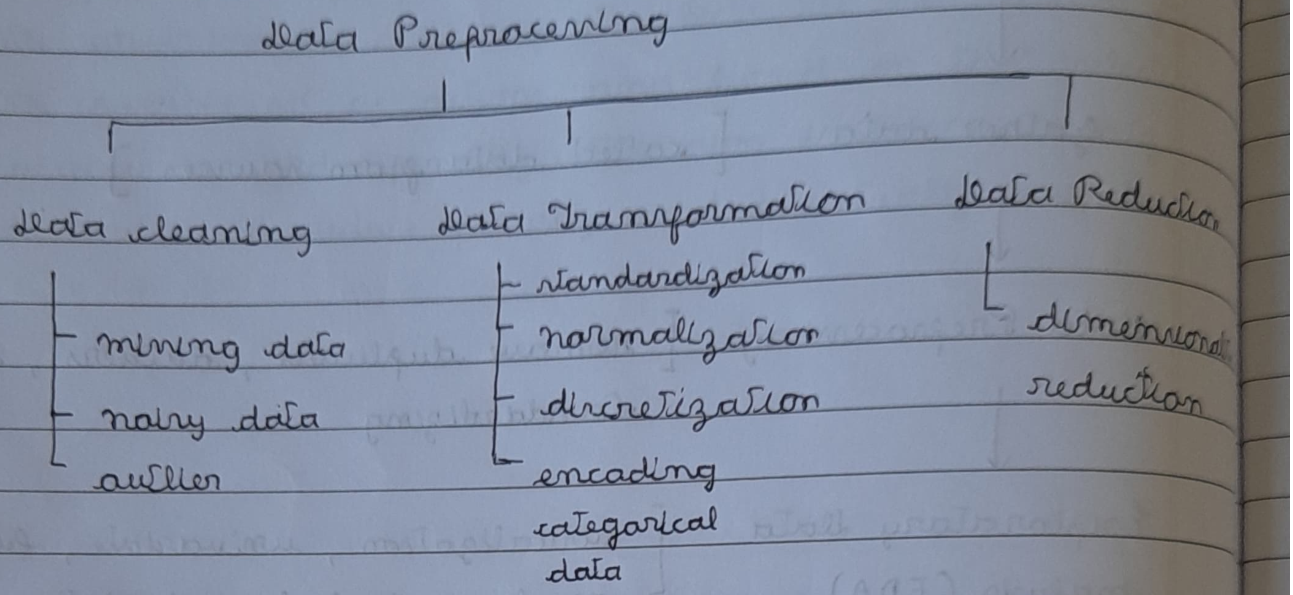
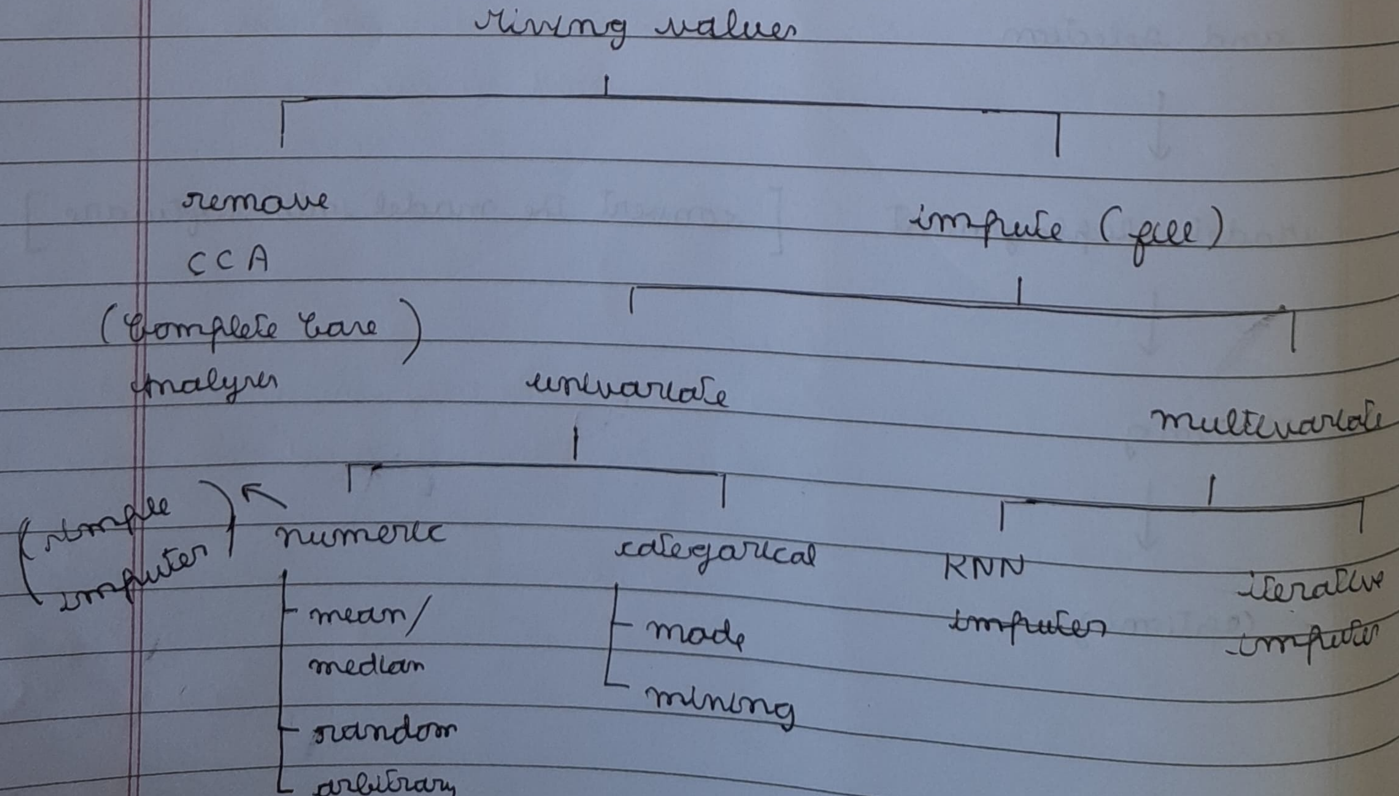
Data Preprocessing

