Apply predictive analytics on multi-omics data with InSyBio Biomarkers

InSyBio
Intelligent Systems Biology

User Manual

July 2024

Insybio Suite v3.3

www.insybio.com

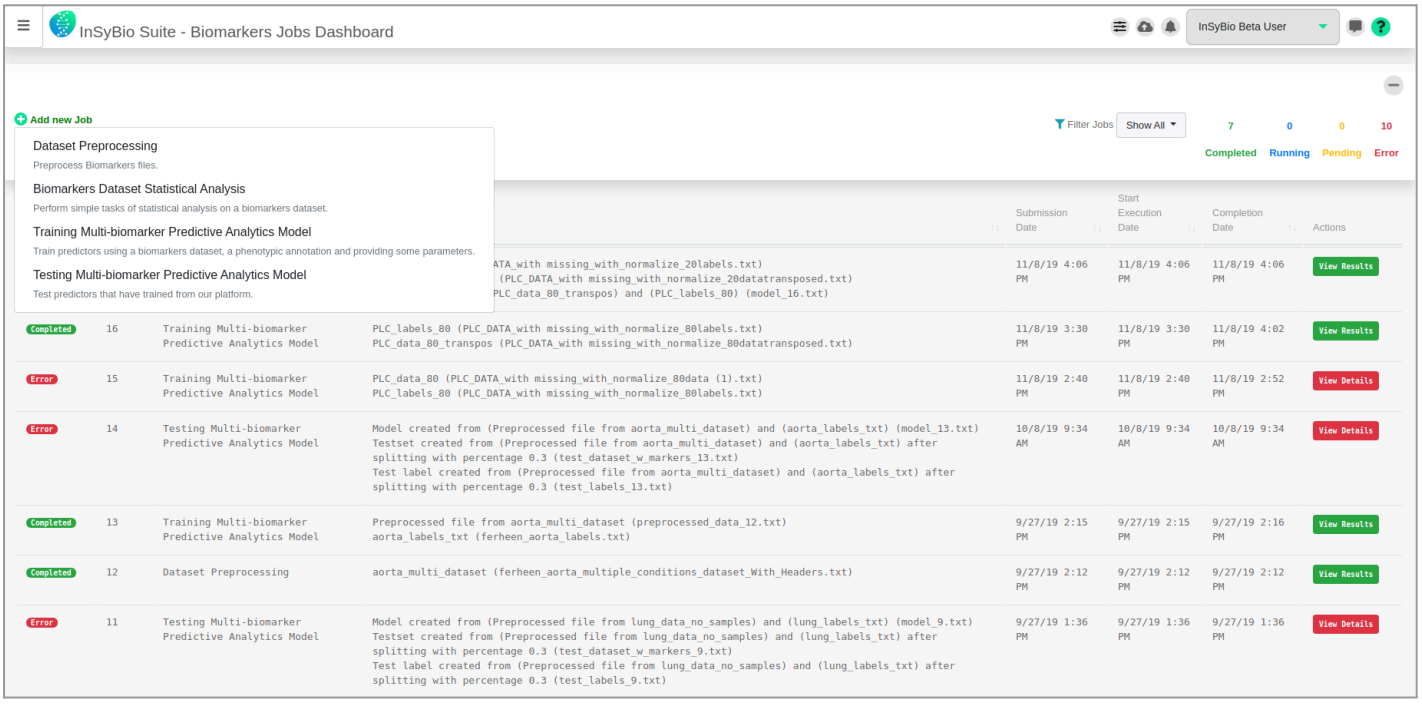
InSyBio Biomarkers

Introduction

Biomarkers is a tool for:

- dataset preprocessing and statistical analysis of omics and multi-omics datasets
- training multi-biomarkers machine learning models for disease diagnosis, prognosis and response to therapies
- applying trained models to new data

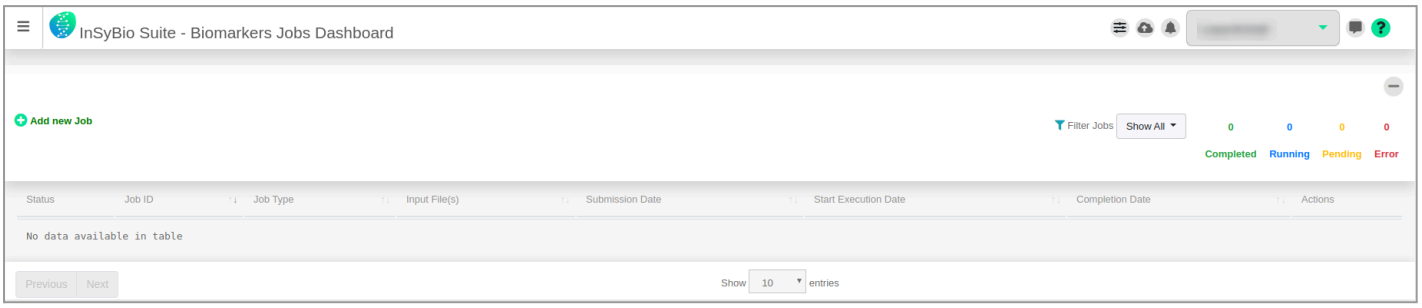
When the user selects the InSyBio Biomarkers tool he will access the Biomarkers dashboard:



The available functionalities in the “Add new Job” button include Biomarkers Dataset Preprocessing, Biomarkers Dataset Statistical Analysis, Training Multi-biomarker Predictive Analytics Model and Testing Multi-biomarker Predictive Analytics Model.

Inside the biomarkers jobs dashboard the user will be able to view his current running, pending, successfully completed and completed with error jobs.

The biomarkers jobs dashboard of a new user will look like the following:

