Assignment 3

Dylan Lawrence

CS 7641

Introduction

For this assignment I have chosen two datasets from the UCIML data repository. The first of these datasets contains the measurements of two species of rice taken from rice grown in Türkiye[Rice]. It contains 7 continuous independent variables, along with the correct species for each observation. The measurements were taken using computer vision strategies, which adds another layer to this dataset. The first features are area and perimeter, which are the number of pixels within or around, respectively, the rice grain. There is also MajorAxisLength and MinorAxisLength, the longest and shortest lines that can be drawn on the grain of rice. This is followed by eccentricity, which is a measure of how round the ellipse with the same moments are the grain of rice is. Next is ConvexArea, the pixel count of the smallest convex shell that can be formed on the grain of rice. The final independent variable is extent, the ratio of the smallest bounding box that can be formed around the rice. The target variable is Class, either Cammeo or Osmancik, two common varieties of rice in Türkiye. This dataset contains 3810 observations.

The second dataset I chose was taken in a lab trying to simulate field detection methods of landmines[Mine Detection]. This dataset contains 3 independent variables, one of which is discrete, presenting unique challenges during dimensionality reduction. The two continuous variables are V (voltage) and H (height). Voltage was measured from a magnetic field sensor that senses magnetic distortion, and outputs a voltage correlated to the strength of the magnetic anomaly created during the observation. Height is the height of the censor, in centimeters, that the sensor was at during measurement. The discrete variable ‘S’ represents the type of soil underneath the sensor during the observation. This variable has 6 possible values: Dry and Sandy, Dry and Humus, Dry and Limy, Humid and Sandy, Humid and Humus, and Humid and Limy. The dependent variable ‘M’, refers to the type of mine being detected. The 6 possible values are: Null, Anti-Tank, Anti-Personnel, Booby Trapped Anti-Personnel, and M14 Anti-Personnel Mine. This dataset contains 338 observations.

Initial Clustering

To begin the analysis we will cluster both of these datasets using Expectation Maximization and Spectral Clustering. For the rice dataset, I will be using euclidean distance. I believe this is the most suitable choice as our data is physical measurements taken in a euclidean space. Manhattan distance was considered as an alternative, however I do not believe it will as accurately capture the correlations between different variables, such as MajorAxisLength and MinorAxisLength. I believe that we will experience some success with both clustering methods on this dataset. However, I believe that due to all of the independent variables being correlated to one another, to different degrees, I will struggle to reach a high level of spatial separation between the clusters if the number of clusters is much larger than the number of possible classes. I hypothesize that there will be improved performance with Spectral Clustering, since it uses eigenvectors to bring the data to a lower-dimensional space. Additionally, I believe that Expectation Maximization will struggle due to the variables not following the normal distribution[Figure 1].

