Homework 2
Problem 2
Ayad Masud
733009045

I certify that I have personally done the coding, generated the figures and written the report without aid from anybody else, and that I have not plagiarized, self-plagiarized, or used AI-generated text. I certify that I have acknowledged any sources I used to complete this assignment. ARM.

1 Part 1: PCA

Figure 1 shows the resultant principal components after running the PCA function from scikit learn on the dataset provided. The first principal component shows the direction with most variance and the second principal components show the second highest direction of variance. The subsequent principal components will show less and less variance. We can see some class separability along the first principal component. Class 3 is well separated from the other classes. Class 2 is somewhat separated and class 1 and 4 have some overlap. We could broadly say that classes 1, 2 and 4 are more like each other than they are similar to class 3. Class 3 is labeled as beef which is protein. The other classes: vegetables, breakfast cereals, and baked goods are not protein food items.

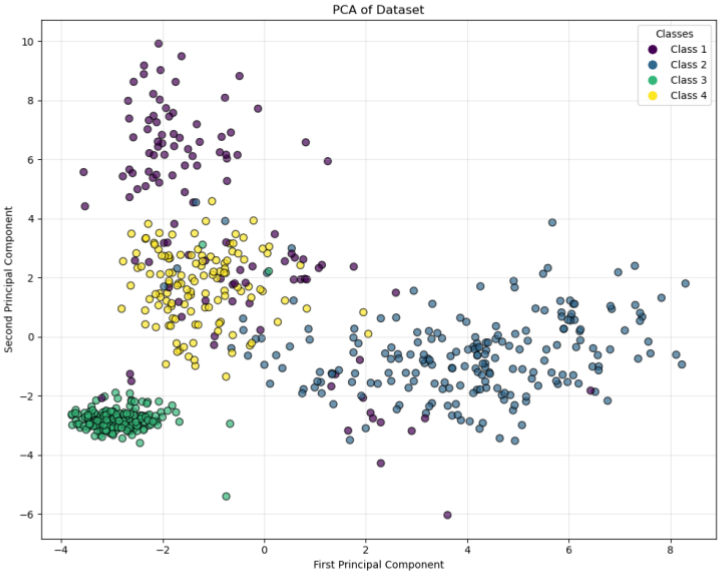

Figure 1. First and second principal components computed after using PCA on dataset.

2 Part 2: PCA Scree Plot

Figure 2 shows the scree plot generated from the above principal component analysis. After computing where the cumulative variance and the 95% variance threshold intersect, a total of 26 principal components are needed to capture 95% of the variance. The scree plot suggests low collinearity because we need many principal components to explain majority of the variance. If majority of the variance was explained by just one or two principal components we would have high collinearity, but in this case, we need 26 principal components.

