Intro to Big Data

BigML platform

Group Information
19127468 - Phan Đức Mạnh
21127078 - Nguyễn Duy Đăng Khoa
21127084 - Lê Xuân Kiên
21127108 - Đặng Hà Nhật Minh

Key points

1. Overview of the BigML platform
2. Data Modeling on BigML
    a. Dataset: Wincosin Breast Cancer
    b. Model: Decision Trees
    c. Model: Logistic Regression
    d. Model: Clustering
    e. Model: Association
3. Conclusion

Overview

— BIGML

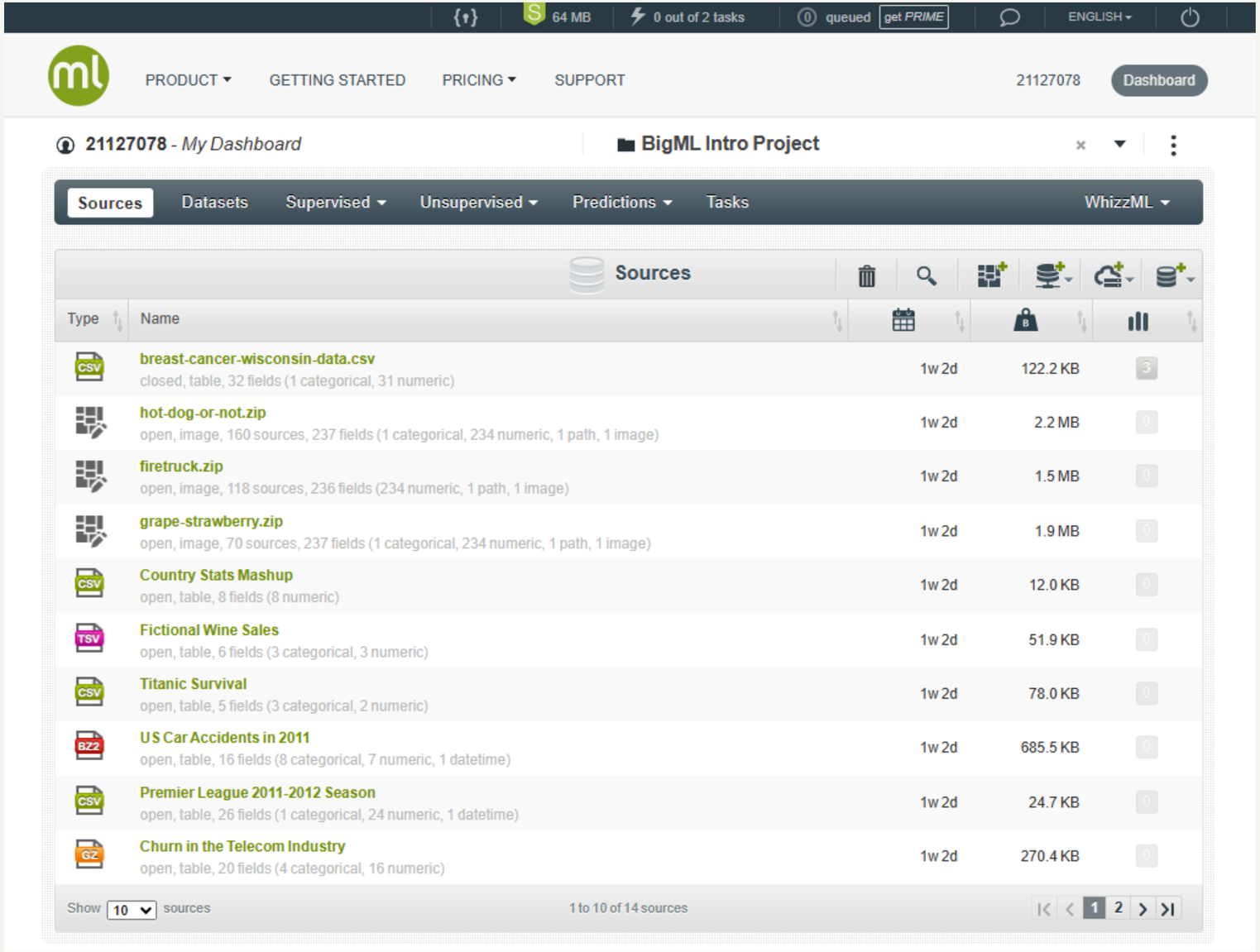
01 **BigML** - a holistic set of tools designed to **simplify** the creation and implementation of machine learning models, expanding accessibility to a wider audience.

02 Founded in 2011 with a clear vision: to make machine learning **easy and accessible** for everyone.

03 BigML can help tackle many problems: Classification, Regression, Time Series, Cluster Analysis, Topic Modeling,...

Some components

- **Sources:** include file formats and upload options, or advanced parsing configuration options.
- **Datasets:** filter, sample, add new fields, or split a dataset into training and test datasets.
- **Supervised, Unsupervised:** Initialize and configure models for training, evaluation and prediction
- **Predictions:** make individual Predictions or generate Batch Predictions for a group of new instances.



BigML vs. others

	Paid/Free	Availability	Deployment	Services	Purpose
bigml	Partially free	Dependent on BigML resource allocation	Cloud-based	Focused on easy and scalable automation of ML tasks	Simplifies machine learning
Google Big Query	Pay-as-you-go	Based on chosen deployment regions and settings	Cloud-based	Part of the GCP ecosystem, easily integrable with other GCP products	Analyzing large datasets of petabytes of data
Keras	Free	Always	Local (cloud configurable)	Has frameworks for deep learning models	Open-source deep learning framework
alteryx	Subscription	Based on deployment (Local-run or rented from Tableau server)	Cloud-based	Focused on providing real-time analytics to users	Drag-and-drop interface, supports building data pipelines

HCMUS

Data Modeling

ON BIGML

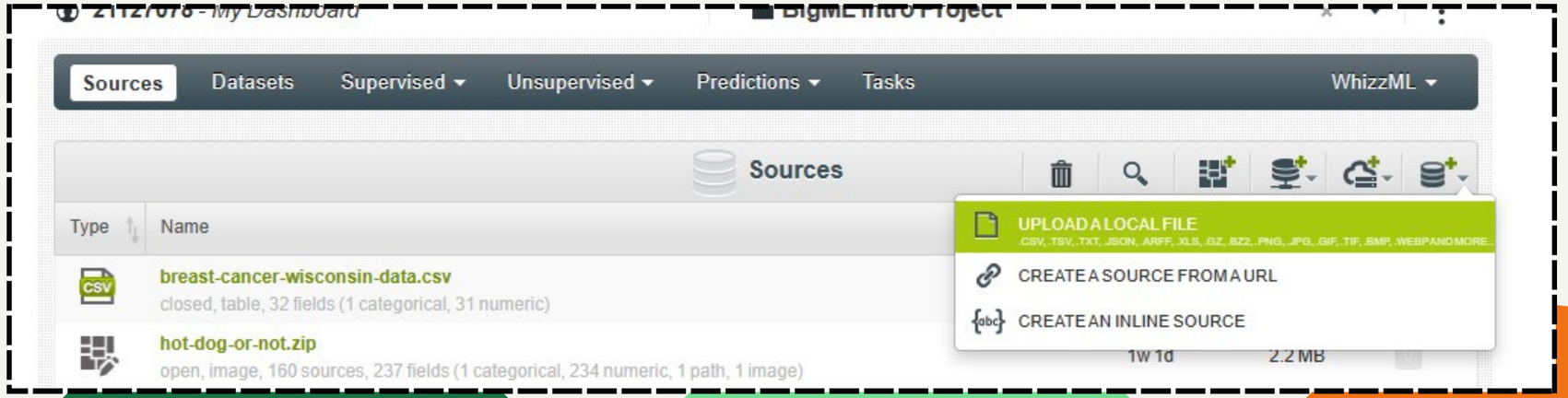
INTRO TO BIG DATA

Introduction to Dataset

BREAST CANCER DATASET (WISCONSIN)