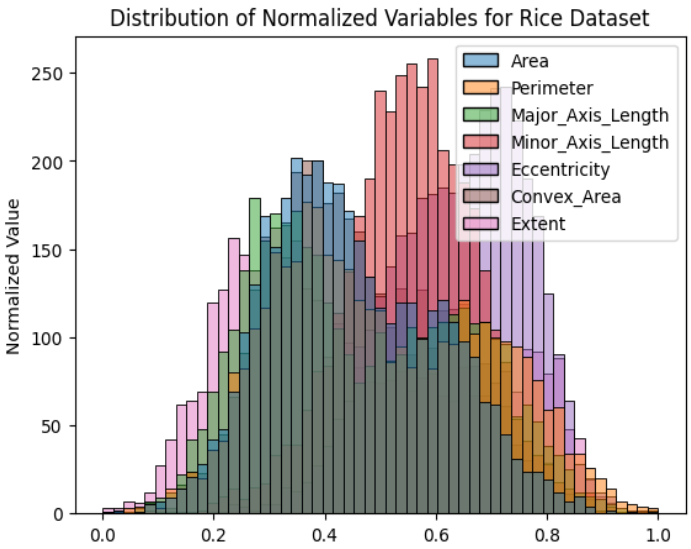

Figure 1: Distributions of Each Independent Variable for the Rice Dataset

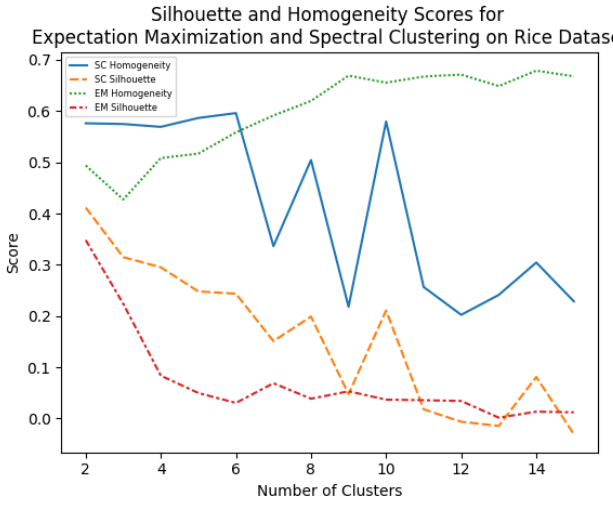

Figure 2: Clustering Results for Rice Dataset

As you can see from my results [Figure 2], my hypothesis mostly stood. However, an interesting note is that the homogeneity of the expectation maximization clusters was able to far exceeded that of the spectral clustering clusters. This leads me to believe that the expectation maximization clusters may be more beneficial when we implement them into a classification problem, despite it's low silhouette scores.

For the mines dataset we will be using nearest neighbors to construct the affinity matrix. This is because our most influential piece of data is the voltage reading from the sensor. It does not have a linear relationship with the soil type. I am hypothesizing that the performance for both clustering methods at this

stage will be poor due to categorical variable having an unknown, but very likely non-linear impact on voltage.

This hypothesis stands when tested with expectation maximization and spectral clustering. While the silhouette scores for both metrics begins at an acceptable value with 2 clusters, it drops off rather quickly [Figure 3]. However, with 4 clusters, spectral clustering manages to maintain a silhouette score that indicates it may have value in later classification models

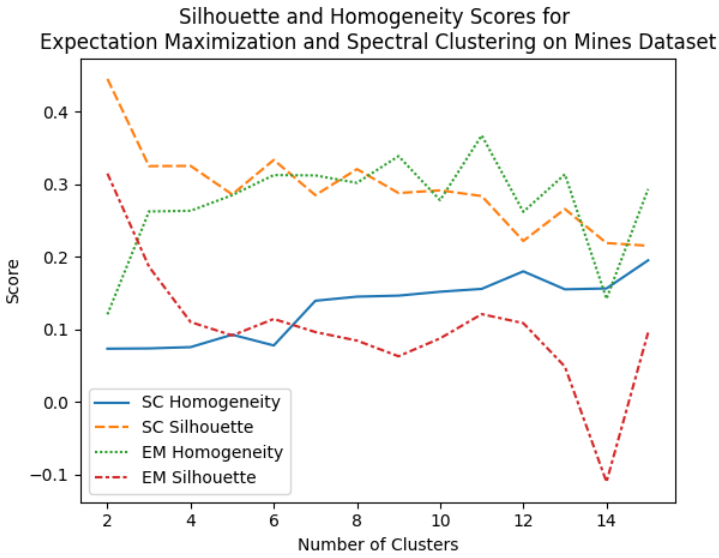

Figure 3: Clustering Performance for Mines Dataset

Dimensionality Reduction

For my own dimensionality reduction algorithm I have chosen to use a multidimensional scaling learner due to its ability to handle non-linear relationships. It also makes cool visualizations.

To begin with the rice dataset, I am hypothesizing that we will be able to generate clusters that will be clean enough to aid in later classification problems. I tested multiple levels of dimension reduction, and have settled on reducing to 3 dimensions, since I believe this will preserve the most data, while eliminating noise, as we should be able to actually capture the length and width of the rice, as well as the area, which should directly relate the volume of the grain. I believe that all 4 algorithms will be capable of doing this as the data is clean and highly correlated.

