Data Cleaning

Data Cleaning are critical step in preparing data for ML.

This process involves removing, correcting data that is inaccurate, incomplete, or irrelevant.

Handle Missing Values

Missing data can occur for various reasons, including human error or system failure. The handling of missing data is crucial, as it can affect the accuracy of the analysis results.

Handling Missing Data

- Numerical
  - Mean | Mean by Class
  - Median
  - Mode

- Categorical
  - Most Frequent

Mean vs. Median

- If there is Outliers → use Median
  (Median does not affected by outliers)

- If there no Outliers → use Mean
  (Mean does affected by outliers)

Outliers Detection

- Outliers are data points that are significantly different from other data points in the dataset. They can occur due to errors in data collection or measurement, or they can represent extreme values in the data.
- Outliers can significantly affect the analysis results, leading to incorrect conclusions and decisions. They can skew statistical measures such as the mean and standard deviation and affect the accuracy of models. Therefore, it's essential to identify and handle outliers in the dataset.

Outliers Detection Techniques (IQR)

Q1 = Q25

Q3 = Q75

IQR = Q3 - Q1

Upper bound = Q3 + (1.5 * IQR)

Lower bound = Q1 - (1.5 * IQR)

Handling Outliers

- Removing outliers from the dataset
- Transform the data using log
- Replace outliers with some representative values such as mean or median

Data Duplicate

- Data duplication occurs when the same data appears more than once in a dataset. It can lead to inaccurate analysis results, and it is essential to identify and remove duplicates

Handle Categorical Data

- Categorical data is non-numerical data that is often used to describe or categorize items. Handling categorical data involves converting it into numerical data that can be analyzed using statistical techniques. Techniques such as one-hot encoding and label encoding can be used to handle categorical data.

Handling Categorical Data Techniques

- One Hot Encoding / Dummy Encoding
- Label / Ordinal Encoding
- Count / Frequency Encoding

One Hot Encoding

Index	Animal
0	Dog
1	Cat
2	Sheep
3	Horse
4	Lion

One-Hot code

Index	Dog	Cat	Sheep	Lion	Horse
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	1
4	0	0	0	1	0

Ordinal Encoding

Original Encoding	Ordinal Encoding
Poor	1
Good	2
Very Good	3
Excellent	4

Count / Frequency Encoding

	Temperature	Color	Target	Temp_freq_encode
0	Hot	Red	1	0.4
1	Cold	Yellow	1	0.2
2	Very Hot	Blue	1	0.1
3	Warm	Blue	0	0.3
4	Hot	Red	1	0.4
5	Warm	Yellow	0	0.3
6	Warm	Red	1	0.3
7	Hot	Yellow	0	0.4
8	Hot	Yellow	1	0.4
9	Cold	Yellow	1	0.2

Data Preprocessing

Data preprocessing is an essential step in data analysis that involves preparing raw data for analysis. This presentation will cover two essential techniques in data preprocessing: data normalization and data transformation.

Data Normalization

Data normalization is a technique used to standardize the data values to a common scale. It helps to eliminate the effects of the differences in the unit of measurement, scale, and range of data. Normalized data is essential for accurate data analysis and modeling.

Data Normalization Methods

- Min-Max Normalization (Min Max Scaler)
- Z-score Normalization (Standard Scaler)

Min-Max Normalization

- Scale the data values between 0 and 1
- Formula:
  X_norm = (x-min(x)) / (max(x)-min(x))

Z-score Normalization

- scales the data values to have a mean of 0 and a standard deviation of 1

- Formula:
  X_norm = (x-mean(x)) / std(x)

Data Transformation

Data transformation is a technique used to modify the data distribution to meet the assumptions of a statistical test or to improve the performance of a model. It involves applying mathematical functions to the data values.

Data Transformation Techniques

- Square-root transformation
- Log transformation → avoid log(0)
- Exponential transformation

Let’s practice with some **CODE**