This dataset is taken from the UCI Machine Learning Repository (Link: https://data.world/health/breast-cancer-wisconsin) by the Donor: Nick Street

Available in csv format
Can be imported to BigML through **Sources**
BigML supports many filetypes and databases

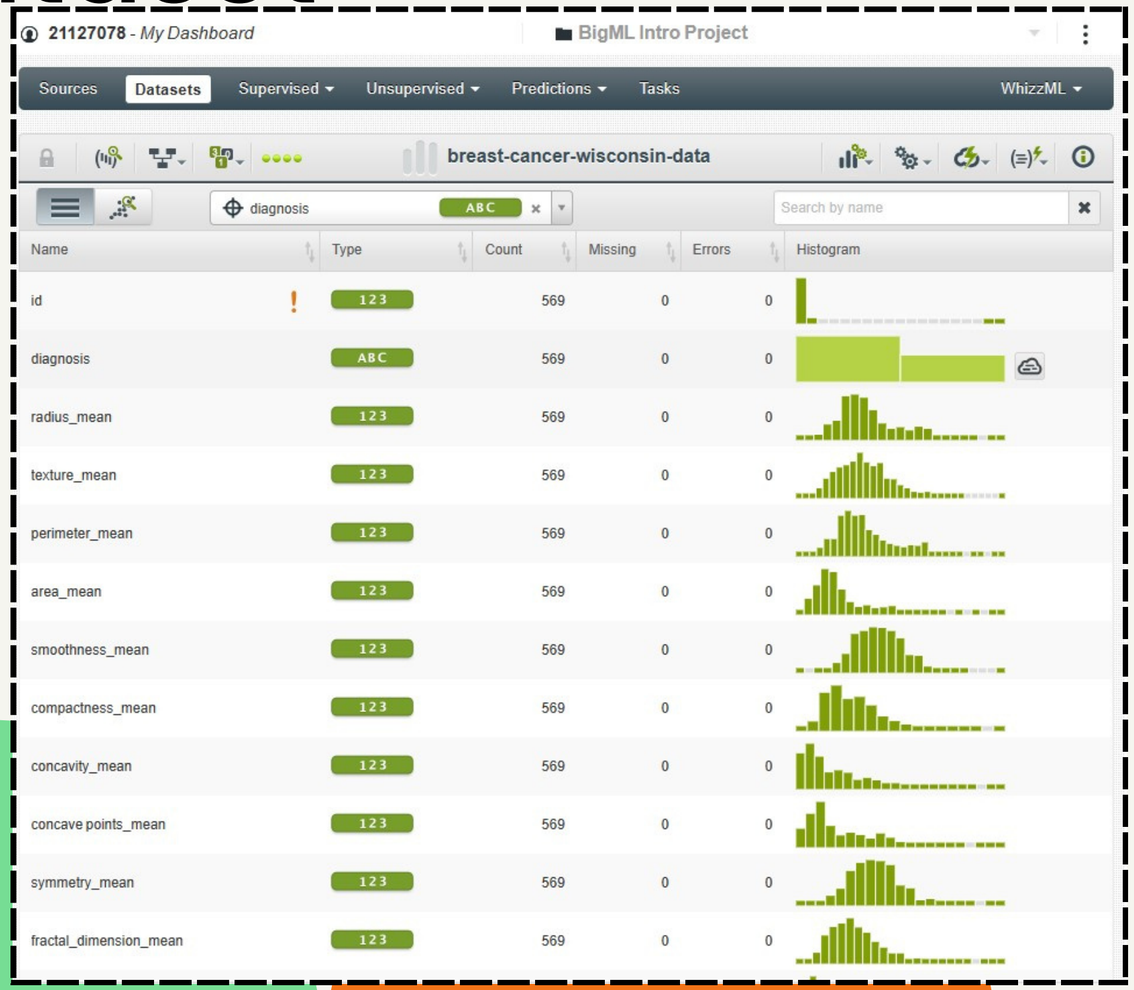


Introduction to Dataset

BREAST CANCER DATASET (WISCONSIN)

Dataset view

- Column name, Data type, Count, missing values,...
- Also provides a scatterplot function
- Action menu: used to interact with the dataset



Name	Type	Count	Missing	Errors	Histogram
id	123	569	0	0	
diagnosis	ABC	569	0	0	
radius_mean	123	569	0	0	
texture_mean	123	569	0	0	
perimeter_mean	123	569	0	0	
area_mean	123	569	0	0	
smoothness_mean	123	569	0	0	
compactness_mean	123	569	0	0	
concavity_mean	123	569	0	0	
concave points_mean	123	569	0	0	
symmetry_mean	123	569	0	0	
fractal_dimension_mean	123	569	0	0	

