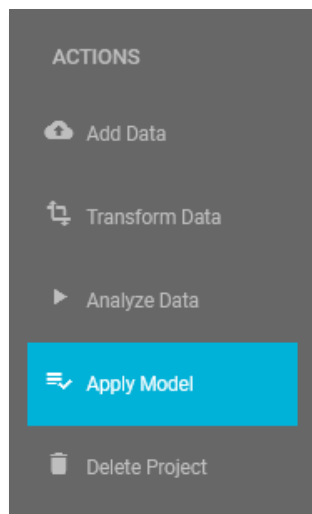
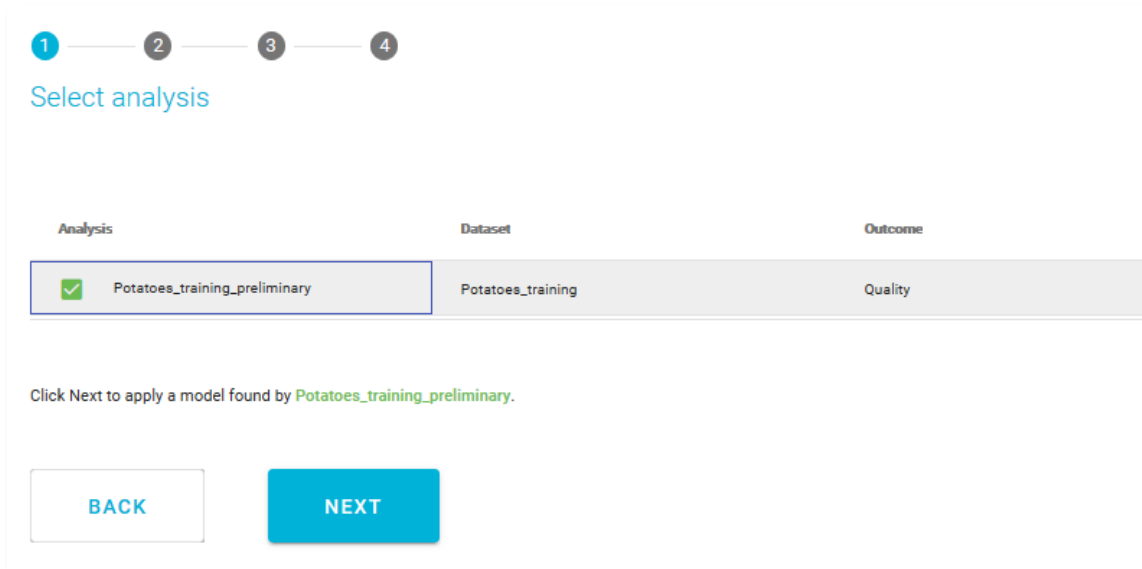


Figure 56 ANALYSIS ACTIONS, Apply Model

There are four steps to the validation process:

1. Select Analysis:

- Check the only analysis option available in the Project, **Potatoes_training_preliminary**, which you just created and reviewed.
- Click **NEXT**.



1 — 2 — 3 — 4

Select analysis

Analysis	Dataset	Outcome
<input checked="" type="checkbox"/> Potatoes_training_preliminary	Potatoes_training	Quality

