**Data Mining**
**Spring Semester 2016-2017**
**2nd Exercise, Delivery Date: 05-06-2016**
**Group Work (2 People)**

**The objective of the task**

The purpose of the work is to understand and explore data input, as well as feature selection. Implementing it work will be done in the Python programming language (like the 1st exercise) with use of tools / libraries: jupyter notebook, pandas, gensim and SciKit Learn.

**Description**

The job is related to borrowing data belonging to two categories (Good / Bad). Your goal is to use the available features in order to create a classification model that you can decide whether a loan should be given to a customer of the bank applying for a loan or not. The available dataset contains a set of columns that relate to the different attributes of the borrowers, while the last column refers to whether the borrower was ultimately good or bad (1 = Good, 2 = Bad). An extensive description of the data is provided by the available document in the following link.

In particular, you will have the following files:
1. Train_set.csv (800 instances): This file will be used to train your algorithms. Also this dataset will use it for evaluation of features and their visualization.
2. test_set.csv (200 instances): This file will be used to do forecasts for new data. Contains all fields of the training file off from the last column, this field will be called to appreciate using it Classification algorithms.

**Download Dataset**

Datasets are available in eclass.

**Visualization of Data**

In this question you should visualize the different feautes of the dataset. In particular, you should add the following in your feature report:
- If the feature is categorical:
  - a histogram for each type of borrower that has been rated Good
  - a histogram for each type of borrower identified as Bad
- If the feature is numerical:
  - a box plot for each type of borrower that has been rated Good
  - a box plot for each type of borrower identified as Bad

Finally, you should describe what you notice from the plots you created. What features?
Expect to be more useful to categorize customers, (Hint) It would be helpful if in the same plot for each feature visualize both Good and them Bad with a different color.

**Classification Implementation**

In this question you should try the following Classification methods:
- Support Vector Machines (SVM)
- Random Forests
- Naive Bayes

You should also evaluate and record the performance of each method using 10-fold Cross Validation using metric accuracy.

**Select Features**
In this question you should evaluate the quality of the available features on categorizing customers as Good or Bad, using the best the classifier you found from the previous question.
More specifically, you need to calculate the Information Gain for each feature. In continue to calculate the accuracy of the classifier by subtracting each feature and each time.

You should present:
● In a plot how average accuracy changes for 10-fold cross-validation as well remove features from the Classifier.
● In a table the feature you chose to remove in each iteration as well the corresponding Information Gain.(Hint) You can use the definition for calculating Information Gain from Wilkipedia from the link below. The numerical attributes you can convert to categorical to that query discriminating them in 5 bins.

**Communication Forum**
Piazza
You should analyze the steps you have taken. Report your not to exceeds 30 pages.

**Output files**
The code should be for the Classification questions you should create the following files
● EvaluationMetric_10fold.csv
● testSet_Predictions.csv

The format of the EvaluationMetric_10fold.csv files is shown below:

| Statistically-Measure | NaiveBayes | RandomForest | SVM |
|---|---|---|---|
| Accuracy | | | |

The format of the testSet_Predictions.csv file, which will contain the categories customers given in the test set is shown below:

| Client_ID | Predicted_Label |
|---|---|
| 1 | Good |
| 2 | Good |
| 3 | Bad |
| ... | ... |
| 10 | Bad |
| ... | ... |

For the "testSet_Predictions.csv" file, the above should be strictly used formatting by separating the two fields with the TAB character ('\ t') and should also in the first line there are two headings

(Client_ID and Predicted_Label) and then your model's predictions on the following lines specifying it client_ID of the client from the test set and the corresponding label.

**About Deliverable**

....