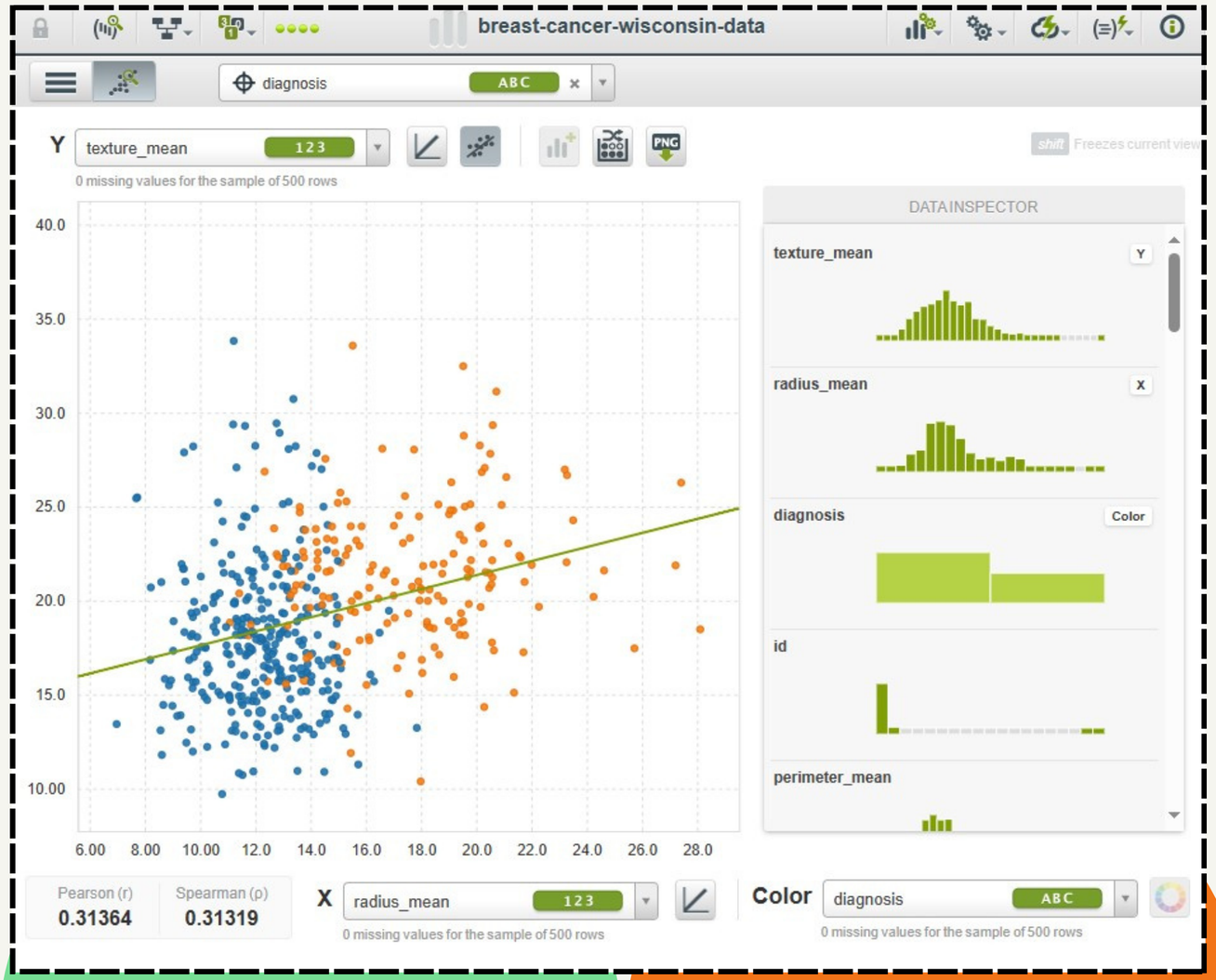
Introduction to Dataset

BREAST CANCER DATASET (WISCONSIN)

Very useful scatterplot function

Supports many operations:

- Log scaling
- Pearson/Spearman correlation coefficient
- Sampling
- Regression line



Model: Decision Tree

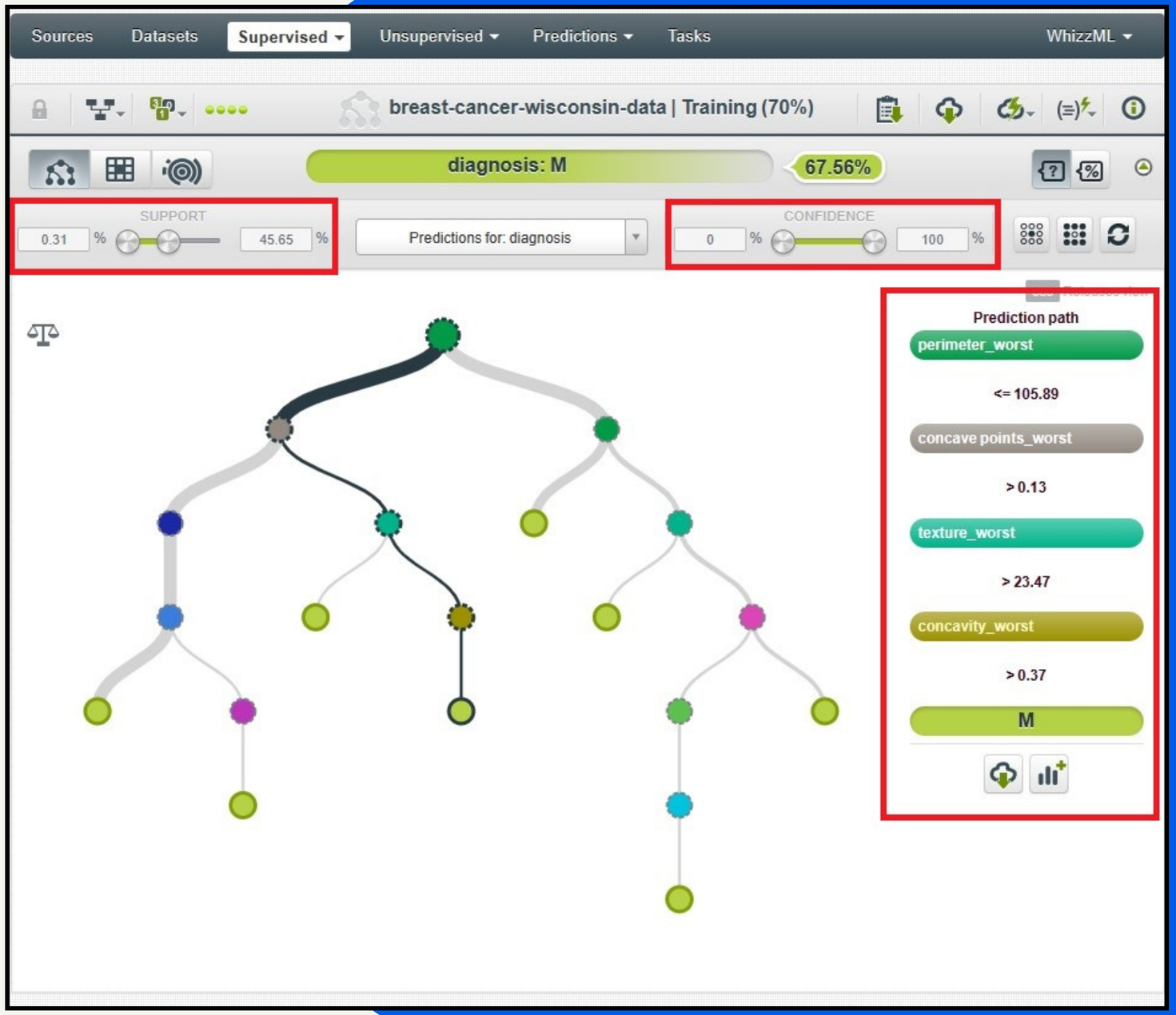
01 CART-styled decision trees for both **classification** and **regression**

02 BigML implements many different measures in building a tree, such as **early splitting, pruning, field importance...**