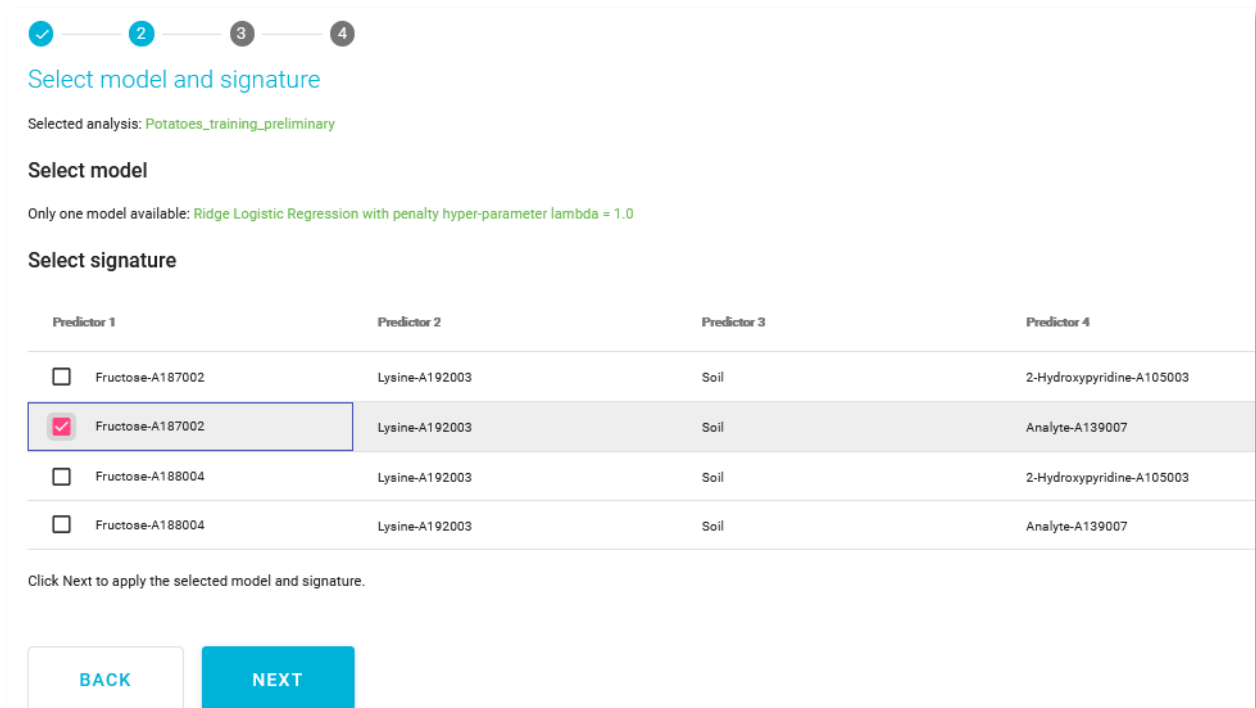
Click Next to apply a model found by Potatoes_training_preliminary.

BACK NEXT

Figure 57 Select analysis

2. Select model and signature:

- **Select model:** There is only model available, which is already selected.
- **Select signature:** Place a check in the second row of the four statistically equivalent signatures.
- Click **NEXT**.



✓ — 2 — 3 — 4

Select model and signature

Selected analysis: Potatoes_training_preliminary

Select model

Only one model available: Ridge Logistic Regression with penalty hyper-parameter lambda = 1.0

Select signature

Predictor 1	Predictor 2	Predictor 3	Predictor 4
<input type="checkbox"/> Fructose-A187002	Lysine-A192003	Soil	2-Hydroxypyridine-A105003
<input checked="" type="checkbox"/> Fructose-A187002	Lysine-A192003	Soil	Analyte-A139007
<input type="checkbox"/> Fructose-A188004	Lysine-A192003	Soil	2-Hydroxypyridine-A105003
<input type="checkbox"/> Fructose-A188004	Lysine-A192003	Soil	Analyte-A139007

Click Next to apply the selected model and signature.

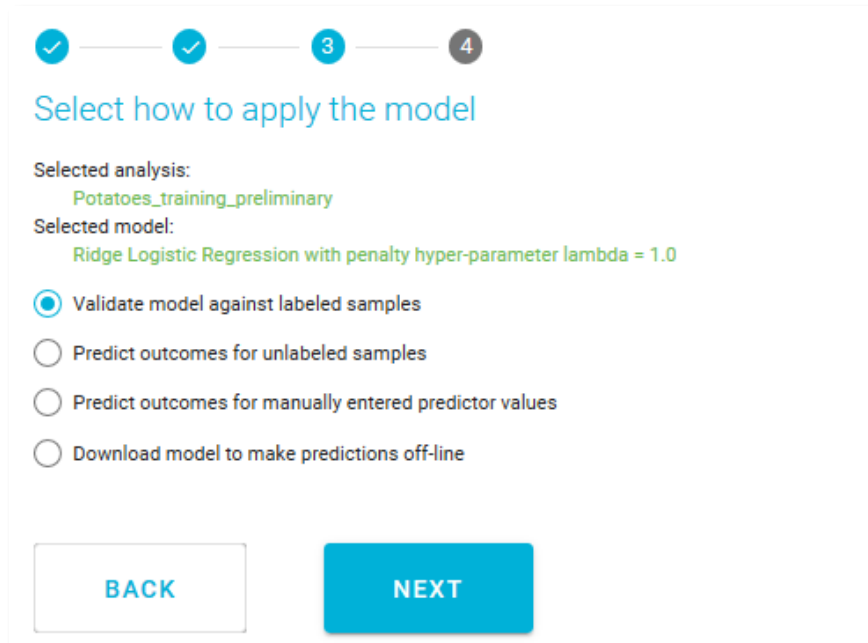
BACK NEXT

Figure 58 Model and signature selection

3. Select how to apply the model: Select the option, **Validate model against labeled samples**.

- Click **NEXT**.

