# Data Mining
## Assignment 1

Ali Gholami

Department of Computer Engineering & Information Technology
Amirkabir University of Technology

*http://ceit.aut.ac.ir/~aligholamee*
*aligholamee@aut.ac.ir*

**Abstract**

In this assignment, several paramount concepts of *Data Analysis* will be explained. we'll discuss the importance of metrics in the first theoretical problem. A quick review on the *Apriori* algorithm for the *Association Rule Mining* will be explained also. We'll also show how *Weka* can be used for *Association Rule Mining*. Furthermore, The effectiveness of *Normalization* concept is proposed. Finally, an *Statistical* point of view will help us to demonstrate and rationalize the relationship between the *Performance* of the *Learning Algorithm* and the amount of *Data* available. A chief section of this assignment is dedicated to solve the *Titanic* problem, which is a great practice of data mining concepts in production. We'll use *Python* programming language and three main libraries; *Scikit-Learn*, *Pandas* and *Numpy* to tackle this problem. The Python implementation of the Titanic problem is provided on a *Jupyter Notebook* attached with this report.

**Keywords.** *Apriori, Association Rule Mining, Normalization, Generalization, Preprocessing, Feature Engineering, Scikit-Learn, Pandas, Numpy, Python 3.5.*

# 1 Performance Metrics Analysis

Given the following *Confusion Matrix* for a prediction about cancer.

| | | Predicted Class | | |
|---|---|---|---|---|
| | | Cancer = Yes | Cancer = No | Total |
| Actual Class | Cancer = Yes | 60 | 290 | 350 |
| | Cancer = No | 150 | 9500 | 9650 |
| | Total | 210 | 9790 | 10000 |

Table 1.1: Confusion matrix of cancer prediction.