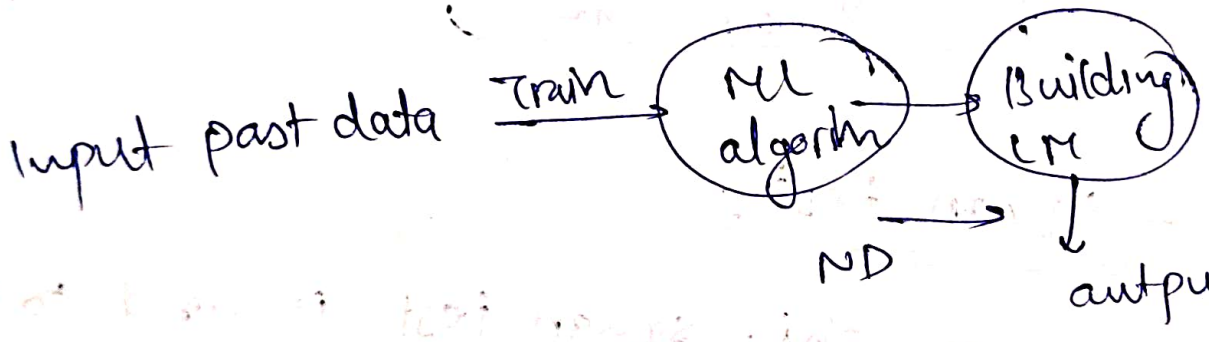
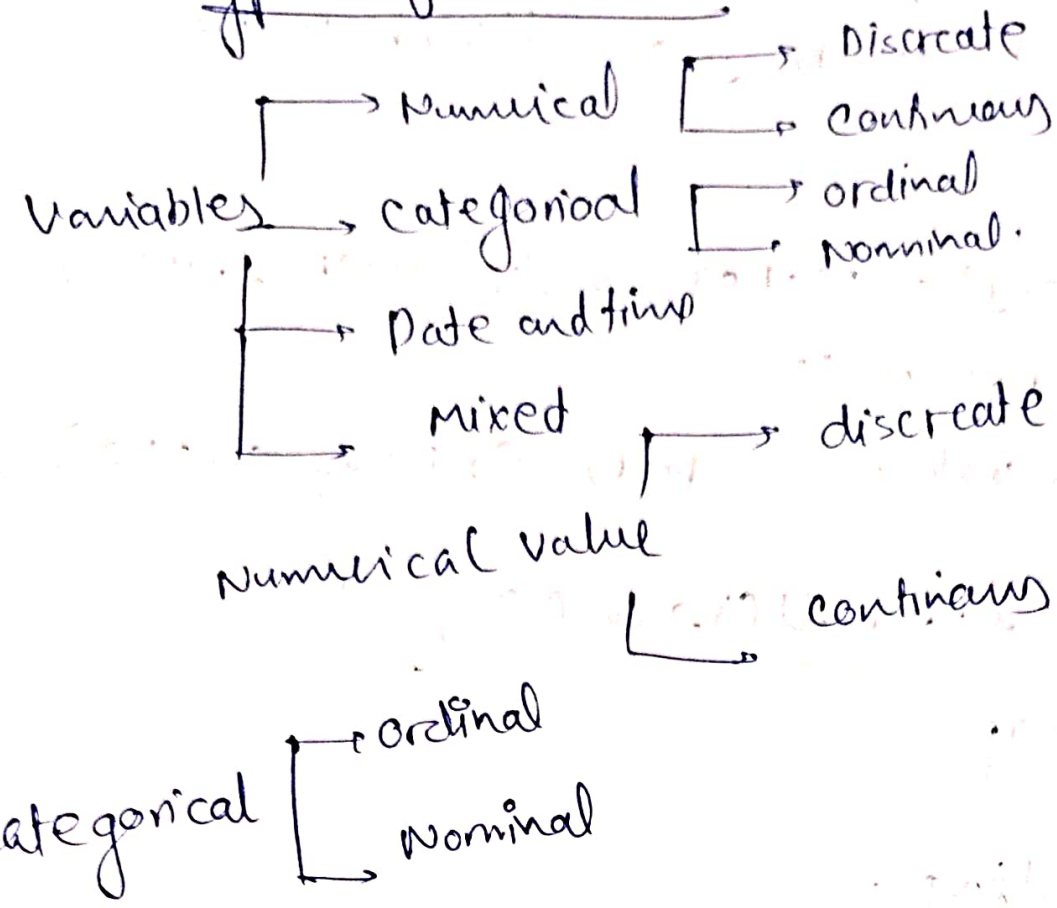
## Machine learning :-