This selection allows JADBio to use the Test data set you created when you split the original dataset into Training and Test datasets. If you choose to **Predict outcomes for Unlabeled samples**, JADBio will give you the option to upload another dataset. The only requirement you will have for this dataset is that it includes the Predictors in the selected signature with the same names as in the training data. If you select to **Predict outcomes for manually entered predictor values**, JADBio will provide a dialog for manually entering values for each of the Predictors in the Signature to generate a prediction. If you choose to **Download model to make predictions off-line** will give you the options to download a standalone version of the chosen model that can be applied on new data on a local machine.



Selected analysis:
Potatoes_training_preliminary

Selected model:
Ridge Logistic Regression with penalty hyper-parameter lambda = 1.0

☒ Validate model against labeled samples

☐ Predict outcomes for unlabeled samples

☐ Predict outcomes for manually entered predictor values

☐ Download model to make predictions off-line

BACK NEXT

Figure 59 Select how to apply the model

4. Select labeled dataset

- Select the dataset: **Potatoes_test**.
- Click **APPLY MODEL**.

✓

✓

✓

4

Select labeled dataset

Selected analysis:
Potatoes_training_preliminary

Selected model:
Ridge Logistic Regression with penalty hyper-parameter lambda = 1.0

Name	Size	Samples	Features	Created
<input type="checkbox"/> Potatoes_quality	841731	478	206	2019-11-29
<input checked="" type="checkbox"/> Potatoes_test	253862	143	206	2020-06-10
<input type="checkbox"/> Potatoes_training	591887	335	206	2020-06-10

Click Apply Model to validate the model against the labeled dataset Potatoes_test.

BACK

APPLY MODEL

Figure 60 Apply Model

JADBio will automatically bring you to the **Applied Models** window, where you can view the results of your Applied Model on the test data set.

Datasets	Analyses	Applied Models	
		<div><div>Search</div><div>Apply to All columns</div></div>	
Dataset	Analysis used	Progress	Actions
Potatoes_Test	Potatoes_training_preliminary	FINISHED (100%)	...

Figure 61 Applied Models

- Click on the **Actions** function, **View results** to open the results window.

Much of this results window is like the original training dataset window, but with some exceptions. In **VALIDATION ACTIONS** sidebar, there are buttons to **Download Predictions** of the samples and **Apply Model** to new samples.

In the main **Applied Model** window, JADBio displays the four features from the selected signature.

- Define **Class High** as the **positive class**, using the radio buttons on the top.

Now, JADBio displays the AUC and the Average Precision of the validation and the train datasets.

Metric	Validation	Train			
	Estimate	Mean estimate	95% confidence interval	Unadjusted estimate	Base line
Area Under the Curve	0.777	0.789	[0.738, 0.834]	0.789	0.500
Average Precision	0.930	0.948	[0.932, 0.962]	0.951	0.822

Figure 62 AUC and Average precision of the validation dataset

The new ROC and Precision recall curves include the results from the original **Train** --- data and from the test data, **Validation**. ---.

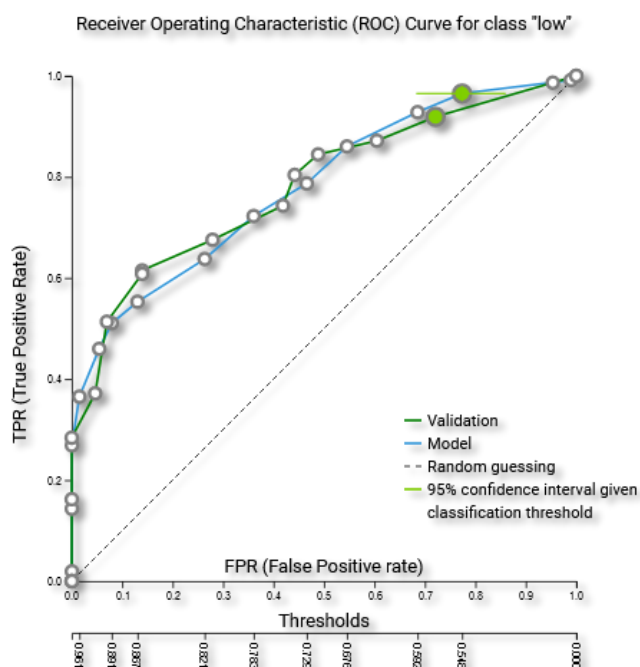


Figure 63 ROC curve for validation (The image used here is from the downloaded SVG, rather than a screen shot).