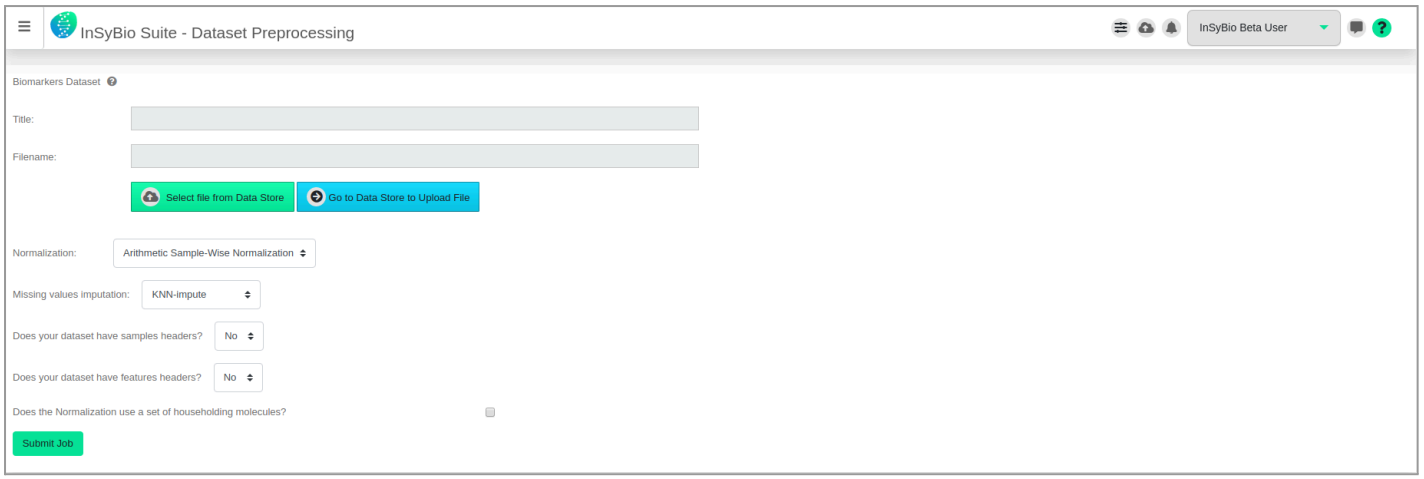


Under the column “Job ID” the table has the identification name of each job. In the column “Job Type” the user will be able to see the type of job, categorized as one of the following: Biomarkers Dataset Preprocessing, Biomarkers Dataset Statistical Analysis, Training Multi-biomarker Predictive Analytics Model and Testing Multi-biomarker Predictive Analytics Model.

In the field “Input File(s)” the user will be able to view the input files of each job. At the field “Submission Date” the dashboard will display the exact date that each job has been submitted. At the field “Start Execution Date” the exact date of the initiation of each job will be displayed. At the field “Completion Date” the completion date of each job will be displayed. Also, at the field “Status”, each job will be either “Completed” or “Pending”, or “Running” or “Error” when an error will have occurred.

Dataset Preprocessing

During preprocessing we filter the dataset, perform normalization, missing values imputation, duplicate measurements averaging and outlier detection with the PCA LOF method.



More specifically, there are two kinds of normalization: arithmetic sample-wise and logarithmic. It should be noted that when the data contain negative numbers the arithmetic normalization should be chosen, since the logarithmic normalization method functions only with non-negative data. If “None” is chosen, no normalization takes place.

There are two kinds of missing value imputation methods as well: average imputation and KNN imputation. Average imputation is a method in which the missing value on a certain variable is replaced by the mean of the available cases[1]. On the other hand, the key idea of KNN imputation is that a point value can be approximated by the values of the points that are closest to it, based on other variables[2]. The above missing values imputation methods are relevant only for cases where a missing value does not imply a quantification value of zero. In such cases, missing values should be replaced with zeros before uploading the dataset. If “None” is chosen then no missing value imputation takes place.

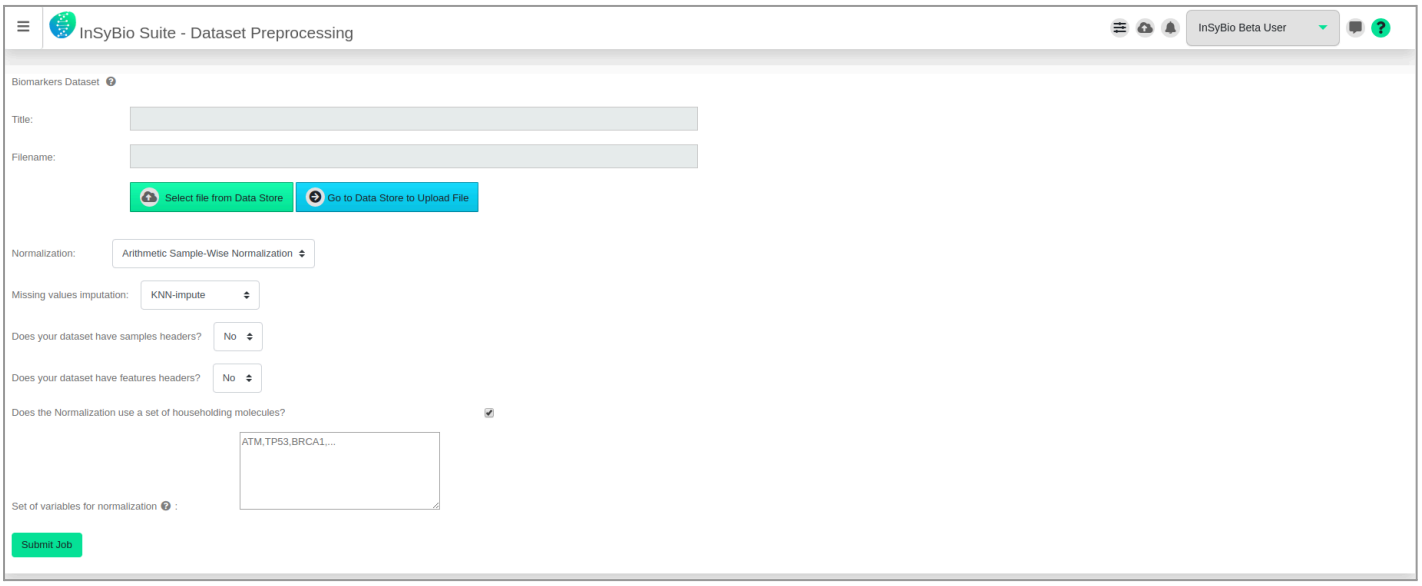
Furthermore, if a missing values imputation method is being chosen instead of “None”, the duplicate measurements will be averaged.

The user will be prompted to upload a biomarkers dataset or select one from the Datastore. A biomarkers dataset should have as rows the biomarkers (molecules, or else features) and as columns the samples. Thus, each cell will contain the value of a biomarker in a sample. Also, the user should provide the information if there are

[1] Single Imputation Methods (iriseekhout.com)
[2] The use of KNN for missing values (towardsdatascience.com)

