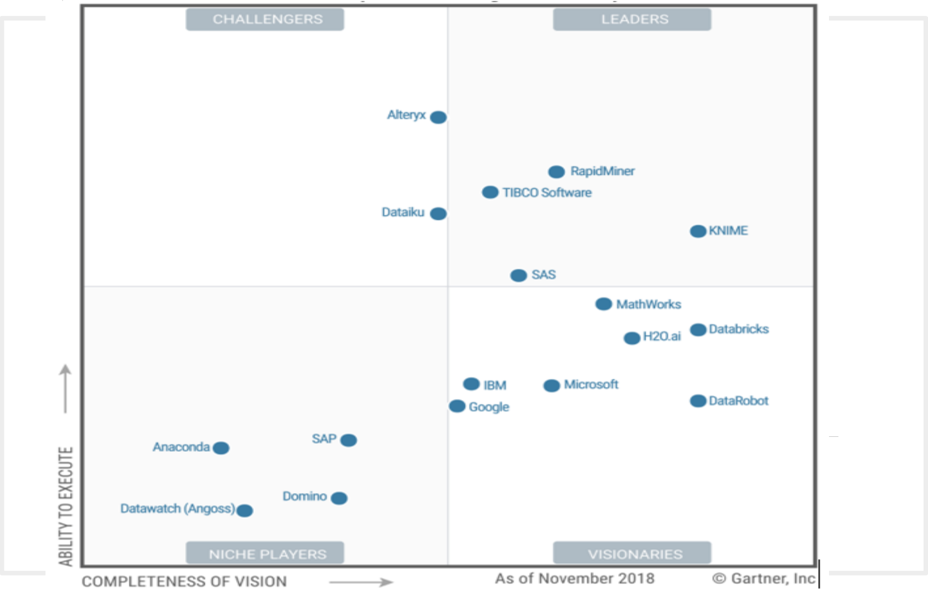
# LEADS 3-Day Camp
# Day 3 Session 2:
# Automated Data Science Tools

Il-Yeol Song, Ph.D.
Professor
College of Computing & Informatics
Drexel University
Philadelphia, PA 19104
Song@drexel.edu
http://www.cci.drexel.edu/faculty/song/

## Quick Survey

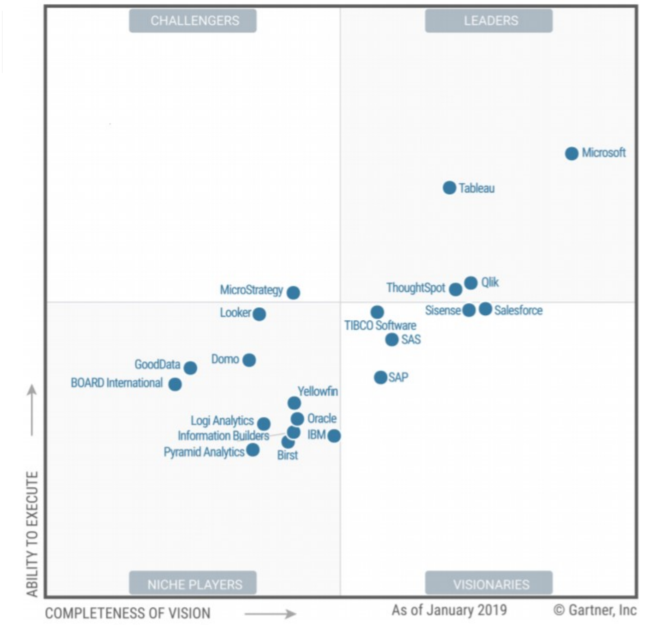
- How many of you are still not comfortable with R coding?
- Do you feel it will be good if there is an automated tool that uses only clicks and drops, without coding?
- Do you believe automated data science tools are possible?

