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
 3   residual sugar        1599 non-null   float64
 4   chlorides             1599 non-null   float64
 5   free sulfur dioxide   1599 non-null   float64
 6   total sulfur dioxide  1599 non-null   float64
 7   density               1599 non-null   float64
 8   pH                    1599 non-null   float64
 9   sulphates             1599 non-null   float64
 10  alcohol               1599 non-null   float64
 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.

```
#getting overall stats about the data
red.describe()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922	46.467792	0.996747	3.311113	0.658149	10.422983	5.636023
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157	32.895324	0.001887	0.154386	0.169507	1.065668	0.807569
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000	3.000000
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	0.995600	3.210000	0.550000	9.500000	5.000000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996750	3.310000	0.620000	10.200000	6.000000
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	0.997835	3.400000	0.730000	11.100000	6.000000
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690	4.010000	2.000000	14.900000	8.000000

For the upcoming week:

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