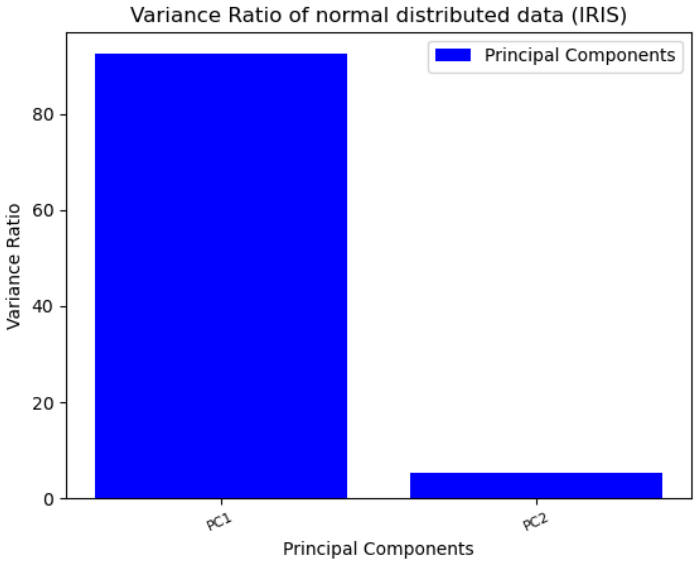
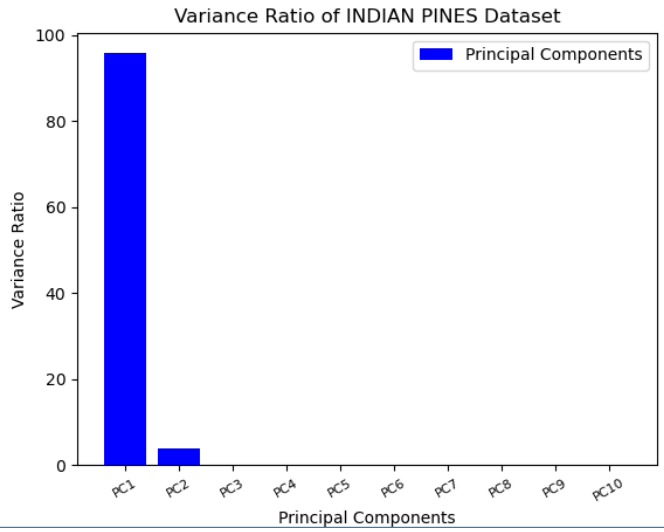
Ian McNichols, CS 488-01, 11/2/2021

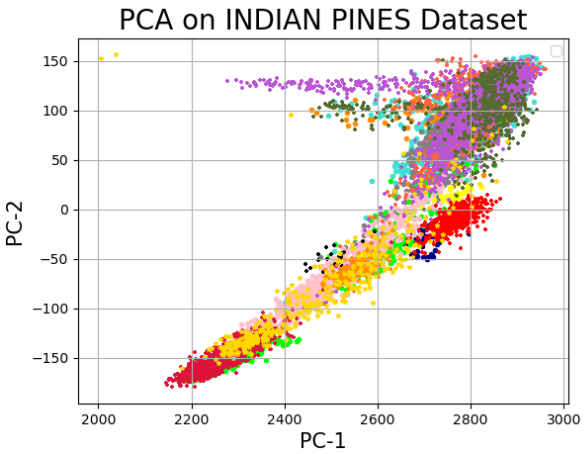
1a)

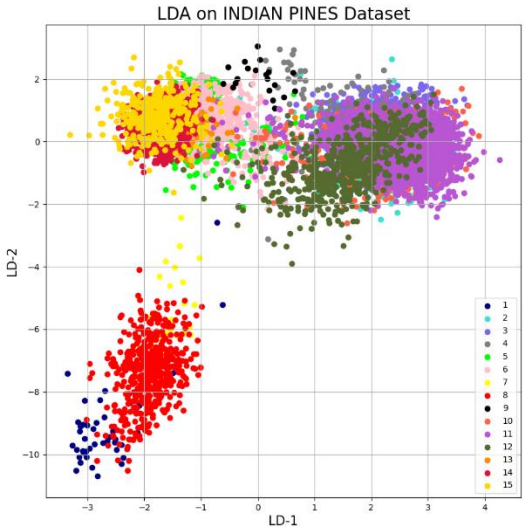
i) Figures 1 and 2 respectively: Plots of variance ratios of the Iris and Indian Pines dataset

