Seeds Exploratory Data Analysis and Classification

Christina Chang

***Abstract* — The purpose of this report is to understand the seeds dataset through visualizations and compare different classification models to see which one performs best on the seeds data. Principal component analysis revealed that most of the variability in the data can be explained by two principal components. Clustering on the data showed that agglomerative hierarchical clustering and k-means clustering have similar results. After comparing the five fold cross validation scores of eight classification algorithms, Multi-Layer Perceptron (MLP), a class of feed forward artificial neural networks, yielded the best results. The final five fold cross validation score of MLP was 91%.**

I. Introduction

The seeds dataset is from the UCI Machine Learning Repository, and it contains 210 observations of three varieties of wheat. The three varieties of wheat are Kama, Rosa, and Canadian. There are 70 observations of each type of wheat. Each observation has measurements of seven geometrical properties: area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient, and length of kernel groove. All of the features are continuous variables.

The motivation of this project is to explore the seeds dataset through figures, dimensionality reduction techniques, and clustering algorithms. Furthermore, different classification models were used to classify the wheat kernel type based on the geometrical properties given in the dataset in order to determine which model achieves the best results. These results may be useful to those who want to solve similar classification problems.

II. Exploratory Data Analysis

A. Dimensionality Reduction using Principal Component Analysis

Principal component analysis (PCA), a dimensionality reduction technique, was performed on the dataset to extract relevant information from high dimensional data [1]. Initially, five principal components were used to represent the data. Figure 1 is a heatmap of the loadings for each of the five principal components. The first principal component has similar loading weights for all features except for the asymmetry coefficient. The second principal component has a large positive weight for the compactness feature and a large negative weight for the asymmetry coefficient feature.

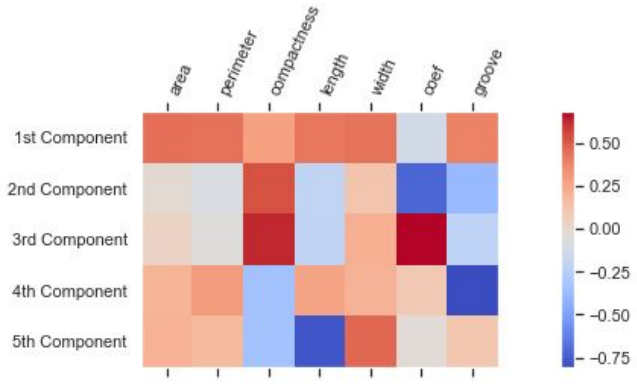

Figure 1: Heatmap of loadings when using five principal components.

Figure 2 is a scree plot which displays the amount of variation each principal component captures from the data [1]. The scree plot shows that the first two principal components are able to explain about 90% of the variation. Therefore, two principal components are sufficient to describe the data.

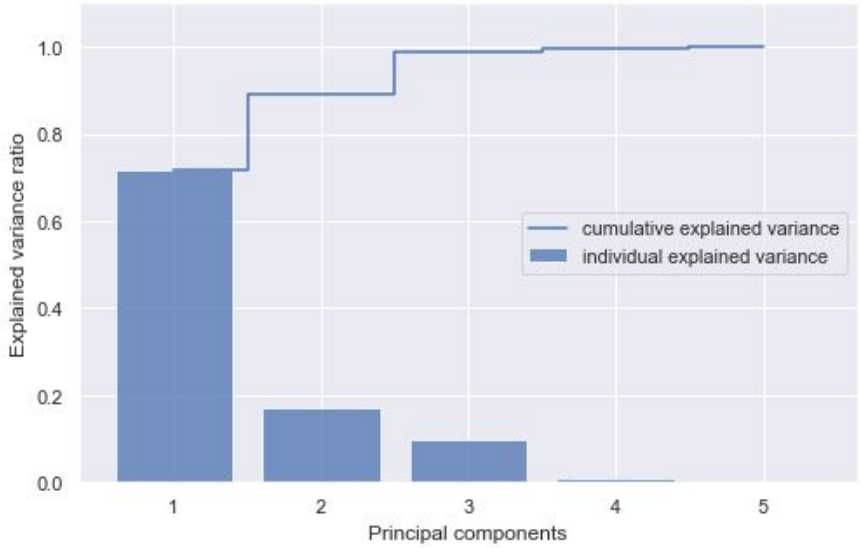

Figure 2: Scree plot using five principal components.

PCA was performed on the dataset using the optimum number of principal components determined from the scree plot (two principal components). Figure 3 shows a scatter plot of the observations using the two principal components and the points are colored using the true labels. Each wheat kernel type is fairly well separated on the plot.