Data preprocessing is used to transform raw data into a useful and efficient format.



Data Cleaning :-

Missing data :



Outlier :

An outlier is a data point that is noticeably different from the rest. They represent errors in measurement, bad data collection, etc.

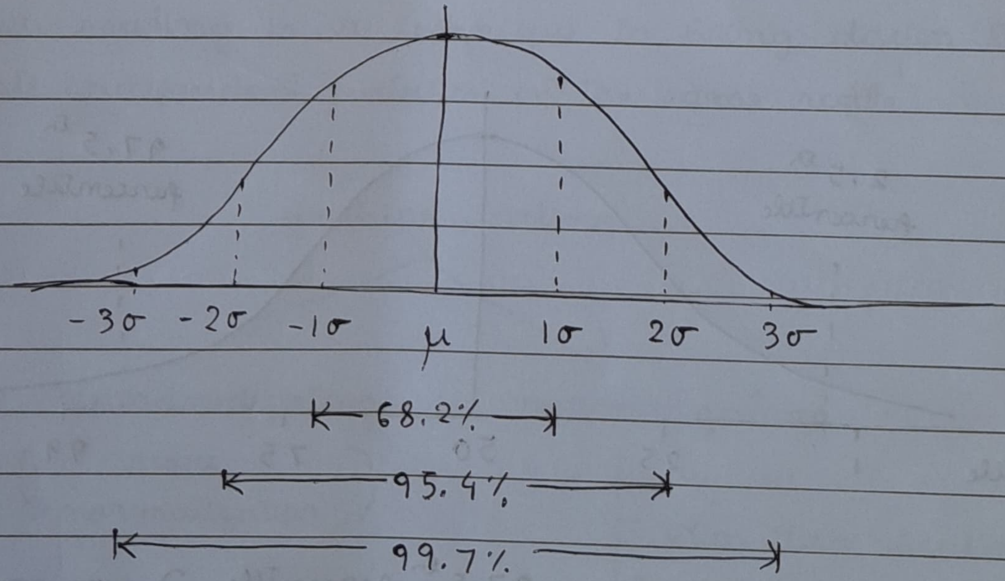
Outliers have major impact on these algorithms –

