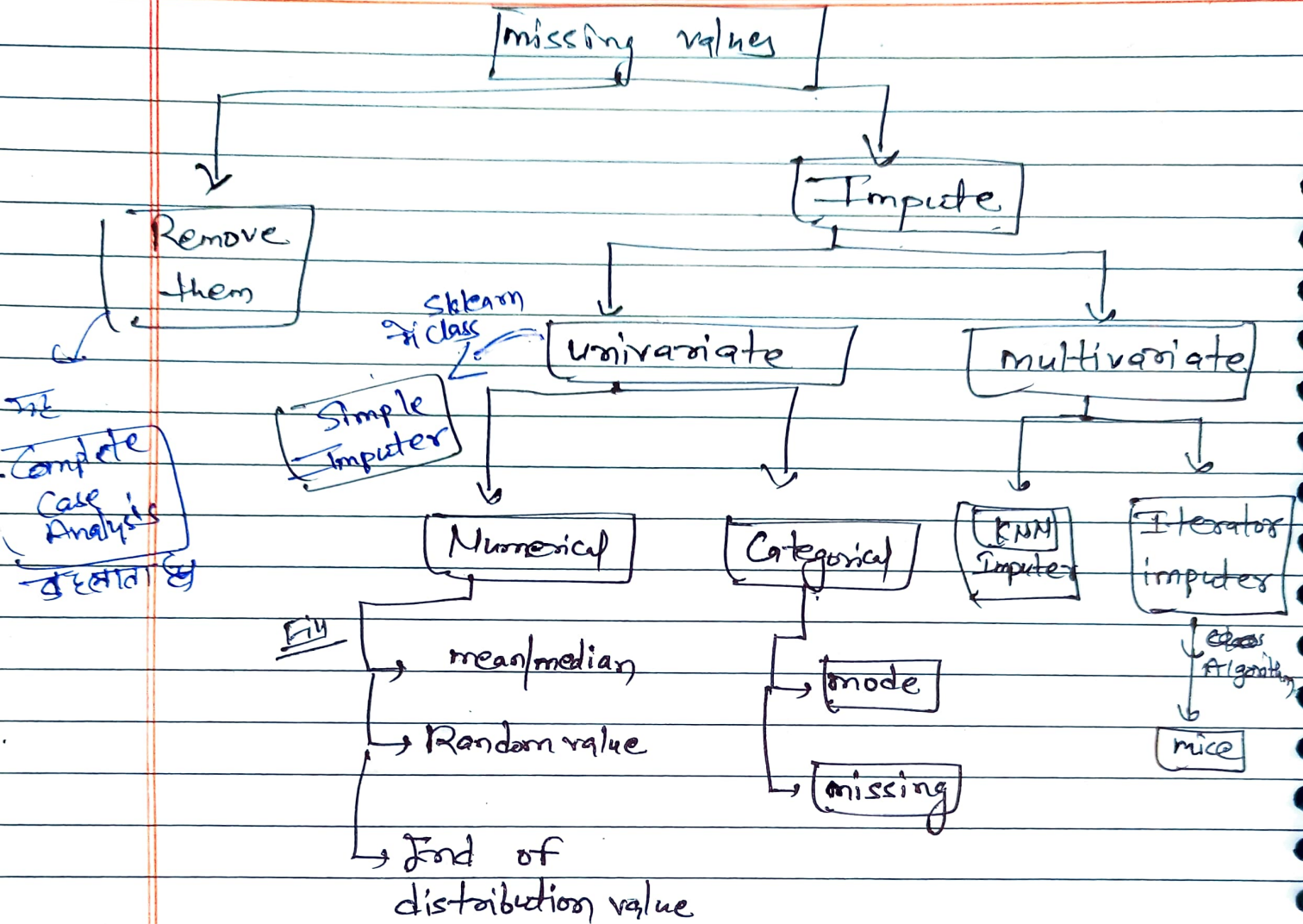


Handling missing Data



Complete case analysis (CCA) also called "list-wise deletion" of cases, consists in discarding observations where values in any of the variables are missing.

Complete case analysis means literally analyzing only those observations for which there is information in all of the variables in the dataset.

⇒ Assumption for CC9 :-

- ① ~~MCAR~~ The data is missing completely at random.

✗

[Age] → 2000x4 → 50 values missing

↳ (950x4)

⇒ Advantage :-

- ① Easy to implement as no data manipulation required.
- ② Preserves variables distributions (if data is MCAR, then the distribution of the variables of the reduced dataset should match the distribution in the original dataset).

⇒ Disadvantage :-

- ① It can exclude a large fraction of the original dataset (if missing data is abundant).
- ② Excluded observations could be informative for the analysis (if data is not missing at random).
- ③ When using our models in production, the model will not know how to handle missing data.

⇒ When to use CCA?

