CA05-A – Logistic Regression

1. The Application

Cardiovascular Disease (CVD) kills more people than cancer globally. A dataset of real heart patients collected from a 15 year heart study cohort is made available for this assignment. The dataset has 16 patient features. Note that none of the features include any Blood Test information.

2. Deliverables

Your job is to:

Part 1: build a binary classifier model to predict the CVD Risk (Yes/No, or 1/0) using a Logistic Regression Model with the best performance possible (deliverable: Notebook)

Part 2: Display the Feature Importance of all the features sorted in the order of decreasing influence on the CVD Risk (deliverable: Notebook)

Part 3: Evaluate the performance of your model (including ROC Curve), explain the performance and draw a meaningful conclusion. (deliverable: Performance outputs in Notebook, explanation and conclusion in Word/PDF document)

3. Data Source and Description

Data File Name: cvd_data.csv

File Location: https://github.com/ArinB/CA05-B-Logistic-Regression/raw/master/cvd_data.csv

NOTE: Use the above EXACT URL in your code as data file location

Data Column (Feature Name) Descriptions:

cvd_4types: Label Column. 0 indicates "No Risk", 1 indicates "Risk Present"
age_s1: Age in Years
race: 1 - White, 2 - Black, 3 – Other
....
....

Get the definition of rest of the 16 features by searching on the feature name at the following web page:

https://sleepdata.org/datasets/shhs/variables