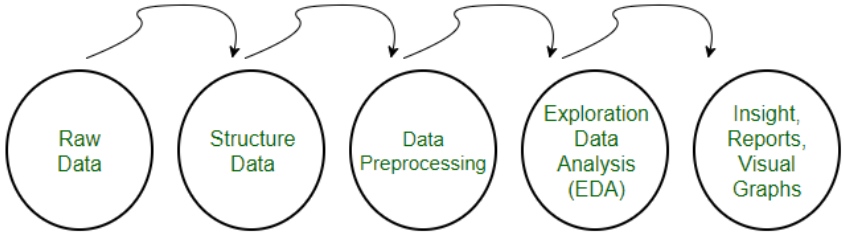
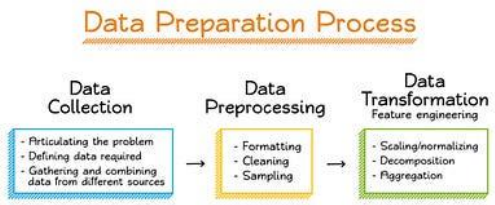
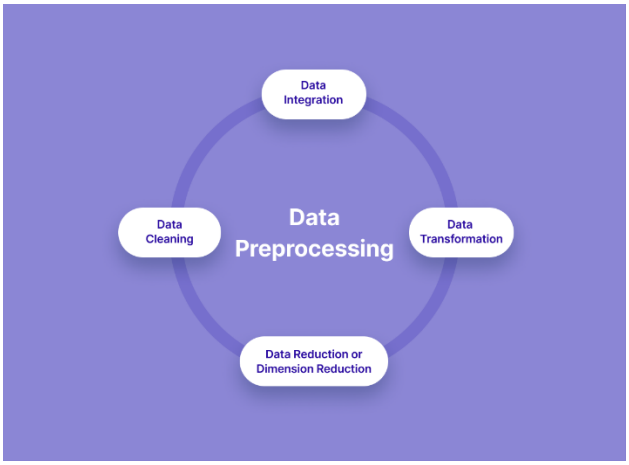
CHAPTER 4

DATA PREPROCESSING

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.







Need of Data Preprocessing

- For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner.
- Another aspect is that the data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithm are executed in one data set, and best out of them is chosen.

Clean data = easy to understand data

Why is data preprocessing important in data analysis and machine learning?

Preprocessing is essential for transforming raw data into a format that is suitable for input into machine learning algorithms. This may involve encoding categorical variables, scaling numerical features, and normalizing the data to ensure consistency and comparability.

How can missing values be handled in data preprocessing?

- **Deletion: This involves removing rows or columns with missing values.**
- **Imputation: This replaces missing values with estimates.**
    - **Mean/Median/Mode Imputation: Replace missing entries with the average (mean), middle value (median), or most frequent value (mode) of the corresponding column. This is a quick and easy approach, but it can introduce bias if the missing data is not randomly distributed.**
    - **K-Nearest Neighbors (KNN Imputation): This method finds the closest data points (neighbors) based on available features and uses their values to estimate the missing value. KNN is useful when you have a lot of data and the missing values are scattered.**
    - **Model-based Imputation: This involves creating a statistical model to predict the missing values based on other features in the data. This can be a powerful technique, but it requires more expertise and can be computationally expensive.pen_spark.**

What are some techniques for outlier detection and treatment?

Outlier can be of two types: Univariate and Multivariate.

- **Uni-varite outliers are outliers in an 1 dimensional space.**
- **Multi-variate outliers are outliers in an n-dimensional space.**

Most commonly used method to detect outliers is visualization.

1.We use various visualization methods, like Box-plot, Histogram, Scatter Plot

2.Use capping methods. Any value which out of range of 5th and 95th percentile can be considered as outlier

3.Data points, three or more standard deviation away from mean are considered outlier

4.Outlier detection is merely a special case of the examination of data for influential data points and it also depends on the business understanding

5.Bivariate and multivariate outliers are typically measured using either an index of influence or leverage, or distance..

Some of the most popular methods for outlier detection are:

- Z-Score or Extreme Value Analysis (parametric)
- Probabilistic and Statistical Modeling (parametric)
- Linear Regression Models (PCA, LMS)
- Proximity Based Models (non-parametric)
- Information Theory Models
- High Dimensional Outlier Detection Methods (high dimensional sparse data)

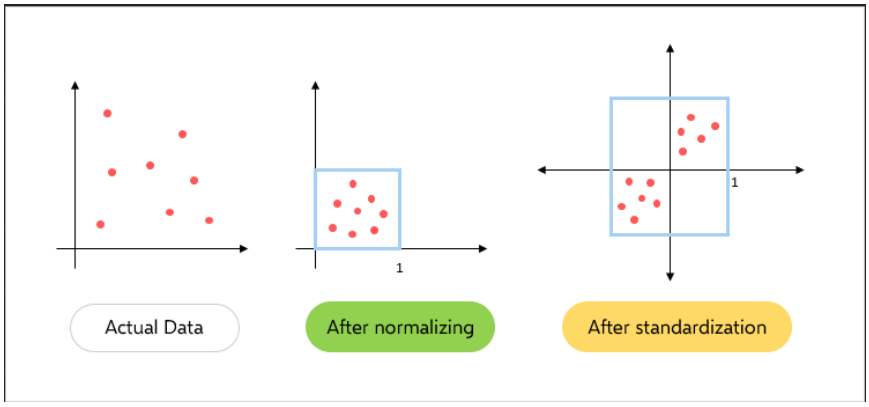
The common techniques used to deal with outliers:

- Deleting observations: We delete outlier values if it is due to data entry error,
- Transforming and binning values: Transforming variables can also eliminate outliers
- Imputing: Like imputation of missing values
- Treat separately: If there are significant number of outliers, we should treat them separately in the statistical model.

Explain the concept of feature scaling and its importance.

Feature scaling is the process of normalizing the range of features in a dataset.

Real-world datasets often contain features that are varying in degrees of magnitude, range, and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling.



- It improves the machine learning model's accuracy
- It enhances the interpretability of data by transforming features on a common scale, without scaling, it is difficult to make comparisons of two features because of scale difference
- It speeds up the convergence in optimization algorithms like gradient descent algorithms
- It reduces the computational resources required for training the model
- For better accuracy, it is essential for the algorithms that rely on distance measures, such as K-nearest neighbors (KNN) and Support Vector Machines (SVM), to be sensitive to feature scales