

CSE523 Machine Learning

Weekly Report - 1

Section - 1

Submitted to faculty: Prof. Mehul Raval

Date of Submission: 11-02-2023

Roll No.	Name of the Student	Name of the Program
AU2040111	Kenil Shah	B. Tech CSE
AU2040215	Yesha Dhivar	B. Tech CSE
AU204087	Anshi Shah	B. Tech CSE
AU2040070	Rahi Shah	B. Tech CSE

2022-2023 (Winter Semester)

Tasks Completed

- 1. Going through the research paper
- 2. Strategies for the model training
- 3 Know and clean the data

Outcomes

Reference Paper:

https://www.sciencedirect.com/science/article/pii/S0167923609001377

Understandings from the paper: The case study was addressed by two regression tasks, where each wine type preference is modeled in a continuous scale, from 0 (very bad) to 10 (excellent). This approach preserves the order of the classes, allowing the evaluation of distinct accuracies, according to the degree of error tolerance (T) that is accepted.

Encouraging results were achieved, with the SVM model providing the best performances, outperforming the NN and MR techniques.

Based on the research about the data, models that can be used and we are planning to use are linear regression, logistic regression, decision tree, K-NN, Random Forest, SVM.

About the data:

Results from shaping the data:

```
#number or rows and columns in the dataset
print("Red Wine Dataset: ", red.shape)
print("White Wine Dataset: ", white.shape)

Red Wine Dataset: (1599, 12)
White Wine Dataset: (4898, 12)
```

There are 11 features and one quality column in both the red-wine and white-wine datasets.

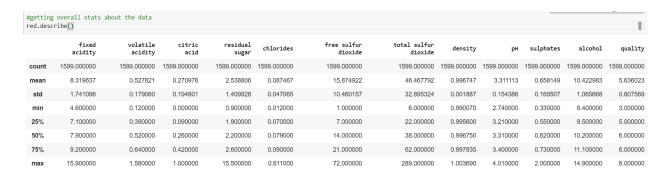
The features include information about fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, density, pH, sulphates, alcohol.

Information about the data:

```
#information about the dataset
red.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
# Column
                       Non-Null Count Dtype
0
    fixed acidity
                        1599 non-null
                                       float64
1
    volatile acidity
                        1599 non-null
                                      float64
2
    citric acid
                       1599 non-null float64
                       1599 non-null float64
3
   residual sugar
4
   chlorides
                       1599 non-null float64
5
   free sulfur dioxide 1599 non-null float64
   total sulfur dioxide 1599 non-null float64
                        1599 non-null float64
                        1599 non-null float64
8
    рН
                       1599 non-null
    sulphates
                                      float64
Q.
10 alcohol
                                      float64
                        1599 non-null
11 quality
                        1599 non-null
                                       int64
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
```

Along with this, we also checked null values in the data, and there were 0 null values found in both of the datasets.

We also got the overall statistics of the data, that helps us getting more sense and knowledge about the spread of the data.



For the upcoming week:

- 1. Data analysis and exploration
- 2. Correlation matrix and heat maps
- 3. More data visualization