Complete Case Analysis

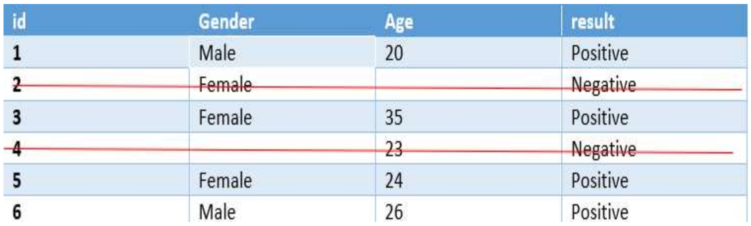
06 July 2023 21:10

Complete Case Analysis(CCA):

CCA is process of carrying out the analysis of the data after **deleting every row** which has **one or more missing values**.
It is also called **"List-wise Deletion" of Cases**.

Complete Case Analysis means literally analyzing only those rows which contain values in all the features in the datasets.

CCA should be used only when **number of missing values is less than 5% of the total values in the corresponding feature.** Otherwise lot of informative observations might get removed.



id	Gender	Age	result
1	Male	20	Positive
~~2~~	~~Female~~		~~Negative~~
3	Female	35	Positive
~~4~~		~~23~~	~~Negative~~
5	Female	24	Positive
6	Male	26	Positive

Main assumption for CCA to be carried out is that the missing data is **Missing Completely At Random (MCAR)**. This means that the probability of a data point being missing does not depend on the observed or unobserved values.

In this case, removing missing value won't change the distribution of the data as missing values are randomly distributed not present systematically.

Advantages & Disadvantages of CCA:

Advantages:-

1. It is **Easy to implement** using **dropna()** in Pandas and no data manipulation is required.
2. If missing data is **MCAR** then It **preserves the distribution** of the features as distribution of features before and after CCA remains almost same.

Disadvantages:-

1. If missing values are in abundance then it **excludes a large portion of the dataset** which might contain some important information. That's why CCA is preferable only when **number of missing values is less than 5% of total values in the feature.**
2. When we use model train on such data in production, then It won't understand how to handle null values.