# BIG DATA PROJECT 2

**Project outline(translated from Greek)**

**Author:**
Benekos Konstantinos, 1067754, up1067754@ac.upatras.gr

March 22, 2024

**Subject 2**

Together with the assignment's statement, three (5) publications (files paper1.pdf, paper2.pdf, paper3.pdf, paper4.pdf, paper5.pdf) are provided. Study these publications, and for each one of them, write a summary.

**Subject 3**

Answer the two questions below:

**I.** In section Lecture 4: Classification(eclass), you can find the file Python-NaiveBayes-SentimentAnalysis.rar that contains Python code performing sentiment analysis on movie reviews using the Naïve Bayes classification algorithm. The compressed file also contains two .csv files containing user reviews. In this topic, you will need to write a program in "R" language that performs sentiment analysis on the data in the IMDBDataset.csv file using the Naïve Bayes algorithm. More thoroughly, it will do the following:

- 1) It will use the reviews in the file IMDBDataset.csv that exists in the compressed file and use them for the training and evaluation of the classifier.
- 2) Perform exactly the same preprocessing of the data as the provided Python code, more specifically: **a)** It will remove from all the words of a review those characters that are not letters or numbers. For this purpose, install the library stringr and make use of the function str_replace_all that it provides. **b)** It will convert all the letters of the text to lowercase. **c)** It will remove from all the reviews the stopwords. To do this, install the R library named tm (Text Mining), which has all the necessary functions for text processing in R. Make use of the English stopwords dictionary. Refer to the library's user manual to select the appropriate function. **d)**It will perform stemming of all the words that exist in the reviews. For this purpose, install the R library Snowball (library name: SnowballC) and refer to the user manual for how to perform stemming of all the words in the English language.
- 3) After the data of the file have been preprocessed in the above manner, the DocumentTermMatrix will be created, using the appropriate function from the tm library. For your assistance, study the article (link) in order to understand the concept of a "DocumentTermMatrix" and how it is used in the tm library.
- 4) It will use 80% of the reviews from the file IMDBData.csv as training data and the remaining 20% as test data.
- 5) It will create a classifier based on Naive Bayes using the training data of the reviews. Install the R library named e1071 which provides a function implementing the Naïve Bayes algorithm. Refer to the user manual to figure which function is appropriate and the arguments it should accept.
- 6) For the evaluation of the classifier, your program should, after training and categorizing the test set, display on the screen only the prediction accuracy on the test dataset.

II. In what way will you use the classifier you have trained in question I) in order to answer the following question: " Do user generated social media comments affect the price of a specific product in the supermarket?”. Your answer should describe a scenario with words, where it is evident how you would use the classifier and the steps and actions you would take in order to answer this question. Your description should cover the following:

- 1. From which sources would you collect data, and what format would said data be in?
- 2. How would you process the data and?
- 3. Which statistic methods/ machine learning algorithms would you use, and why?

In your description, do not worry about technical issues (you do not need to mention specific tools). You can mention any other aspect you deem appropriate. To get a better idea of how to describe such a scenario, you can refer to the section titled ’How does it work?’ from the following article:link.

**Subject 4**

From the page (link), download the Mushroom dataset, which contains the characteristics of various mushroom species and indicates whether they are edible or poisonous. Carefully read the information on the aforementioned page, especially the section ’Attribute Information,’ which explains how the values in the Mushroom dataset should be interpreted.

The following are requested:

- 1) Write a program in R and Python that builds a decision tree from the Mushroom dataset and predicts whether a mushroom is edible or poisonous. For building the decision tree in the R environment, use the rpart package. The relevant manual for the rpart package of R is provided along with the assignment. Since this package is not pre-installed in R, it should be installed and used using the library() command of R.For training and testing your model, you should create two random samples of size equal to 80% and 20% of the size of the initial dataset respectively (i.e., 80% of the initial dataset should be used for training and the remaining 20% for testing). Your code for building the decision tree should include commands for this separation. Your code should visualize the decision tree that has been created and display the labels on the nodes. Additionally, your programs should display the confusion matrix and the accuracy of the model on the screen. In your response, include the code in R and Python that you have created.
- 2) For the first 30 records of the Mushroom Data Set (and only for those), calculate manually, using the appropriate types, the entropy gain of the attribute ”habitat,” establish if the classification attribute is the one that indicates whether the mushroom is edible or poisonous.
- 3) Write a program only in Python, which implements classification with the Naïve Bayes algorithm for whether the mushroom is edible or poisonous.