Develop Capabilities and Talents
Not-for-Profit University

American University of Madaba
Faculty of Information Technology
Dept. of Computer Science

Fundamentals of Artificial Intelligence

Project II: Titanic Passengers Survival Prediction January 19, 2022

Dr. Iman Abu Hashish

1 Objectives

The objective of this assignment is to get familiar with formalizing a given problem as a machine learning problem and follow the typical steps for building a machine learning system.

2 The Titanic Dataset

The problem to be formulated is predicting which passengers survived the Titanic shipwreck by building a machine learning model. Refer to the datasets provided on eLearning. You will find two datasets as follows:

1. `train.csv`: consists of *891* data points and *12* variables, one of which, i.e., `Survived` is the variable we want to predict.
2. `test.csv`: consists of *418* data points and *11* variables. The test set does not have a column for `Survived`, as it is the target variable.

A description of the *12* variables is provided as follows:

a. `PassengerId`: the passenger's ID number.

b. `Survived`: indicates whether the passenger survived or not – *0 = No, and 1 = Yes.*

c. `Pclass`: passengers' class – *1 = 1st class, 2 = 2nd class, and 3 = 3rd class.*

d. `Name`: passengers' names.

e. `Sex`: passengers' sex.

f. `Age`: passengers' age.

g. `SibSp`: number of siblings and/or spouses abroad.

h. `Parch`: number of parents and/or children abroad.

i. `Ticket`: ticket number.

j. `Fare`: passengers' fares.

k. `Cabin`: cabin number.

l. `Embarked`: port of embarkation – *C = Cherbourg, Q = Queenstown, and S = Southampton.*

3 Steps for ML System Development

To address the problem, follow the typical steps for ML system development as follows:

a. **Data collection**: given the provided dataset, formulate the problem as a machine learning problem.

b. **Features engineering**: check for missing values, handle categorical variables, add new attributes if needed, and make sure the dataset is clean and ready before moving on to the next step.

c. **Exploratory data analysis and visualizations**: provide summary statistics and at least four visualizations to explore the data and gain an understanding of it.

d. **Model selection and training**: given the nature of the data, implement a classification tree using the cleaned training set, make sure to implement K-fold cross validation to avoid overfitting. Using the test set, predict which passengers survived the Titanic shipwreck.

e. **Performance measurement**: calculate the accuracy score of your tree to see how well it predicts.

4 Deliverables

Each group has to submit a brief report containing:

- A problem formulation.
- The results you get after implementing each step as explained.
- A functional documented implantation in Python.
- A list of references for the resources you have used, if any.

5 Deadlines and Assessment

Each group must submit their work, no later than January 27, 2022. The assessment is based on a 10-minute discussion with each group.