03 Many configuration options: **missing splits, node threshold, predefined weights,...**

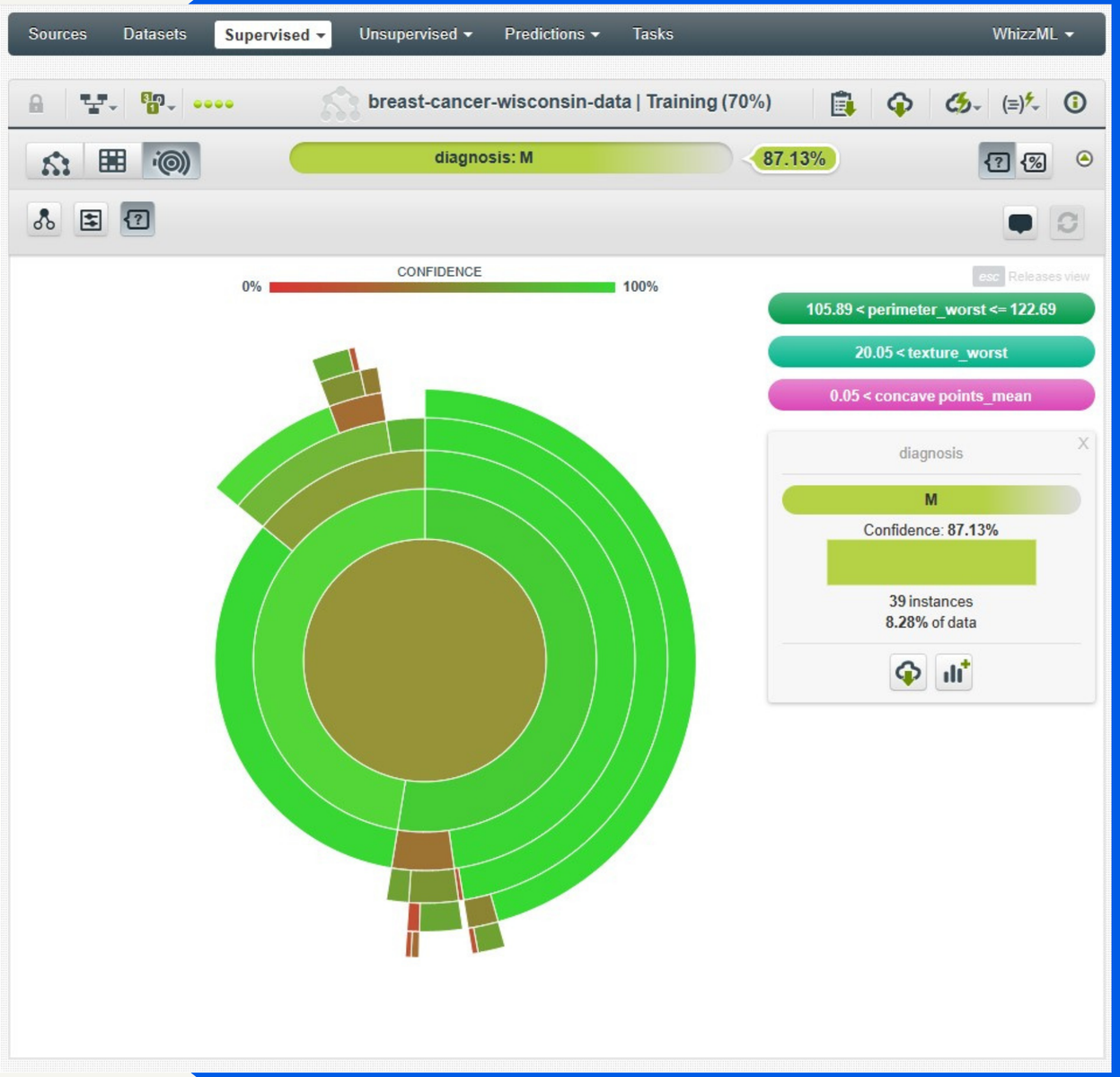
Visualization: Tree

- **Prediction path:** Hover over a node to view its path
- **Two metrics:** confidence and support
- **Filters:** threshold, frequent patterns, rare interesting patterns.



Visualization: Sunburst

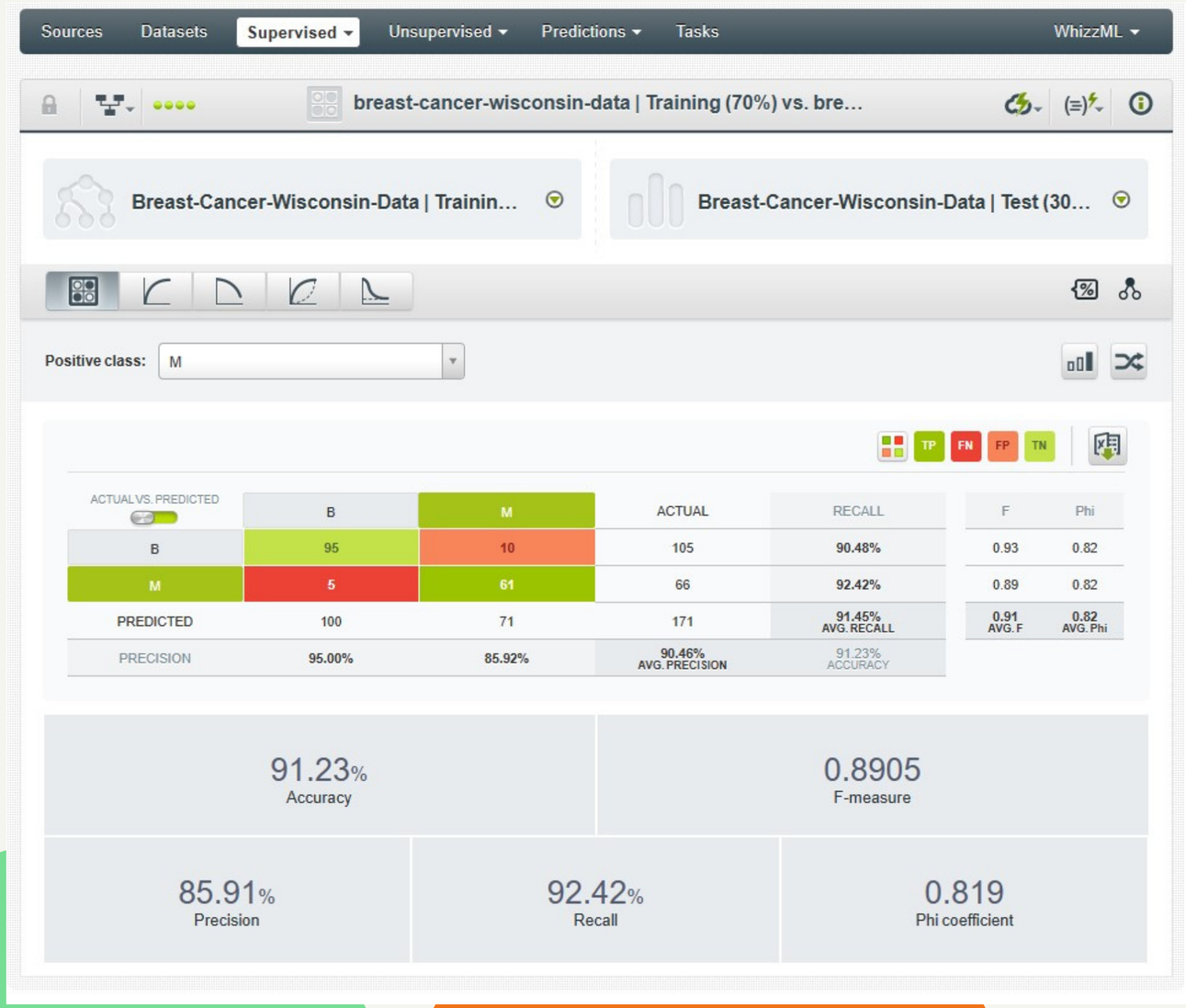
- "Top-down" visualization of the model



Evaluation

ON THE TEST SPLIT

For classification, BigML creates a confusion matrix and calculates many useful metrics: **Accuracy, Precision, Recall, F-measure, Phi coefficient.**
For this dataset problem, we should monitor the **precision** metric.
BigML can also plot the **ROC curve, precision-recall curve**



Model: logistic regresion

01 BigML's logistic regressions including how they can be created with 1-click, all configuration options available, and the different visualizations provided by BigML

