Title: Clustering Techniques to Determine Origins of Wine

By: Mikey Joyce

Tasks:

In this project, my aim is to utilize unsupervised learning techniques to determine which wines are similar to each other in hopes that the clusters determined separate the different cultivars that created the wine. This is an important topic for machine learning to be able to process because this application could be applied to anthropology researchers. If an anthropology researcher has a dataset of the chemical residues left in old wine containers, they could use the technology I am proposing to train the unsupervised algorithms on and hopefully be able to determine which wines are similar which could then be used by their expert knowledge to determine which civilization the wine container belonged to.

Data:

The dataset I am going to utilize is a wine dataset for clustering found on Kaggle (https://www.kaggle.com/datasets/harrywang/wine-dataset-for-clustering). This dataset is a 13 feature dataset that has many features relating to the chemical makeup of the wine. Some of the features of the wine relate to the color, but I may end up dropping these features from the final feature vector because my aim would be to make a system that is useful for anthropologists/archaeologists and since the wine would already be dried up when the researchers get their hands on it, they would not have access to the color features.

Approach:

The first thing I will do will be exploratory data analysis, since it is best practice to get familiar with the data before trying to manipulate it. Since this dataset is somewhat high dimensional feature vector, I will most likely opt for a dimensionality reduction technique such as t-SNE. It is expected this won't give the most crisp of clusters, so my last step will be to implement the final clustering algorithm. The algorithm I am wanting to utilize is a possibilistic clustering algorithm from the following paper:

- https://ieeexplore.ieee.org/abstract/document/227387

Timeline:

One week for each step. EDA 1 week, dimensionality reduction 1 week, final algorithm 1 week, and lastly crafting the report and presentation 1 week. This should take 4 weeks in total.