① MCAR

② generally अगर $[5\% < \text{data}]$ तो ही CCA Apply करेंगे

③ अगर $[95\% < \text{data}]$ missing values हैं तो इस data को Row and Column से Remove कर देंगे

⇒ Handling missing Numerical Data :-

Numerical Data

Univariate imputation

↓
अगर किसी ~~Row~~ Row में कोई value missing है तो ~~अगर~~ उसी Particular row की help से अपने values fill कर रहे हैं, तो इस Process को Univariate imputation कहते हैं।

Multivariate imputation

↓
अगर हम किसी Row के missing values को आस-पास के Rows की help से fill कर रहे हों, तो इस Process को Multivariate imputation कहते हैं।

Ex)

1	2	3	4
⊖			

↓
missing and fill with second row

Ex)

1	2	3	4
⊖			

missing but fill with 1st or 3rd or 4th row

Univariate imputation :-

(i) Mean/Median Imputation :-

⇒ अगर हमारा Data का distribution Normal है तो हम Mean को use में लेंगे।




⇒ यदि हमारा Data थोड़ा बहुत skewed है तो हम Median को use में लेंगे।



⇒ Benefits of this imputation :-

- ① Simp. easy way process
- ② जब हम हमारा m.l. model, server पर भी ~~load~~ deploy करते हैं तो बहुत easy होता है इसे recreate / reuse करना।
- ③ अगर (5% < missing values) तो उस Condition में use ना करें। और Production में भी use ना करें।

⇒ Disadvantages :-

- ① जब हम missing values को mean/median से replace करते हैं तो उस Process में distribution का shape change हो जाता है। Ex 
- ② Outliers आ जाते हैं। अर्थात कुछ extra Outliers आ जाते हैं जो कि Outliers नहीं हैं। But हमारा system उसे Outlier समझने लग जाता है।

- (3) माना Age नाम का Column है जिसमें missing values हैं, ~~लेकिन~~ इस Age Column का जो Relationship है बाकि Column के साथ, वह भी change हो जाता है। इस ~~process~~ को Phynomina को Covariance / Correlation में Changes कहा जाता है।

When to use?

- (i) When data is missing Completely at Random (MCAR).
(ii) 5% > missing data हो तब use करें।

(3) Arbitrary Value Imputation :-

Categorical data \rightarrow NA \rightarrow missing

Numerical में भी use कर सकते हैं।

\Rightarrow Arbitrary Value Imputation में हम Difference create करते हैं। जैसे जैसे observation का value पर data है वैसे जैसे data नहीं है।

\Rightarrow Benefit

easy to apply

\Rightarrow Disadvantage :-

- (1) graph display हो जाता है।
- (2) variance change हो जाता है।
- (3) Covariance / Correlation with other column में change हो जाता है।

NOTE \Rightarrow Arbitrary method ko use krna hai jisse Data is not missing at random.

(4) End of Distribution Imputation :-
इसमें हम किसी की Arbitrary value को missing value से replace kr dete hai

(i) अगर हमारा Data Normal Distributed है तो हम $(mean + 3\sigma)$, $(mean - 3\sigma)$ से value को Replace kr dete hai

(ii) यदि Data Skewed है तो हम (IQR Proximity) use में लेते हैं। ऐसी Problems में हम $[(Q1 - IQR \times 1.5)]$ से value को Replace kr dete हैं। या $[(Q3 + 1.5 IQR)]$ से Replace kr dete हैं।

जहाँ $IQR = Q3 - Q1$

$Q3 = 75\%$

$Q1 = 25\%$

\Rightarrow Benefits :-

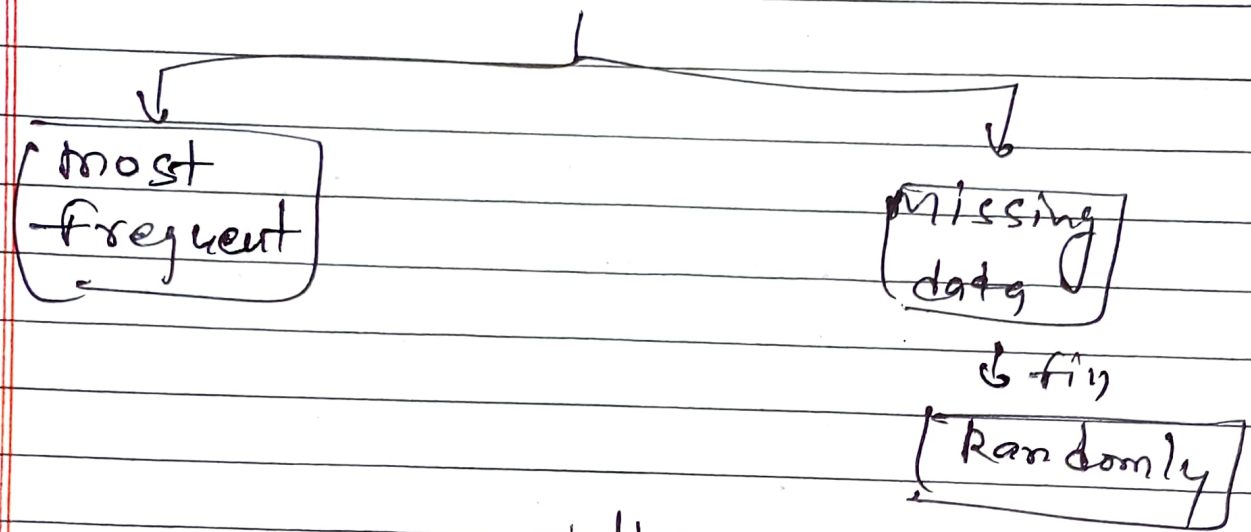
① easy to apply

② Disadvantages :-

① graph distplot में अंतर है
② variance change हो जाए

NOTE \Rightarrow End of Distribution Imputation ko use krne में Data is not missing at random (NMAR).

Handling Categorical Missing Data :-



NOTE :-

