1. **Problem statement**

* Prediction of Heart Disease

1. **EDA ( Data Information ) (10)**
2. How many features do you have?
3. How many observations do you have?
4. What is the data type of each feature?
5. Do you have null values?
6. What is the distribution of each variable?
7. Do there appear to be outliers?
8. Are the max/min values reasonable for the variables? Do you see any values that look like errors?
9. What is the mean for each variable? What do the means tell you about your dataset as a whole?
10. Is there any duplicated row?
11. Can you drop the rows with null values without it significantly affecting your analysis?
12. Looking at the distributions of the variables, can you justify filling in the missing values with the mean or median for that variable? (We use mean for Numerical Data with no Outliers, Median for Numerical Data with Outliers and we use Mode for Categorical data)
13. Do you have outliers (represented as dark circles on the boxplots) in your variables?
14. Which variables are most correlated with your target variable? (If applicable)
15. Is there multicollinearity? (Two features that have a correlation > 0.8) How will this affect your model?
16. Do you have variables that represent the same information? Can one be dropped?
17. **Data Cleaning ( Feature Engineering and Find Outlier ) (5)**
18. Do you have outliers (represented as dark circles on the boxplots) in your variables?
19. Why do you think you have outliers?
20. Do the outliers represent real observations (i.e. not errors)?
21. Should you exclude these observations?
22. What method would you use for feature engineering?
23. **Modeling (3)**
24. What method would you use for model selection?
25. Which are algorithms that can be used for a classification problem?
26. How can you avoid overfitting your model?
27. **Model Analysis (5)**
28. Which classification algorithm gives you the best accuracy?
29. Which classification algorithm gives you the better ROC curve?
30. Which classification algorithm gives you the better precision?
31. Which classification algorithm gives you the better recall value?
32. Check Whether the data is highly imbalanced or not?
33. **Deployment**

* Streamlit