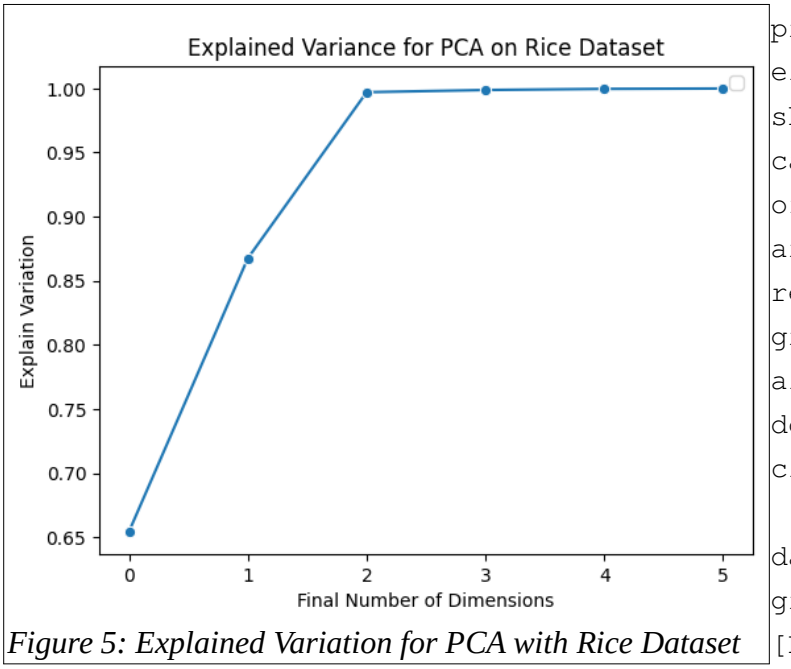

Figure 5: Explained Variation for PCA with Rice Dataset

The reduction in each dataset provides clean visual groups for all 4 algorithms [Figure 4]. Principal component analysis produced the separation that appears to be most clean to me. The explain variance for PCA with 3 dimensions rounds to .997

[Figure 5]. This supports my hypothesis that I will be able to cleanly capture the data in 3 dimensions.

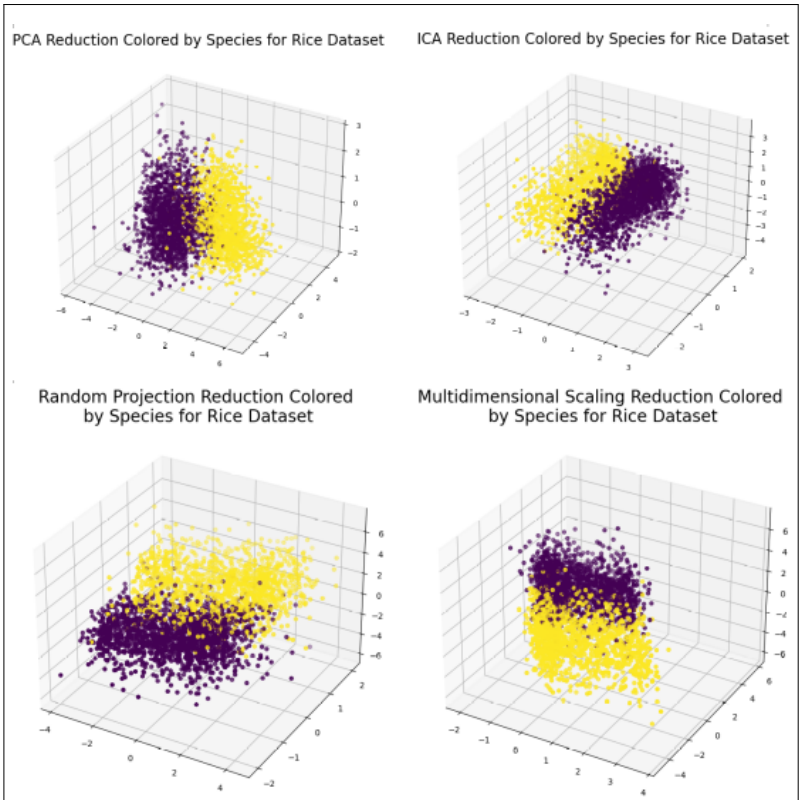

Figure 4: 3D Visualizations of Dimensionality Reduction on Rice Dataset

We will be splitting rice into 2 dimensions to try to "capture S and H in one dimension". Since these two variables will work in tandem to affect voltage, I am hypothesizing that we will be able to capture the data in these 2 dimensions, and see improved clustering results in our next run. I believe that principal component reduction and random projection will be the most effective methods for this, as the categorical data is difficult to handle when combined with continuous data in dimensionality reduction.

My hypotheses for the mines dataset was mostly correct, however the groups are not as distinct as I would like. Additionally, PCA did not perform much better than the other algorithms, although random projection did perform the best. An interesting note about the random projection results is that you can notice that many observations with an M value of 1 (no mine), are separated from the rest of the observations [Figure 5]. This still provides value if we can determine whether or not a reading is a mine.