- Linear Regression
- Logistic "
- Adaboost
- Deep Learning

Techniques for outlier detection ~~and/removal~~ –

i) Z-score treatment –

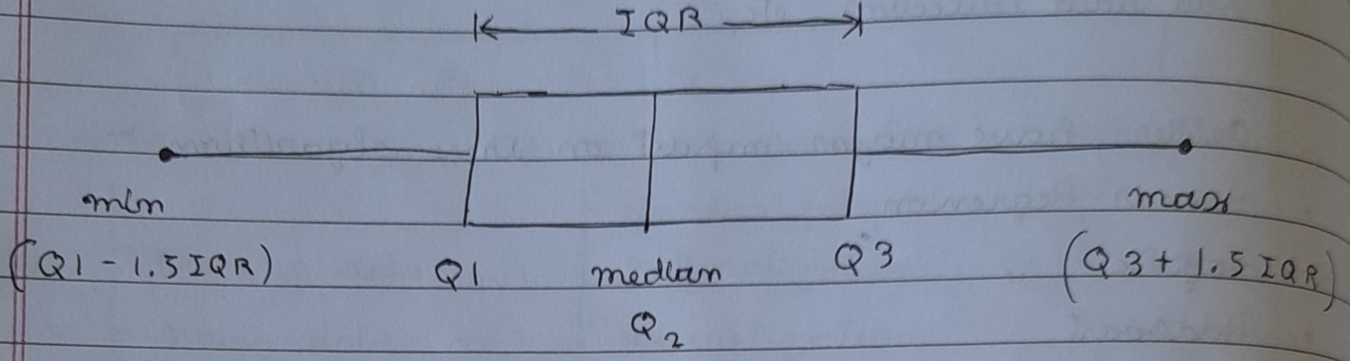
(Applied for normally distributed data)



Data points that are > (μ + 3σ)
< (μ − 3σ) } are considered outliers

ii) IQR-based filtering -

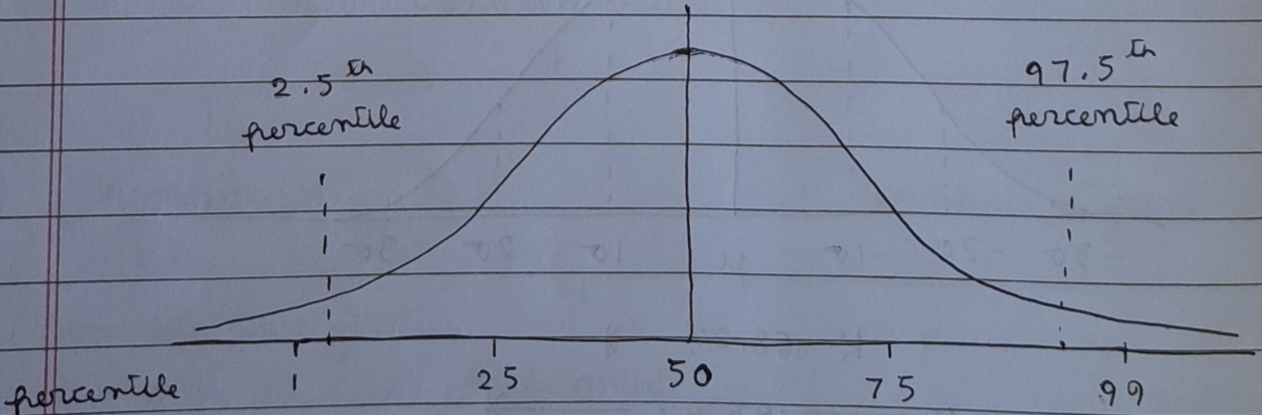
(Applied for skewed data)



