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 2023**		
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* *US Grading scale*		

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 in Python, 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, that is all experiments performed and all conclusions should be reported. If you need to exceed the word limit, use an appendix. c) an oral presentation d) code in GitHub
Team size: three persons

		Person being rated		
		Person-1	**Person-2**	**Person-3**
Person doing the rating	**Person-1**	1.25	1	0.75
	Person-2	1.10	1.10	0.80
	Person-3	1	1	1
Average Rate		**1.12**	**1.03**	**0.85**
Individual score (project grade: 80%)		**89.6**	**82.4**	**68**

Grading: peer-review

Teams consist of two or three persons. Group members will be asked to rate the relative contribution of themselves and the other group member(s). The ratings provided by each member must add up to the number of persons the group consists of (see example above). These ratings will be taken into account in the final grading of the project for each individual.

Example: In a group consisting of three members, each member provides a rating of all group members. As this is a three-member group, the ratings provided by each member add up to 3.00. A rating of 1.00 means that the person in question did exactly as much as expected of him/her. A rating that is less than 1.00 means that the person in question did less than expected, whereas a rating that is greater than 1.00 means that this person's contribution was greater than expected.

General Instructions

To carry-out the classification, clustering, and regression task you may 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. (If not stipulated by the instructions)
e. Use multiple classifiers and evaluate the parameters of each classifier: Try at least the following: Decision Trees, 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, and agglomerative (hierarchical) clustering.
g. In regression: Try at least linear regression, and polynomial regression. Explore regularization.
h. Evaluate
    a. the performance of each classifier: at least provide F1 measure, precision, recall and ROC curves (if applicable) and AUC.
    b. clusters based on criteria such as silhouette, and inertia.
    c. regression based on criteria such as the R score and others.
i. Observe findings and draw conclusions.
j. Future work: Also include things you might try/consider in the future.

1. Clustering: Market Segmentation: Unsupervised learning (25%)

The data set refers to clients of a wholesale distributor.
https://archive.ics.uci.edu/ml/datasets/Wholesale+customers

Summarize the data set by discovering clusters, evaluate and characterize them.

2. Regression (25%)

It is up to you to choose a regression problem, 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/ .

The data set should present some challenges, e.g., size, missing values, or categorical features.

3. Predicting outcome (25%)

Consider the following data set which is about predicting the annual income of a person (<50K or >50K): UCI Adult data set (https://archive.ics.uci.edu/ml/datasets/adult).

I. Prediction
   a. Training can be done on the adult.data
   b. The evaluation (i.e. testing) should be done on the adult.test

II. Explainable AI (XAI): The explainability of the machine learning models has become very important. In this task you are required to do some research on the Shapley Additive Explanation (SHAP) or the Local Interpretable Model-Agnostic (LIME). Examine the results of explainability on the current data set on the best model found in step I.

Data set description

>50K, <=50K: **class**
Features
age: continuous
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
fnlwgt: continuous
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
education-num: continuous
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
sex: Female, Male
capital-gain: continuous
capital-loss: continuous
hours-per-week: continuous
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands

1. Report Quality (15%)

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.

2. 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