Reichman University

Location: Zoom or Room

Days and Time:

Manas: Tuesday 11:15-8:45 16:15-13:45

CS: Thursday 10:30-8:00

Economics: Thursday 16:15-13:45

CLASSIFICATION with DECISION TREE

Dr. Gail Gilboa Freedman

In the forthcoming sessions, you will continue learning how to **solve the problem of CLASSIFICATION for data on FLIGHTS' DELAY with the DECISION TREE algorithm**
In the current lesson, we will implement it with the BigML platform (instead of Python)

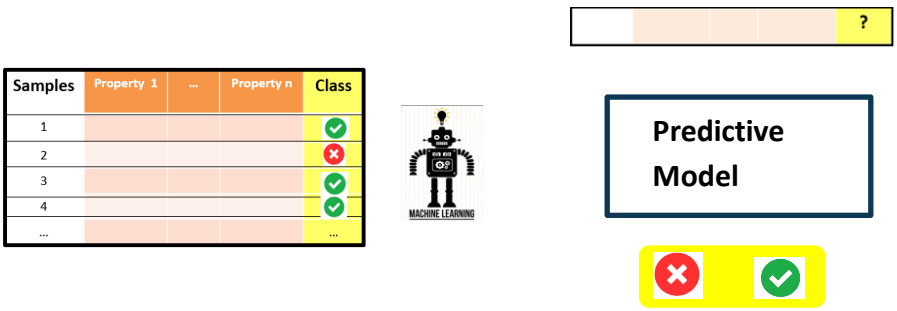
material: 'DS Decision Tree flights data.csv'

Key terms: BigML; Entropy; confidence level; Wilson score; support measure

I. Problem: CLASSIFICATION

a) **Definition (and schematic illustration) of the problem:**

The Classification Problem is to find an appropriate class for a new object, according to historical classifications of some other objects.



b) Algorithm of the hour: **Decision Tree** Classifying observations based on a set of decision rules.

Gini (of a random variable)

represents the probability of a predictor being classified incorrectly when selected randomly

$$1 - \sum_{i=1}^{n} proportion_i^2$$

Gini (of each predictor in the dataset)

used as an index for selecting the predictor for splitting a node.

c) It is the sum of the children's weighted Gini indices

(there is a child for each of the *m* predictor's categorical values)

Formula:

$$\sum_{j=1}^{m} \left(1 - \sum_{i=1}^{n} proportion_i^2 \right) \frac{group\_size}{total\_samples}$$

For each predictor, calculate its score. Then, select the predictor with the minimal score.

Reichman University

Location: Zoom or Room
Days and Time:
Manas: Tuesday 11:15-8:45 16:15-13:45
CS: Thursday 10:30-8:00
Economics: Thursday 16:15-13:45

Gini manual example:

Calculate the Gini score for 'Outlook'

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Solution:

- For each of the (3) categorical values of 'outlook', calculate its weighted Gini:

Sunny

Play	Not-Play
2/5	3/5

$$\left(1-\left(\left(\frac{2}{5}\right)^2+\left(\frac{3}{5}\right)^2\right)\right)*\frac{5}{14}$$

Overcast

Play	Not-Play
4/4	0/4

$$\left(1-\left(\left(\frac{4}{4}\right)^2+\left(\frac{0}{4}\right)^2\right)\right)*\frac{4}{14}$$

Rain

Play	Not-Play
3/5	2/5

?

- Sum them all.

Reichman University

Location: Zoom or Room
Days and Time:
Manas: Tuesday 11:15-8:45 16:15-13:45
CS: Thursday 10:30-8:00
Economics: Thursday 16:15-13:45

II. Data Y : the flight delayed problem

The data in 1 file:
'DS Decision Tree flights data.csv'

Each row represents a flight and characterizes by 7 features (columns):

'CARRIER', 'DAY_OF_WEEK', 'DEP_TIME', 'DEST', 'ORIGIN', 'WEATHER', `'Delayed'`

The flight delayed problem is to produce a grouping of data points with the purpose of predicting whether a flight will be delayed

III. Tools: BigML

bigml gallery

Location: Zoom or Room

Days and Time:

Manas: Tuesday 11:15-8:45 16:15-13:45

CS: Thursday 10:30-8:00

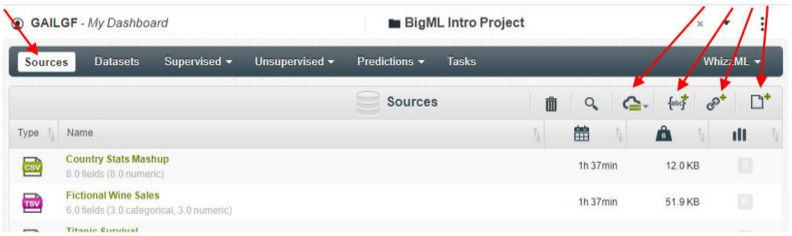
Economics: Thursday 16:15-13:45

IV. Practice: Solving X for Y with Z Step by Step:

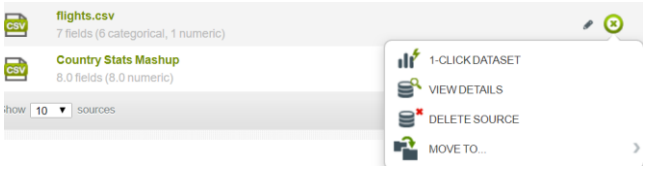
1. Data preparation

USE BIGML SOURCES

UPLOAD YOUR DATA

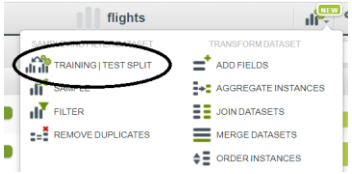


Create dataset

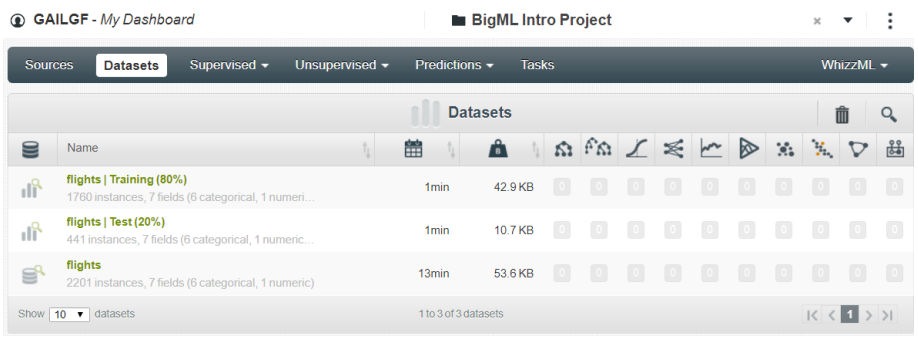


Hover over Histogram,

Push "training | test split", and "create training | test"



In the upper menu select “Datasets”, then select the training file you just created.



Location: Zoom or Room

Days and Time:

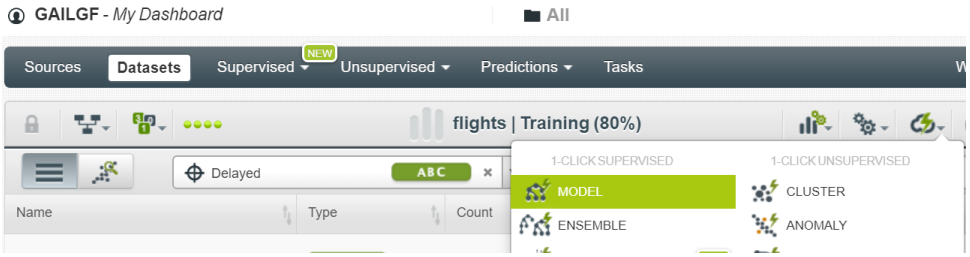
Manas: Tuesday 11:15-8:45 16:15-13:45

CS: Thursday 10:30-8:00

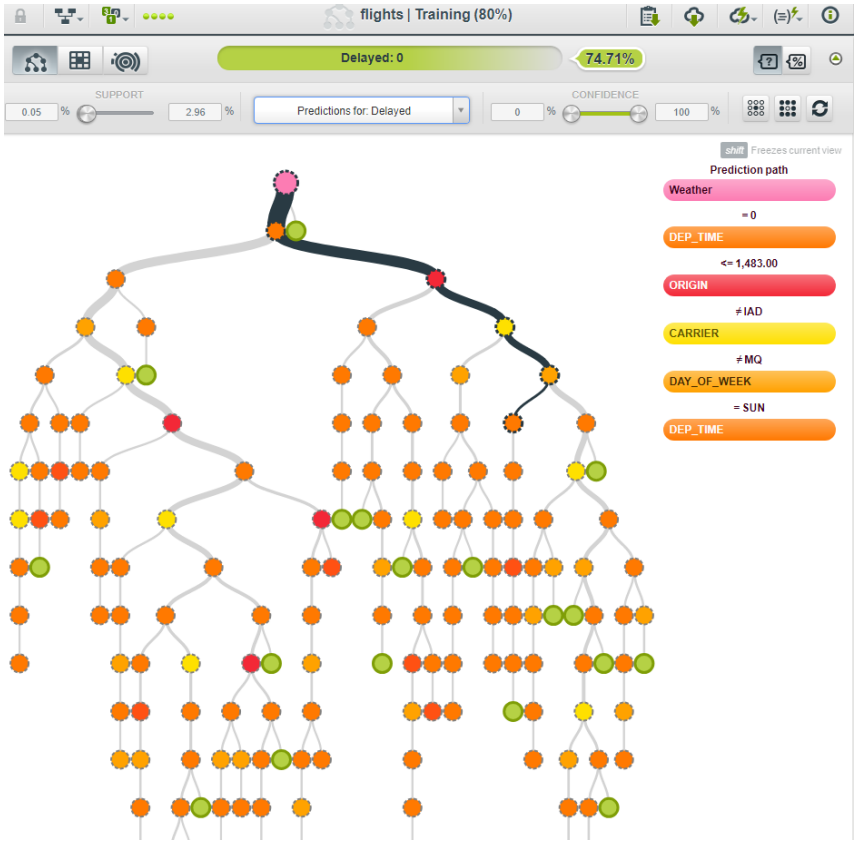
Economics: Thursday 16:15-13:45

2. Modeling

Hover over the “cloud and lightning” button to select “model”



This is what should appear on your screen

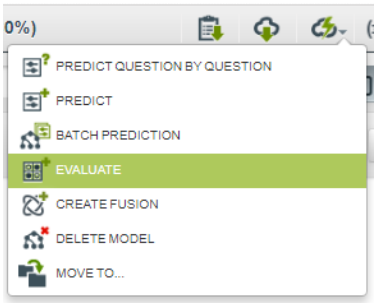


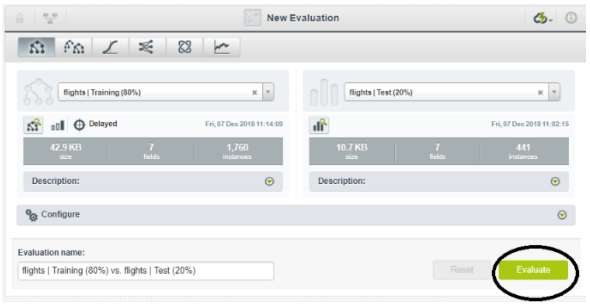
Location: Zoom or Room
Days and Time:
Manas: Tuesday 11:15-8:45 16:15-13:45
CS: Thursday 10:30-8:00
Economics: Thursday 16:15-13:45

3. Performance evaluation

Hover over the "cloud and lightning" button to select "EVALUATE"

When the "New Evaluation" window is open, push the "Evaluate" button.





You will see evaluation measures, including the confusion matrix learned in the previous lesson.

4. Visualization

You cut the depth of the tree.

You may toy with:

a) The confidence level is about the probability that the prediction at a leaf matches the class. It takes into account the distribution of the classes and the number of instances (using the lower end of the Wilson score interval at 95% confidence).

b) Support measure (that represents the proportion of samples in a leaf out of the total set).

V. Summary and Discussion

Location: Zoom or Room
Days and Time:
Manas: Tuesday 11:15-8:45 16:15-13:45
CS: Thursday 10:30-8:00
Economics: Thursday 16:15-13:45

We learned about the Gini index that is used by the decision tree algorithm, as a criterion for selecting which predictor is best to be selected as the splitter of a node. We re-solved the delayed flight problem of classifying flights by similarity in terms of their characteristics. We utilized Bigml to implement the Decision Tree and Confusion matrix evaluating our model, respectively.

VI. I AM VERY CURIOUS!

- The official blog of the company
- bigml introductory education videos