# Magic Quadrant for Data Science and Machine-Learning Platforms (2019)



https://www.gartner.com

# Magic Quadrant for Analytics and BI Platforms (2019)



Source: Gartner (February 2019)

https://www.gartner.com

## 19 Data Science and Machine Learning Tools for people who Don’t Know Programming

**ML Tools**

- RapidMiner
- DataRobot
- BigML
- Google Cloud AutoML
- MLBase (Berkeley)
- WEKA
- Driverless AI
- Microsoft Azure ML Studio
- MLJar
- Amazon Lex (Chatbot builder)
- IBM Watson Studio
- KNIME
- Pure Predictive
- Logical Glue
- FeatureLab
- MarketSwitch
- Automatic Statistician

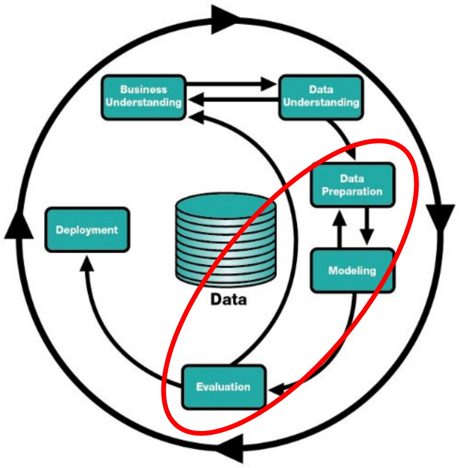
**Data Preparation Tools**

- Paxata
- Trifacta

https://www.analyticsvidhya.com/blog/2018/05/19-data-science-tools-for-people-dont-understand-coding/

## Automated DS Tools

- DS Lifecycle



ML Tasks

Sub-tasks of Data Science including

- Preparation (feature extraction, etc.)
- Modeling (algorithm selection, hyper-parameters tuning, etc.)
- Evaluation (Validation)

# Future of Automated ML Platform

- Industry pushes automated DS/ML platforms
- *Automated tools support data scientists, but do not replace data scientists*
  - Frees data scientists from the burden of repetitive and time-consuming tasks of
    - developing ideas on where to start,
    - preprocessing,
    - selecting models and parameters,
    - evaluating models and output, and
    - optimizing models.

# Automated DS Tools

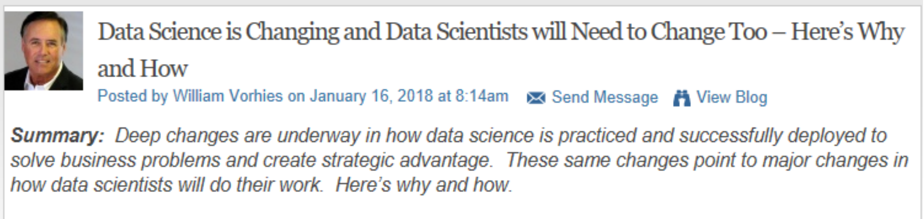
- *Automated ML is not the same as Automated DS*
- “*In industry, data scientists will be evaluated on the value added to the business, rather than algorithm accuracy. A project with 99% classification accuracy, but that isn't deployed in production, is bringing no value to the company.”*

  **-- Sandro Saitta**

Source: **Data Science Automation: Debunking Misconceptions**
https://www.kdnuggets.com/2016/08/data-science-automation-debunking-misconceptions.html

# Automated DS Tools

- Steps that are difficult to fully automate:
  - Defining problems to solve
  - Getting data and integrating data
  - Exploring data
  - Deploying the project
  - Debugging and monitoring solutions



Data Science is Changing and Data Scientists will Need to Change Too – Here's Why and How

Posted by William Vorhies on January 16, 2018 at 8:14am

***Summary:*** *Deep changes are underway in how data science is practiced and successfully deployed to solve business problems and create strategic advantage. These same changes point to major changes in how data scientists will do their work. Here's why and how.*

- Gartner says that by 2020 more than 40% of data science tasks will be automated.
- Advanced analytic platforms are becoming one-stop automated system with full support of analytics lifecycle

- Algorithm Selection and Tuning Will No Longer Matter
- Data Prep will be Mostly Automated
- The ability to correctly understand the business problem and translate that into a data science problem will be key.
- Feature Engineering and Model Validation Become a Focus
- Data Science will Increasingly be a Team Sport