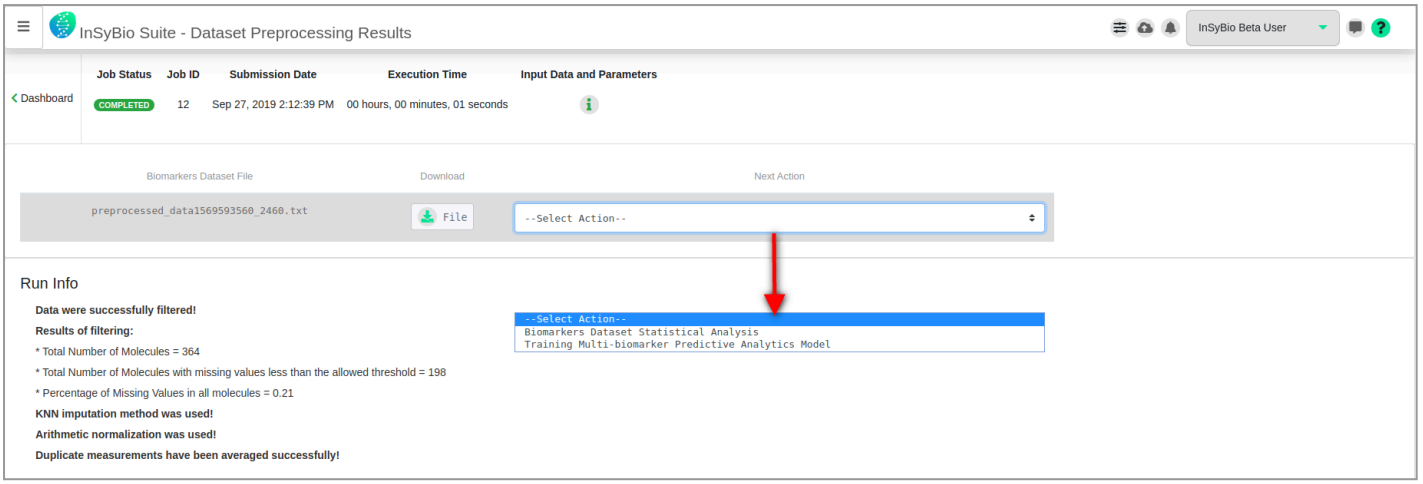
headers inside the dataset (i.e. a samples header, features header, or both or none). To do this you'll have to choose "Yes" or "No" as answers to the questions "Does your dataset have sample headers?" and "Does your dataset have feature headers?". Additionally, the dataset should have its values separated with tabs (have TSV format).

The user can also provide a set of household molecules (biomarkers) to perform geometric normalization of the data additionally to the normalization method already being performed. To do that the user will have to tick the box beside the question "Does the Normalization use a set of householding molecules?", and insert the names of molecules in the box that appears, separated with commas or a new line. Householding molecules should represent a large abundance in the dataset, not have missing values and should not present high variability among the examined phenotypes.



Complex combinations of the above preprocessing methods can be conducted by applying these analysis steps sequentially (e.g. first logarithmic normalization and then arithmetic normalization).

After dataset preprocessing finishes successfully, by selecting "View Results" from the Biomarkers dashboard page the user will view a page like the following:



In the above page the user will be able to view the Input Data and the Parameters that he chose for the specific job, that is the description of the input file and its name, the normalization method chosen, the missing values imputation method chosen and the set of variables chosen for normalization. Additionally the user will be able to view the Biomarker Job Information, which includes the submission, start, execution and completion date, the total runtime (execution time) and the status.

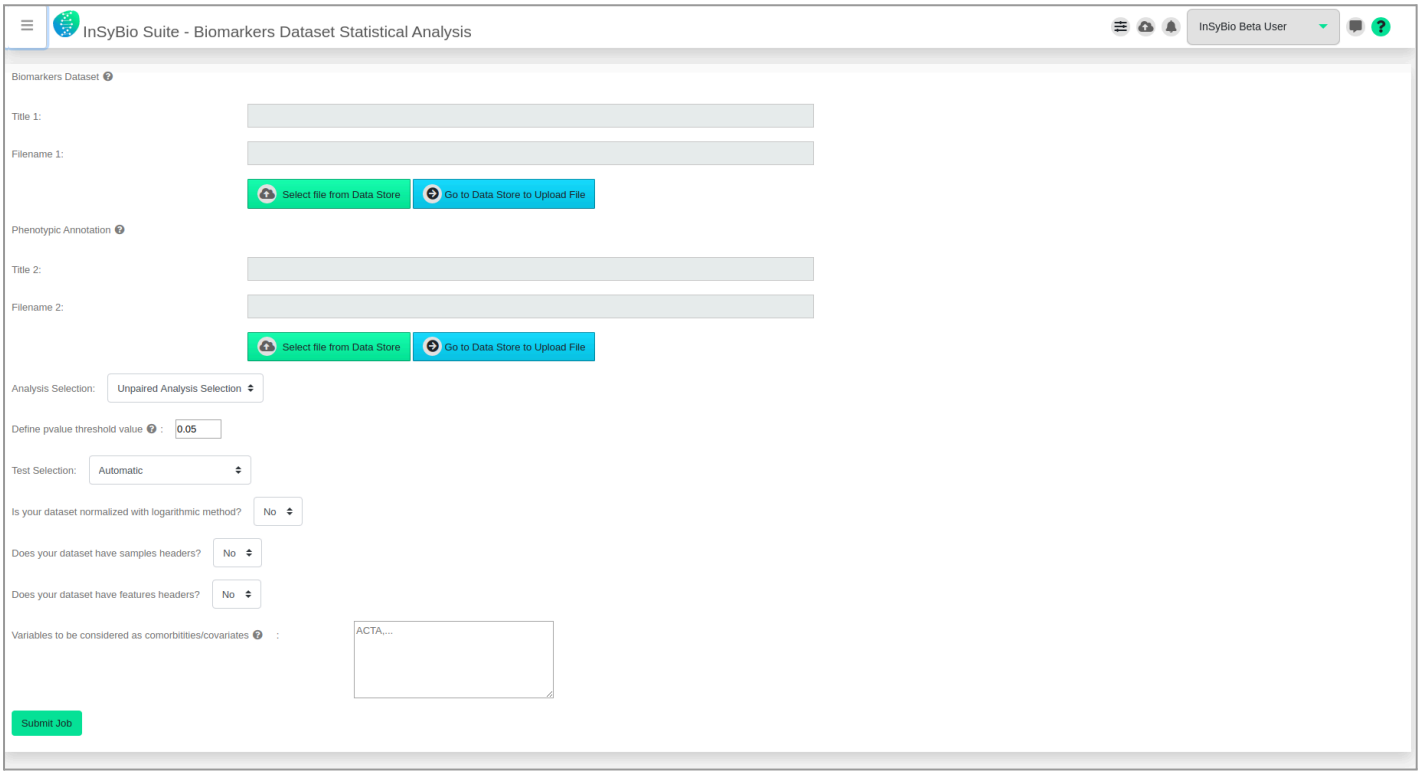
From the results, the user will be able to download the resulting preprocessed file and choose one of the two actions as a next step.

From this point the actions that the user will be able to select from are to perform statistical analysis of the dataset ("Biomarkers Dataset Statistical Analysis") or to train a predictive analytics model ("Training Multi-biomarker Predictive Analytics Model").