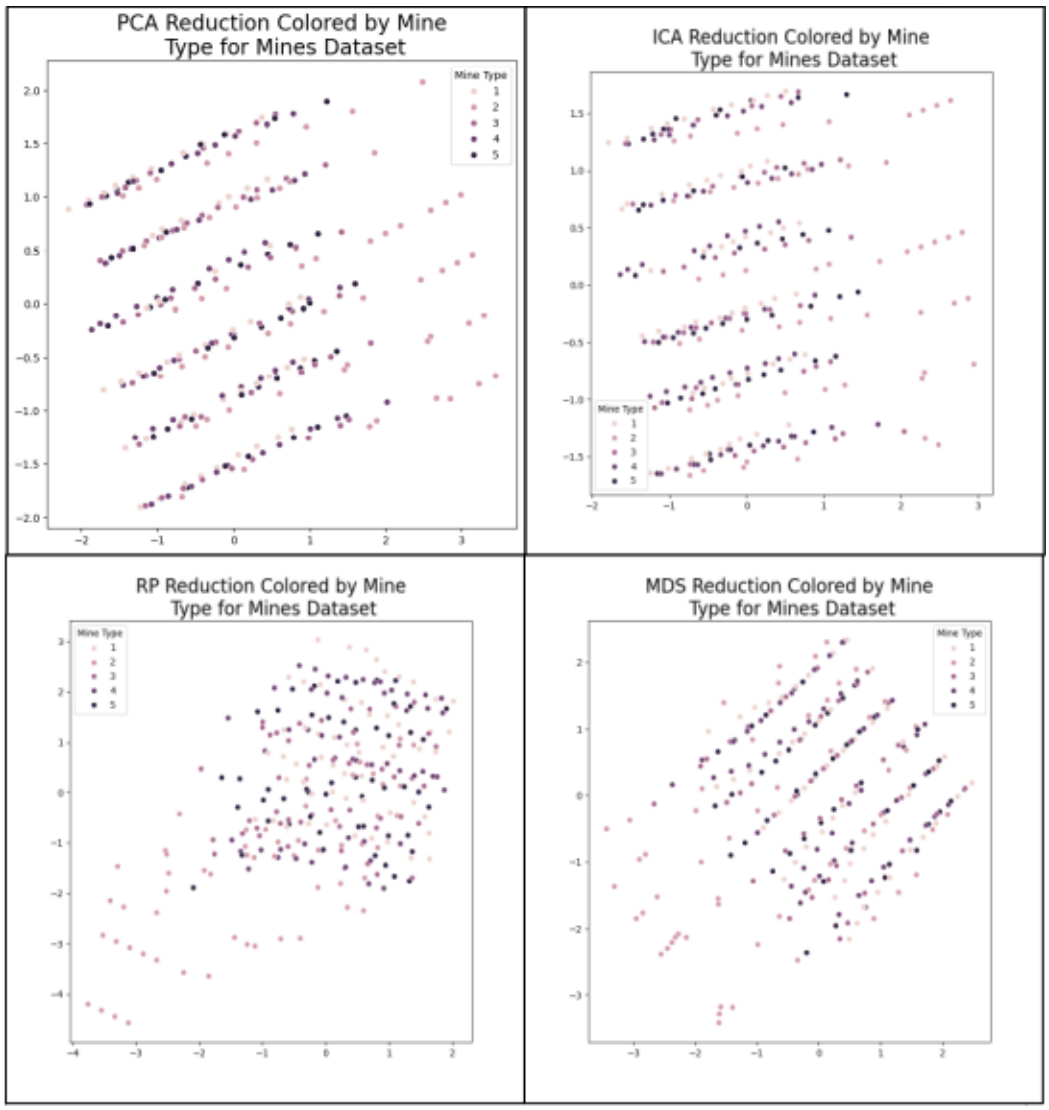

Figure 6: Dimensionality Reduction on the Mines Dataset

Reclustering

For reclustering I have chosen to work with the random projection data as it appears to be messy when used with the rice data, but appears to be the cleanest option on the mines data. I am hypothesizing that performance will be worse on the rice dataset due to information loss that can be seen [Figure 4]. However, I expect better performance on the mine dataset, as we can see several groups forming [Figure 5]. I believe that the mines dataset will perform better with spectral clustering due to it falling far from a normal distribution, and that the rice dataset will perform better with expectation maximization for the opposite reason.