Figure 3: PCA plot using two principal components.

B. Dimensionality Reduction using t-SNE

Besides PCA, t-Distributed Stochastic Neighbor Embedding (t-SNE) is another technique for dimensionality reduction. t-SNE attempts to find a two-dimensional mapping based on probabilistic distributions and it is oftentimes a better technique for visualizations [2]. Figure 4 is the scatterplot of the

data using t-SNE and the points are colored by the true labels. The classes are well separated and there is less overlap between the groups when compared to PCA.

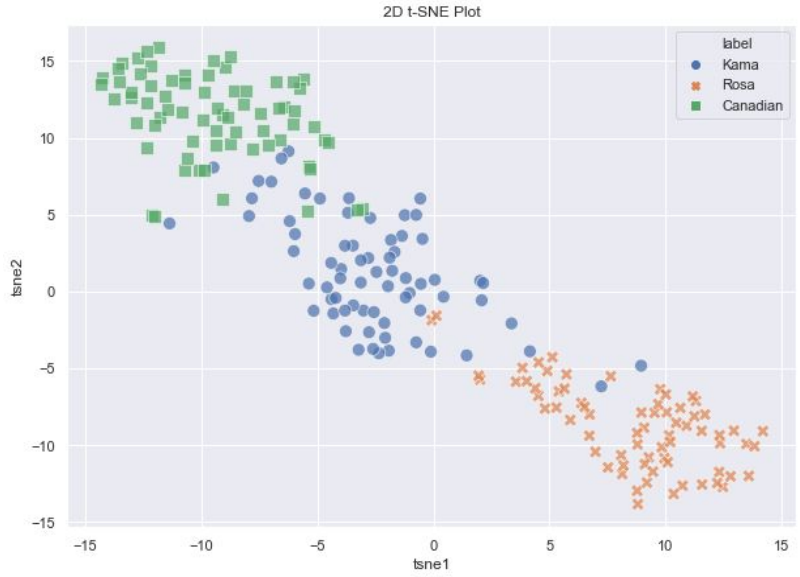

Figure 4: t-SNE plot of the data.

C. *Clustering*

Clustering is a form of unsupervised learning that can be used to explore the groupings in the features space of the data [3]. The results from hierarchical clustering and k-means clustering will be compared.

Agglomerative hierarchical clustering is a method of clustering where each observation starts as their own cluster and clusters are combined until there is a single cluster [3]. Figure 5 is a dendrogram which shows each cluster and heights reflect the distance or dissimilarity between clusters. The dendrogram illustrates the three clusters as three branches. The break between the two red clusters occurs at a smaller vertical distance when compared to the break between the red and green cluster. This suggests the green cluster is quite different from the two red clusters.

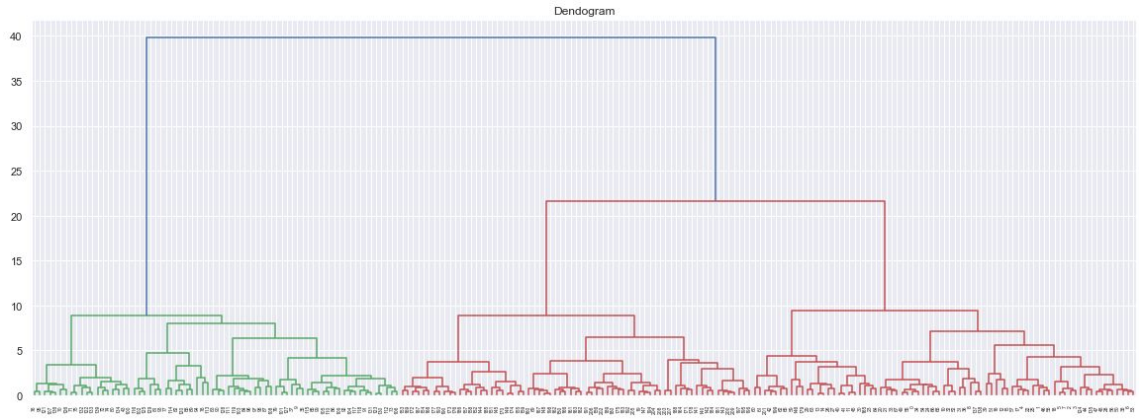

Figure 5: Dendrogram of agglomerative hierarchical clustering.

Figure 6 is a clustermap of the data. A clustermap is a heatmap with dendrograms of the hierarchically clustered observations and the features. Each column was standardized by subtracting the minimum and

dividing by the maximum. The ‘centroid’ method was used for clustering the data. There are some areas on the plot with the same color, like the dark areas and the light areas of the plot. This suggests that the group of observations is correlated with the corresponding group of columns.