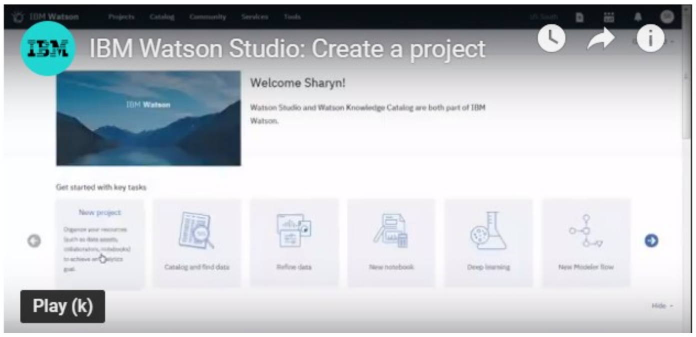
Source: https://www.datasciencecentral.com/profiles/blogs/data-science-is-changing-and-data-scientists-will-need-to-change-

# A Summary of Automated DS Tools

- *Automated ML is not the same as Automated DS*
- *DS tools will become increasingly automated*
- *Automated tools support data scientists, but do not replace data scientists*
  - Frees data scientists from the burden of repetitive and time-consuming tasks of developing ideas on where to start, preprocessing, selecting models and parameters, evaluating models and output, and optimizing models.
- *Citizen Data Scientists will use automated tools*
- *Some tasks difficult to automate include:*
  - Ability to translate business questions into data science problems are critical
  - Feature engineering and model validation are becoming more important

# Watson Analytics

IBM Watson Studio



IBM Watson Studio: Create a project

Welcome Sharyn!

Watson Studio and Watson Knowledge Catalog are both part of IBM Watson.

Organize your resources (such as data assets, collaborators, notebooks) to achieve an analytics goal.

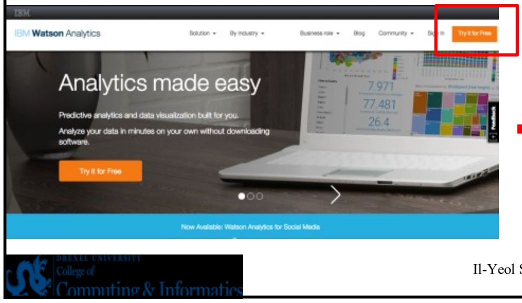
## Hands-on Experience with Watson Analytics

- Access to Watson Analytics

http://www.ibm.com/analytics/watson-analytics/

Step 1: register a Watson Analytics account

Upload your IRIS data set.



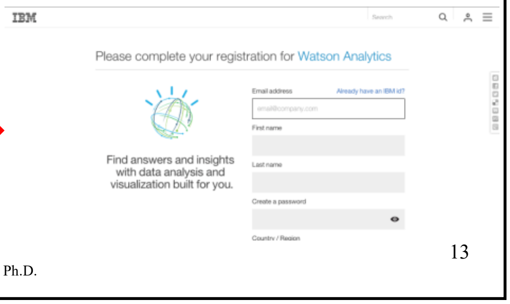
Analytics made easy

Predictive analytics and data visualization built for you.

Analyze your data in minutes on your own without downloading software.

Try it for Free

Now Available: Watson Analytics for Social Media



Please complete your registration for Watson Analytics

Find answers and insights with data analysis and visualization built for you.

Email address | Already have an IBM id?

email@company.com

First name

Last name

Create a password

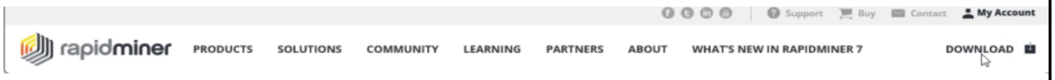
Country / Region

## Hands-on Experience with RapidMiner

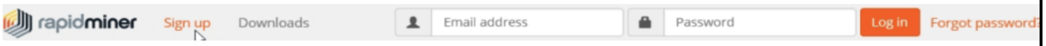
Follow these instructions to download RapidMiner Studio:

1. To download the application, go to the RapidMiner website (https://rapidminer.com/products/studio/)

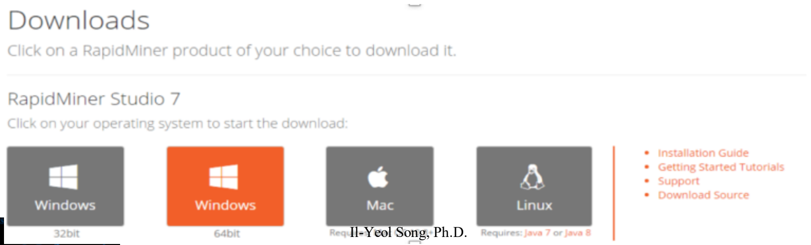
2. Click the **Download** button in the upper right corner.