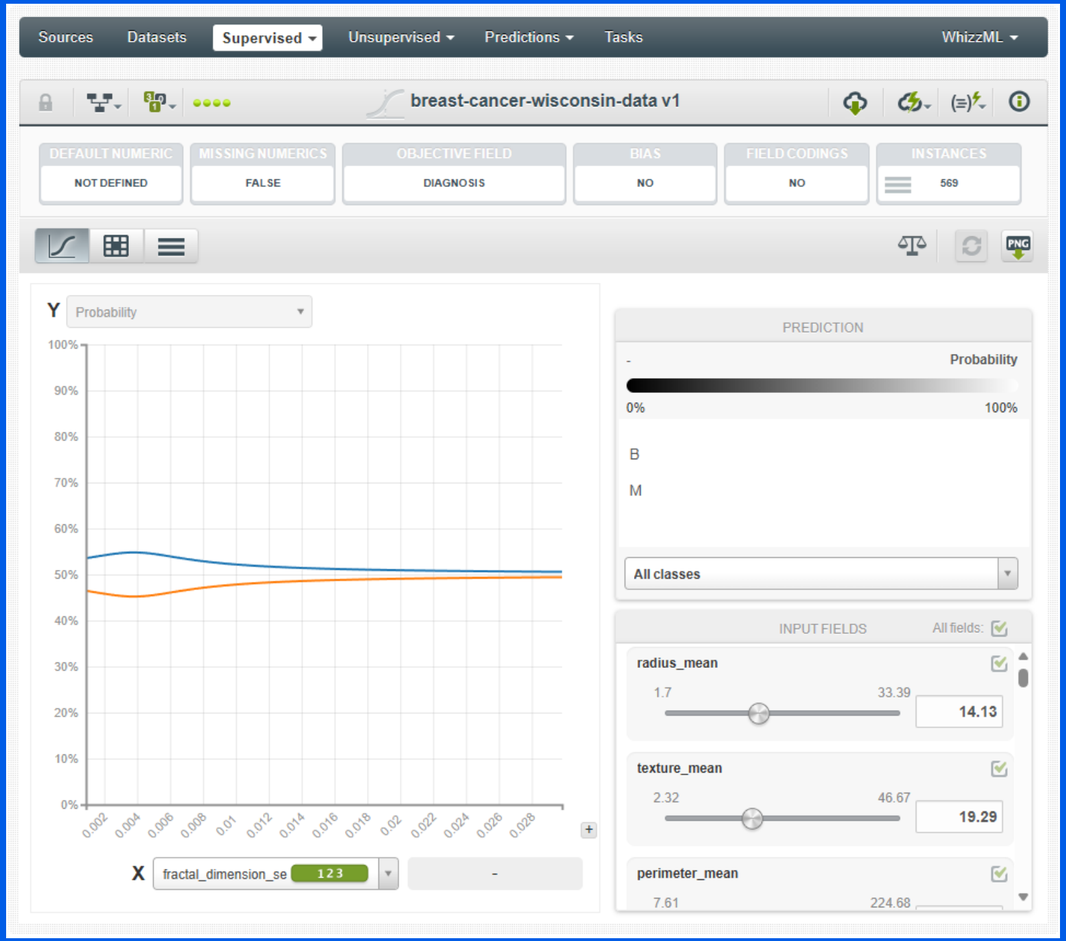
02 You can make predictions for single instances or for many instances in batch

03 You can also export your logistic regressions in different formats to make local predictions faster at no cost

Visualization: Chart

The chart view is composed of three main parts: the CHART itself, the PREDICTION legend, and the INPUT FIELDS form.

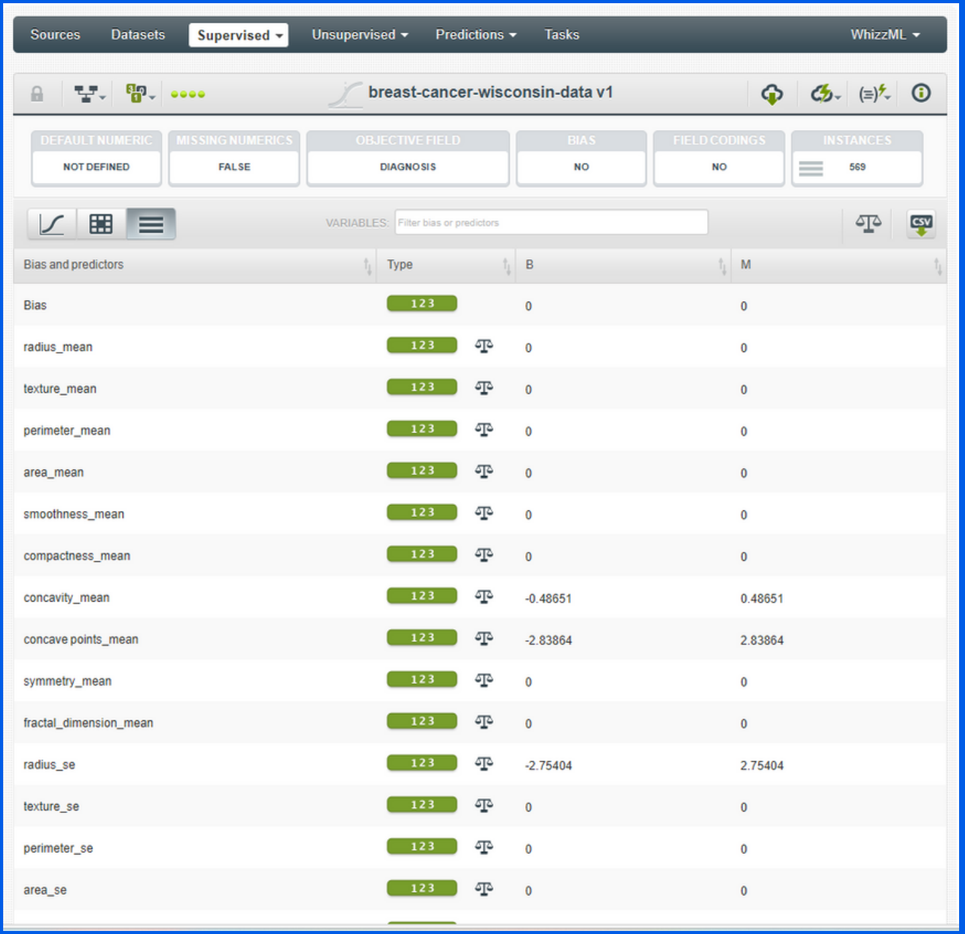
- **The Chart:** allows you to view the impact of the input fields on the objective classes predictions.
- **The PREDICTION legend:** allows you to visualize the classes represented in the chart along with their corresponding colors.
- **The INPUT FIELDS form:** You can config ure the values for any numeric, categorical, text or items field.



Visualization: Table

The main goal of the logistic regression algorithm is to learn the coefficients of the logistic function for each of the dependent variables, i.e., for each of the input fields. A different set of coefficients is associated with each class of the objective field.

BigML allows you to inspect the learned coefficients for each one of the input fields in the coefficient table. The table columns represent the objective field classes while the table rows represent the input field variables and the bias (a.k.a. intercept term) of the logistic regression. In the first row you will always find the Bias coefficients.

