IST 707 Applied Machine Learning

By Prof. Kelvin King

Assignment 4

**Clustering**

Submitted by:

Samarth Sandesh Mengji  
SUID: 718473878

NetID: [smengji@syr.edu](mailto:smengji@syr.edu)

Date of Submission: 10/1/2023

Syracuse University – School of Information Studies

Part 1: Table of Contents

1. Introduction
2. Methodology
   1. Data Preprocessing
3. Results and Analysis

3.1. PCA

3.2. K-Means Clustering

3.3. Hierarchical Agglomerative Clustering (HAC)

1. Conclusion

Part 2: Table of Contents

1. Data Preprocessing
2. K-means and HAC for Age and Income
3. K-means and HAC for Age and Children
4. K-means and HAC for Income and Children
5. K-means and HAC using all features using PCA

Part 3: Table of Contents

1. Introduction
2. Data Preprocessing
3. Association Rule Discovery
4. Introduction

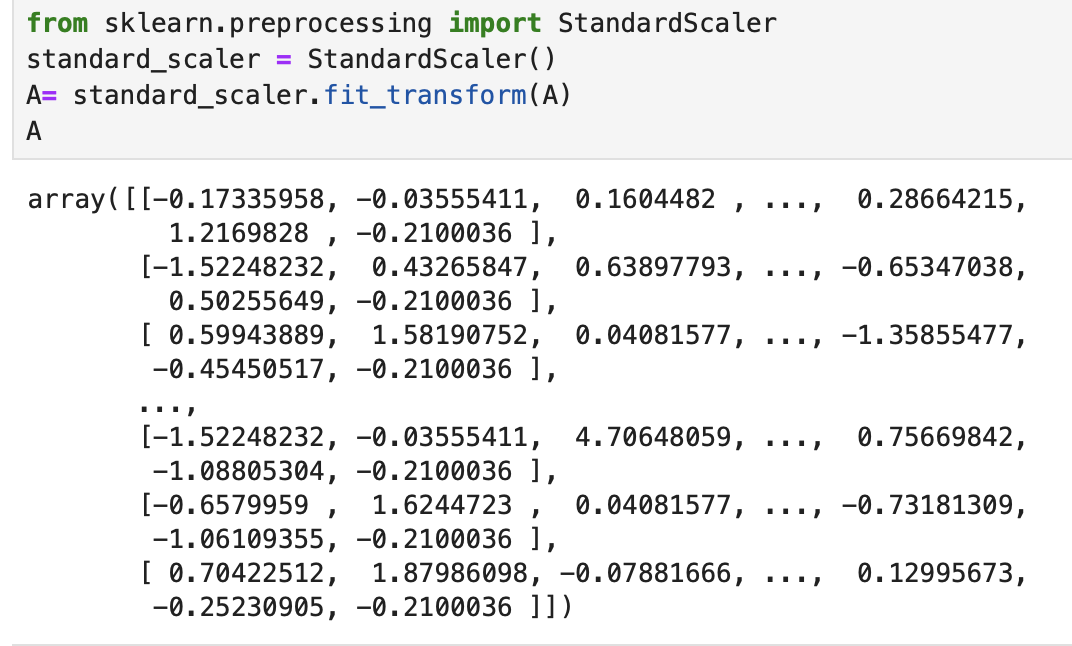
In this report, our objective is to determine the most likely authors of the disputed Federalist papers, which were published anonymously. There are 85 essays out which 11 essays have disputed authorship.

We will apply clustering methods – K-means and HAC based on the frequency distribution of function words in the essays.

1. Methodology
   1. Data Preprocessing

The dataset provided had no missing values. We dropped two columns ‘author’ as it was the target feature and ‘filename’ as it had no use in our analysis. Then we transformed our data in a standard form by using standard scaler to prepare the data for PCA.

These were the result of data after standard scaler:



1. Results and Analysis

3.1. PCA

A chart with colorful dots

Description automatically generated

From the above PCA visualization, we can see that different authors have been assigned different colors and disputed articles which don’t have any authors are in blue color. Each data point represents an article, and the colors tell us which author it belongs to. Based on the visualization from PCA, we can see that most of the articles fall in the Madison cluster. We can conclude that most articles fall under Madison with a couple outliers which can be joint authorship.

3.2. K-means clustering

A graph with many dots

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

From the above visualizations and K-means cluster results, we can conclude say that most of the disputed articles belong to the Madison cluster 1.

3.3 Hierarchical Agglomerative Clustering (HAC)

A screenshot of a computer screen

Description automatically generated

From the above dendrogram, we can again conclude that most of the disputed articles belong to Madison cluster.

1. Conclusion

We performed PCA, K-means and HAC clustering algorithms and concluded that 10/11 articles belong to Madison.