3. Log in if you haven't already. See next slide if you need to create an account. If you just want to download RapidMiner Studio without logging in, click on **Downloads**.



4.Click on your preferred operating system to begin the download. Your current operating system will be highlighted.



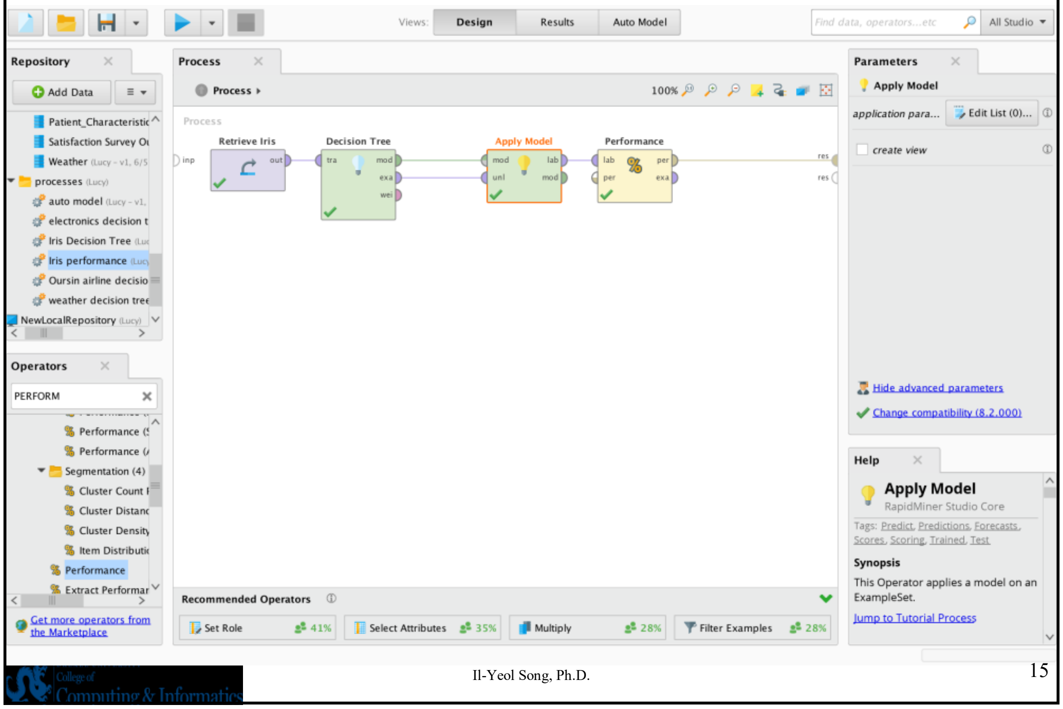
Downloads

Click on a RapidMiner product of your choice to download it.

RapidMiner Studio 7

Click on your operating system to start the download:

# RapidMiner Process



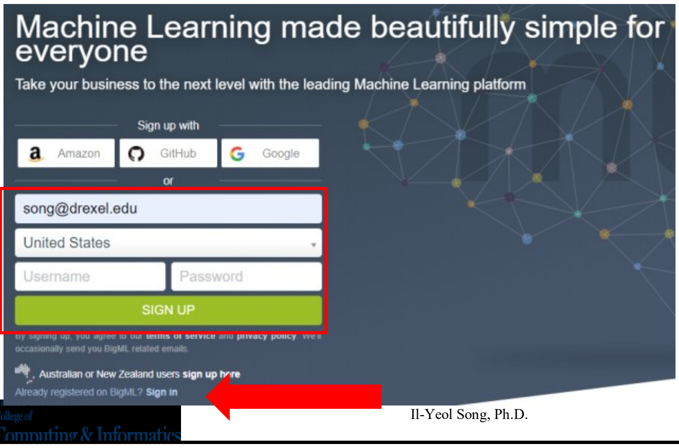
Synopsis

This Operator applies a model on an ExampleSet.

