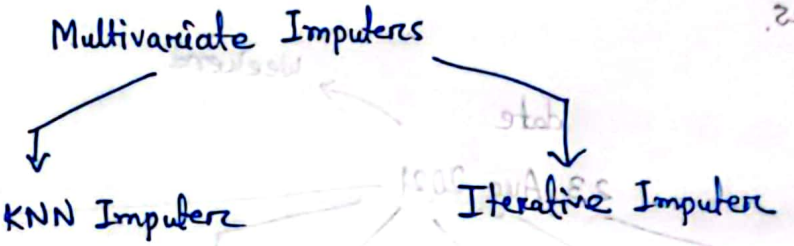
Univariate Feature missing value imputers: Already noted.

Multivariate Features missing value imputers:

~~KNN Imputers~~

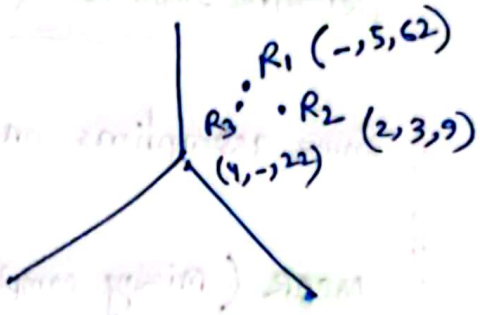
In univariate imputation, we just imputed the missing value with those techniques which are mostly bound to that specific column values.
But in multivariate imputation, we can take the help of other rows and columns as well to fill in the missing values of a particular column.

There are two types of multivariate imputers we will study




KNN Imputer: The basic concept of KNN Imputer is, suppose, feature 1 column has a missing value. To fill up that missing value we will take another row value whose Euclidian distance, is the smallest, and that row has to contain value in it's Feature 1 column.

| | Feature 1 | Feature 2 | Feature 3 |
|---|---|---|---|
| R1 | — | 5 | 62 |
| R2 | 2 | 3 | 9 |
| R3 | 4 | — | 22 |




So, we will find the euclidian distance using the coordinates (row values) and then we will take the value from that row whose distance is smallest and impute that value to the current row.

Advantage and Disadvantage:

1) More accurate

Disadvantage:

1) More no of calculation
2) You have to put your X-train in the server because if any missing value comes from user input, then it needs other rows for calculation and doing imputation.

## Iterative Imputer : (MICE Algorithm)

Some assumptions on Missing data

**MCAR** (Missing completely at random) → The data that are missing, they are actually not collected. That's why they are missing.

**MAR** (Missing at Random) → While collecting data from people one org, they intentionally didn't provide any data or data was optional so some people didn't provide it. That's why they are missing.

**MNAR** (Missing not at Random) → Data are missing because they are removed consciously.

Among the three, MAR missing values can be imputed by its other colum values. We use MICE Algorithm, when we are sure that the missing values of our data is MAR.

**Advantage:**

→ Quite Accurate

**Disadvantage:**

→ As you are predicting missing values by using an algorithm, so it will be time consuming.

→ Training dataset has to keep on the server to predict missing values given by user input