Machine Learning module – Python Lab – Exam 11/09/2020

Find the best classification scheme for the included dataset.
The solution must be produced as a Python Notebook.
The notebook must include appropriate comments and must operate as follows:

0. read a market basket database from the csv file provided and generate a dataframe **basket** of boolean values with one row per transaction and one column per distinct item of the database; the dataframe values must be True if a distinct item is contained in the transaction
    - the file contains one transaction per line, the first element is the number of items in the transaction, followed by the items of the transaction, and then a variable number of empty fields
    - the field names in the first row of the csv file are not relevant
1. ignore the transactions containing a single item (2 points)
2. the column names of the output dataframe are the distinct items (2 points)
3. show the first five rows of the output dataframe (1 point)
4. show the number of transactions and of distinct items (1 point)
5. find a value of min_support such that the apriori algorithm generates at least 8 frequent itemsets with at least 2 items (5 points)
    - output the result with the message below
    - min_support: 0.xxxx - number of itemsets with at least 2 items: nn)
6. find the minimum metric threshold such that at least 10 association rules are extracted from the frequent itemsets found (5 points)
    - use "confidence" as metric and output the line below:
    - Metric: "confidence" - min_metric: 0.xxxx - Number of rules: n
7. print the first 10 rules found, sorted by descending confidence and support (3 points)
8. plot confidence and support for all the sorted rules found (3 points)
9. scatter plot the rules by confidence and support, labelling the points with the index value of the corresponding rule (hint https://stackoverflow.com/questions/14432557/matplotlib-scatter-plot-with-different-text-at-each-data-point) (3 points)

Quality of the code: (6 points)

- Include appropriate comments with reference to the numbered requirements
- Useless cells, pieces of code and non-required output will be penalized
    - Remove the code you use for testing and inspecting the variables during the development
- Naming style of variables must be uniform and in English
- Bad indentation and messy code will be penalized

Additional directions, the assignments not compliant with the rules below will not be considered

1. The notebook name must be ***emailusername.ipynb*** in lowercase letters
    a. E.G. if your email is mario.rossi45@studio.unibo.it the notebook filename will be mario.rossi45.ipynb
2. The first cell must contain the student first name, last name and email
3. The solution must directly access the data in the same folder of the notebook
4. Upload the notebook only to EoL

Cooperative work will be **heavily sanctioned**
The candidate can freely access the manuals available online.
The candidate can freely access the teaching materials available in the course website, including the available examples of python notebooks.