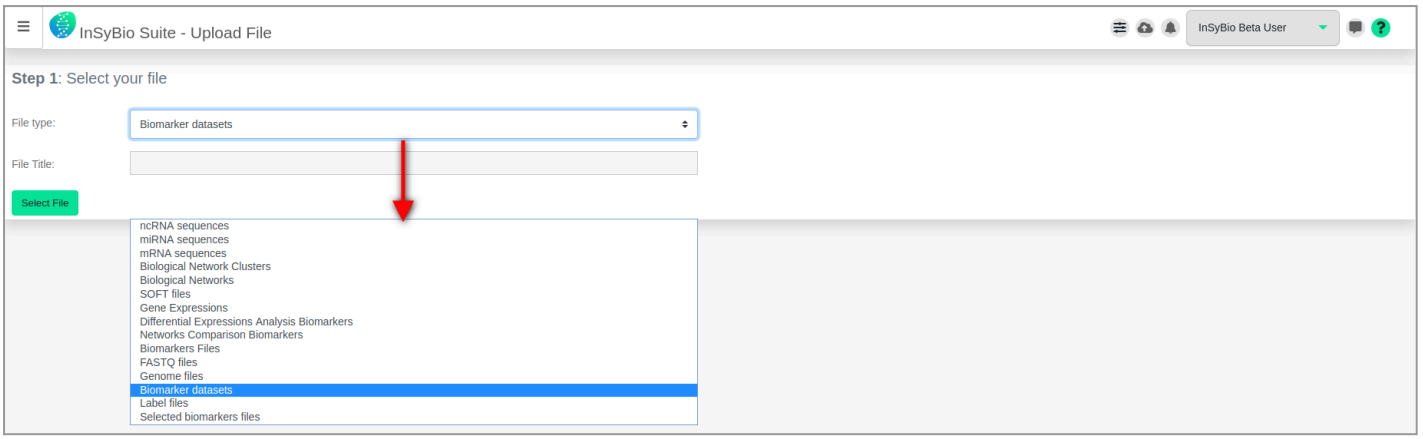
Biomarkers Dataset Statistical Analysis

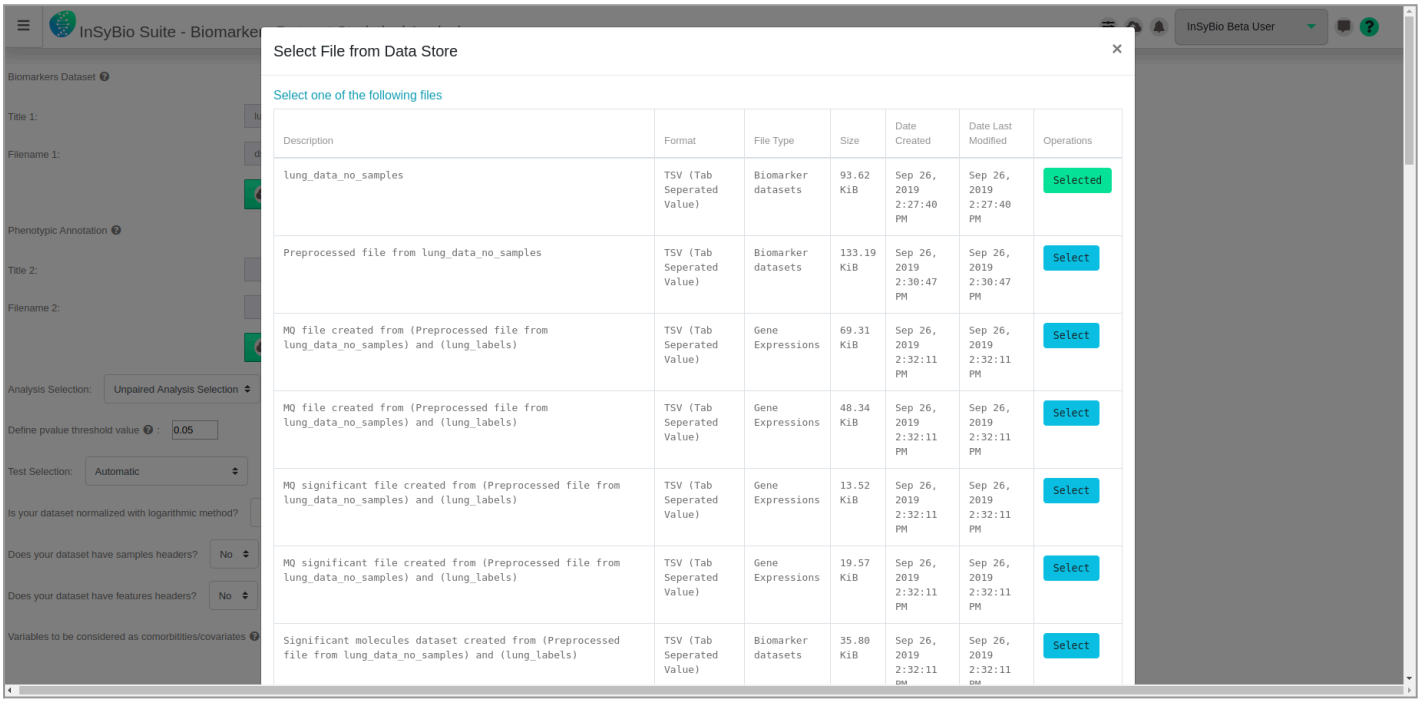
This page allows the user to perform simple tasks of statistical analysis on a biomarkers dataset including: Differential Expression Analysis, Heatmap construction and spearman correlation table construction.

Only variables annotated as genes/transcripts/proteins will be used for differential expression analysis. If a user has uploaded a phenotypic annotation file with more than two columns then multiple tasks will be created with one column of the phenotypic annotation per file. Every phenotypic column can take two or more values. Please note that these analyses are intended for classification problems, where phenotypic columns represent distinct groups or classes.



If the user visits the statistical analysis page after executing the Dataset Preprocessing step, the Biomarkers Dataset file will be pre-chosen for him. Otherwise, he'll have to upload a file on Datastore by selecting "Go to Data Store to Upload File" by selecting as filetype "Biomarker datasets", and then select the uploaded by pressing "Select file from Data Store", as shown in the examples below:





Afterwards, the user will have to either upload or select from the Data Store the phenotypic annotation which pairs the biomarkers dataset. If he uploads the labels file, the user should select as filetype “Label files”, as shown in the next image:

The labels (each one corresponding to a sample) should be all in one row, separated by tabs.

Later on, the user will have to select the type of analysis to be made on the inserted dataset. There are two types of statistical analysis, paired and unpaired analysis.

Afterwards the user will have to insert the p-value threshold value, which is recommended to be 0.05.

Then, the user will choose the kind of test to be performed on the selected dataset: automatic, parametric Ebayes Test Selection, parametric 2-sided Students T-test (or One-way ANOVA Test Selection) or non-parametric Kruskal Wallis (or Mann Whitney Test Selection) test. If the automatic version is chosen, then our algorithm will decide which test to run: the parametric or non-parametric.

