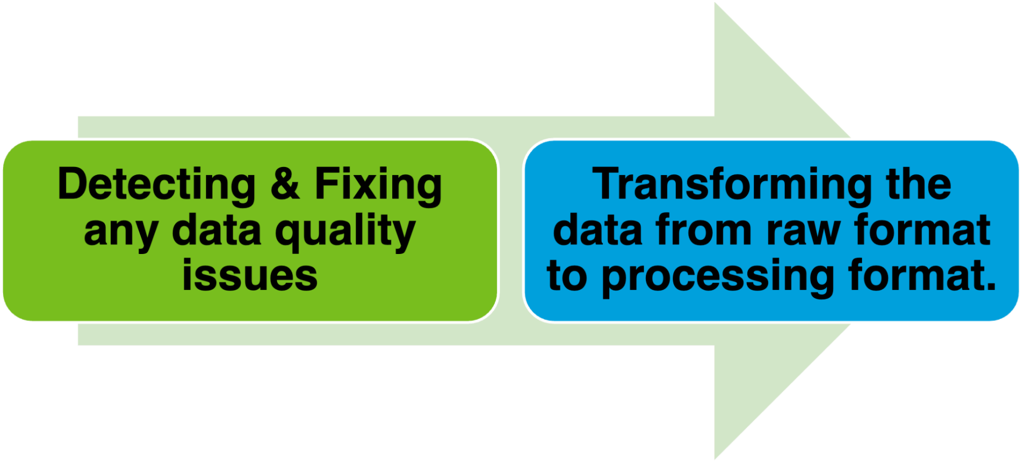
Assignment - 5

Data Preparation Steps

Harsh Siddhapura

1230169813

02/02/2024



Detecting & Fixing any data quality issues

Transforming the data from raw format to processing format.

Data preparation is a crucial phase in the data analysis pipeline, involving two main steps: detecting and fixing data quality issues, and transforming the data from its raw format to a format suitable for processing. Below are the details of each step and the importance of each:

Step - 1: Detecting & Fixing Data Quality Issues

1. **Handling Missing Values:** Identify and handle any missing values in the dataset. This can involve imputing missing values, removing rows/columns with missing data, or using more sophisticated methods depending on the context. Missing data can lead to biased or inaccurate analysis. Addressing missing values ensures a more complete and reliable dataset.

2. **Handling Duplicates:** Identify and remove any duplicate rows in the dataset. Duplicate entries can distort statistical analysis and lead to incorrect conclusions. Eliminating duplicates ensures data accuracy.

3. **Dealing with Outliers:** Detect and handle outliers using appropriate methods, such as truncation or transformation. Outliers can significantly impact statistical measures and machine learning models. Addressing outliers improves the robustness of analysis.

4. **Standardizing Data Formats:** Ensure consistency in data formats (e.g., date formats, categorical variables). Consistent formats facilitate easier analysis and prevent errors that may arise from varied data representations.

5. **Handling Inconsistent or Erroneous Data:** Identify and correct any inconsistencies or errors in the data, ensuring data accuracy. Inconsistent or erroneous data can lead to misleading results. Cleaning such issues enhances the reliability of the dataset.

Step - 2: Transforming Data

1. **Feature Engineering:** Create new features or transform existing features to better represent the underlying patterns in the data. Well-engineered features can improve the performance of machine learning models and provide better insights.
2. **Normalization/Scaling:** Standardize or normalize numerical features to a common scale. It ensures that all features contribute equally to the analysis, especially in algorithms sensitive to scale.
3. **Encoding Categorical Variables:** Convert categorical variables into numerical representations using techniques like one-hot encoding. Many machine learning algorithms require numerical input. Encoding ensures categorical variables can be used in these models.
4. **Data Aggregation:** Aggregate data at a higher level, such as summarizing daily data to monthly data. Aggregation can reveal higher-level trends and patterns, reducing data complexity.
5. **Handling Text and Time Data:** Preprocess and transform text data using techniques like tokenization and stemming. Handle time data appropriately, extracting meaningful features. Text and time data often require specialized preprocessing for effective analysis.

Importance

1. **Enhanced Analysis Reliability:** Clean and well-prepared data ensures that analysis and modeling results are more reliable and accurate.
2. **Improved Model Performance:** High-quality data, with appropriate transformations, leads to better-performing machine learning models.
3. **Time and Resource Efficiency:** Proper data preparation reduces the time and resources required for subsequent analysis and modeling.
4. **Facilitates Insights Discovery:** Well-prepared data makes it easier to uncover insights and patterns, contributing to better decision-making.