# Data Mining

## Azure Machine Learning Homework

**Due**: 27/12/2016 23:59

**Instruction - Read Carefully!**

**Handing in:** You must hand in the home works by the due date and time by compressing your file in a .zip or .tar.gz file with all your answers and subject [Azure ML Homework] and send it to the e-mail:

martini.1722989@studenti.uniroma1.it

The solution must contain the screenshot requested at the end of this file.

For any other information feel free to contact us.

**Task.** Given the set of recipes, you are required to build an experiment able to classify a recipe in "Vegetarian" or "Not Vegetarian".

1. CREATE NEW DATASET. Create a new Azure Machine Learning Dataset importing the csv file.

2. CLEANING MISSING DATA. Remove rows that contains empty column.

3. PREPROCESS. Do stemming, remove stop words, … (You are free to decide whatever rule to apply).

4. FEATURE EXTRACTION. Ingredients are grouped together in one column, turn it into a set of features. Use **Extract N-Gram Features** module.

5. COLUMN SELECTION. Only columns that are result from the feature extraction must be trained.

6. SPLITTING DATASET. Split your Dataset into Train and Test set, to train and to test your Binary Classifier and plot the result of your Classifier.

7. INITIALIZATION MODEL: select a classification model (for instance, *Two Class Support Vector Machine Module*).

8. SCORE MODEL: compute the score of the model selected.

9. EVALUATE MODEL. Provide evaluation result of your model.
   **Suggestion:** apply before a column selection to have this result cleaner (otherwise all the feature will be shown).

10. As you can see, *Evaluate Model* takes in input two *Scored Dataset to compare*. In the same experiment, add a further classification model and compare it with the previous one.

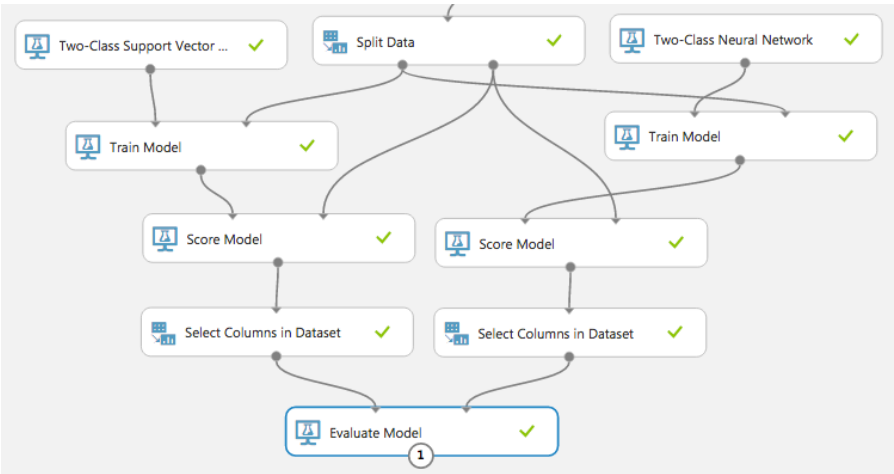


*Figure 1. Hint*

**OUTPUT you must provide:**

- *Screenshot of the trained example after a complete run (each module must be checked, for instance* 

*)*
- *Screenshot of the evaluation results:*
  - *ROC plot*
  - *PRECISION/RECALL plot*
  - *LIFT plot*
  - *Statistics in the bottom of the page (True Positive, False Negative, Accuracy, …)*