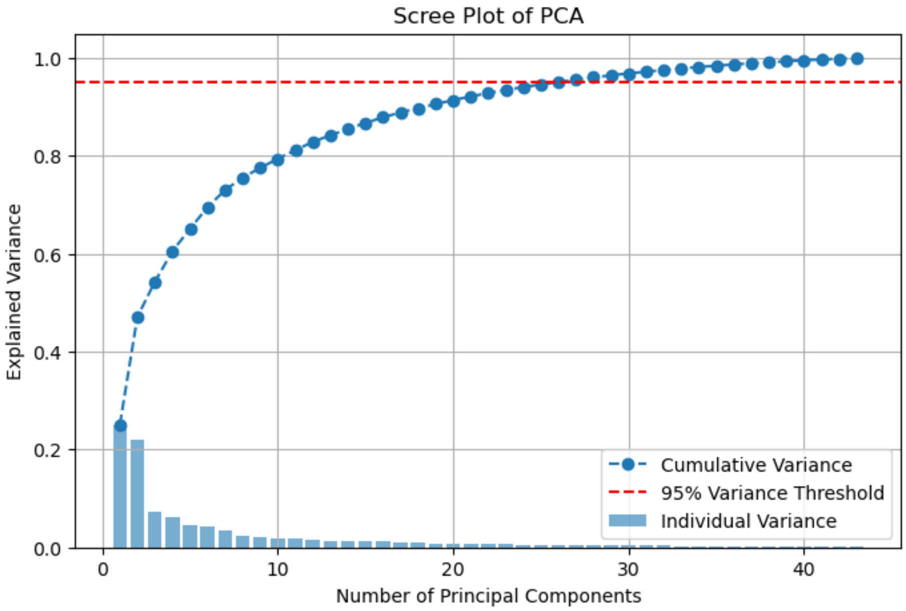

Figure 2. Scree plot of PCA on dataset

3 Part 3: Fisher's LDA

Figure 3 shows the Fisher's LDA projection of the training data. We can see separation between classes 2 and 3 from classes 1 and 4. Classes 1 and 4, however, is not separated and have significant overlap. Since classes 1 and 4 overlap, LDA is not completely effective, and a different dimension reduction technique may be needed. Figure 4 shows the LDA projection of the validation data. We can see a similar structure as the projection of the training data. Because we have similar structures after projecting both the training and validation data, we can say LDA for this dataset generalizes quite well.

