Wine quality prediction and anomalies detection in the wine quality

Link: http://archive.ics.uci.edu/ml/datasets/Wine+Quality

Dataset provider – UCI machine learning repository

About organization and its product: The dataset deals with various features related to the chemical composition of a wine from Portugal. Vinho verde is a unique product from the Minho (northwest) region of Portugal. The drink has Medium in alcohol and it is particularly known for its freshness (specially in the summer).

Goals of the project:

- To try to predict the wine quality from the given data and hence, determine data's predictive strength to be able to predict wine quality
- Identifying and predicting anomalies
- Conducting a full data analysis on the data covering every expect of the data

Project timeline

A. *First chapter includes preprocessing, visualizing data and doing basically everything to get it ready for unsupervised learning*

1) **Importing Libraries**: Importing libraries like numpy, panadas, scipy etc
2) **Loading dataset and analyzing basics**: In the Starting, I will do basic tasks such as:-
    - Loading my dataset
    - Determining the size of dataset
    - Determining their variable types
    - Exploring the first few rows to get a glimpse of the data structure.
3) **Basic Statistics of the data (Univariate Statistics):** This part will deal with the basic univariate statistics. This will include determining things such as:-
    - Analyzing measures of central tendency (mean, median, mode) to understand the data's central values.
    - Evaluating dispersion through standard deviation and variance.
    - Visual univariate statistics will be performed to analyze the distribution of all the features by plotting their distributions.
    - For further strengthening the claims of distribution using Q-Quantile plots and Shapiro Wilk test.
4) **Real pre-pre-processing**: This part of my analysis will deal with some basic problems with the data such as:-
    - check missing values and dealing with missing values as per the distribution of the feature
    - Analysing the outliers for which methods like boxplot, z-score and IQR can be used and further & dealing with outliers as per the results.
    - Then, the standardisation or normalisation is done for further processing.

5) **Bivariate Analysis**: It refers to analysis of 2 features together. It includes:-

- Exploring correlations between pairs of features using scatter plots or heatmaps.
- Calculating correlation coefficients, considering Pearson, Spearman, or Kendall methods based on data characteristics.
- Visualizing bivariate relationships to identify potential patterns or trends.

B. *Second chapter revolves around dimensionality reduction*

Dimensionality reduction: Using various dimensionality reduction methods to curb the curse of dimensionality due to large number of features - feature selection and feature extraction.

- Under Feature Selection, certain features will be by selected such as selecting only one feature from all highly correlated features.
- Under Feature extraction, techniques like PCA will be used depending on the linearity of the data.

C. This chapter deals with the applying unsupervised methodologies to draw a proper inference from the data.

- Determine the number of clusters to use for clustering
- Applying clustering algorithms
- Validating clustering results – it can be achieved mainly in 2 ways:-
- Using various internal indices methods to evaluate the quality of clustering
- Use of external indices method to strongly evaluate the clustering algorithms

Name – Ajeesh garg

Course - IMAPP