# Hands-on Experience with BigML

- Access to BigML: https://bigml.com/

Step 1: Create your own ID and PW. In order to get a free ID for one year, you must use an email ID with **.edu** domain, not gmail or other private IDs.


Machine Learning made beautifully simple for everyone

Take your business to the next level with the leading Machine Learning platform

song@drexel.edu

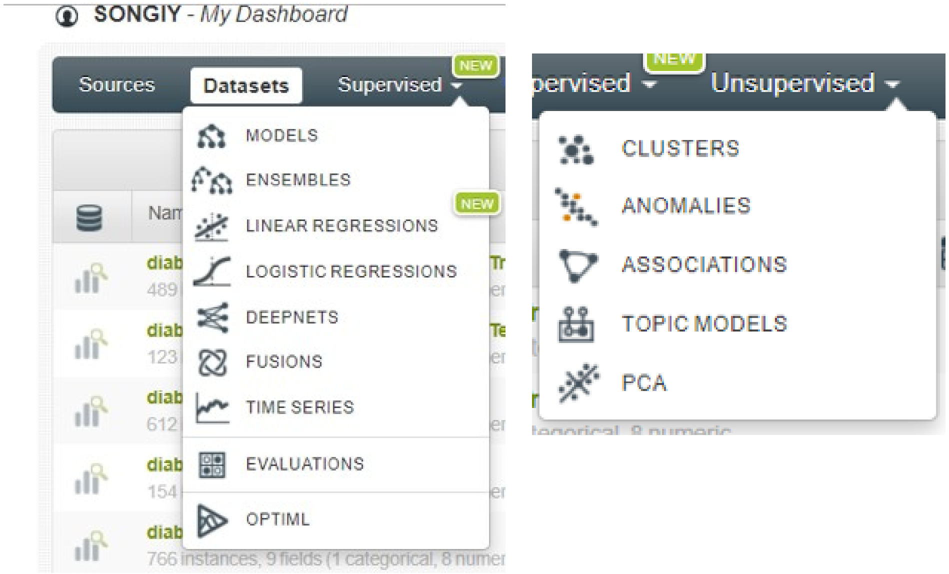
United States

By signing up, you agree to our terms of service and privacy policy. We'll occasionally send you BigML related emails.

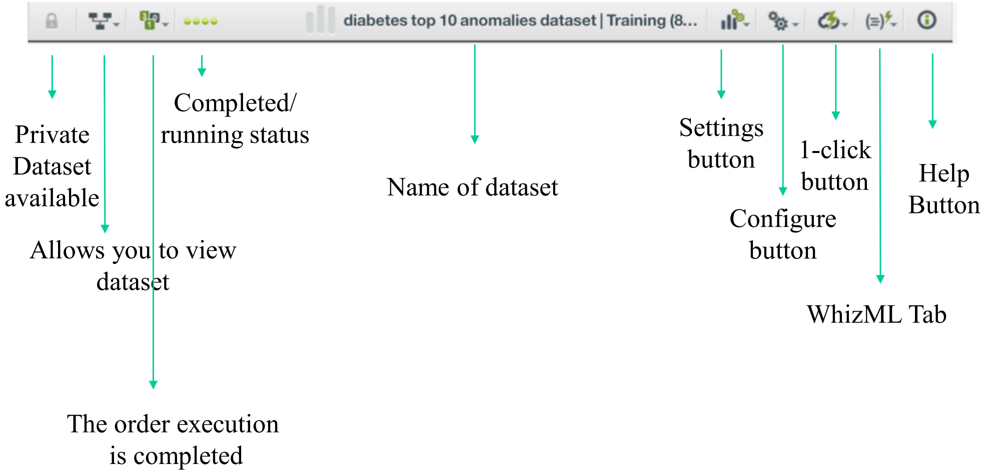
Australian or New Zealand users sign up here

Already registered on BigML? Sign in

# ML Methods supported in BigML



# BigML Dashboard



Private Dataset available

Allows you to view dataset

The order execution is completed

Completed/ running status

Name of dataset

Settings button

Configure button

1-click button

WhizML Tab

Help Button

## Diabetes Data File Download and Loading

- Downloaded a diabetes data from the UCI dataset:

  https://www.kaggle.com/uciml/pima-indians-diabetes-database

- Or Google “pima indian diabetes data set”
- You will find the file description and can download the file