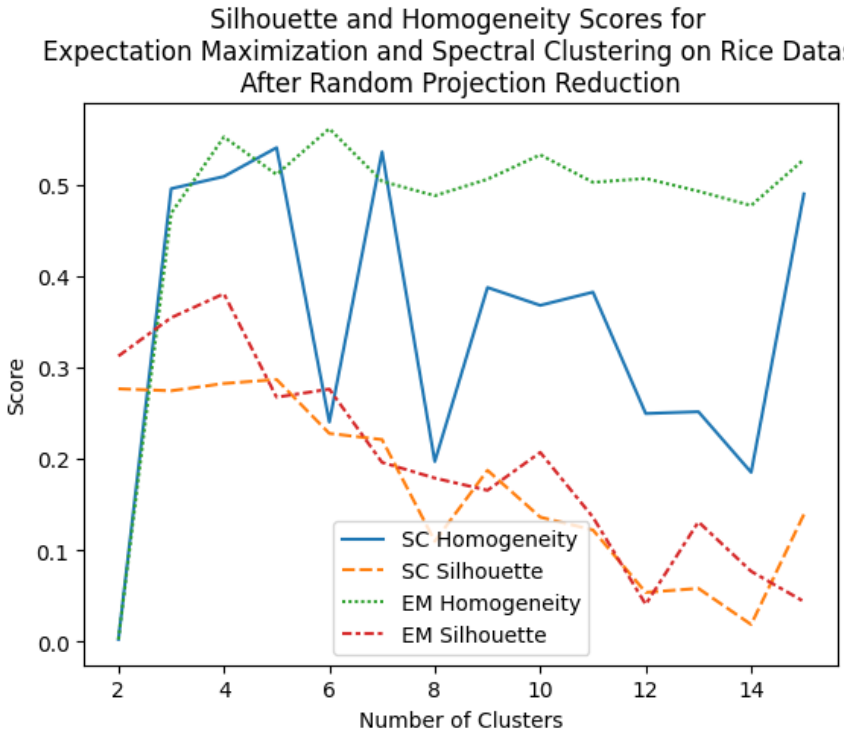

Figure 7: Clustering Performance on Rice Dataset after Dimensionality Reduction with Random Projection

Interestingly, the clustering performance on the rice dataset began to converege. For some cluster n values, Expectation Maximizations exceeds Spectral Clustering by varying amounts, but for about half of the values they are essentially even [Figure 7]. This trend carried over in mines to an even greater extent [Figure 8]. However both clustering algoirthms saw a notable performance improvement. The improvement on the mines dataset was much larger, possible due to eliminating a large portion of the noise in the data caused by the S column being categorical and having an unknown affect on V.

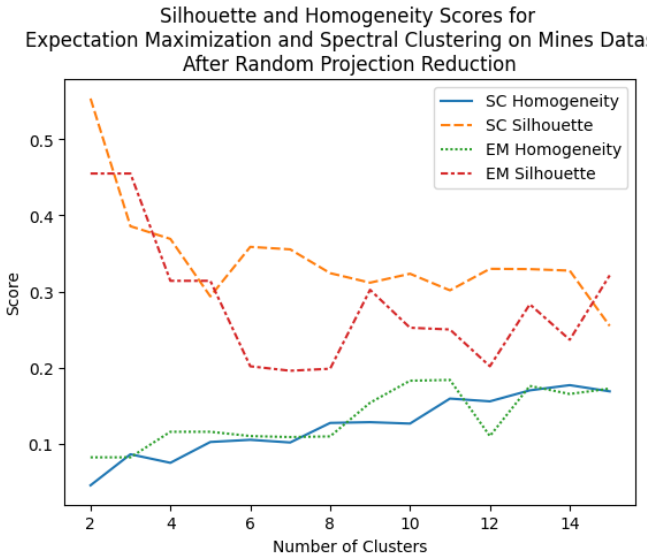

Figure 8: Clustering Performance on Mines Dataset after Dimensionality Reduction with Random Projection

