CS5701(Quantitative Data Analysis)

Lab 7: Part 2

The aim of this part of the lab exercise is to give you practical experience in Logistic regression analysis using R Studio and an R Notebook.

1 Logistic Regression

In this lab we will be using the Titanic dataset (https://www.kaggle.com/c/titanic/data). This data us used as a benchmark for many competitions. There are two versions: the first csv (titanic-analysis.csv) contains a subset of the columns and the second one has all columns (titanic-all-cols.csv). The second data set is the one used for the *Titanic: Machine Learning from Disaster* Competition.

The data includes the following variables:

- PassengerId
- Survived (1= survived, 0 = died)
- Pclass - Passenger's travel class
- Name
- Sex
- Age in Years
- SibSp - number of siblings / spouses aboard the Titanic
- Parch - number of parents / children aboard the Titanic
- Ticket - ticket number
- Fare - passenger fare
- Cabin - cabin number (if known)
- Embarked - the port of embarking (C = Cherbourg, Q = Queenstown, S = Southampton)

1. Load the `titanic-analysis.csv` data into an R notebook.
2. Explore the data numerically and graphically. Confirm the variables that are categorical and numerical/continuous and that R has read them in appropriately.
3. Is there a relation between the likelihood of surviving and the gender, age and fare paid?
4. What are the odds ratio for Male vs Female passengers?
5. OPTIONAL: Load the `titanic-all-cols.csv` into your R notebook. Explore this additional data, and see if you can enrich your Logistic Regression model.