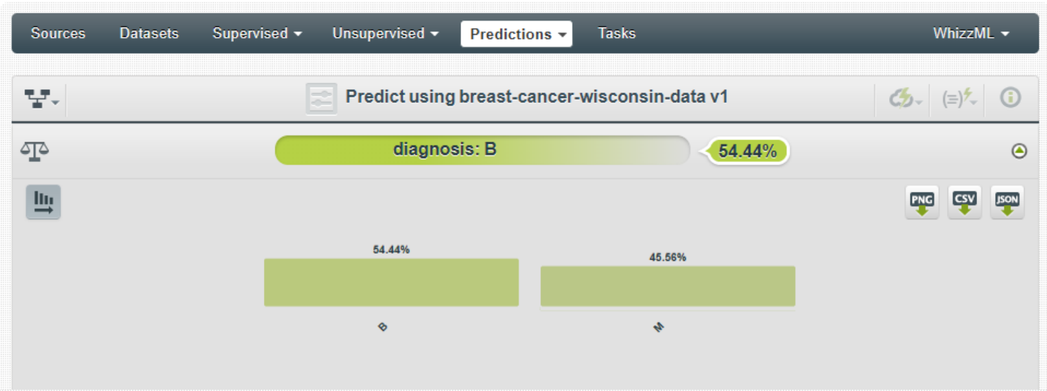


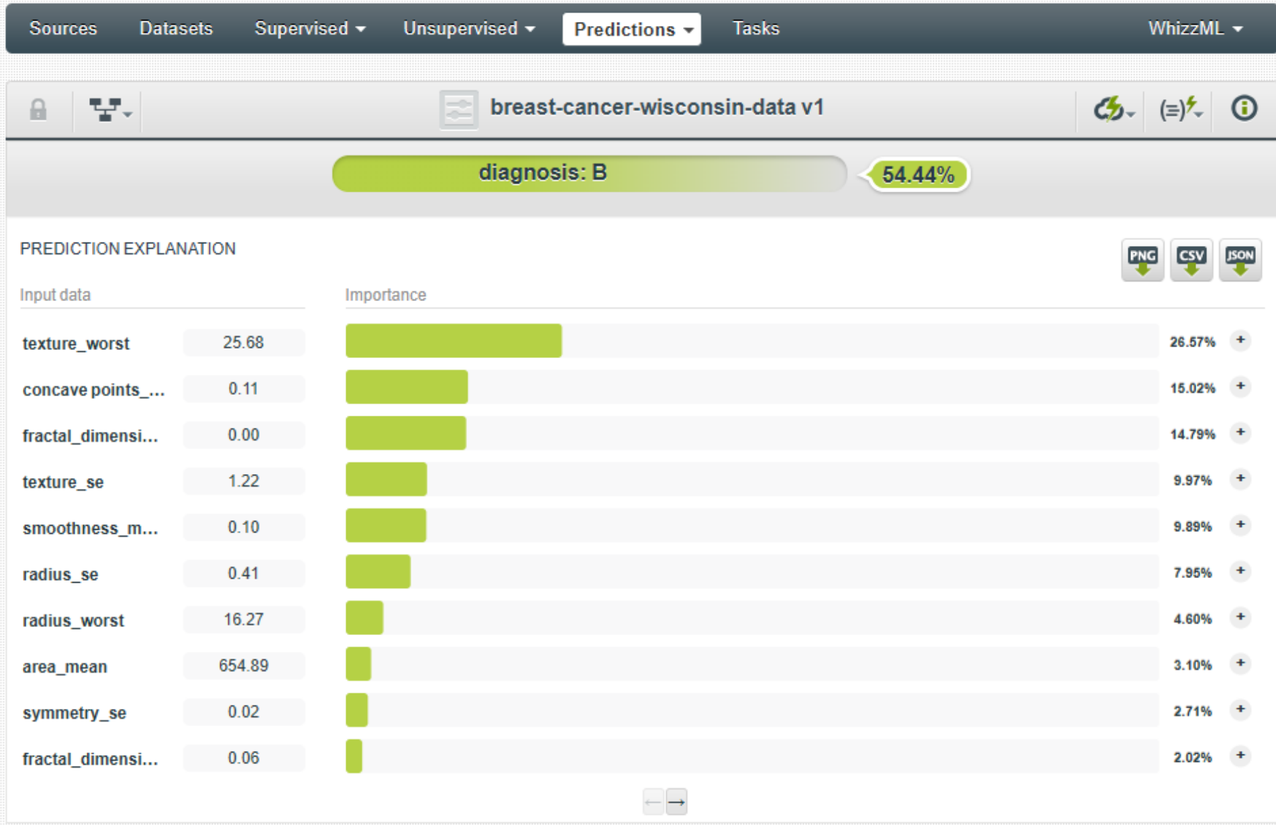
Bias and predictors	Type	B	M
Bias	123	0	0
radius_mean	123	0	0
texture_mean	123	0	0
perimeter_mean	123	0	0
area_mean	123	0	0
smoothness_mean	123	0	0
compactness_mean	123	0	0
concavity_mean	123	-0.48651	0.48651
concave points_mean	123	-2.83864	2.83864
symmetry_mean	123	0	0
fractal_dimension_mean	123	0	0
radius_se	123	-2.75404	2.75404
texture_se	123	0	0
perimeter_se	123	0	0
area_se	123	0	0

Prediction

Get the prediction at the top of the view along with the predicted class probability. BigML predictions are synchronous, i.e., when you send the input data, you get an immediate response. Moreover, single predictions from the BigML Dashboard are performed locally, so unless you save your prediction, it will not consume any credits and it will be updated instantly when you change your input values.

The prediction explanation represents the most important factors considered by the logistic regression in a prediction given the input values. Each input value will yield an associated importance. The importances across all input fields should sum 100%.





Model: Clustering

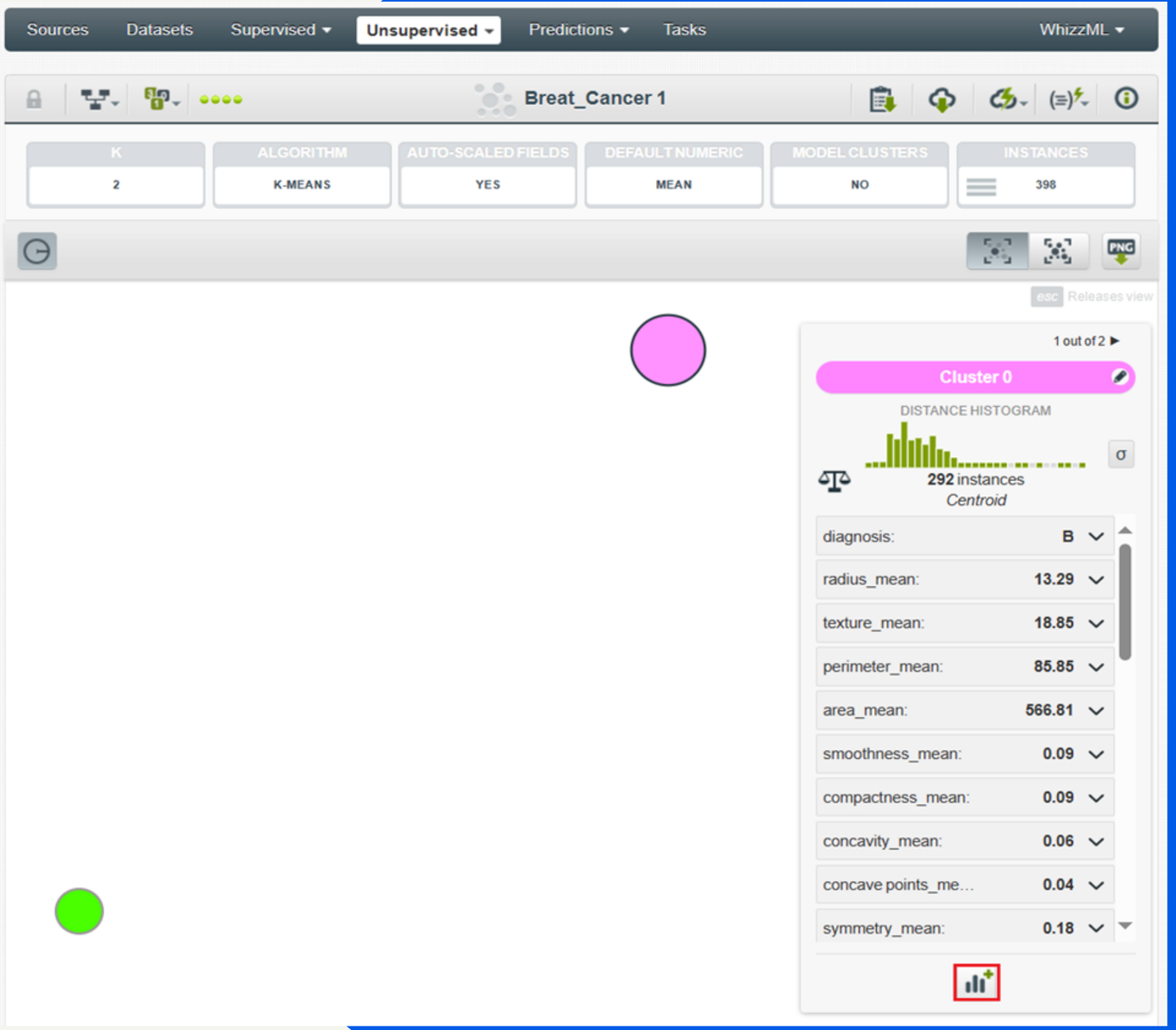
01 BigML clusters use optimized versions of **K-means** and **G-means** algorithms to group the instances. Each cluster group is represented by its center (or centroid).