JADBio also displays the confusion matrix for the Validation and the Model (training) datasets and the 13 threshold dependent metrics.

Percentages		Heatmap	
Validation		Predicted Class	
True Class		Class "high"	Class "low"
	Class "high"	0.063	0.162
	Class "low"	0.063	0.712
Model		Predicted Class	
True Class		Class "high"	Class "low"
	Class "high"	0.041 [0.024,0.056]	0.137 [0.111,0.162]
	Class "low"	0.029 [0.017,0.042]	0.793 [0.763,0.823]

Figure 64 Confusion matrix with percentages for the validation and training datasets

- Click on the **Analysis Visualization** tab.

Here, the **Supervised Principal Component Analysis** displays the segregation of the low and high class test data based on the model from the training data.

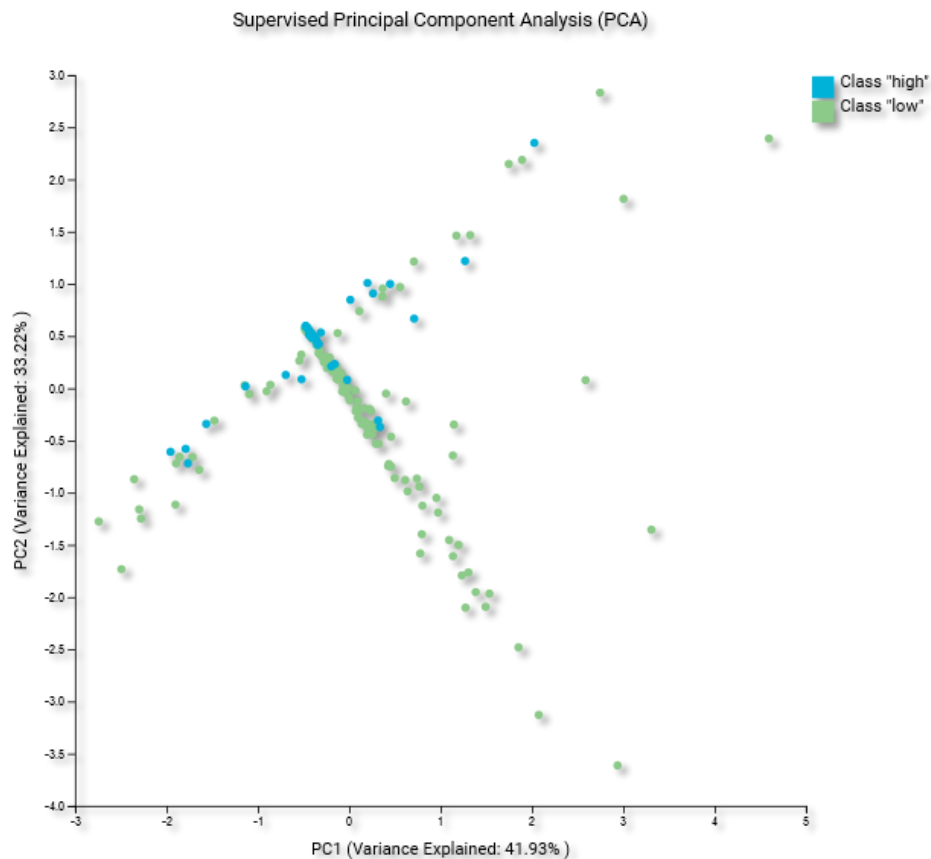


Figure 65 PCA for Applied Model (The image used here is from the downloaded PNG, rather than a screen shot.)

Note of appreciation to JADBio users: We constantly make changes in the software and do our best to update these materials, but you may notice some differences. We welcome your feedback on how to make this more useful for you and requests for future tutorials.