Project topic:

Algorithm optimization and data imputation techniques in cancer prediction

Dr Amol Shinde

Description about the problem :

- Cancer is a group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body. Cancer is usually diagnosed in advance stages.
- Investigations required to diagnose cancer are costly.
- So if routine investigation can predict the presence of cancer ,it would be of great help for the patients.
- My aim is to develop and optimize the accuracy of models that can predict the occurrence of cancer with the help of metabolic parameters that can be used as screening methods.

Aim

1. To explore machine learning algorithms that may be useful to predict cancer presence.
2. To optimize the accuracy of classifiers used in prediction of cancer with the help of metabolic parameters.
3. To explore data imputation techniques to solve the issue of missing data.

Methodology :

1. Sampling and partitioning of data will be done.

   - Train and Test set.

2. For each variable, its median value, interquartile range, means and standard errors, the z-values, the p-values and the odds ratios will be obtained.

3. Univariate analysis :

- to assess the diagnostic value of each parameter mentioned

4. Multivariate analysis is performed :

   - Predictors are combined

   - ROC analysis will be done

- In this following things will be done:

  a) Evaluation of role of each variable :

  - Variable importance plot will be built
  - It provides a list of the most significant variables in descending order by a mean decrease in Gini coefficient.

  b) Various Classifiers will be tried

  c) Accuracy of the models will be checked :

  - Evaluation metrics like Logarithmic Loss, Confusion Matrix, Area under Curve and F1 Score will be explored

Output will be in terms of

1. Accuracy of developed models.

2. Specificity and Sensitivity of cancer prediction by models developed in this project.

3. Data imputation techniques will be discussed –

- a. Deleting Rows
- b. Replacing With Mean/Median/Mode
- c. Assigning An Unique Category

THANK YOU