getting comp learn and act like human do and improve their learning over time in autonomous fashion.




supervised  unsupervised  Reinforcement

## Types of variables




## Data cleaning :

Data cleaning is a process of preparing data for analysis / ML/DL by removal or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted.

## Missing value :

- mandatory and non-mandatory
- This data is stored as blank content

↳ Removal of Missing value is important beoz it donot work on blank value.

↳

dataset = pd.read_csv(r"Address.csv")

data.head(3)

Null value → NaN (Missing content)

dataset.isNull().sum()

↳ heatmap

## Encoding :-

one-hot encoding :- one-hot encoding is a technique used to convert categorical variables into a format that can be provided to machine learning algorithms. Each category of a variable is transform.

↳ one-hot encoding is used when there is no-ordinal relation ship.

## Dummy variables :

Dummy variables are a variation of one-hot encoding. when performing regression analysis, one of the categories in a categorical feature is left out to avoid the dummy variable trap.

Similar to one-hot encoding but removing one category to serve as baseline.

## Label encoding :

Label encoding is used to convert categorical text value to number. This is generally used for ordinal data, where there is a clear order or ranking.

Low = 1

Median = 2

High = 3

The problem arise when there is no ordinal relationship b/w categories, as Machine may interepret the no. as ordinal.

## Ordinal Encoding :-

Ordinal encoding is a technical used when there is a natural order b/w the categories. Each category is assigned an increase integer based on order.

To assign no. to categories, reflecting their order

## Outliers :-

Definition :- An outliers is a data points that is significant different from the rest of data. It can affect the model's performance and lead to skewed results

Why to remove outliers :- outliers can distort statistical measures like mean and standard deviation and affect algorithms like regression, which assume normal distribution

## outliers removal using IQR:

IQR (Interquartile Range): measure of variability, defined as the difference between 75th percentile and 25th percentile (Q3 - Q1)

IQR = Q3 - Q1

lower bound → Q1 - 1.5 * IQR

upper bound → Q3 + 1.5 * IQR

## outliers removal using z-score:

A statistical measurement that describe a value's relationship to the mean of a group of value. It measure how many SD a data point from the mean.

$$z = \frac{(x - \mu)}{\sigma}$$

→ data point with a z-score less than -3 or greater than +3 are consider outliers.

# Feature Scale (Normalization)

Normalization scale the data to a fixed range, typically [0,1] which is useful when the value need to constrained

$$X_{new} = \frac{(X - X_{min})}{X_{max} - X_{min}}$$

## Handling duplicate data :-

Definition :- Duplicate data refer to identical or almost identical row of data in a dataset, which can lead to bias in machine learning model.

## Replace and data type change :-

Replacing refers to the process of handling missing or erroneous data by replacing it.

# Feature selection techniques :-

Feature selection techniques are used to select important features that contributes the most to the output, improving model performance and reducing overfitting

## A) forward Elimination :-

Process :- start with no variables in model, and add on features at a time. After adding each feature, evaluate the Model's perfomance. stop adding variable when performance doesn't improve.

## B) Backward elimination :-

Process :- start with all the feature and it eratively remove the least significant feature, one at a time, based on some criteria. continue until no further features can be removed without harming the model's performance.