Deree
School of Graduate and Professional Education

THE AMERICAN COLLEGE OF GREECE
Pierce Deree Alba
1875

Module	**ITC 6003 – APPLIED MACHINE LEARNING**		
Term	**WINTER SEMESTER 2020**		
Assessment	**PROJECT**	**Weight**	**50%**
Duration			
Deliverables	1. *Report in TurnitIn* 2. *Code in Blackboard* 3. *An oral examination/ presentation of your work* 4. Code in GitHub		
Method of Submission	*TurinitIn, Blackboard, GitHub*		
Deadline:	*13th Week* *Grading US system*		

The rules of academic ethics apply when taking this assessment, including the requirement that you produce work ***without improper or unauthorized assistance*** *from anyone.*

General Instructions

Your project involves a series of experiments, observations coming out of the experiments, and drawing conclusions. Essentially you will collect data (or they will be provided by the instructor), then a programming language will be used (you are encouraged to use python) along with the appropriate libraries to process the data. Tables, diagrammes, and data visualizations are essential for presenting your findings.

Deliverables: a) code in blackboard, along with instructions for running it b) a report of 3000±500 words that will present your findings, and will be submitted at Turnit-in. The report must be self-contained. If you need to exceed the word limit use an appendix. c) an oral presentation d) code in GitHub
Team size: three persons

Grading: A grade will be assigned to the project, but it will be divided among the team member according to peer assessment.

General Instructions

To carry-out the classification, clustering or regression task you need to consider the following steps:

a. Data description & Visualization that aids the comprehension of the problem
b. Data pre-processing
c. Data/feature selection/evaluation
d. Decide how to split the data between training and data set
e. Use multiple classifiers and evaluate the parameters of each classifier: Try at least the following: Support Vector Machines (linear, and non-linear), Decision Trees, Naïve Bayes, and one based on ensemble learning (especially consider the Random Forests) and Neural Networks.
f. Use clustering algorithms, evaluate parameters. Try at least the following: k-means, DBSCAN, gaussian Mixtures, agglomerative (hierarchical) clustering
g. In regression: Try at least linear regression, polynomial regression and a regression algorithm of your choice. Explore regularization.
h. Evaluate
    a. the performance of each classifier: at least provide F1 measure, precision, recall and ROC curves (if applicable)
    b. clusters based on criteria such as silhouette, and inertia
    c. regression based on criteria such as the R score and others
i. Observe things and draw conclusions
j. Future work: Also include things you might try/consider in the future

1. Classification: Predicting arrhythmia type (20%)

Source data & description: http://archive.ics.uci.edu/ml/datasets/Arrhythmia

The aim is to distinguish between the presence and absence of cardiac arrhythmia and to classify it in one of the 16 groups.

2. Clustering: Market Segmentation: Unsupervised learning (20%)

Source data & description: https://archive.ics.uci.edu/ml/datasets/Wholesale+customers

Discover clusters, evaluate and characterize them.

3. Regression (20%)

It is up to you to choose regression task, but you should inform the instructor and get approval for it.

Indicative data sources: Kaggle.com , https://www.analyticsvidhya.com/ , https://github.com/awesomedata/awesome-public-datasets , https://www.openml.org/ .

4. Scaling-up: Predicting buys (20%)

Source data & description: https://2015.recsyschallenge.com/challenge.html Your task is to predict whether a user will buy a product or not based on his/her online behavior, and in particular his/her clicks during a session. There are two files:

yoochoose-clicks.dat - Click events. Each record/line in the file has the following fields:

- Session ID – the id of the session. In one session there are one or many clicks.
- Timestamp – the time when the click occurred.
- Item ID – the unique identifier of the item.
- Category – the category of the item.

yoochoose-buys.dat - Buy events. Each record/line in the file has the following fields:

- Session ID - the id of the session. In one session there are one or many buying events.
- Timestamp - the time when the buy occurred.
- Item ID – the unique identifier of item.
- Price – the price of the item.
- Quantity – how many of this item were bought.

The Session ID in yoochoose-buys.dat will always exist in the yoochoose-clicks.dat file – the records with the same Session ID together form the sequence of click events of a certain user during the session. The session could be short (few minutes) or very long (few hours), it could have one click or hundreds of clicks. All depends on the activity of the user.

Tasks to perform

1. Build a data set that can be used in classifier to decide whether someone will buy or not.
2. Preprocess the data & perform classification.

3. Report Quality (10%)

The quality of report is based on many factors including: organization of the material, presentation of data, experiments, models, evaluation, drawing conclusion using various aids such as tables, diagrammes, equations etc., and references if applicable.

4. Oral Presentation (10%)

During the presentation each group will present their work in a comprehensive manner and will be called to answer questions regarding their work.

Grading scale: US System

	GP	Letter	US
Excellent	4.00	A	90+
Very good	3.70	A-	86-89
Very good	3.50	B+	81-85
Good	3.00	B	73-80
Satisfactory	2.50	C+	64-72
Satisfactory	2.00	C	51-63
Fail	0	F	<50

The rules of academic ethics apply when taking this assessment, including the requirement that you produce work ***without improper or unauthorized assistance*** *from anyone.*