Part 3:

1. Introduction

In this part of the assignment, we are performing association rule mining on the bank dataset to determine the top 5 association rules about people who have high chances of buying PEP. This report provides insights which will help our client to identify the right audience and potential PEP buyers to perform further action.

1. Data Preprocessing

There were two things which were done to clean the data. The field ‘id’ was removed as it had no meaningful information. After that, numeric fields were discretized to convert them into nominal attributes.

Bins were created from continuous numeric data. This was done for ‘age’, ‘income’ and ‘children’. 4 bins for age, 5 for income and 2 for children.

1. Association Rules

A screenshot of a computer code

Description automatically generated

These were the 5 rules found in our analysis:

**Rule 1**

Support: 0.055

Confidence: 0.702

Lift: 1.538

**Explanation:**

People of the age 50 and above with children have a 70.2% chance of buying PEP. This can be aligned with real world understanding that old people seek more financial security for their family

**Recommendation:** Target Marketing to people in this category. Marketing of financial security for them and their children.

**Rule 2**

Support: 0.058

Confidence: 0.660

Lift: 1.446

**Explanation:** Males with income in the range of 40-60k have a 66.03% chance of buying a PEP. People in these categories have a hunger to achieve their financial goals and looking to invest.

**Recommendation:** Company should target theses people in such a way that it will help them achieve their financial goals. Target marketing and personalized advertisement should be done.

**Rule 3**

Support: 0.058

Confidence: 0.660

Lift: 1.446

**Explanation**: Males with income in the range of 40-60k who also have a savings account have a 66.0% chance of buying a PEP. People in these categories display that they might have natural tendency to save, and PEP can be one way of it.

**Recommendation**: Company should target theses people in such a way that it will help them with their savings strategy or a new way to save. Target marketing and personalized advertisement should be done.

**Rule 4**

Support: 0.057

Confidence: 0.642

Lift: 1.405

**Explanation**: People with income 40-60k and own a car have a 64.2% chance of buying a PEP. Owing car signifies a level of financial stability

**Recommendation**: Company can create marketing campaigns to people in these categories justifying PEP is a smart choice since they many financial responsibilities including a car.

**Rule 5**

Support: 0.057

Confidence: 0.642

Lift: 1.405

**Explanation**: People with income range in 40-60k, owns a car and has a savings account have 64.2% chance of buying a car. These people already have some financial knowledge and discipline.

**Recommendation**: PEP can be useful to help them with their goals and strategy to invest.

Part 2:

1. Data Preprocessing

Before clustering, we conducted essential preprocessing steps:

1. Handling missing values. None found.
2. Drop id as it was not informative
3. One-hot encoding of categorical variables.
4. Data standardization to ensure variables have the same scale.
5. PCA to plot all features
6. K-means and HAC for Age and Income

A comparison of a number of clusters

Description automatically generated with medium confidence

From the above graph, we can see that it is hard to extract any insights. We will further explore more variables for better understanding.

1. K-means and HAC for Age and Children

A screenshot of a graph

Description automatically generated

From the above graph, we can see that it is hard to extract any insights. We will further explore more variables for better understanding.

1. K-means and HAC for Income and Children

A comparison of a graph

Description automatically generated with medium confidence

From the above graph, we can see that it is hard to extract any insights. We will further explore more variables for better understanding.

1. K-means and HAC with all features using PCA

A comparison of colored dots

Description automatically generated with medium confidence

From the above graph, we can see that there 3 categories of people. Cluster 0 and 2 of K-means and HAC have high variance suggesting that people from this cluster are most likely to buy PEP while cluster 0 and 2 of k-means and HAC are least likely to buy PEP. Further analysis on this is needed to understand the people in these clusters. These graphs are not enough to address the business questions.

Reference:

https://seaborn.pydata.org/generated/seaborn.boxplot.html https://seaborn.pydata.org/generated/seaborn.countplot.html https://seaborn.pydata.org/generated/seaborn.histplot.html https://matplotlib.org/stable/api/\_as\_gen/matplotlib.pyplot.pie.html https://matplotlib.org/stable/api/\_as\_gen/matplotlib.pyplot.subplots.html https://conx.readthedocs.io/en/latest/MNIST.html https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch\_openml.html https://builtin.com/data-science/step-step-explanation-principal-component- analysis#:~:text=Step%201%3A%20Standardization&text=So%2C%20transforming%20the%20d ata%20to,transformed%20to%20the%20same%20scale. https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html https://matplotlib.org/stable/api/\_as\_gen/matplotlib.pyplot.colorbar.html https://pandas.pydata.org/docs/reference/api/pandas.get\_dummies.html https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html https://stackoverflow.com/questions/32857029/python-scikit-learn-pca-explained-variance- ratio-cutoff

<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html>

<https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html>

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.dendrogram.html>

I certify that this assignment represents my work. I have not used any unauthorized or unacknowledged assistance or sources in completing it, including free or commercial systems or services offered on the internet.