Dimensionality Reduction Report II

## Dimensionality Reduction Tests

We addressed the task by first testing the various Dimensionality Reduction methods using the Decision Tree as a classifier. After obtaining the results which we will discuss shortly, we have tested various classification methods (namely, Decision Tree, KNN, Naïve Bayes and Logistic Regression) on two-component PCA (next chapter).

We started with the **Variance Threshold** method. We realized that, as expected, the different threshold parameters yielded different numbers of dimensions. Namely, in terms of integers, results of the parameter search can be observed in the adjacent table.

|  |  |
| --- | --- |
| Dataset Shape | Integer Variance Threshold |
| (8143,5) | 0 |
| (8143,4) | 1 |
| (8143,3) | 2 |
| (8143,2) | 31 |
| (8143,1) | 37926 |

The results of the Variance threshold analysis are exactly the same in terms of precision, accuracy and recall as if the dataset dimensions were not reduced at all when the threshold is anywhere between 1 and 37925 (which yields a number of dimensions between 2 and 4). However, the results significantly worsen for number of dimensions equal to one which is achieved with a threshold of 37926 onwards(Figure 1).

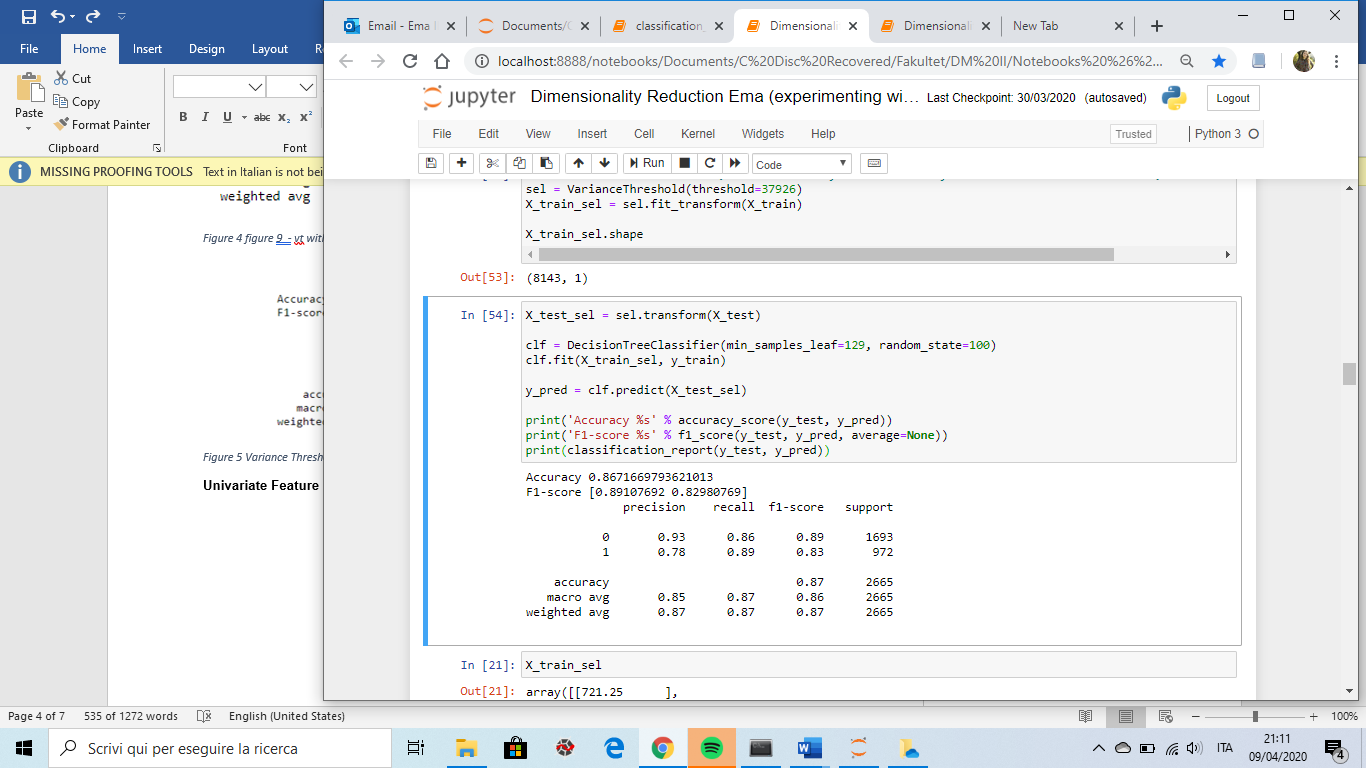


Figure 1

**Univariate Feature Selection, Singular Value Decomposition** and **Principal Component Analysis.** As the parameter we inserted numbers from k=1 to k=4 number of dimensions, and observed the results. Interestingly enough, the results in terms of accuracy, precision and recall were all the same for different k number of dimensions, but changed with respect to each of these three DR methods. They however change for each k when other classifiers are applied, such as KNN, Linear Regression and Naïve Bayes. For example when we employ UFS with Linear Regression or Naïve Bayes, the results stay the same for k=1&2, but change for k=3&4. It is also important to mention that the UFS method selects Light when the k=1, Light and CO2 when the k=2, Light, CO2 and temperature when the k=3, Light, CO2, temperature and the humidity ratio when the k=4.Singular Value Decomposition and PCA, however, completely alter the attributes and yield a new single-or multi dimensional array (depending on the k given) instead of selecting the best attributes.

The results obtained by the UFS and RFE analysis are identical to the dataset which is not reduced. It is interesting to observe that the RFE method delineated ‘Light’ as the only attribute worth keeping. The results obtained by the PCA are reported in figure 2 and the results obtained by SVD are reported in figure 3. Finally, we can conclude that, using the Decision Tree classifier, none of the methods yielded better results than the non-reduced dataset did. In order to complete the analysis, we should reproduce the results using all the four classifiers, but for the sake of keeping to the page limit of the report, we will not do that.

