Arianna Brisco

CYBR 304

Written Analysis

May 15, 2025

i. Data Information

The data used for the project comes from the classic wine tasting dataset from the UCI Machine Learning Repository. The dataset was retrieved from the UC Irvine Machine Learning Repository website (https://archive.ics.uci.edu/dataset/109/wine). A description of the dataset from the website states "These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines."

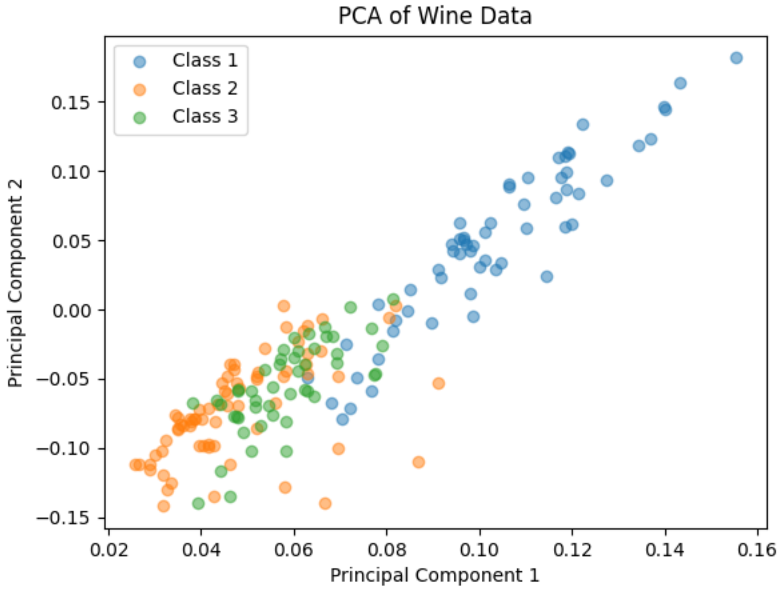
The dataset has 178 instances and 13 features, plus one class label. The wines are in one of three classes and the features are all regarding the chemical makeup of the wine. The features are as follows: 1) Alcohol 2) Malic acid 3) Ash 4) Alcalinity of ash 5) Magnesium 6) Total phenols 7) Flavanoids 8) Nonflavanoid phenols 9) Proanthocyanins 10) Color intensity 11) Hue 12) OD280/OD315 of diluted wines 13) Proline.

1. Question

Can the wine dataset be reduced from 13 features to 2 features and be able to be clustered to clearly differentiate the three classes form one another using a custom PCA function?

2. Analysis

The result of the analysis is shown in the graph below. As the graph demonstrates, there is some clustering shown for class one, but still not a full clean clustering with minimal outliers, and there is nearly full integration of classes two and three. This does however accomplish the endeavor of simplifying the data and making it easier to model and be understood at a glance. While the analysis shows the dataset cannot be properly analysized with this method, it acts a quick way to get a bit of understanding and insight into the model before performing a more complex machine learning analysis. Additionally, the results from the PCA grant insight into what models may be useful for accurately analyzing the data.



3. Answer

No, the wine dataset cannot be reduced to two features and be differentiable. Whilst class one is a distinct cluster, classes two and three and cannot be easily separated.

4. Model Used

The model used for this research is the Principal Component Analysis, or PCA, model. PCA is a model used to reduce the dimensionality of data whilst keeping the most important features of the data intact. PCA can remove correlated data that are redundant in the analysis process. By performing principal component analysis, the most important features are maintained and called principal components. Additionally, PCA allows for easier visualization of data by reducing the dimensions from a higher number to a lower number such as two or three, which are much more easily visualized by the human brain. This is important as it leads to greater and easier understanding. PCA can be used in tandem with other analysis models and is commonly a way to preprocess big data. Additionally, the model was used due to its prior existence from the unit one summative program assignment. This allowed for building upon previous work and deeper learning.