University of Macedonia

Department of Applied Informatics

7Thesemester | Academic Year: 2022 – 2023

# Machine Learning

## 2theAssignment – Classification problems

Having completed your studies, you interview for a position with the subject "data analytics & financial risk estimations associate" in an international financial institution.

As part of the evaluation, you are handed an excel file (Dataset2Use_Assignment2.xlsx) that contains financial indicators and some other information about a number of Greek companies.

Your data is:

1. The performance indicators of the companies (columns A to H)
2. Three binary indicators of activities (columns I, J, K)
3. The state of the company (1 all good, 2 has declared bankruptcy)
4. The year to which the above figures relate.

The point is to create the best possible model for the given problem.

The code you submit must implement the following:

1. To read the data from the excel file. Notes: a) the Pandas library will come in handy. b) you will need to upload the excel to your drive and access it from there.

2. To print on the screen per year the following data:
    a. Number of healthy and bankrupt businesses
    b. The min, max, average value for each indicator
3. Normalize the data in the interval [0.1] and divide them intotrain and test sets. The train/test data ratio is determined by you.
4. To print on the screen how many bankrupt and how many healthy companies there are in a) train and b) test set.
5. Implement (train) four of the following models:
    a. Linear Discriminant Analysis
    b. Logistic Regression
    c. Decision Trees
    d. k-Nearest Neighbors
    e. Naïve Bayes
    f. Support Vector Machines
6. It will print confusion matrices in both train and test set.

7. It will calculate the performance of the trained models in both the train and the test set, based on the following metrics: Accuracy, Precision, Recall, F1 score.
8. It will pass the results to a csv file. Each line will have the following elements:

Classifier Name | Training or test set| Number of training samples | Number of non-healthy companies in training sample | TP | TN | FP | FN | Precision | Recall | F1 score | Accuracy

a. Classifier Name: the name of the classifier
b. Training or test set: report if the line results refer to training or evaluation data
c. Number of training samples: Number of data in the train set
d. Number of non-healthy companies in training sample: How many bankrupt companies the training set had.
e. TP: Number of true positives (insolvent company that the model took as insolvent).
f. TN: Number of true negatives (healthy companies that the model identified as healthy)
g. FP: number of false positives
h. FN: number of false negatives
i. Precision: the known metric
j. Recall: the familiar metric
k. F1 score: the familiar metric
l. Accuracy: the familiar metric

Using excel and the csv file from question 8, make graphs that show which is the best model based on the F1 score. How is the large difference in results between train and test sets justified?

You will collect all the results and submit them in a report format. Information on the structure of the report can be found in the "Instructions" section.

**Instructions:**

A. Assignments are in groups of up to four (4) people. Each person can submit work to only one group at a time.

B. Assignments should be uploaded to eClass in a zip (not rar) file by the deadline. No extension will be granted.**Caution**: Each group assignment will be submitted by only one group member (you choose who/who).

C. Each assignment must be accompanied by:

• One and only one file.py will contain the answers to the queries

• One**report**inpdf with the following information: o

- Cover: 1 page, includes the details of the students in the group, course name, date, department and other relevant details.
- o Summary table of contents, images, and other graphics that cite in the report.
- o Introduction section: 1 page, describe the problem (*without* copying exactly the speech of the exercise)
- o Methods applied: from 2 to 10 pages, describe the methods you used and list the relevant results. Make sure it's clear which question you're referring to.
- o Conclusions: 1 page, based on the results what do you recommend, which model performs better, what could be done to further improve the performance.
- o The report will contain graphical representations of all kinds and tables of evaluation of the results that must be accompanied (each) by at least one paragraph with commentary.

**Make sure that:**

• The code must necessarily be accompanied by appropriate comments.

• An editorial and spelling check has been done on the report you will submit.

• The sentences should be understandable and short in length.

• Images to***do not***have arisen fromprint screen. If the program creates an image, save it normally (jpg or png) before using it.

• The graphs should include names on the axes and a legend. Purpose is to understand what it shows, at a glance.

• If something is not specified, you have the right to make any implementation that suits you. Make sure you can explain exactly what you did in the code.

• The libraries you will use *must* be able to be installed via it pip.

• The code *must* run onGoogle Colab.

Delivery Due Date: **January 9, 2023**. No extension will be granted.