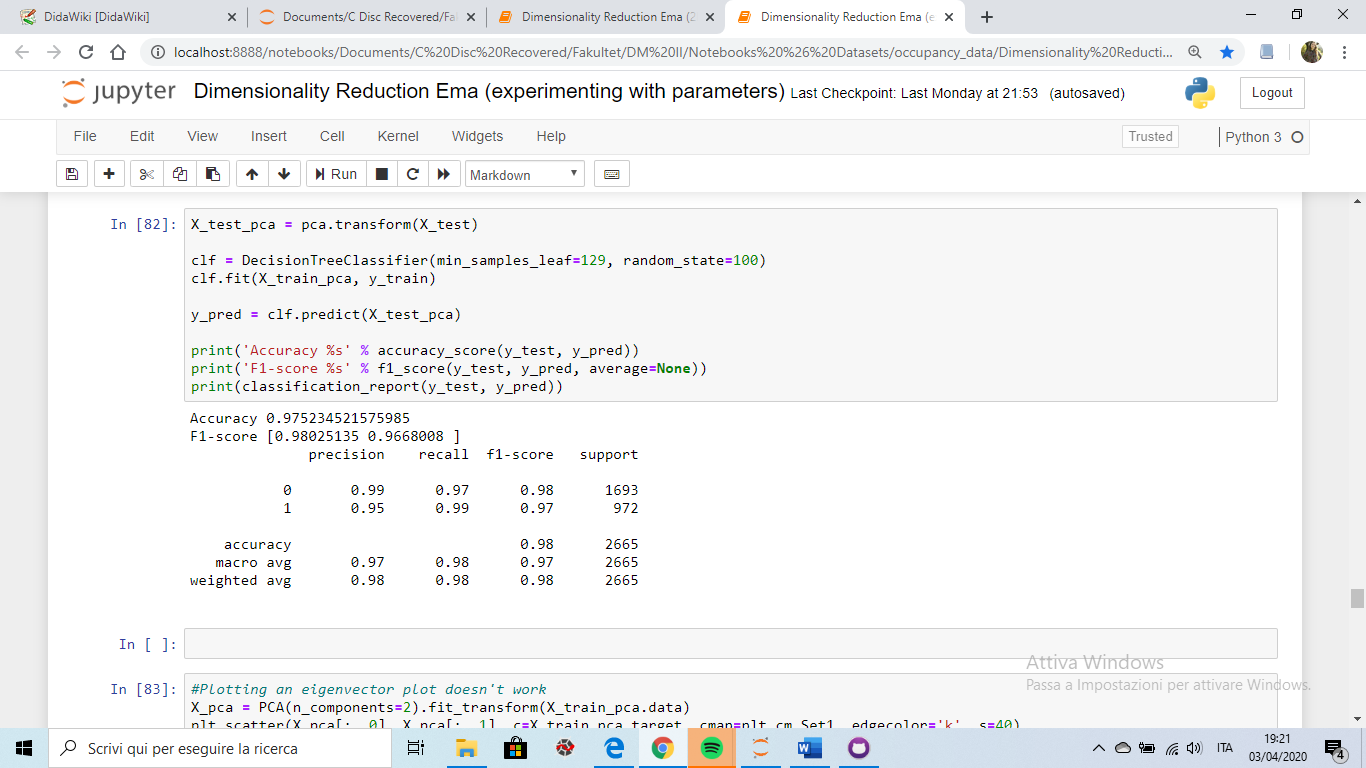


Figure 2

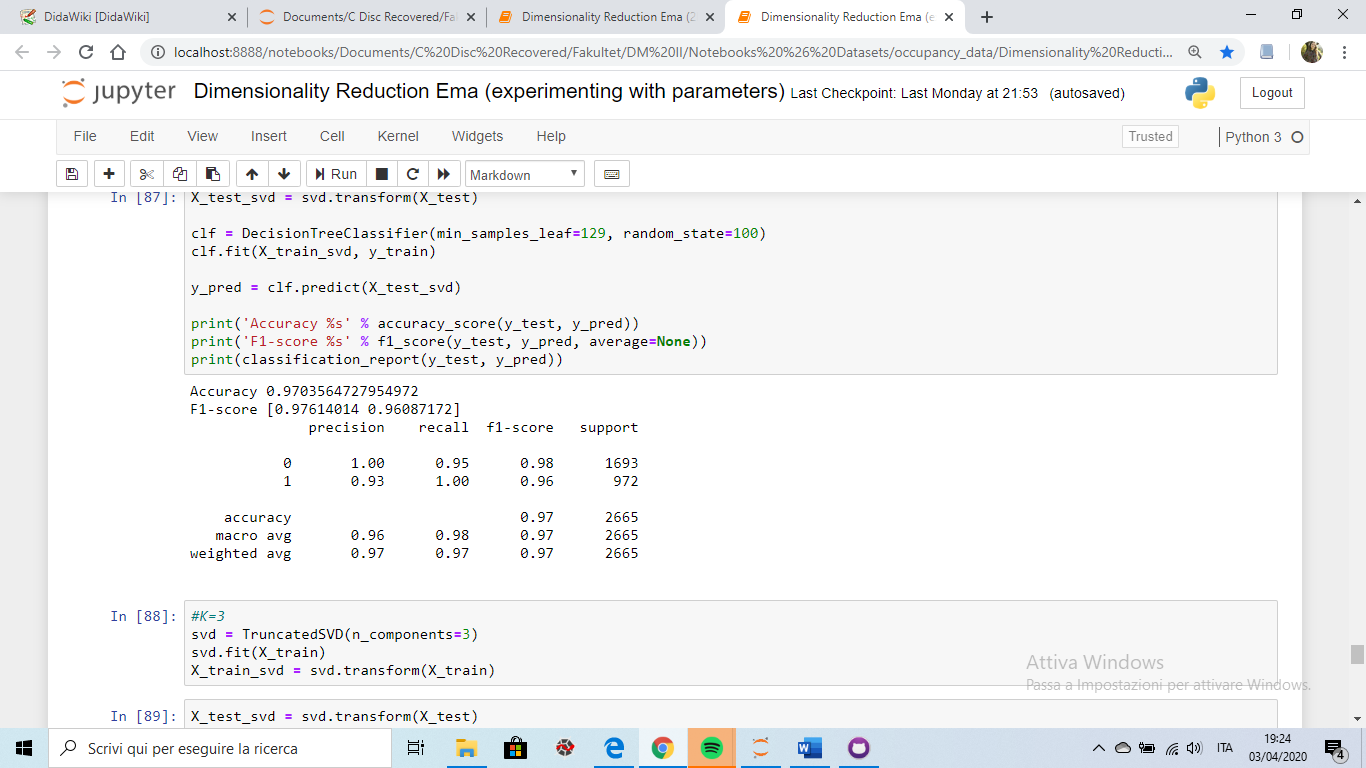


Figure 3

## PCA two component analysis tested on different classifiers

In addition to experiments on different dimensionality reduction techniques, we tested the PCA analysis with two components (two dimensions) on Decision Tree (Figure 2), KNN(Figure 4), Naïve Bayes(Figure 5), and Logistic Regression(Figure 6). The results obtained for the decision tree with two components are slightly worse with respect to the results obtained using the complete dataset. Accuracy with 2-component PCA for KNN remained the same, F1 with respect to the class 0 slightly improved whereas the F1 for the class 1 slightly worsened. The results worsened for the NB and, most importantily, improved for the Logistic Regression when using only two components in the PCA with respect to applying the Logistic regression to all the attributes in the dataset. Interestingly enough, Logistic regression with a two component PCA analysis delivers precision, accuracy and recall results identical to the decision tree when applied to the whole dataset. We tested the results on the second test set (datatest2.txt) and the results improved even further (Figure 10).