data points that are > (Q3 + 1.5 IQR) } are considered outliers
< (Q1 − 1.5 IQR)

iii) Percentile method -

(Applied for other distributions)



data points that lie above 97.5th percentile } are considered outliers
below 2.5th percentile

In this method, we can decide any percentile threshold value.

Techniques for outlier treatment -

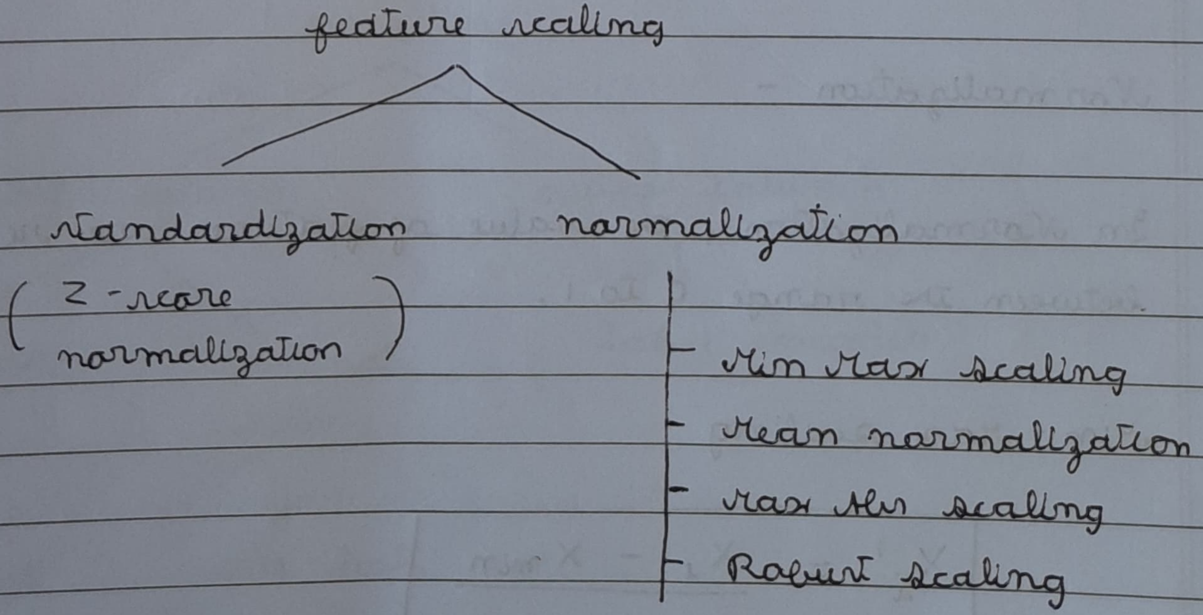
i) Trimming - Here the outlier rows are removed from the dataset

ii) Capping - The outliers below the lower limit are replaced with the lower limit value and the outliers above the upper limit are replaced with upper limit value.

Data Transformation :-

Feature Scaling :

Feature scaling is a technique to bring down the values of all independent features on the same scale.



i) Standardization -

Standardization makes the value of features have mean = 0 and standard deviation = 1

$$\boxed{x_i' = \frac{x_i - \bar{x}}{\sigma}}$$

x_i' → transformed value

$\bar{x}$ → mean

σ → standard deviation

Standardization is done for the following algorithm -

- K - means
- KNN
- Principal Component Analysis (PCA)
- Logistic Regression
- Gradient Descent
- Artificial Neural Network

ii) Normalization -

In Normalization, the value of features are rescaled between the range 0 to 1.