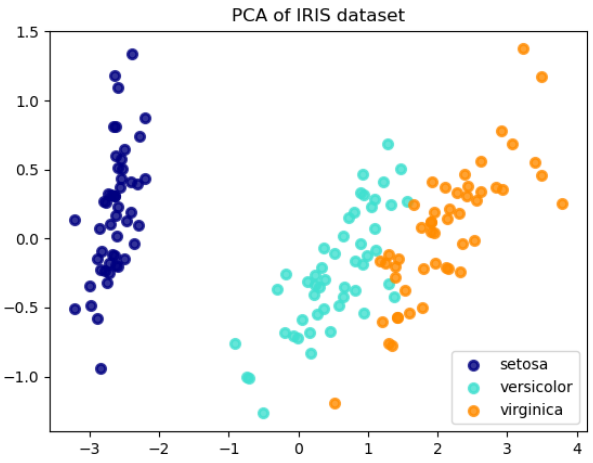


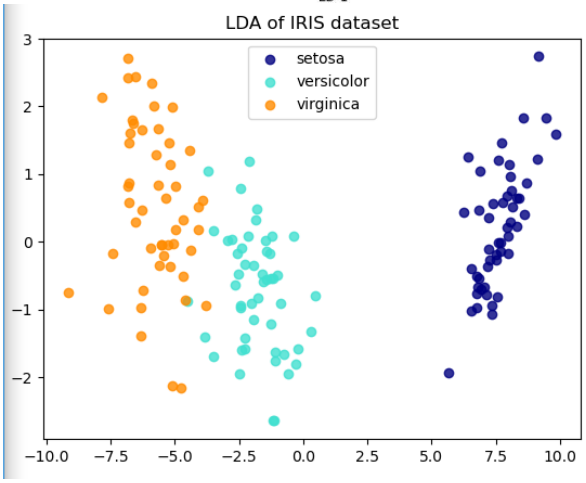


ii) Figures 3, 4, 5, and 6 from top left to bottom right. Plots of 2D PCA and LDA dimensionality reduction on the Indian Pines and then Iris datasets









Ian McNichols, CS 488-01, 11/2/2021

1b)

Pines Dataset:

The HSI for the Indian Pines location had 202 layers of pixels for each pixel of ground truth, and only 150x150 pixels of ground truth. To a person, that's basically impossible to visualize the differences between each type of ground pixel. There's too many dimensions to the data to be understandable or visualizable. By performing dimensionality reduction and feature extraction we're able to see only the more important features and dimensions that are affecting the category a pixel is defined as. It unclutters the data so people don't have to see all the noise in the background and can instead view the distinction between classes based on the most important features. Reducing to only 2 dimensions here for display did overly restrict the distinction between classes, because there is a large amount of overlap in both the PCA and LDA plot. The classes have different trends, but it's difficult to separate which exactly is which, especially in the clumps of data. My takeaway here is that there is not a high level of data separability, because we were unable to cluster. For the Pines dataset I used K=10 for both PCA and LDA. That was the initial value in the demo video for running PCA, and it seemed to produce viable results so I kept it. The LDA seems to have clumped the different ground types better into clusters. I think because LDA is supervised, it was better able to handle the large amount of data and classifications. It's goal is to maximize feature separability, so it makes sense that it would separate the different classes into clusters better, whereas PCA might be a more accurate depiction of the similarities of the classes.

