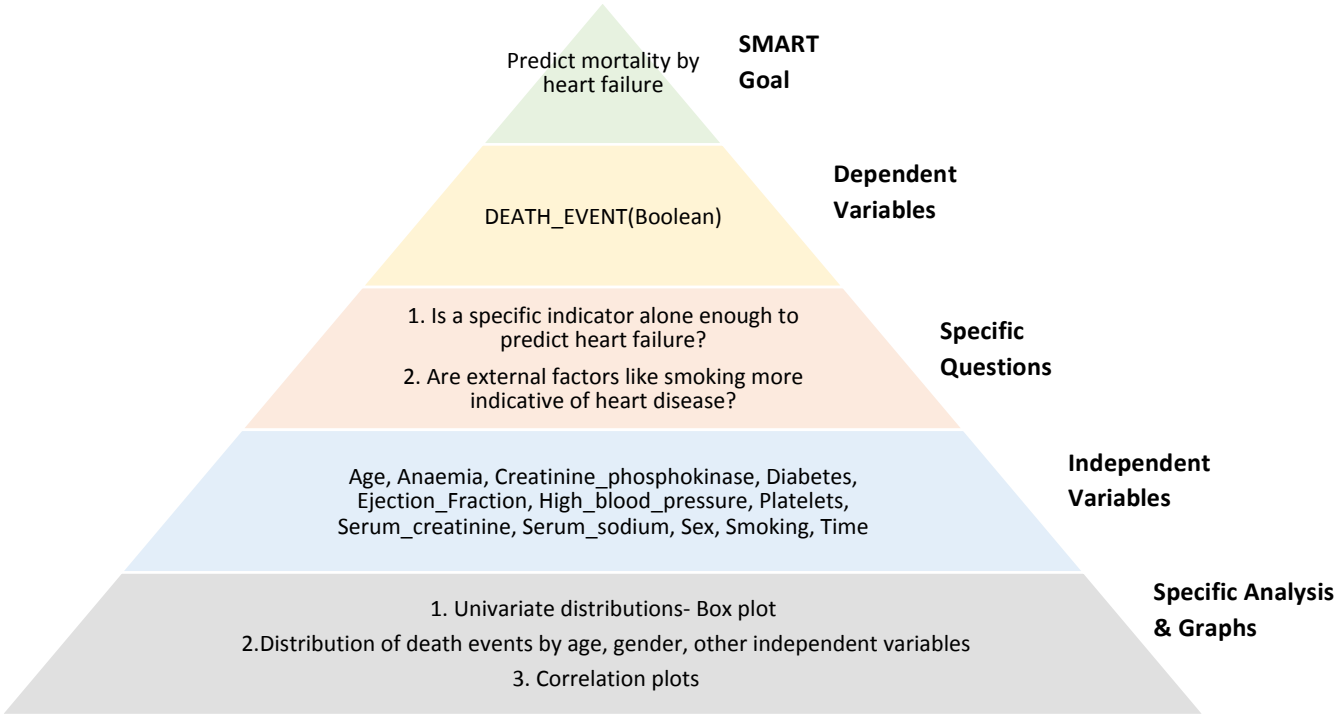
26:630:670 Multivariate Analysis Fall 2020

Assignment 1

1. GitHub link with the appropriate repository.
2. GitHub repository should have the appropriate name reflecting the dataset being used and what is the purpose of the repository.
3. It should briefly describe the problem statement for the dataset
4. It should contain information about the team members.
5. It should also have a dataset dictionary - A data dictionary contains information about the data set.
6. The data dictionary would be written using the GitHub markdowns.
7. Please upload PDF file for the assignment.

Github Link - https://github.com/ag77in/HeartFailurePrediction-MVA

Initial Draft of SPAP – Version 1.0



Assignment 2

1. This assignment is to determine KPI (Key performance indicator) to measure success for the main question of your project
2. Document the hypothesis for each of the questions raised for your project.
3. Update SPAP details in your GitHub profile and submit a GitHub link on the blackboard.

1. KPI Discussion

Our SMART goal for this exercise is to predict mortality by heart failure in patients in the dataset provided. There is no ambiguity in goal so our dependent variable is well-defined.

The primary KPI to measure success for this would be the accuracy of a machine learning model one can develop to predict the Dependent Variable of DEATH_EVENT (0 or 1).