- Min - max scaling

$$\boxed{X_i' = \frac{X_i - X_{min}}{X_{max} - X_{min}}}$$

- Mean normalization

$$\boxed{x_i' = \frac{x_i - \bar{x}}{x_{max} - x_{min}}}$$

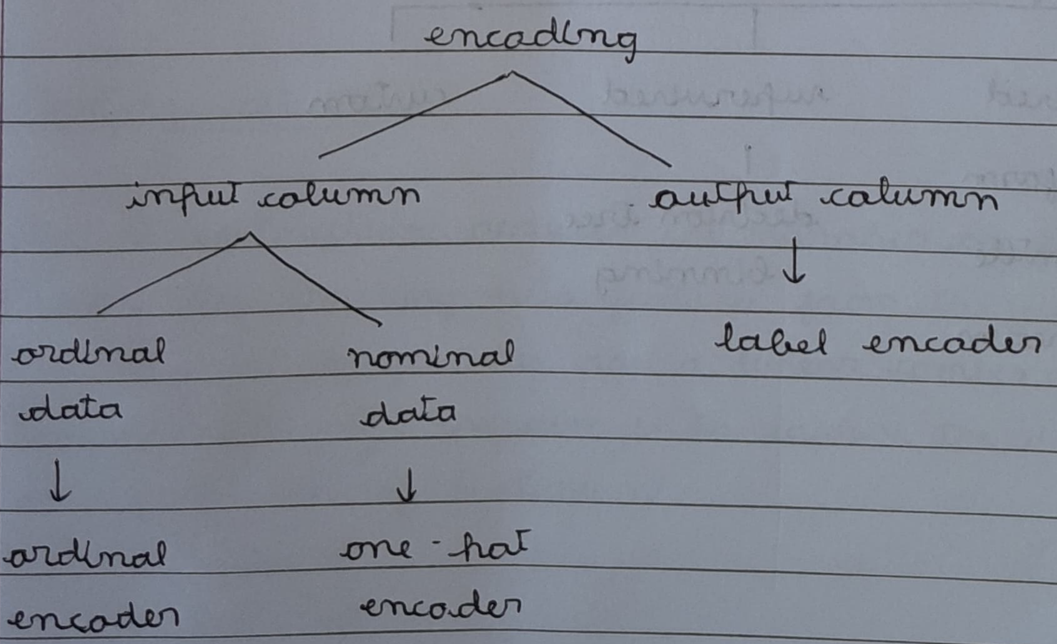
- Max - Absolute scaling (used in case of sparse data)

$$\boxed{x_i' = \frac{x_i}{|x_{max}|}}$$

- Robust scaling (used when data has outliers)

$$\boxed{x_i' = \frac{x_i - x_{median}}{IQR}}$$

Encoding categorical data :