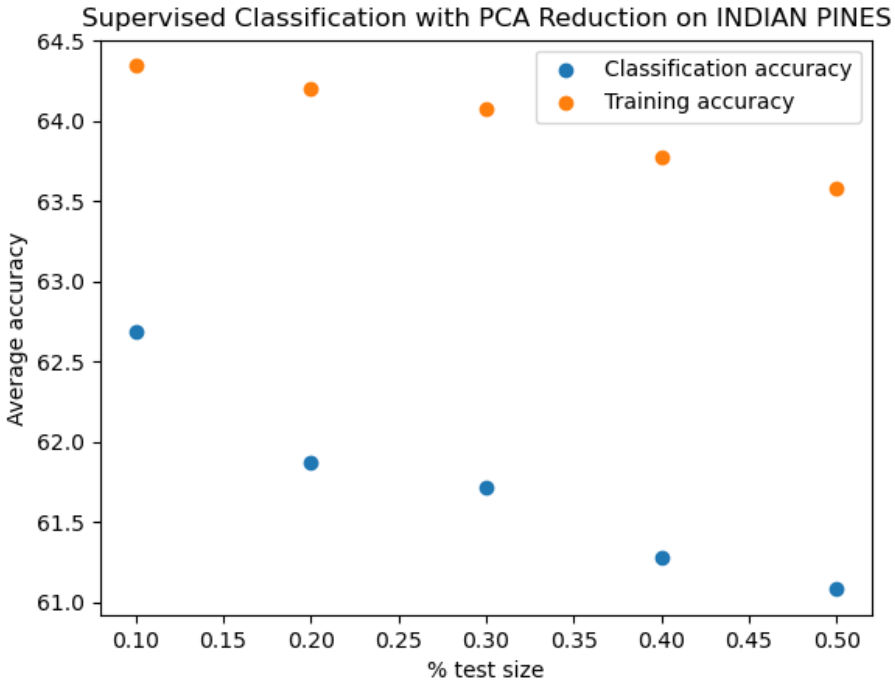
Iris Dataset:

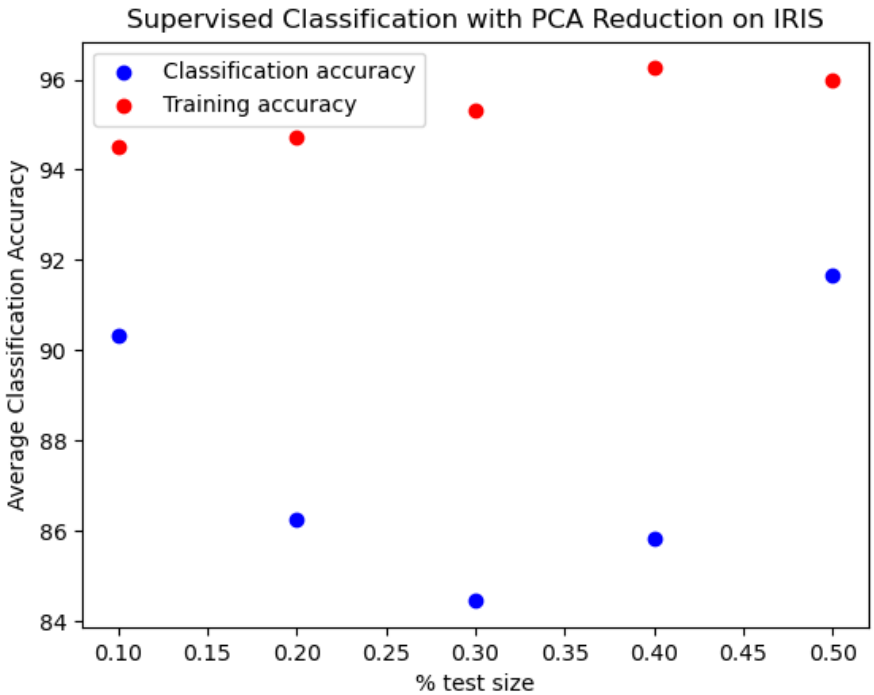
The Iris dataset only has 4 features, so it's almost overkill to run dimensionality on it when you could probably classify the different classes by hand. However, it does enable you to display it in a very concise manner that's easy to understand. On the Iris dataset the clusters are almost entirely separate, so we can infer that there is high data separability and difference between the 3 classes. I chose K=2 for this dataset because that is the highest value allowed for LDA (the minimum of classes or features -1). I used that value for both PCA and LDA so there would be fewer differences in how they were run. It's hard to say which worked better on the Iris dataset because both gave very similar results. However, the LDA had fewer overlapping points between the Versicolor and Virginica so I would argue that it did better at dimensionality reduction.