## Data Description

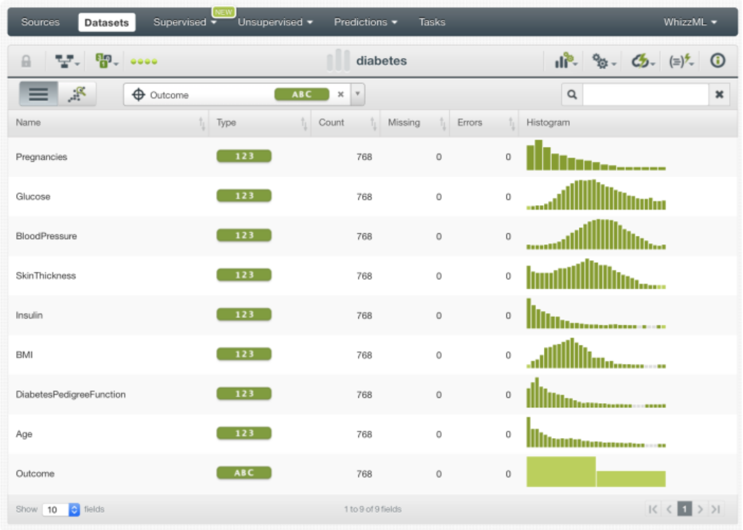
✓ The Diabetics dataset consists of **9** variables

✓ There are a total of **768 rows**.

- **Pregnancies**: Number of times pregnant
- **Glucose**: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- **BloodPressure**: Diastolic blood pressure (mm Hg)
- **SkinThickness**: Triceps skin fold thickness (mm)
- **Insulin**: 2-Hour serum insulin (mu U/ml)
- **BMI**: Body mass index (weight in kg/(height in m)^2)
- **DiabetesPedigreeFunction**: Diabetes pedigree function
- **Age**: Age (years)
- **Outcome**: Class variable (0 or 1) 268 of 768 are 1, the others are 0

## Diabetes Data Description

- ✓ The Diabetics dataset consists of **9** variables
- ✓ There are a total of **768 rows**.

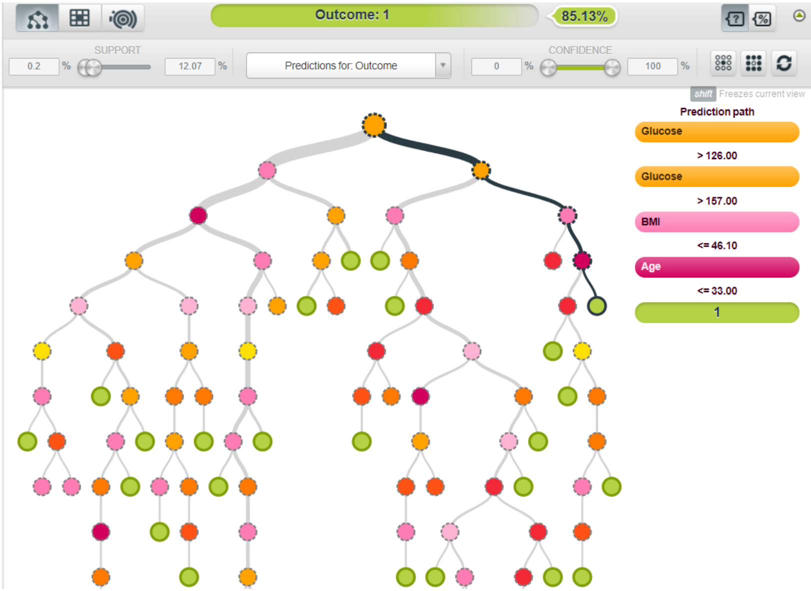


| Name | Type | Count | Missing | Errors | Histogram |
|---|---|---|---|---|---|
| Pregnancies | 123 | 768 | 0 | 0 | |
| Glucose | 123 | 768 | 0 | 0 | |
| BloodPressure | 123 | 768 | 0 | 0 | |
| SkinThickness | 123 | 768 | 0 | 0 | |
| Insulin | 123 | 768 | 0 | 0 | |
| BMI | 123 | 768 | 0 | 0 | |
| DiabetesPedigreeFunction | 123 | 768 | 0 | 0 | |
| Age | 123 | 768 | 0 | 0 | |
| Outcome | ABC | 768 | 0 | 0 | |