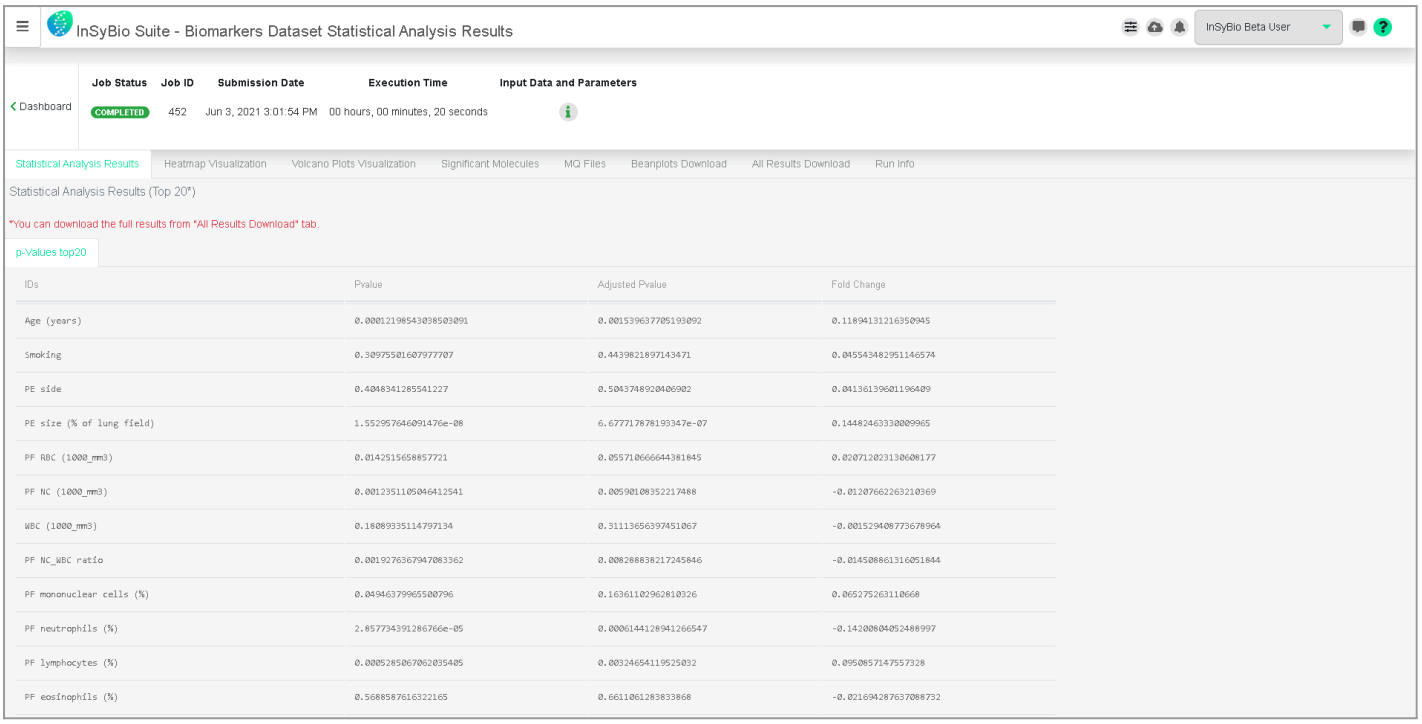
Afterwards the user will have to indicate if his dataset was preprocessed using logarithmic normalization or not.

Also, the user will have to insert the information regarding the headers of the input dataset. If the dataset has a sample header, “Yes” should be chosen as an answer to the first question, else “No”. If the dataset has a features (biomarkers) header, “No” should be chosen as an answer to the second question.

Finally, the user can also provide a set of household molecules (biomarkers) to perform geometric normalization of the data additionally to the normalization method already being performed. To do that the user will have to insert the names of molecules in the box that appears, separated with commas or a new line.

Householding molecules should represent a large abundance in the dataset, not have missing values and should not present high variability.

When the job is completed in the biomarkers dashboard, the user will be able to select to “View Results” getting to a page like the following:

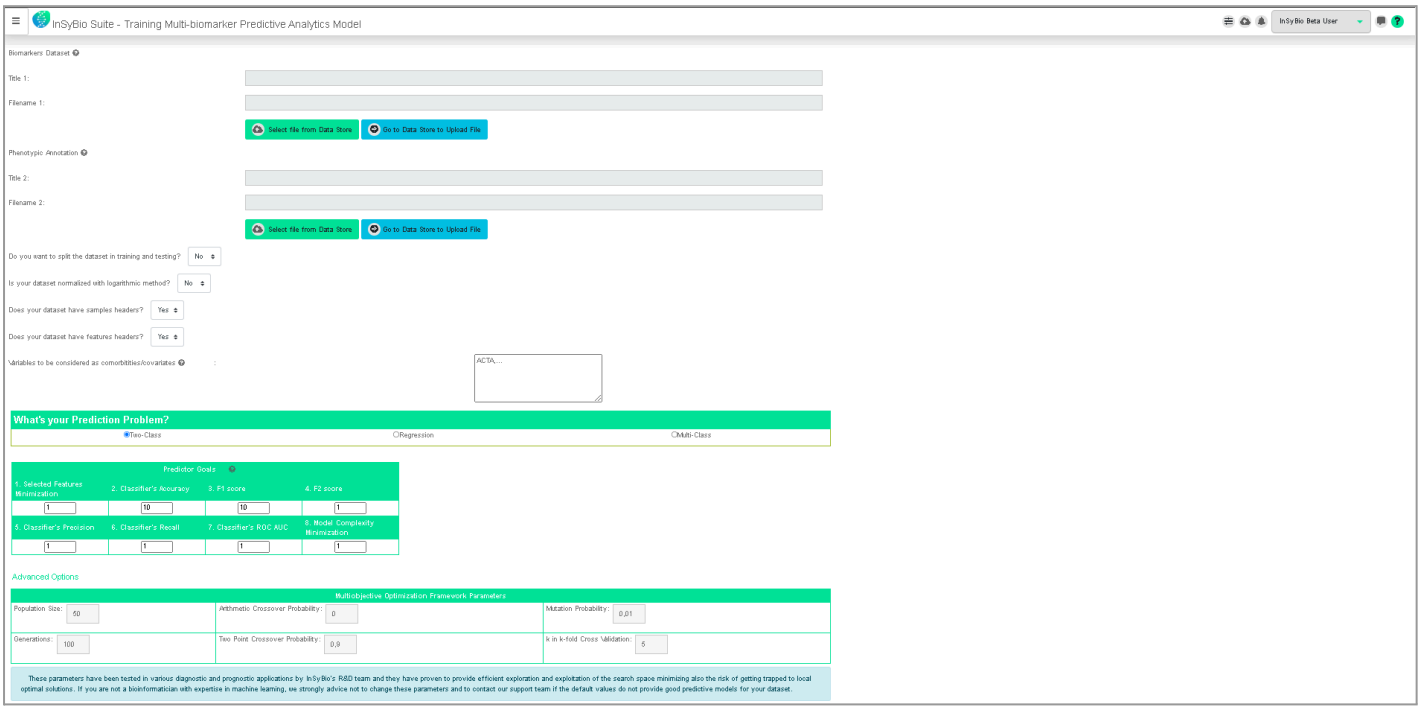


IDs	Pvalue	Adjusted Pvalue	Fold Change
Age (years)	0.000121985430385030 91	0.001539637705193092	0.11894131216350945
Smoking	0.309755016079777 07	0.4439821897143471	0.045543482951146574
PE side	0.4048341285541227	0.5043748920406902	0.041361396011196409
PE size (% of lung field)	1.552957646091476e-08	6.677717878193347e-07	0.14482463330009965
PF RBC (1000_mm3)	0.0142515658857721	0.055710666644381845	0.020712023130608177
PF NC (1000_mm3)	0.001235110504641254 1	0.00590108352217488	-0.012076622632103 69
WBC (1000_mm3)	0.18089335114797134	0.31113656397451067	-0.001529408773678964
PF NC_WBC ratio	0.001927636794708336 2	0.008288838217245846	-0.014508861316051844
PF mononuclear cells (%)	0.049463799665500796	0.16361102962810326	0.065275263110668
PF neutrophils (%)	2.857734391286766e-05	0.000614412894126654 7	-0.142008040524889 97
PF lymphocytes (%)	0.000528506706203540 5	0.003246541195250 32	0.095085714755732 8
PF eosinophils (%)	0.5688587616322165	0.6611061283833868	-0.021694287637088732

In the results page the user will be able to view the Statistical Analysis Results (the top 20 features), the Heatmaps, the Volcano plots, the significant molecules, the molecular quantification (MQ) files and he'll be able to download the Beanplots and all the resulting files. Finally, in the last tab the run information will be displayed.

Training Multi-biomarker Predictive Analytics Model

This page allows users to train their own predictors using a biomarkers dataset, a phenotypic annotation and by providing some parameters.



Firstly the user will have to input a preprocessed biomarkers dataset, selecting it from the Data Store or by uploading it. Then the user will have to insert a phenotypic annotation file.

Later on, the user will have to indicate if his dataset was preprocessed using logarithmic normalization or not.

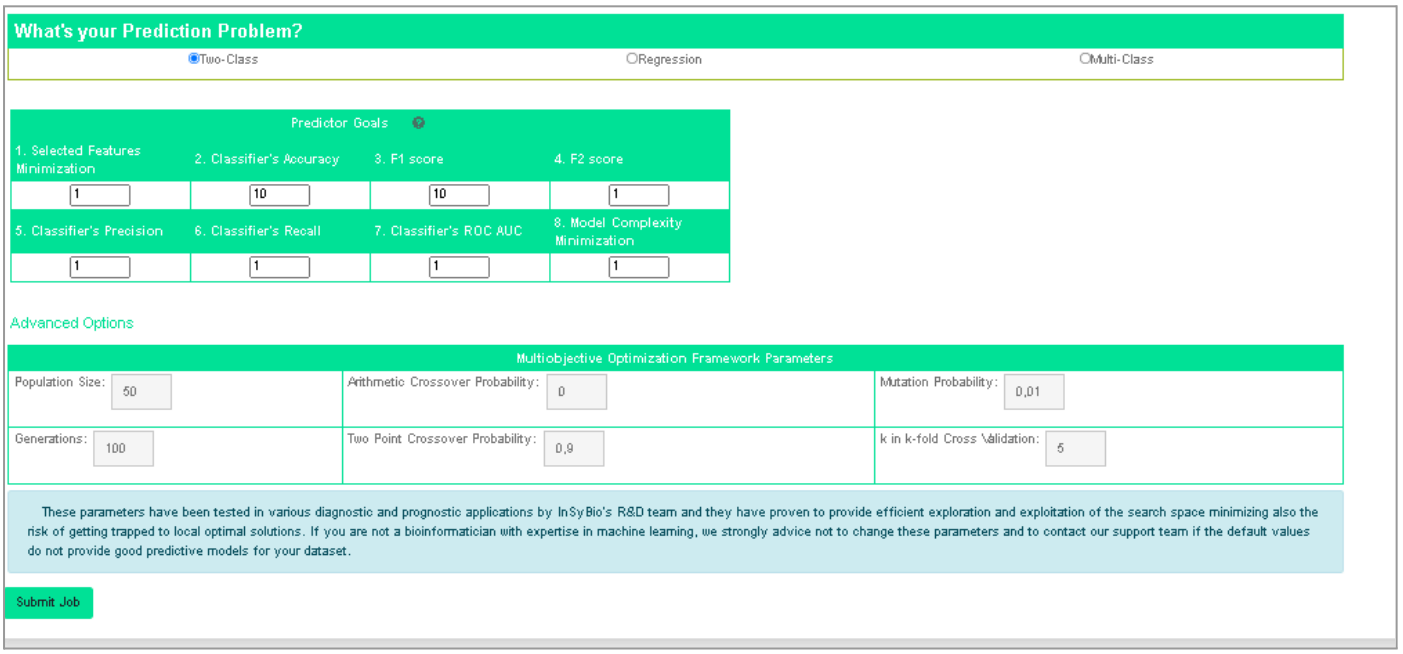
Afterwards, the user will have to choose to split or not the input dataset to training set and test set, by selecting the percentage ("Filtering percentage") of the original file

that will be the test set. If chosen, the model will be trained using the training set, and the test set will be stored in Data Store for later use.