Ian McNichols, CS 488-01, 11/2/2021

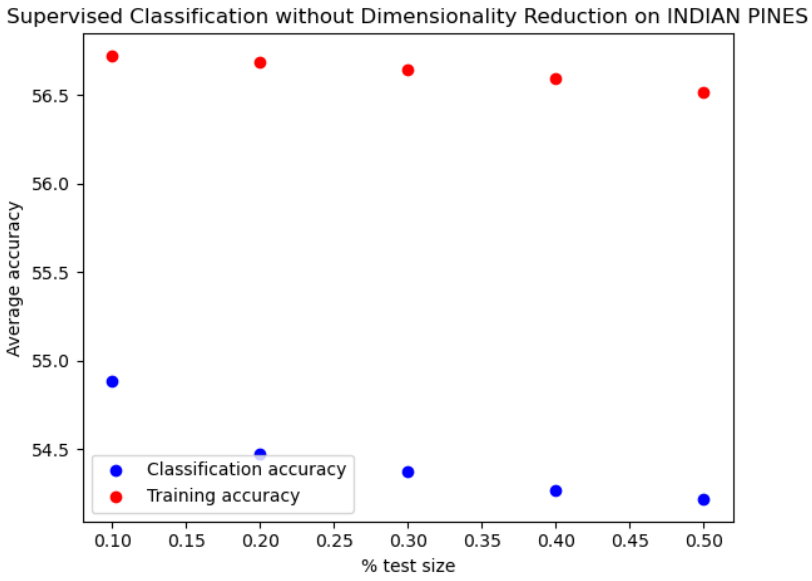
2a)

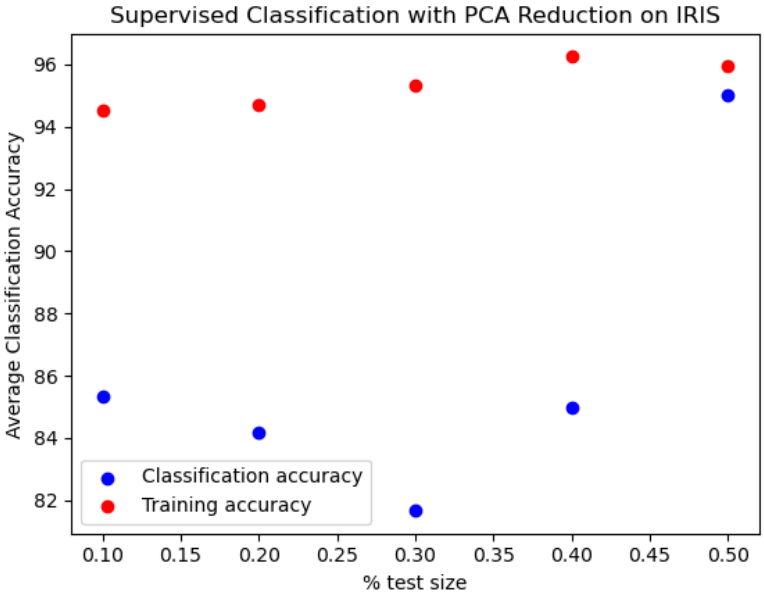
i) Figures 7 and 8 from top to bottom, plots of average model classification and training accuracy with respect to test size on the Iris dataset





ii) Figures 9 and 10 from top to bottom, plots of average model classification and training accuracy with respect to test size on the Indian Pines dataset





iii) Figure 11, a tabularized class-wise classification accuracy chart for 30% training size over all methods, for both with and without PCA dimensionality reduction for the Indian Pines and Iris datasets