Neural Network

For the neural network portion of this assignment I have decided to use the rice dataset reduced with multidimensional scaling, and PCA. I chose PCA due to the clean visual split it has when reducing the rice dataset to three dimensions. Additionally, since I am able to calculate the explained variance and know essentially all of it is captured with 3 dimensions. Since I had to modify the neural network slightly for this new problem, I decided to run the neural network on the original data after standardizing it. I did not include loss for this because for all 5000 epochs it stayed above 40. As you can see, with the original data accuracy stays flat at around .46, worse than guessing randomly. This is likely due to the large amount of statistical noise found in the original data as mentioned earlier. This supports my original hypothesis that after dimensionality reduction there would be a substantial increase in classification model performance due to decreased statistical noise. The data reduced by MDS slightly edges out PCA in accuracy, but loses slightly in loss. The performance difference is pretty negligible however. After applying the clustering algorithm, loss skyrocketed, and accuracy plummeted. I believe that this is due to the problem being one of binary classification, without distinct groups within each binary group.

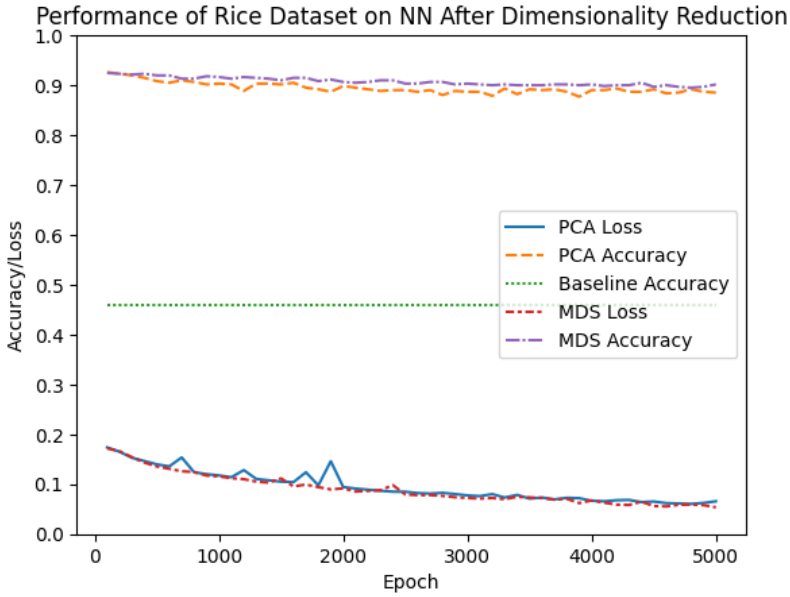

Figure 8: Neural Network Performance On Dimensionally Reduced Data

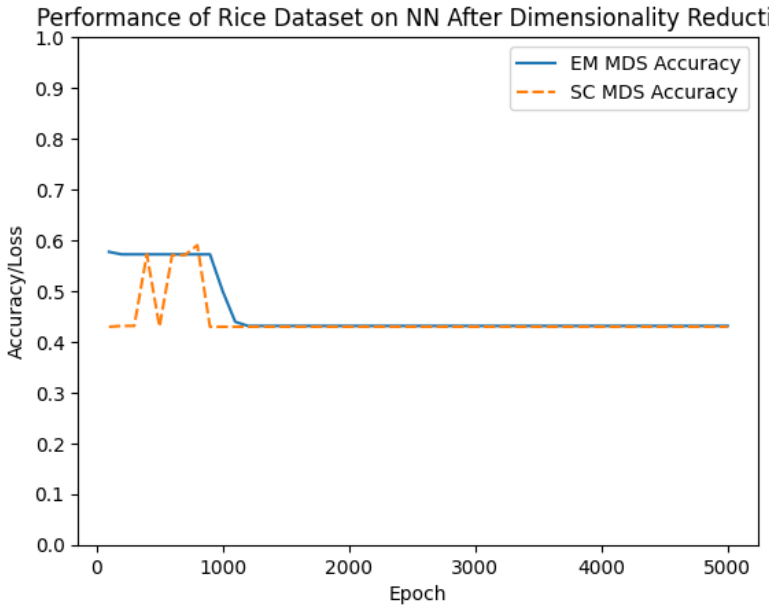

Figure 9: NN Performance After Clustering Dimensionally Reduced Data

Bibliography

Rice: Ilkay Cinar, Murat Koklu, Classification of Rice Varieties Using Artificial Intelligence Methods, 2019
Mine Detection: Cemal Yimaz, Hamdi Tolga Kahraman, Salih Soyler, Passive mine detection and classification method based on hybrid model,