Data Pre-Processing

Hazim Fitri

Contents

Introduction

The process of improving or correcting the data is what is called data pre-processing

Common problem in a data =>

- Missing data
- Inconsistent data
- Too many attributes (variables) that are almost the same
- Outlier

Data Pre-Processing methods =>

- **Data Integration** : combine data from multiple sources, add new attributes, removing inappropriate attributes
  - Inconsistent attribute names
  - Inconsistent attribute values
- **Data Cleaning** : manage missing data, correct inconsistent data, manage outliers
- **Data Reduction** : reduce the dimension or the amount of the data
  - Dimension reduction
    - Principal Component Analysis (PCA)
    - Remove inappropriate vairables
  - Numerosity reduction
    - Parametric model (regression, log-linear, distribution model)
    - Non-parametric model (histogram, clustering, resampling)
- **Data Transformation** : scaling, discretizing, normalizing

Quality of data =>

- Accurate
- Complete
- Unique
- Consistent
- Timeliness
- Trusted
- Interpretable