SSE 2024 Data Analysis and Mining

Assignment 2: Data Pre-processing

Objectives

The objective of this assignment is to familiarize students with the basic steps of data preprocessing, including data cleaning, data integration, data transformation and data reduction. Through hands-on practice, students will deepen their understanding of the importance of data preprocessing and master the basic skills of data preprocessing.

Datasets

The dataset used in this assignment is the "Wine Quality" dataset from the UCI Machine Learning Repository. This dataset contains the physicochemical properties and quality ratings of wine, suitable for data preprocessing and subsequent analysis.
Dataset Link: Wine Quality.

Tasks

- **Data Cleaning**
  1. Missing Value Handling: Identify and deal with missing values in the data.
  2. Duplicate Data Handling: Identify and deal with duplicate values in the data.
- **Data Integration**
  1. Combine data with the same attributes from different sources. For this assignment, calculate the "total acidity," which is the sum of "fixed acidity" and "volatile acidity," and add it as a new column to the dataset.
- **Data Transformation**
  1. Normalization: Normalize the "quality" data to the [0,1] range.
  2. Discretization: Discretize the continuous attribute "fixed acidity" into three levels: "low," "medium," and "high."
- **Data Reduction:**
  1. Feature Selection: Use Analysis of Variance (ANOVA) to select the top three features that have the most significant impact on the quality rating of wine.

Submission Requirements

- **Submission Format:** Submit the assignment as a ZIP file named *studentID_name_hw2.zip*, where the name should be in Chinese. The file should include the following contents.
  1. **Code**: Submit a complete code file that implements the above tasks.
  2. **Report**: Submit a report briefly describing the data preprocessing steps you took, an analysis of the results of data preprocessing.
- **Submission Date:** 2024/3/30, 23:59:59.
- **Note:** Please submit to Canvas.

Reference Materials

- Pandas Documentation: For data processing Pandas Documentation
- Scikit-learn Documentation: Provides methods for data preprocessing, feature selection, dimensionality reduction, etc. Scikit-learn Documentation