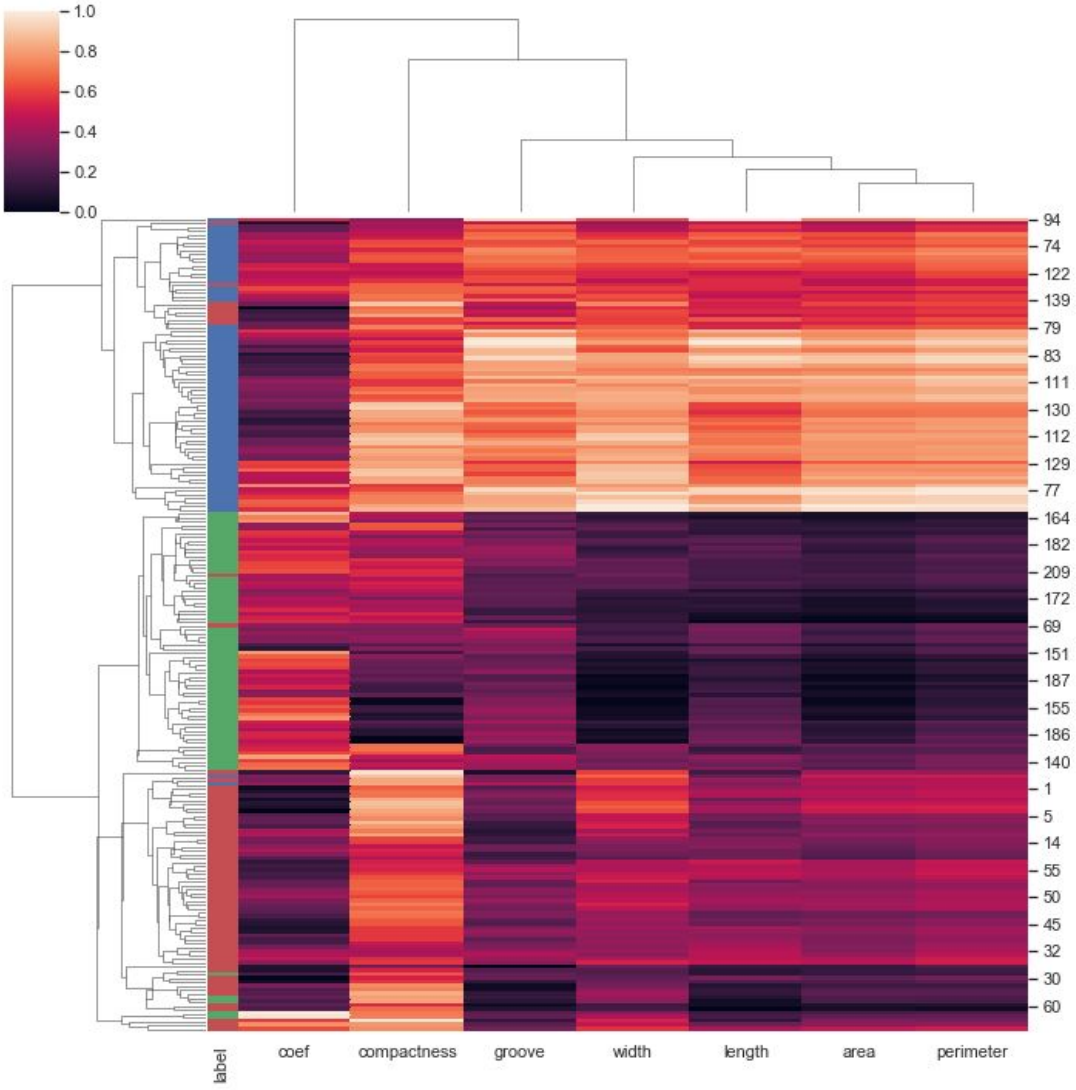

Figure 6: Clustermap of seeds data.

Additionally, k-means clustering can be used to explore groupings in the data. This algorithm searches for a predetermined number of clusters within a dataset by utilizing the arithmetic means of clusters and also distances [3]. K-means clustering was performed with K assigned to three.

Figure 7 is the clustering results from agglomerative hierarchical clustering and k-means clustering. The color represents the cluster that the observation was assigned to and the shape represents the true label of the data point. The hierarchical clustering and k-means clustering yield similar results on this dataset. Also, for both clustering algorithms, each cluster is dominated by a single wheat kernel type. This implies that these two clustering algorithms were able to find clusters whose centers are recognizable wheat kernels, even though the algorithms were not given any information on the true labels.

