25-11-2024

Training Day – 49

November 25, Monday

- *Topic:* **Summary of Key Learnings**
  - Documented techniques learned over the past weeks.
  - Example: Listed best practices for data cleaning and visualization.

1. *Data Cleaning Techniques*

Handling Missing Data:

Imputation: Filling missing values using mean, median, or mode (for numerical data) or the most frequent value (for categorical data).

Removal: Dropping rows or columns with too many missing values.

Interpolation: For time series or sequential data, missing values can be interpolated based on surrounding data points.

Example:

```
df.fillna(df.mean(), inplace=True)    # Impute missing values with column mean
```

Data Transformation:

Normalization/Standardization: Scaling numeric data to a standard range, often required for machine learning models.

Log Transformation: Used to deal with skewed distributions by applying a logarithmic scale.

Categorical Encoding: Converting categorical variables into numeric formats using one-hot encoding or label encoding.

Example:

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
df['scaled_column'] = scaler.fit_transform(df[['column']])
```

Outlier Detection and Removal:

Z-Score Method: Identifying and removing data points that deviate significantly from the mean (e.g., z-scores greater than 3).

IQR Method: Removing data points outside the interquartile range (Q1 - 1.5 * IQR, Q3 + 1.5 * IQR).

Example:

```
from scipy import stats
df = df[(np.abs(stats.zscore(df['column'])) < 3)]    # Remove outliers based on Z-score
```

2. Combining Multiple Datasets

Concatenation: Combining datasets vertically (stacking rows) or horizontally (adding columns) using `concat()`.

Merging: Joining datasets based on common columns or indices using `merge()` (similar to SQL joins).