ESE 417 Final Project Classification of Wine Quality by Machine Learning

1st Haoyu Quan
Mckelvey School of Engineering
Washington University in St. Louis
St. Louis, MO
quanhaoyu@wustl.edu

2nd Anny Qiao
Mckelvey School of Engineering
Washington University in St. Louis
St. Louis, MO
a.qiao@wustl.edu

2nd Bruce Li
Mckelvey School of Engineering
Washington University in St. Louis
St. Louis, MO
liyifei@wustl.edu

I. INTRODUCTION

Machine learning is a branch of artificial intelligence that enables machines to learn from data and make predictions or decisions without being explicitly programmed. It finds applications in a wide range of domains such as image recognition, natural language processing, and predictive modeling. In this project, we aim to use supervised machine learning algorithms to classify a wine quality data set based on 11 given features. While wine was once viewed as a luxury good, it is now enjoyed by a wider range of consumers. Quality evaluation is a crucial part of the certification process and can help to identify influential factors during the wine production process [citation 1]. In our project, these factors will be used as features for our machine learning algorithm.

The primary objective of this project is to build a machine learning model that can accurately predict the quality of wine into 10 different levels with the given data. Furthermore, we aim to compare the performance of different machine learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Random Forest to determine the most effective algorithm for this task. To improve the accuracy of our model, we have implemented various methods of data cleaning, chosen optimal weight factors, and tuned other hyperparameters.

After multiple testing and trial and error iterations, the model was able to achieve an accuracy rate of over 60 percent. Overall, this project demonstrates the effectiveness of supervised machine learning algorithms in predicting the quality of wine based on various influential factors during the production process.

II. METHODS

A. Exploratory Data Analysis and Cleaning

B. Grid Search and Weighted Features

C. SVM

Support Vector Machines (SVM) is a supervised learning algorithm that is used for classification tasks. It constructs a hyperplane or a set of hyperplanes in a high-dimensional space, which can be used for classification. The goal of SVM is to maximize the margin between the classes, which results in a more robust classifier.