CS 454
Introduction to Machine Learning and Artificial Neural Networks

Prof. Dr. Ethem ALPAYDIN
2022 Spring

Project Progress Report
Wine Quality Prediction

Team D
Onur Alaçam – S014958
Tuna Tuncer – S018474
Uygar Kaya – S015570

May 9, 2022

1. What have we done so far?

*We downloaded the **Wine Quality** dataset from the Kaggle. The dataset includes 6497 rows and 13 columns. After pulling the raw data with the help of Pandas, we made the Exploratory Data Analysis - EDA, after examining the raw data in this analysis, we visualize the data with the help of the Matplotlib library. After making EDA analysis, we put the data into pre-process operations, firstly we Imputed the missing values by taking their mean, then applied the One-Hot Encoding method to convert the categorical data to numerical data and we performed some pre-processing steps to apply the classification method. After these steps, we converted the data into two separate csv files for both regression and classification.*

So far, we implemented the Linear Regression and Logistic Regression algorithms using Scikit-Learn Library. 4 different DataFrame were created for both linear regression and logistic regression algorithms. In df1, we use all the data that we pre-process. In df2, we use 6 different dependent variables data with the highest correlation with the independent variable. In df3, we use 6 different dependent variables data with the lowest correlation with the independent variable. In df4, we use 6 different dependent variables data that have one highest and one lowest correlation with the independent variable. Currently, in Linear Regression and Logistic Regression, the best result is df1, where we use all Dataframe. While we got a 29% score value, 55% MSE, and 57% MAE value in Linear Regression, we reached a 93% Accuracy value in Logistic Regression.

When we did a literature search, we realized that the result obtained in logistic regression gave higher accuracy than the result obtained in the article. According to the article [1], accuracy is reported as 76%.

2. What are we planning to do next?

After pre-processing the raw data, we are planning the implementing Multilayer Perceptron and K-Nearest Neighbor algorithms by using the different csv files we obtained for regression and classification. The predictions for each algorithm will then be analyzed to see how similar to the facts they are while computing their individual confusion matrices and accuracies. The collected data will be compared to recent studies in the field.

We plan to stick to our Project Proposal, but the linear regression model results we obtained did not satisfy us, so we are considering using an outlier in the data pre-processing.

In addition, we plan to divide the data into train, test, and validation sets in the algorithms we will implement, so we plan not to drop duplicate values since the number of rows in the data we have is low, but when we drop duplicate values and try in the algorithms we implement, we reach a score value of 31% with a 2% increase in linear regression, while we reach a % accuracy value 91% in logistic regression with a 2% decrease. Therefore, we are undecided as to whether we should generate a new dataset, so we plan to meet to TA.

3. References to Related Work

[1] *Wine quality detection through machine learning algorithms. IEEE Xplore. (n.d.). Retrieved May 7, 2022, from https://ieeexplore.ieee.org/abstract/document/9009111*

4. Appendix

You can find the source codes written in this project from the links below.

1. *https://github.com/UygarKAYA/WineQualityPrediction*