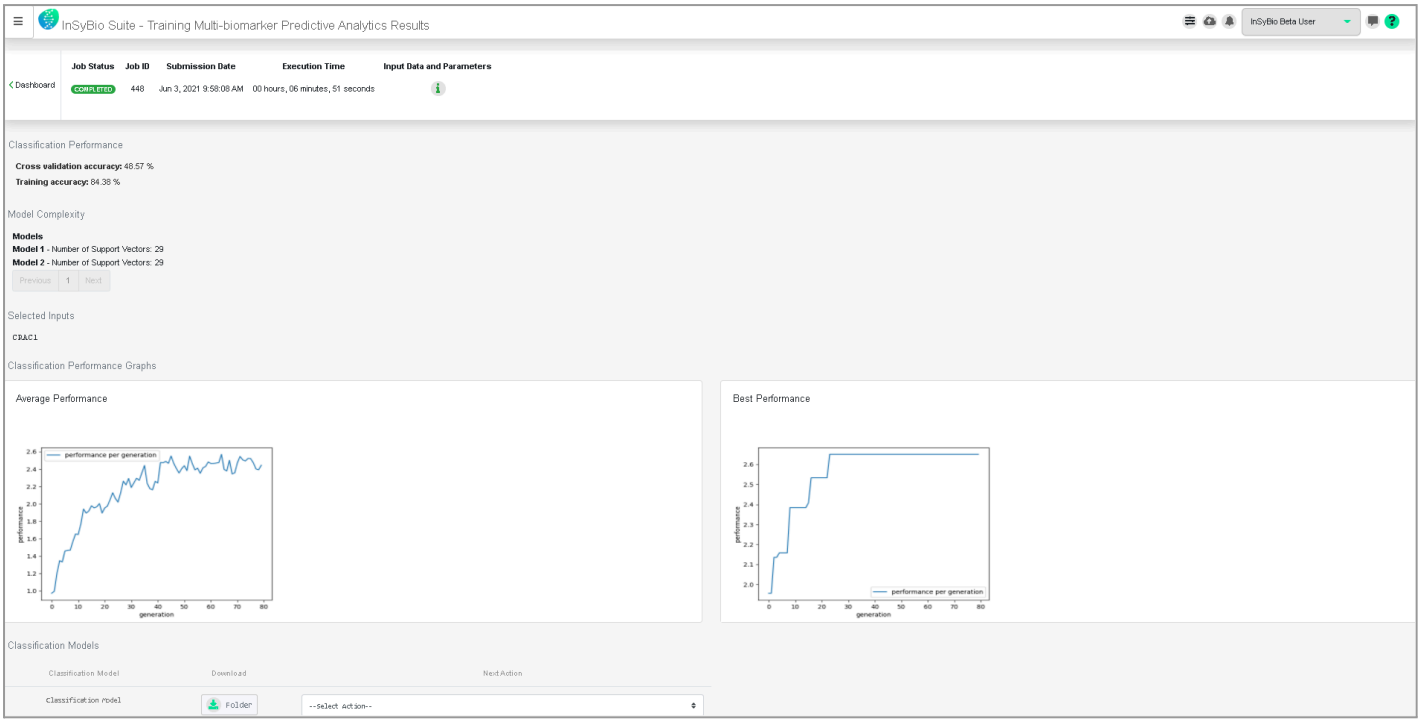
Then the user will have to insert the information regarding the headers. That is to inform the application if the original dataset has sample headers or feature headers. Optionally, the user will have the option to insert the names of the features that will be used for normalization.

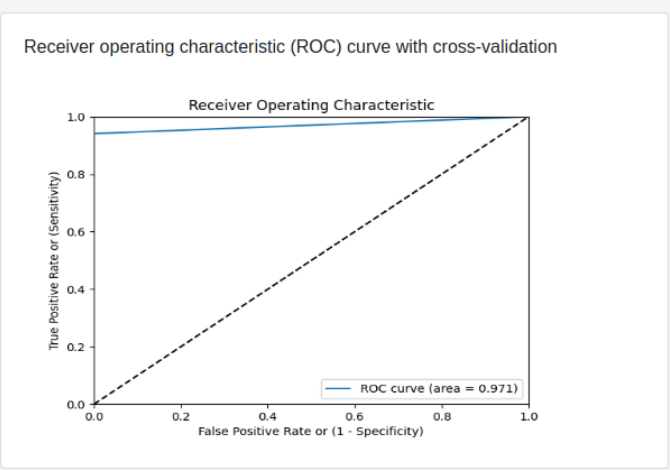
Eventually, the user chooses the kind of prediction problem he has at hand. Later, he can alter the weights of the predictor goals. It is advisable to use the default values. The higher the weight, the more significant the goal.

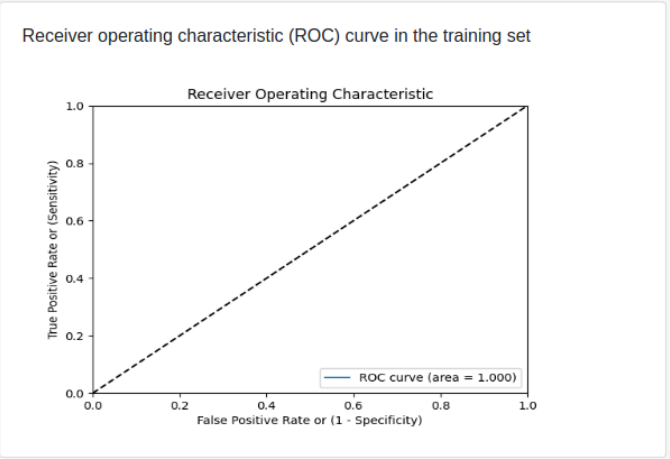
Finally, the user can alter the multi-objective optimization framework parameters by pressing the button “Show Advanced Options”. Those are the population size, the number of generations, the arithmetic crossover probability, the two point crossover probability, the mutation probability and the number of folds k for the cross validation.



The result page will have the following form:







As shown, the user will be able to see the input files that he's inserted, the type of prediction problem and the biomarker job information (submission date, start execution date, completion date, execution time and status).

Also the user will be able to view the classification (or regression) performance of the cross validation and of the training set. For the two-class prediction problem the user will be able to see the accuracy, sensitivity, and specificity. For the multi-class prediction problem the user will be able to see only the accuracy and for the regression prediction problem the user will be able to view the root mean square error.

Additionally, the complexity of every model, which is the total number of support vector machines, the number of trees for RandomForest and the number of neurons for the CNN and also, the average and the best performance of the trained model are being displayed.

Two more Roc Curves for the cross validation and the training set are also produced and displayed. An Roc Curve (Receiver Operating Characteristic Curve) is a graph showing the performance of a classification model at all classification thresholds using the True and False Positive Rates.

From this point the user will be able to continue by choosing as a next action the "Testing Multi-biomarkers Predictive Analytics Model".

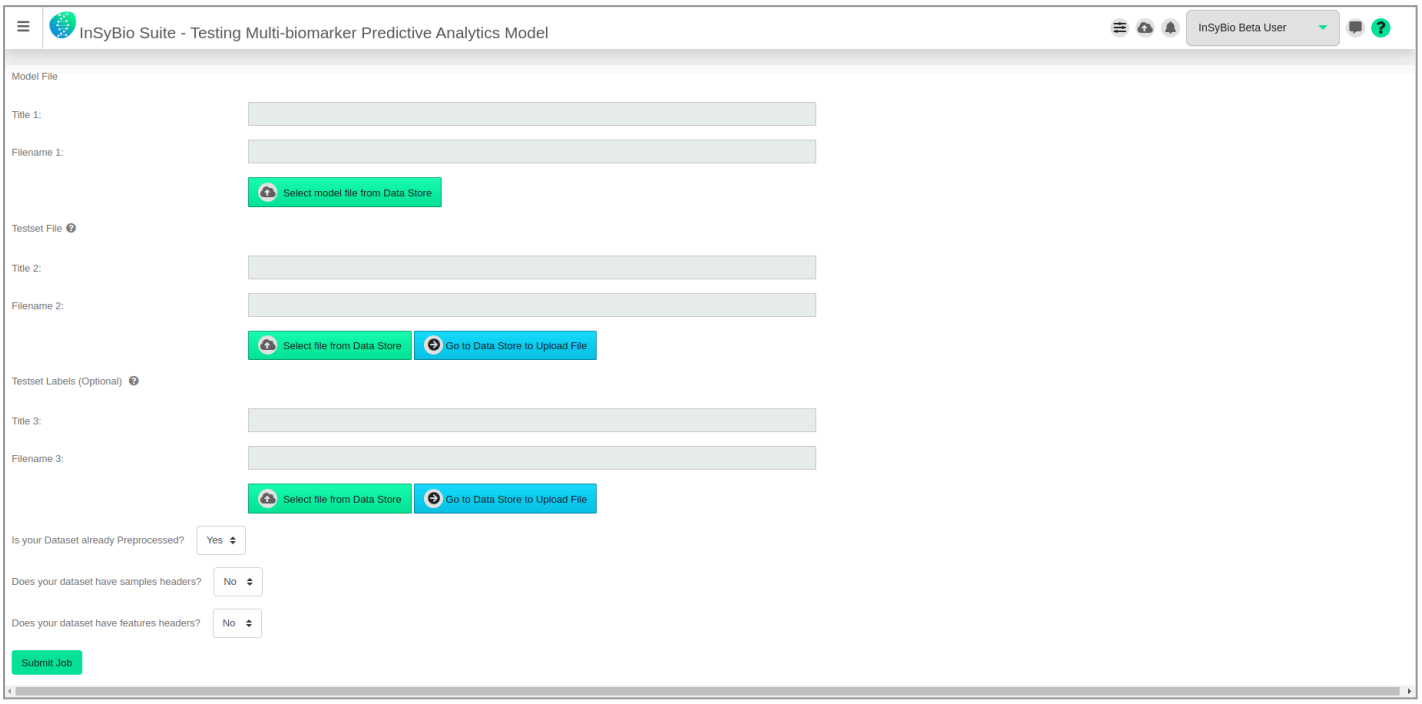
Testing Multi-biomarker Predictive Analytics Model