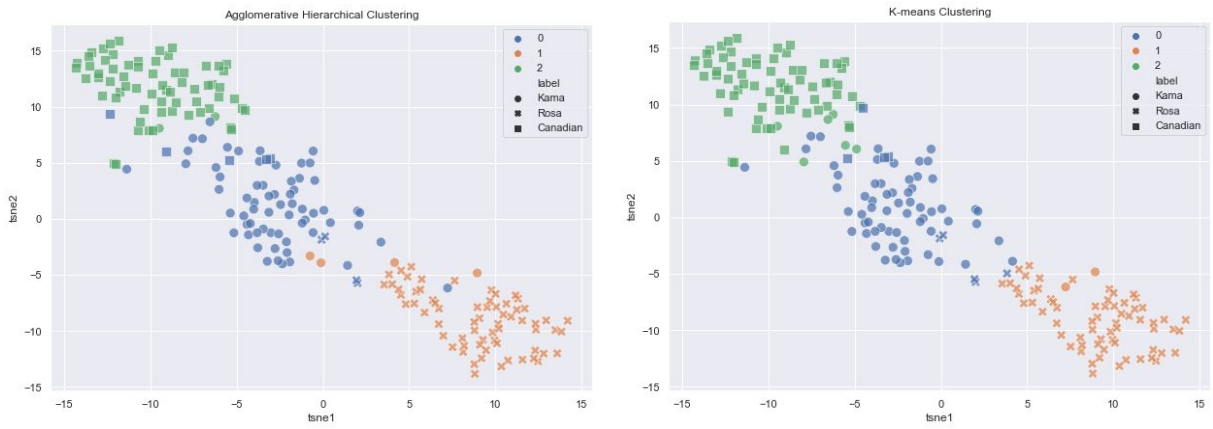

Figure 7: Results from agglomerative hierarchical clustering and k-means clustering.

III. Modeling

There are many types of classification models that can be used to classify data. A comparison between the performance of several classifiers on this dataset is useful when trying to understand which classifier is best for a particular dataset. The models were used to assign the wheat kernel type (Kama, Rosa, or Canadian) based on seven attributes: area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient, and length of kernel groove. The classification algorithms used were Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Decision Tree, Random Forest, AdaBoost, XGBoost, and Multi-Layer Perceptron.

For data preprocessing, the features were scaled because some classification methods are distance based. The scaling method used was standardization.

The data was split into testing and training sets, with a test size of 30%. Stratified sampling was used to ensure that the train and test sets have approximately the same number of classes.

Table 1 has the mean and variance of the five fold cross validated accuracy scores for each classification model.

Model	**Mean 5 Fold CV Score**	**Variance 5 Fold CV Score**
Logistic Regression	0.90	0.09
K-Nearest Neighbor	0.88	0.09
Support Vector Machine	0.89	0.11
Decision Tree	0.90	0.03
Random Forest	0.88	0.12
AdaBoost Classifier	0.73	0.15
XGBoost Classifier	0.90	0.06
Multi-Layer Perceptron	0.91	0.08

Table 1: Mean and variance of five fold cross validated scores.

Based on the results from Table 1, Multi-Layer Perceptron (MLP) performs slightly better than the rest of the classifiers with a mean cross validation score of 91%. Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Decision Tree, Random Forest, and XGBoost had similar performances between 88% to 90%. The AdaBoost classifier performed the worst out of all models, with a five fold cross validation score of 73%.

IV. Conclusion

Methods such as PCA and t-SNE are helpful for visualizing datasets with many features. However, the t-SNE plot seemed to group the observations into their respective wheat kernel class better than the PCA plot. Additionally, the results from clustering the data reveal that unsupervised learning methods have the potential to extract information and group the data into reasonable clusters.

The classification modeling results reveal that MLP, a class of feedforward artificial neural networks, achieved the highest accuracy among all compared classification algorithms. The worst accuracy was from the AdaBoost classifier. Note that the results in the report may have been affected by the suboptimal choice of parameters, since the parameters were not tuned in each model. Thus, in the context of this dataset, it seems that deep learning algorithms have better accuracy when compared to other models.

References

[1] G. Kanyongo, "Determining The Correct Number Of Components To Extract From A Principal Components Analysis: A Monte Carlo Study Of The Accuracy Of The Scree Plot," *Journal of Modern Applied Statistical Methods*, vol. 4, iss. 1, article 13, May 2005.

[2] L. Matten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, November 2008.

[3] M. Kaushik and B. Mathur, "Comparative Study of K-Means and Hierarchical Clustering Techniques," *International Journal of Software & Hardware Research in Engineering*, vol. 2, iss. 6, June 2014.