Pines	30% Test size				
	Model:	Accuracy:	Training Accuracy:	Sensitivity:	Specificity:
	KNN_pca	70.57705	80.58028131	0.77777778	0.99962049
	KNN	72.98668	82.07515119	0.75	0.99962908
	Gaussian_pca	64.61636	64.82978868	0.90909091	0.99862322
	Gaussian	30.18389	30.63124278	1	0.93061224
	SVM-Poly_pca	51.87064	50.98865258	0	1
	SVM-Poly	56.8643	56.60800435	0	1
	SVM-RBF_pca	59.84464	59.90351294	0	1
	SVM-RBF	57.46671	57.2603112	0	1
Iris	30% Test size				
	Model:	Accuracy:	Training Accuracy:	Sensitivity:	Specificity:
	KNN_pca	95.55556	93.27272727	1	1
	KNN	95.55556	95.27272727	1	1
	Gaussian_pca	93.33333	87.63636364	1	1
	Gaussian	93.33333	93.54545455	1	1
	SVM-Poly_pca	51.11111	94.18181818	0	1
	SVM-Poly	55.55556	94.36363636	0	1
	SVM-RBF_pca	86.66667	95.27272727	1	1
	SVM-RBF	93.33333	98.09090909	1	1

2b)

Pines Dataset:

With the Pines dataset there was a definite, if slight, improvement of data analysis following the PCA dimensionality reduction compared to running models on the raw data. It increased the classification accuracy from 2-34 percent depending on the model. The effect on the sensitivity and specificity are less noticeable. In some cases it improved, some cases it worsened, and most it remained the same regardless of PCA. The best supervised classification model for the Pines dataset is highlighted in Figure 11, the K-Nearest Neighbors model trained after running PCA dimensionality reduction. In fact, the second best model was the K-Nearest Neighbors without PCA dimensionality reduction. The closest different model was 8 percent behind, the Gaussian NB model with PCA. The support vector machine models did much worse than the KNN and Gaussian, hovering in the 50%'s. This makes sense, because KNN models work better with more complex problems, and HSI with 200+ bands is a very complex problem with 11 different classes. However, the KNN does begin to overfit as seen in the 10% difference between its training accuracy and classification accuracy in Figure 11. A larger dataset or larger training percent would probably increase how well it works. The SVM's probably struggle because there is less data separability, so they aren't able to find distinct lines between classes. The Gaussian model

performed the worst, at about 30%. I think the data is not conducive to a Gaussian model because the different classes are not following a statistical curve. There are classes very similar to each other, like different kinds of crops, and classes very different from each other. To me, it seems like this would make Gaussian classification difficult because the features will lean heavily towards multiple classes in a similar way, so they're not all distinct and equally different.

Iris Dataset:

A lot of the discussion of the Iris dataset analysis hinges on the fact that it is a very small dataset with very few features, which makes it prone to overfitting. Running dimensionality reduction is useful if there is too much noise and a model will struggle to pick up on the important features. With a dataset that only has 4 features, dimensionality reduction will take away important and useful features, making it more difficult for supervised models to learn. Because of this, running the PCA before training models decreased model classification and training accuracy or left it the same in every single case on the 30% test size, and in most averages of all the models over each test size. Sensitivity and specificity were again not affected by running PCA. Another interesting point is the overfitting. The SVM with a poly kernel severely overfits the data at 30% test size, with a 94% training accuracy and a 51-55% validation accuracy. It's higher after running PCA. The SVM with an rbf kernel overfits as well but to less of an extreme. You can see it as well in the average accuracies from 20-40% training sizes, where the training accuracy is about 10% above the classification accuracy. These solutions might be overly complex for such a simple dataset. The K-Nearest Neighbors once again performs the best on the 30% training size, winning by 2% over the SVM with rbf kernel on PCA data and the Gaussian model on both the PCA and raw data. I was surprised by this because I figured for a simpler dataset the SVM's would work better. The RBF kernel was close without PCA, so the kernel must have a large effect on how an SVM works. Perhaps a simpler kernel like linear would produce better results. The Gaussian model was also a high performer, which makes sense with what I said earlier on the Pines dataset, that it would work better on a more evenly distributed dataset. Overall, it seems like the K-Nearest Neighbors model is the most effective model for datasets in this range.