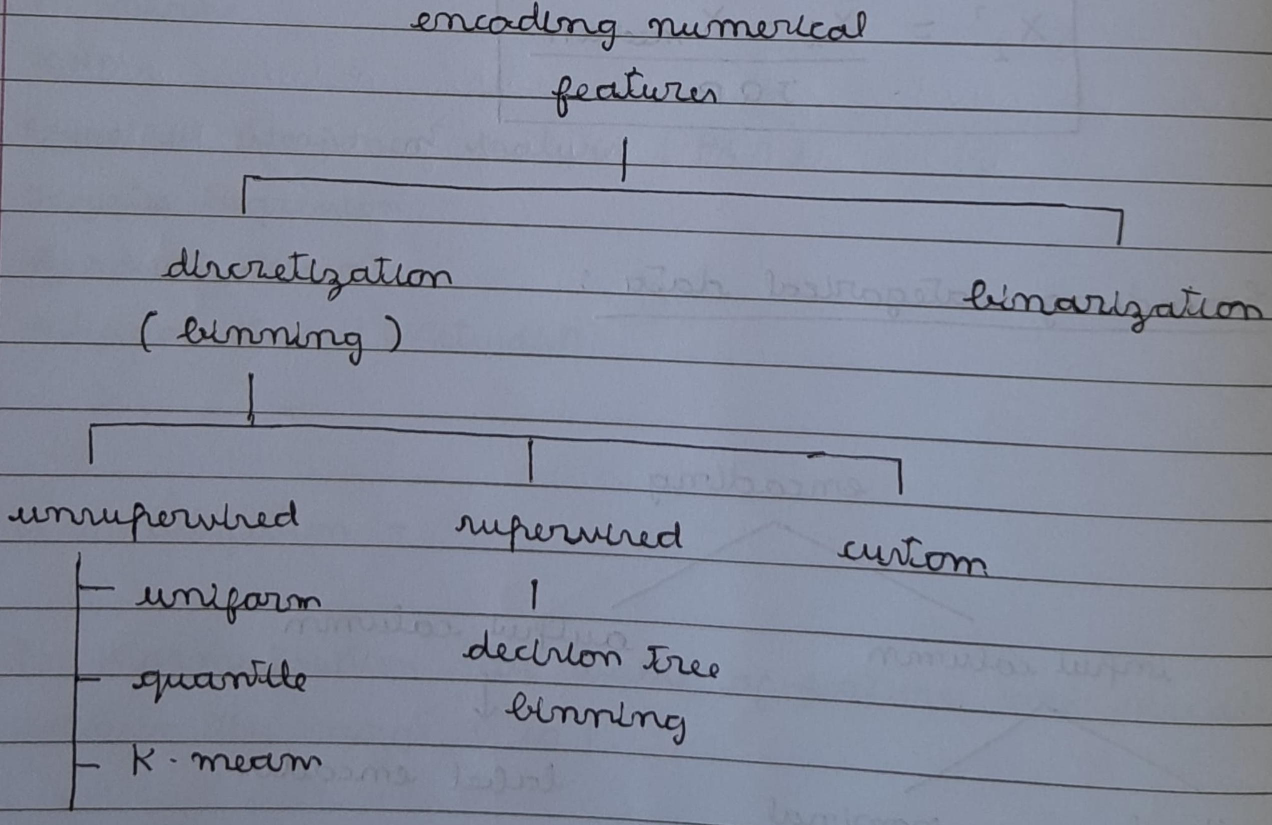


Encoding categorical data is the process of converting categorical data into integers so that it can be fed to the ML model.

Categorical data is of 2 types -
ordinal - there is a specific order
nominal - " " no " "

Encoding numerical features :

Encoding numerical features is the process of converting continuous numerical column to categorical columns.



i) discretization - is the process of converting continuous variables into categorical variables by creating a set of contiguous intervals that span the range of the variable values.

ii) Binarization - is the process of converting continuous variables into binary numbers

□ Data (dimensionality) reduction :-

dimensionality reduction —

- Feature extraction (before modelling)
  algo : PCA, ICA, LDA
- Feature selection (after modelling)
  algo : Pearson's correlation coeffcient, chi-squared test

Feature extraction :

Feature extraction reduces the number of features in a dataset by creating new features from the existing ones. The new set of features are a linear combination of the orginal features. The aim is to capture the data pattern with fewer no. of features.

LDA (Linear Discriminant Analysis)	PCA (Principal Component Analysis)
• used for supervised models.	• used for unsupervised models
• describes direction of maximum separability.	• describes direction of maximum variance
• requires class label info to fit()	• doesn't require class label info to fit().

Feature selection:

Feature selection reduces the no. of features by selecting ~~a sub~~ a subset of features from the dataset