02 All BigML field types are valid inputs for **Cluster Analyses,** although there are a few caveats.

03 BigML provides several strategies for dealing with them, or those instances may also be excluded entirely when computing the clusters.

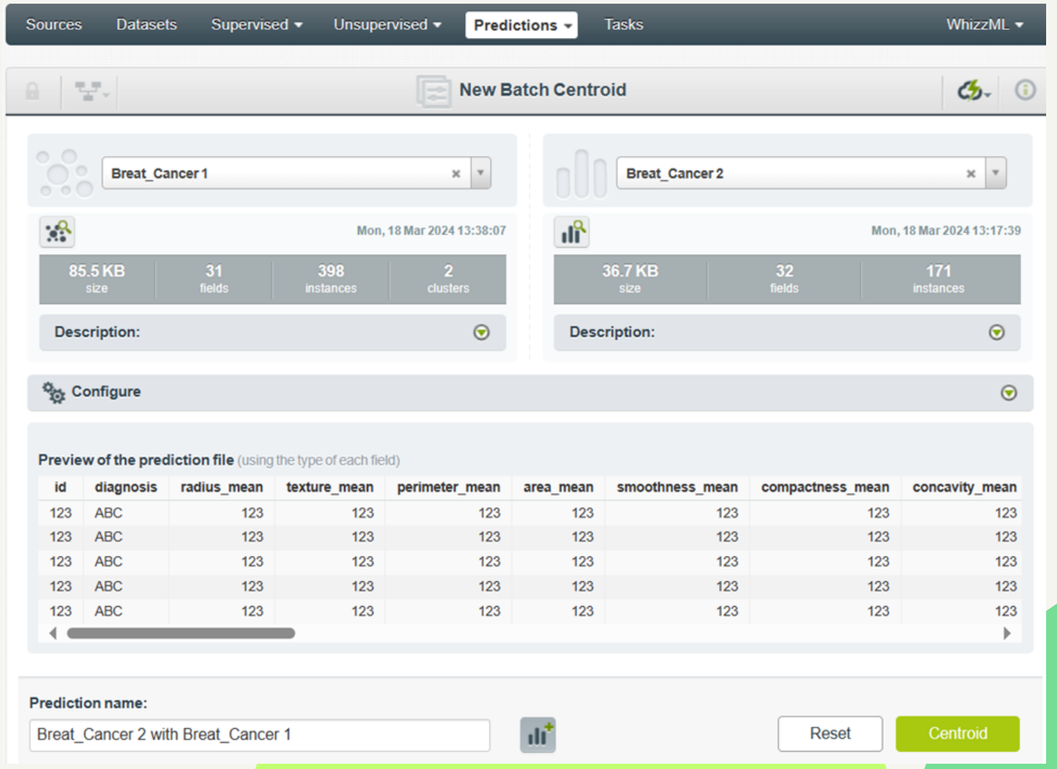
Visualization: Clustering

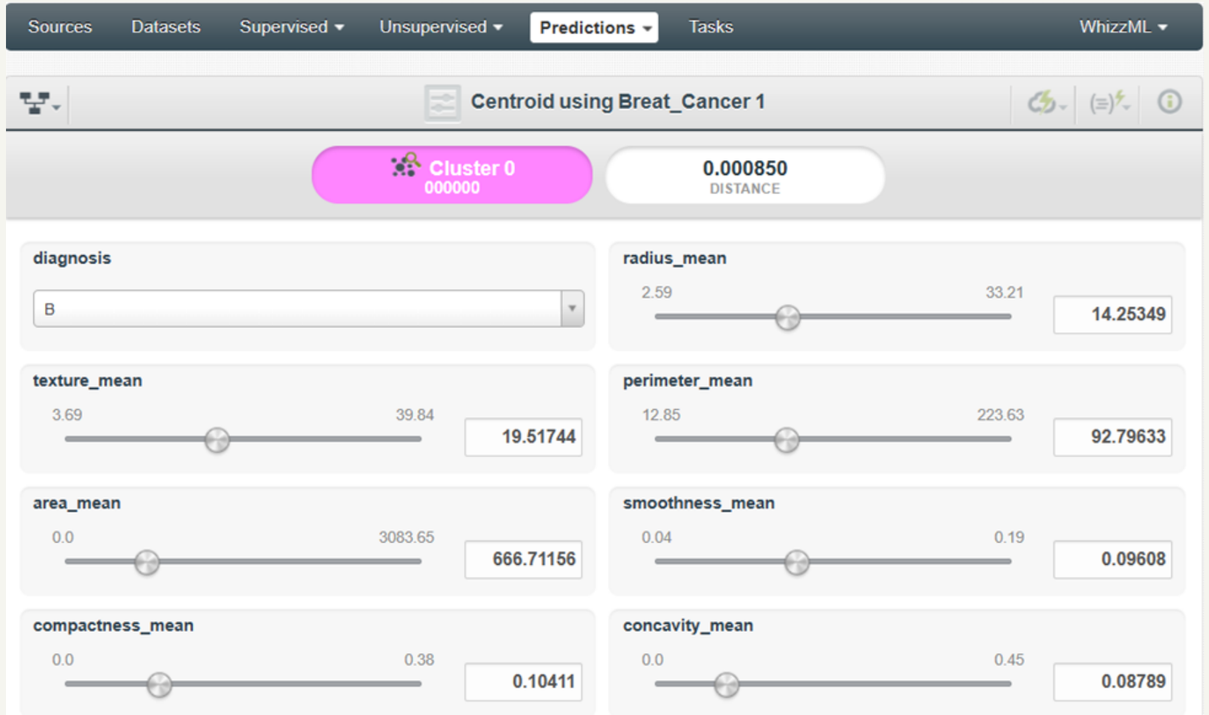
- **Cluster information:** Click any planet to view its information
- **Size differences:** Express the correlation of data size among clusters.
- **dataset generating:** Click the graph icon to generate a new dataset.



Feature: Batch Centroids

- Batch all centroids to another similar dataset and add a label field





Feature: Add centroids

- Add a centroid to the clustering model, with information can be adjusted

Model: Association

01 Association Discovery is an unsupervised machine learning method based on rules for uncovering relationships between variables in large datasets. This technique aims to identify statistically significant rules, moving beyond simple correlations to reveal complex rules. Statistically significant association rules help answer questions like which products are often bought together or what might be a user's next action.

02 Beyond market basket analysis and next best offer, this technique is applied in recommender systems, cross-sell/upsell analysis, marketing campaign analysis, web usage mining, digital forensics, continuous production, bioinformatics, and many other scientific applications.