This page will allow the users to test the predictors that they have trained in the previous "training" step.

The first input is the model file which will have been autocompleted from the previous step.

The second input file is the test dataset, which can be preprocessed or not preprocessed. If the user has chosen in the training step to split that input dataset, the

form of the test dataset will have been completed with the produced test dataset. In such a case the user should insert that the input dataset doesn't have sample headers but it has featur headers.



The third input are the testset labels, which is optional. If the user inserts the testset labels then he'll receive as an output along with the predicted labels the performance metrics of the prediction.

The user may choose to test a non-preprocessed testset. Then he'll have to choose as normalization and missing imputation methods those methods that he'd chosen to preprocess the training dataset. He may also insert a set of variables for normalization, to use the geometric normalization method.

Is your Dataset already Preprocessed? No

Normalization Method Arithmetic

Missing Imputation Method Average

Set of variables for normalization ?: ACTA,...

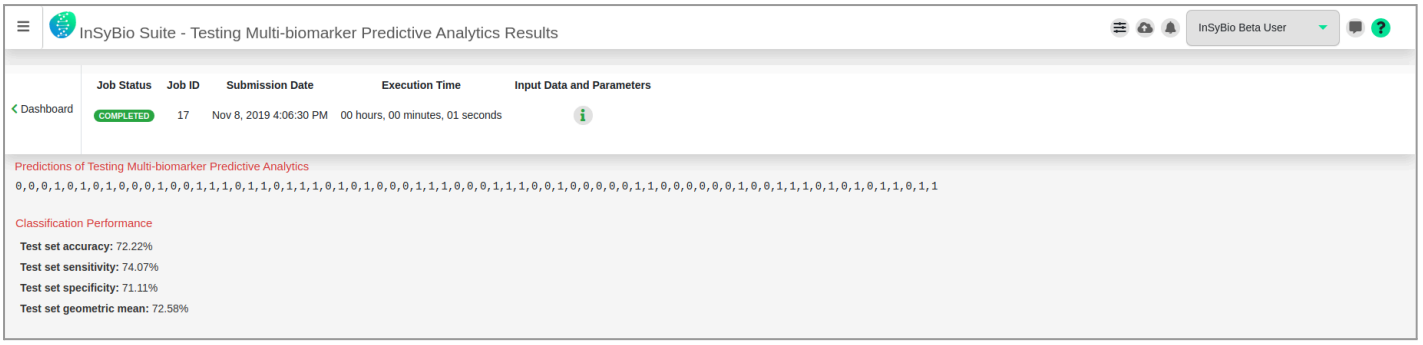
Does your dataset have samples headers? No

Does your dataset have features headers? No

Submit Job

It should be noted that the input dataset must have the format of the previous functionalities, that is it should have as rows the features and as columns the samples.

The results the user will be getting are the following. Firstly he will view the input data and the chosen parameters, such as the model file, the test dataset file, the labels file, and the type of the problem. Then he'll be able to view the biomarker job information, such as the submission date, the start execution date, the completion date, the execution time, and the status (completed or not). Finally he'll view the predicted labels and the performance metrics. For the two classification problems (two-class, multi-class) he'll view the test set accuracy, specificity, sensitivity and the geometric mean of specificity and sensitivity. For the regression he'll view the testset mean squared error and the test set squared correlation coefficient. The above are being displayed in the following image as an example.


InSyBio Suite - Testing Multi-biomarker Predictive Analytics Results
InSyBio Beta User
Dashboard
Job Status
Job ID
Submission Date
Execution Time
Input Data and Parameters
COMPLETED
17
Nov 8, 2019 4:06:30 PM
00 hours, 00 minutes, 01 seconds
Predictions of Testing Multi-biomarker Predictive Analytics
0,0,0,1,0,1,0,1,0,0,0,1,0,0,1,1,1,0,1,1,0,1,1,1,0,1,1,0,1,0,1,0,0,1,1,1,0,0,1,1,1,0,0,1,0,0,0,0,1,1,0,0,0,0,0,1,0,0,1,1,0,1,0,1,0,1,1,0,1,1
Classification Performance
Test set accuracy: 72.22%
Test set sensitivity: 74.07%
Test set specificity: 71.11%
Test set geometric mean: 72.58%

How to get InSyBio Biomarkers

To request a free one month license of InSyBio Suite please email us at info@insybio.com .

To purchase InSyBio Biomarkers commercial version 3.3 please contact us at sales@insybio.com.

About Us

InSyBio Ltd is a bioinformatics pioneer company (www.insybio.com) in personalized healthcare, that focuses on developing computational frameworks and tools for the analysis of complex life-science and biological data in order to develop predictive integrated biomarkers (biomarkers of various categories) with increased prognostic and diagnostic aspects for the personalized Healthcare Industry.

InSyBio Suite consists of tools for providing integrated biological information from various sources, while at the same time it is empowered with robust, user-friendly and installation-free bioinformatics tools based on intelligent algorithms and methods.

COPYRIGHT NOTICE

External Publication of InSyBio Ltd - Any InSyBio information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the InSyBio Ltd. A draft of the proposed document should accompany any such request. InSyBio Ltd reserves the right to deny approval of external usage for any reason.

Copyright 2024 InSyBio Ltd. Reproduction without written permission is completely forbidden.