Om Shale
Group 7
2103163

THADOMAL SHAHANI
TSEC
ENGINEERING COLLEGE

Experiment No: 1

Aim: Select appropriate dataset and perform data preprocessing steps:
i) Imputation
ii) Anomaly detection
iii) Standardization
iv) Normalization
v) Encoding

Theory:

i) Imputation

⇒ Imputation is a process of replacing missing or incomplete data with substituted values.

- Missing data can arise from various issues such as data corruption or human error.
- Handling missing data correctly is crucial to avoid bias and inaccuracies in ML model.
- Common imputation techniques involves:

a) Mean/Median Imputation
⇒ Replace missing values with mean or median of the column. This is suitable for numeric data.

b) Forward/Backward Fill
⇒ Use the previous or next observation to fill missing values, which is useful for timeseries

ii) Anomaly detection

=> Anomaly detection is the process of identifying unusual patterns that do not conform to expected behavior, often referred to as outliers.

- Anomalies can indicate errors, rare events or important insights.

iii) Standardization

=> Standardization involve rescalling the feature so that they have a mean of 0 and a standard deviation of 1.

- This is essential for algorithms that assumes normally distributed data or that are sensitive to the scale of the features.

- The formula of standardization is:-

$$\text{Standardized Value} = \frac{x - \text{mean}(x)}{\text{std}(x)}$$

iv) Normalization

=> Normalization scales the data to a fixed range, typically [0, 1] or [-1, 1].

- It is especially useful for algorithms that requires bounded input values.

- The Min-Max Normalization formula is :

$$\text{Normalized value} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

v) Encoding

⇒ Encoding converts the categorical data into numerical format so that machine learning algorithm can process it.

- Common encoding techniques are :-

a) Label Encoding

⇒ Convert each unique category into a unique integer.

b) Ordinal Encoding

⇒ Used for categorical features that have a natural order, such as "low", "medium", "High".

- Each category is assigned an integer based on its order.

A

27/9/24