Show 10 fields

1 to 9 of 9 fields

## Decision Tree Model



Outcome: 1 85.13%

SUPPORT 0.2 % 12.07 %

Predictions for: Outcome

CONFIDENCE 0 % 100 %

Prediction path

Glucose

> 126.00

Glucose

> 157.00

BMI

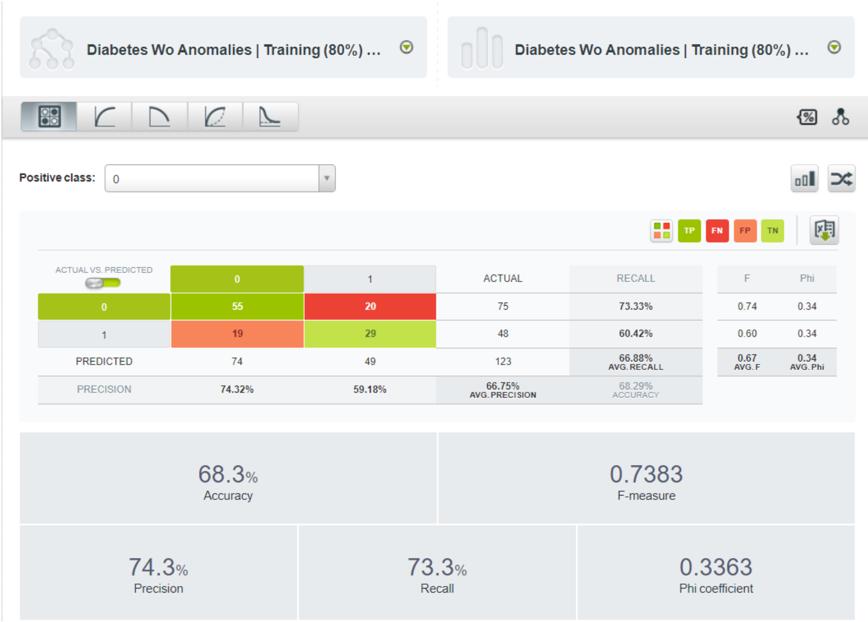
<= 46.10

Age

<= 33.00

1

# Confusion Matrix



Diabetes Wo Anomalies | Training (80%) ...
Diabetes Wo Anomalies | Training (80%) ...
TP
FN
FP
TN

Positive class: 0

| ACTUAL VS. PREDICTED | 0 | 1 | ACTUAL | RECALL | F | Phi |
|---|---|---|---|---|---|---|
| 0 | 55 | 20 | 75 | 73.33% | 0.74 | 0.34 |
| 1 | 19 | 29 | 48 | 60.42% | 0.60 | 0.34 |
| PREDICTED | 74 | 49 | 123 | 66.88%<br>AVG. RECALL | 0.67<br>AVG. F | 0.34<br>AVG. Phi |
| PRECISION | 74.32% | 59.18% | 66.75%<br>AVG. PRECISION | 68.29%<br>ACCURACY | | |

68.3% Accuracy

0.7383 F-measure

74.3% Precision

73.3% Recall

0.3363 Phi coefficient

## YouTube Tutorial Links on BigML

- **BigML Interface**
  https://www.youtube.com/watch?v=6xbNpILmQYo
- **Anomaly Detection:**
  https://www.youtube.com/watch?v=a5Q7b4e7lqg&list=PL1bKyu9GtNYHAk0PUojkLYZzASoYVcsTQ&index=13
- **Model Link (Decision Tree):**
  https://www.youtube.com/watch?v=hnt7z24wvxs&list=PL1bKyu9GtNYHAk0PUojkLYZzASoYVcsTQ&index=5
- **Ensemble Link:**
  https://www.youtube.com/watch?v=zqFj6l2WZCU&list=PL1bKyu9GtNYHAk0PUojkLYZzASoYVcsTQ&index=7
- **Evaluation:**
  https://www.youtube.com/watch?v=cPErxYP9CmQ&list=PL1bKyu9GtNYHAk0PUojkLYZzASoYVcsTQ&index=11