UNIVERSITY OF PIRAEUS
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGIES DEPARTMENT OF DIGITAL SYSTEMS
"DATA ANALYSIS"

Data Set Categorisation

DESCRIPTION

In this assignment you are required to **apply** one or more classification algorithms to a given dataset and write a technical report, according to the specifications given below.

The categorisation of the dataset will be based on:

- the categorisation methods you have learned in the course
- a library of categorization algorithms (e.g., in Python with Scikit-Learn, in Java with the WEKA API) that has implemented categorization algorithms
- other techniques/algorithms you may find in external sources, such as books, articles, etc. (*all external sources used should be mentioned in terms of methods and/or code*).

DATA SETS

The datasets to be used for this work are (mostly) provided by the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.html).
Each student will use one of the following data sets, ***based on the last digit of*** his/her ***Registration Number:***

(Registration number) mod 10	Data set	Link
0	Pima Indians Diabetes Dataset	https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.names https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.csv
1	Sonar	https://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+(Sonar,+Mines+vs.+Rocks)
2	banknote authentication	http://archive.ics.uci.edu/ml/datasets/banknote+authentication
3	Ionosphere	https://archive.ics.uci.edu/ml/datasets/Ionosphere
4	wheat seeds	http://archive.ics.uci.edu/ml/datasets/seeds
5	Car Evaluation	https://archive.ics.uci.edu/ml/datasets/Car+Evaluation
6	Dermatology	https://archive.ics.uci.edu/ml/datasets/dermatology
7	Ecoli	https://archive.ics.uci.edu/ml/datasets/Ecoli
8	AutisticSpectrum Disorder Screening	https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Children++
9	HIGGS Data Set	https://archive.ics.uci.edu/ml/datasets/HIGGS

TECHNICAL REPORT STRUCTURE

All students will use the same template for writing the technical report,which is available here: https://www.acm.org/binaries/content/assets/publications/word_style/interim-templatestyle/interim-layout-.docx

- Writing language: greek
- Maximum size: 6 pages

The technical report will be structured in the following main sections:

- a **Summary**,
- a short **Introduction** explaining the purpose of the work,
- **a** brief **description of the dataset**,
- **pre-processing** steps that may have been necessary,
- description and documentation of the **categorisation algorithm**(s) you applied (why these?),
- the **methodology** you applied (separation into test and control set, cross-validation, avoiding overfitting, etc.)
- the **Experimental Evaluation of the** categorization results (in the form of charts showing some metrics, such as Accuracy, precision, recall, F1 measure, etc.), where you can change the algorithm parameters,
- **Conclusions of** the study, and
- **Literary sources**