**Department of Computer Science**
**University of Cyprus**

# MAI643 Artificial Intelligence in Medicine

Spring Semester 2024

Project Assignment 1

Weight: 35% (of the total mark)

**Description:**

Develop a medical decision support system, applying for a disease diagnosis, following all the steps of a data analytic approach: data exploitation – acquisition, pre-processing, model development for classification, model evaluation and validation with medical literature. The methods will be applied on a medical dataset which will be selected from the list of open datasets given below.

**Assessment Evaluation:**

The project requirements are the following:

- Follow all the steps specified in the sequel for the corresponding chosen dataset using Python or R (use existing libraries).
- Submit in Moodle, the word/power point document and source code files requested for each deliverable.
- You will work in a group of three people.

**List of Open Datasets:**

1. **Structured Tabular Dataset:**
   Cervical cancer dataset - http://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29
2. **Longitudinal Dataset:**

   Myocardial infarction complications Data Set - http://archive.ics.uci.edu/ml/datasets/Myocardial+infarction+complications

3. **Image Dataset:**

   SPECTF Heart Data Set (already pre-processed) - http://archive.ics.uci.edu/ml/datasets/SPECTF+Heart

See below the different milestones of the project along with deliverables, deadlines and mark breakdown.

## Deliverable 1: Brief problem description

## Deadline: 6/2/2024 [20%]

Word document including:

1. Project Title
2. Brief description of the problem – description of the medical domain and the problem (2-3 paragraphs)
3. Database/dataset description – as provided in the given link and cited papers
4. 5-10 references should be given (hint: identify similar papers/reports published relevant to the dataset, the medical domain and/or methods to be investigated. You can find some papers cited in the dataset link. Follow IEEE format for the references (https://ieeeauthorcenter.ieee.org/wp-content/uploads/IEEE-Reference-Guide.pdf))

## Deliverable 2: Explore the data

## Deadline: 20/2/2024 [25%]

Word document including:

1. Project Title
2. Data understanding, i.e., % of missing data, identification/understanding of features, #classes, imbalanced data (tables, graphs, or short description)
3. Data pre-processing steps followed (data cleaning, encoding, scaling, feature selection, dimensionality reduction, imbalanced data handling)
4. Results from the data pre-processing (i.e., list of selected features, % of records from each class)
5. Algorithms/methods to be used (simply provide **two** potential machine learning algorithms you expect to investigate)

## Deliverable 3: Model the data

## Deadline: 12/3/2024 [25%]

Word document including:

1. Project Title
2. Methodology

2.1 Training/Evaluation Data Sets – Splitting method

2.2 Description of the selected **two** models (implementation details description and assumptions taken e.g., KNN, logistic regression, decision trees, random forest, Naïve Bayes, SVM, ensemble methods, etc.)

2.3 Model Evaluation Metrics Description (Accuracy, Sensitivity or Recall, Specificity, Precision, F1 Score, and any other measure needed including any imbalanced evaluation metrics) (to be reported for both the training and testing sets)

3. Results from both models (tables/figures) – not necessarily to be complete at this stage

## Deliverable 4: Communicate and Validate the Results

## Deadline: 16/4/2024 [25%]

**Submission:** Create a github project and share it with the instructor: kalia.orphanou@gmail.com. Add the following on your github project:

a. **Final Report** including:
   1. Project Title
   2. Final parts of Deliverables 1-3
   3. Results
      3.1 Proposed best model/models to be used
      3.2 Tables/figures – results of both models after given feedback
   4. Discussion

   4.1 Main findings

   4.2 Comparison of results with literature [prepare a table comparing the main findings of your work with other studies applying the same dataset based on features selected, algorithms used, performance metrics, etc.]

   4.3 Discuss/validate the results based on the literature related to the selected medical domain

   4.4 Discuss any limitations of this study

b. **Source code** of your implementation along with instructions of how to run your program and comments on your code.

## Deliverable 5 [5%]

## Deadline: Week 13

A short presentation in the class (~15 minutes) including:

- Brief description of the medical domain

- Brief description of the dataset
- Point out the process/techniques used for data cleaning and development of classification models
- Presentation of the results in tables and comparison with other works (use tables/graphs)
- Interpretation of the results to medical experts (i.e., doctors, nurses, oncologists…)