**Wine Quality:**

**A Regression and Machine Learning Proposal**

BY

Christopher Benjamin Scott II

Final Project for CSC 433 Machine Learning

Bachelor of Science in Computer Science

Department of Computer Science

Mississippi College

April, 2022

***Proposal:***

Often times the quality of a product is determined by a few very definable features, such as that of a pencil. The list of features that determines such an object is slim and easily agreed upon; however, there are other things in life that are far more complex to measure the quality of. In this proposal of research, I am to tackle one of these complex items: wine.

My project seeks to answer the question of what makes quality wine. In the wine industry, several features are notable in the measurement of the score of wine, these features are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, and alcohol. This brings the total number of attributes to eleven plus the output (quality). Unfortunately, data points that would establish additional features such as the grapes used are private information that was not made available in the data-set according to the authors, so it is important to note that although the features discussed are of note, other features may play a more significant role. The quality of the wines in question was all determined by wine experts and the data set, which was made public for research use, comes from the University of Minho. It is important to note that the output is based upon “sensory data” from a scale of 0 (awful) to 10 (magnificent) as noted in the data-set notations (Cortez et al.). The wine selected is all from the “Vinho Verde” wine. My project is to determine which of the input variables are the most relevant to the quality of a wine. The benefits of my project are it will be of use in the determination of what factors winemakers should focus upon to establish a finer quality product for their customers and their brand. I will aim to achieve this with the use of regression modeling using Random Forest Regression, Multiple Linear Regression, and a Support Vector Regression.

Ultimately, my aim is to prove features of notable importance for the quality of wine, through the use of regression modeling and feature importance techniques.

## *Works Cited*

"API Reference". *Scikit-Learn*, 2022, [https://scikit-learn.org/stable/modules/classes.html#module-](https://scikit-learn.org/stable/modules/classes.html" \l "module-)sklearn.svm.

"Feature Importances With A Forest Of Trees". *Scikit-Learn*, 2022, https://scikit-learn.org/stable/auto\_examples/ensemble/plot\_forest\_importances.html.

"Feature Scaling | Standardization Vs Normalization". *Analytics Vidhya*, 2022, <https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning->normalization-standardization/.

"1.4. Support Vector Machines". *Scikit-Learn*, 2022, https://scikit-learn.org/stable/modules/svm.html.

Brownlee, Jason. "How To Calculate Feature Importance With Python". *Machine Learning Mastery*, 2022, <https://machinelearningmastery.com/calculate-feature-importance-with->python/.

Cortez, Paulo et al. "Modeling Wine Preferences By Data Mining From Physicochemical Properties". *Decision Support Systems*, vol 47, no. 4, 2009, pp. 547-553. *Elsevier BV*, https://doi.org/10.1016/j.dss.2009.05.016.

Gupta, Kaustubh. "Plotting Visualizations Out Of Pandas Dataframes - Analytics Vidhya". *Analytics Vidhya*, 2021, <https://www.analyticsvidhya.com/blog/2021/06/plotting-> visualizations-out-of-pandas-dataframes/.