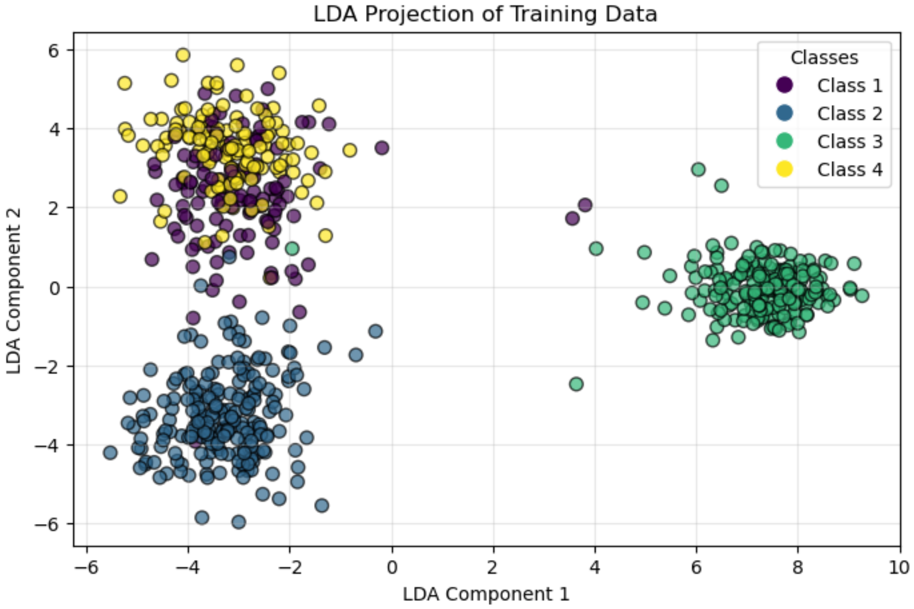

Figure 3. Fisher's LDA projection of training data with color coded labels

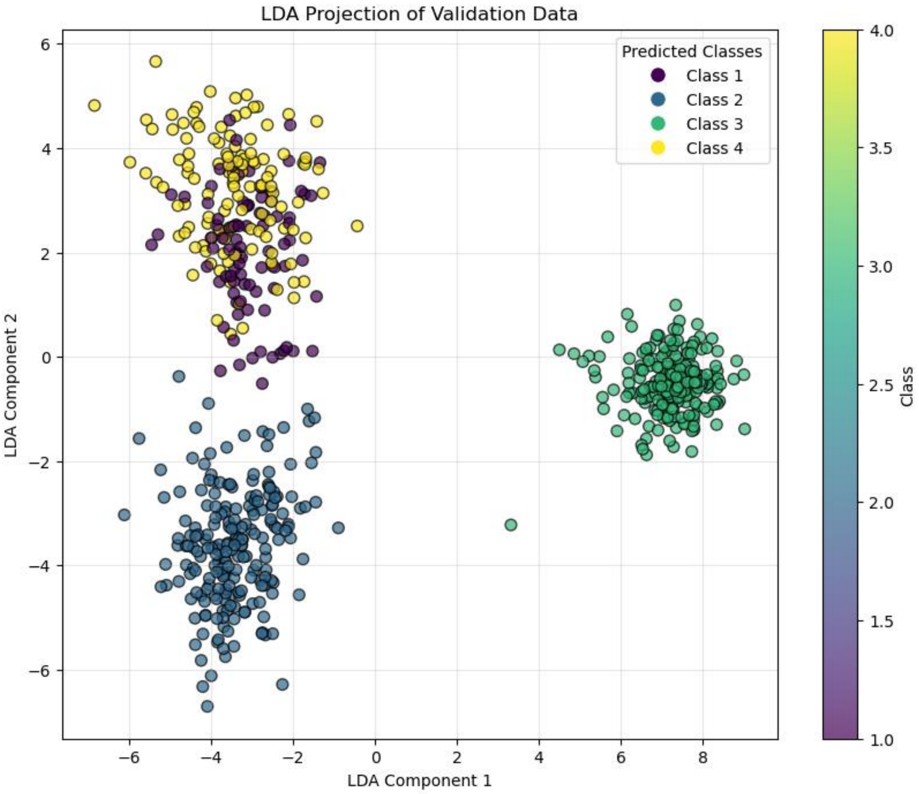

Figure 4. Fisher's LDA projection of validation data with color coded labels

4 Part 4: Fisher's LDA Scree Plot

Figure 5 is a scree plot corresponding to the Fisher's LDA that was done earlier in this assignment. After computing the intersection between the cumulative variance and the 95% variance threshold, 2 principal components are needed to explain 95% of the variance. Because the first principal component explains around 75% of the variance, this suggests that the dataset has high collinearity between features.

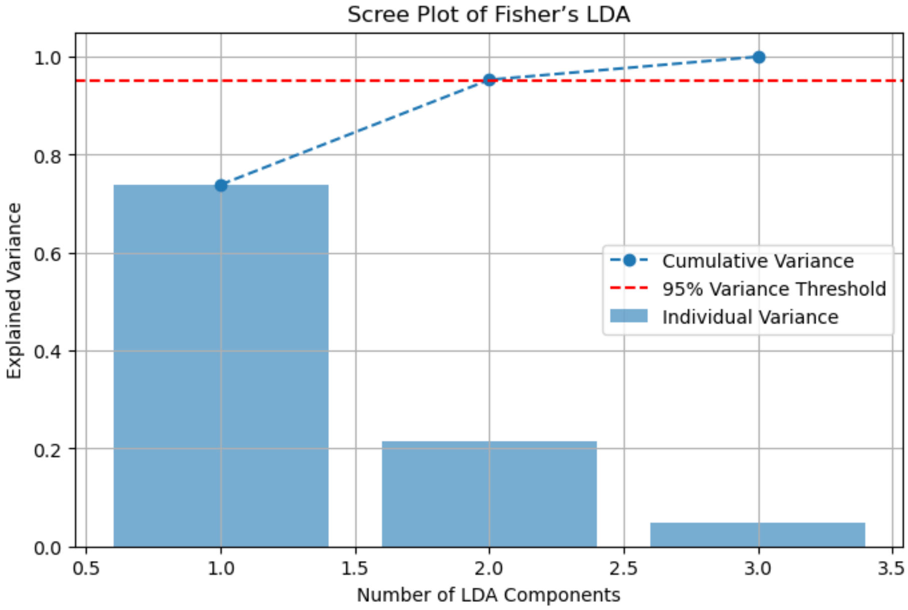

Figure 5. Scree plot of Fishers LDA

5 Resources used to achieve this goal

Canvas: Homework template, Lecture Slides

Python Libraries: NumPy, pandas, matplotlib, scikit learn

6 References

- "Comparison of LDA and PCA 2D Projection of Iris Dataset — Scikit-Learn 0.21.3 Documentation." *Scikit-Learn.org*, 2019, scikit-learn.org/stable/auto_examples/decomposition/plot_pca_vs_lda.html.
- scikit-learn. "Sklearn.decomposition.PCA — Scikit-Learn 0.20.3 Documentation." *Scikit-Learn.org*, 2009, scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html.