Below are some accuracy measures we will look at to understand the predictive power of this classification model –

- Confusion matrix (classification accuracy) - Primary KPI
    - We will look at total correct predictions / total input samples

$$Accuracy = \frac{True\ Positive + TrueNegative}{Total\ Input\ Sample}$$

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

- AUC – Area under curve
  - AUC is the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example (note: in our case a positive event will be the death of a patient denoted by '1' while the sruvival of a patient is a negative example denoted by '0'). This will require us to plot the True Positive rate (y-axis) vs False positive Rate (x-axis). It will take a range of [0,1]

- Precision – To understand of the predicted death events, how many did actually occur in the data
  - Formula is denoted by

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positives)}$$

- Recall – To understand of the actual death events, how many did we predict accurately
  - Formula is denoted by

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negatives)}$$

- F1 score – It is the harmonic mean of precision and recall
    - Formula is denoted by

$$F1\ Score = 2 * \frac{1}{(1/precision + 1/recall)}$$

These measures are important because we may have imbalance data and hence would want to move our proability threshold of classifier while having a good balance of our precision and recall

In order to choose between models, we can also look at measures like –

- AIC (Akaike Information criterion)
- BIC (Bayesian Information criterion)

Both of the above measures take the form –

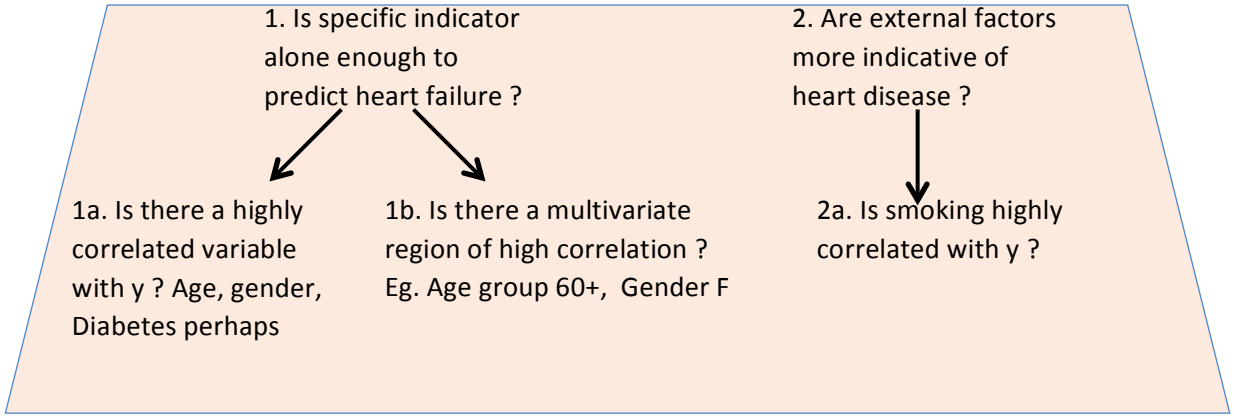
$$AIC = 2k - 2\ln(\hat{L})$$

$$BIC = k\ln(n) - 2\ln(\hat{L})$$

Here k is # parameters in model, and $\hat{L}$ is the maximum value of the likelihood function, n is the number of data points in sample. We prefer the one with minimum AIC / BIC which serves to show model parsimony (lesser number of parameters for a higher goodness of fit)

2. Document hypothesis for questions raised in your project –

Let us re-visit the specific questions and note the hypothesis we can raise in our problem –



Hypothesis in detail –

Univariate hypothesis -

- Is gender a good indicator of death event ?
  - Our hypothesis is that gender would be somewhat indicative of heart failure but not as much as other events
- Is age a good indicator of death event ?
  - Our hypothesis is that age would be a good indicator of heart failure
- Is diabetes a good indicator of death event ?
  - Our hypothesis is that individuals with diabetes would be more likely to have heart failure than those who didn't
- Is anaemia a good indicator of death event ?

    - Our hypothesis is that individuals who are anaemic and have reduced Red blood cells counts would be more likely to have heart failure than those who don't
- Is high blood pressure a good indicator of death event ?
    - Our hypothesis is that individuals with high blood pressure are morely to have heart failure than those who don't
- Is smoking a good indicator of death event ?
    - Our hypothesis is that individuals who smoke are more likely to have heart failure than those who don't
- We can construct similar hypothesis for creatinine_phosphokinase, ejection_fraction (does lower indicate higher likelihood ?) , serum_creatinine, serum_sodium, platelets i.e do their levels indicate a higher death event
    - Our hypothesis is that mose of these would be good indicators
- Is a longer follow up time a good indicator of death event ?
    - We would expect longer follow up times to be more indicative

Bi-variate/ Multivariate hypothesis –

- Is age and gender a good indicator of death event ?
    - Our hypothesis is that there could be a range of age and gender where death event are more likely however it could be based on the chosen sample
- Is age and diabetes a good indicator of death event ?
    - Our hypothesis is that there could be a range of age and diabetic individuals where death event are more likely
- Does gender and levels of creatine_phosphokinase / serum creatinine indicate death event better ?
    - We would expect one of male or females to have higher levels of enzyme or creatinine in blood as opposed to the other indicating a higher chance of heart failure
- We can similarly construct some more more hypothesis for age & anaemia, serum creatinine & serum sodium

Visualization –

Given we have all information on data, we will not require any color coding however we can decide on what chart is needed for our specific hypothesis

- Uni-variate analysis –
    - Histogram for distribution of age etc.
    - Box plot (for outlier analysis)
- Bi-variate analysis -
    - Bar chart
    - Pie chart (to see Gender vs death event etc.)
    - Box plots
- Multi-variate analysis -
    - Heat Map (Correlation analysis)

With this information, let's update our **SPAP version 1.1**

Second Draft of SPAP – Version 1.1