A table that summarizes all the rules discovered

breast-cancer-wisconsin-data

ASSOCIATIONS (K)	ITEMS	SEARCH STRATEGY	MIN. SUPPORT	MIN. CONFIDENCE	INSTANCES
100	37	LEVERAGE	10.7210%	42.9250%	569

ITEMS: Filter items or values

LEVERAGE 7.87%

Association — 100% OF INSTANCES

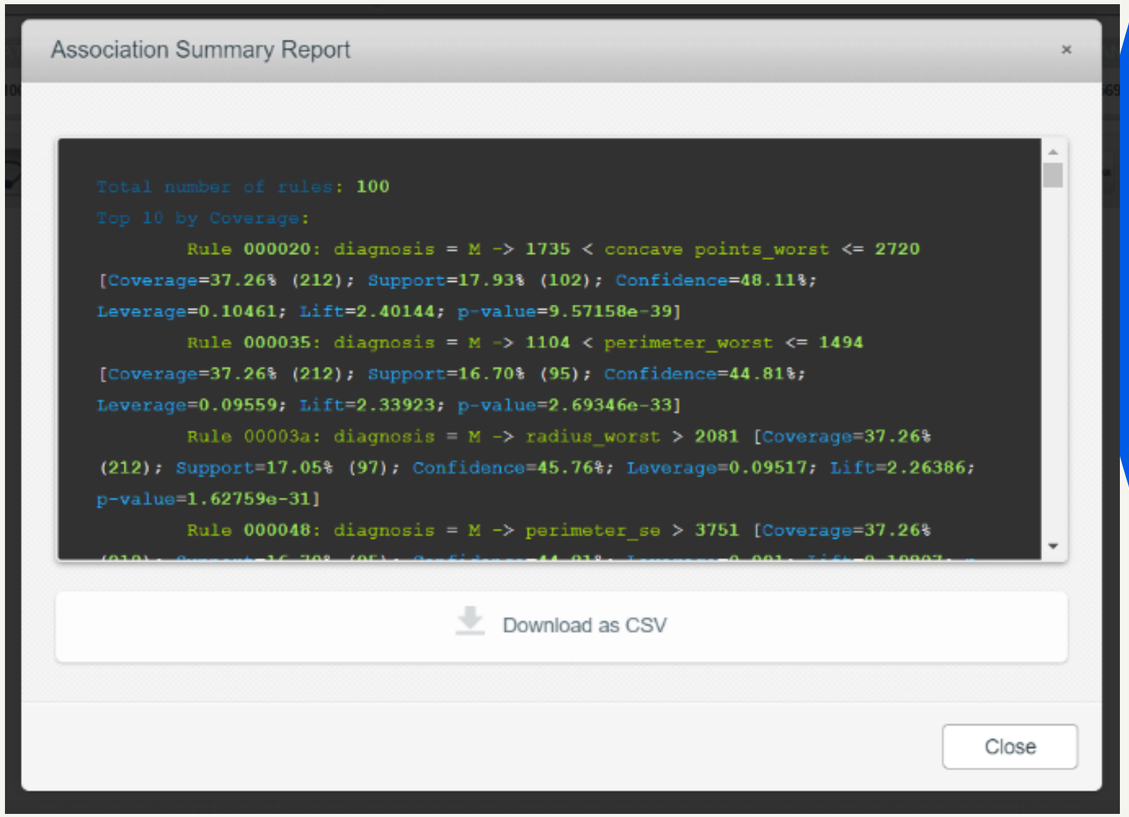
Intersection if unrelated — 100% OF INSTANCES

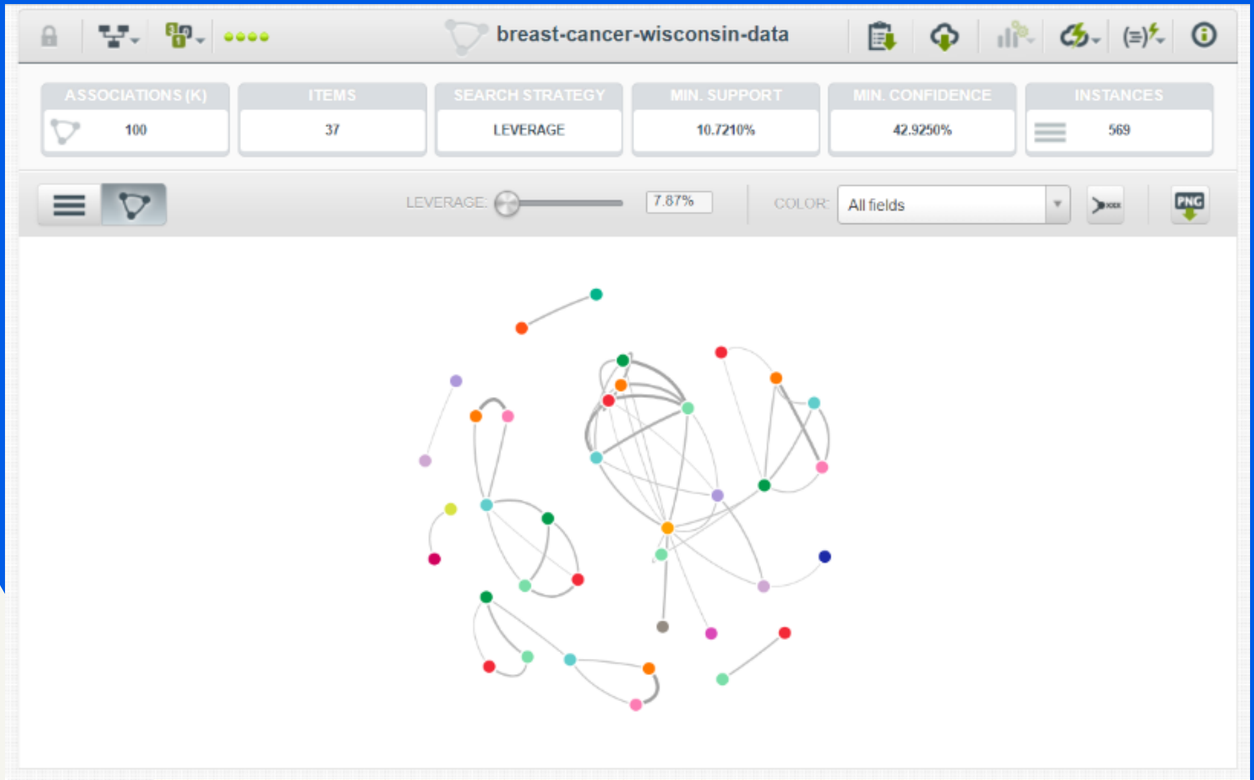
Antecedent Intersection Consequent

Antecedent and consequent occur together 12.68% more often than if they were statistically independent.

Antecedent	Consequent	Coverage	Support	Confidence	Leverage	Lift
3795 < area_mean <= 4956	1114 < radius_mean <= 1271	20.2110%	16.6960%	82.6090%	12.6820%	4.1597
1114 < radius_mean <= 1271	3795 < area_mean <= 4956	19.8590%	16.6960%	84.0710%	12.6820%	4.1597
4956 < area_mean <= 6328	1271 < radius_mean <= 1446	19.6840%	16.6960%	84.8210%	12.6490%	4.1251
1271 < radius_mean <= 1446	4956 < area_mean <= 6328	20.5620%	16.6960%	81.1970%	12.6490%	4.1251
1503 < area_worst <= 3584	1494 < perimeter_worst <= 7153	20.3870%	16.5200%	81.0340%	12.4360%	4.0446
1494 < perimeter_worst <= 7153	1503 < area_worst <= 3584	20.0350%	16.5200%	82.4560%	12.4360%	4.0446
1190 < perimeter_mean <= 6108	radius_worst > 2081	19.8590%	16.1690%	81.4160%	12.1550%	4.0283
radius_worst > 2081	1190 < perimeter_mean <= 6108	20.2110%	16.1690%	80.0000%	12.1550%	4.0283
1190 < perimeter_mean <= 6108	1503 < area_worst <= 3584	19.8590%	16.1690%	81.4160%	12.1200%	3.9936

Associations Chart View and The association summary report





Thank you for listening.

WE WELCOME YOUR QUESTIONS.