Finally, in the Figure 7 we can observe the two dimensional representation of the dataset obtained with the PCA. Figure 8 shows the same, except for the fact that the training set has been reduced to 1000 samples instead of 8143, and the trend can thus be observed better. It is also interesting to observe the fraction of variance with respect to each of the 5 attributes (date excluded) in Figure 9.

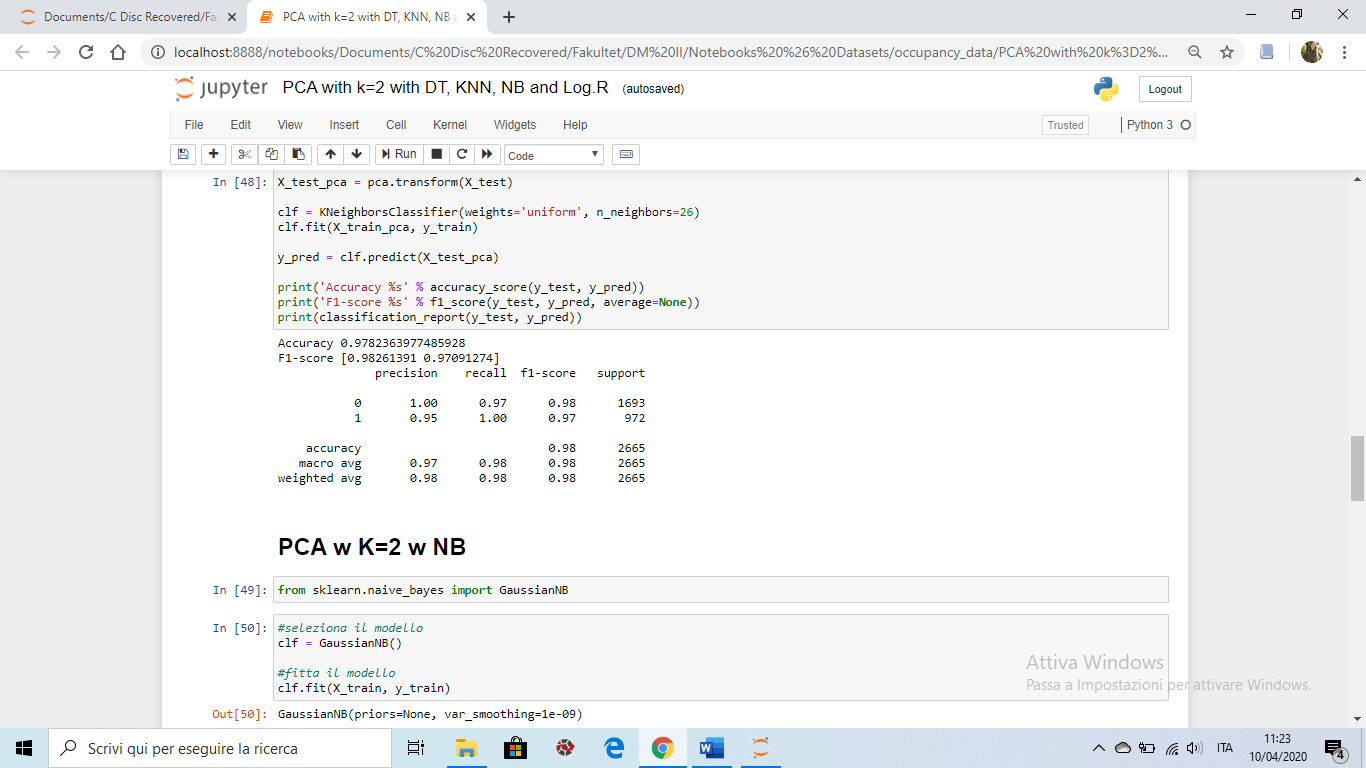


Figure 4

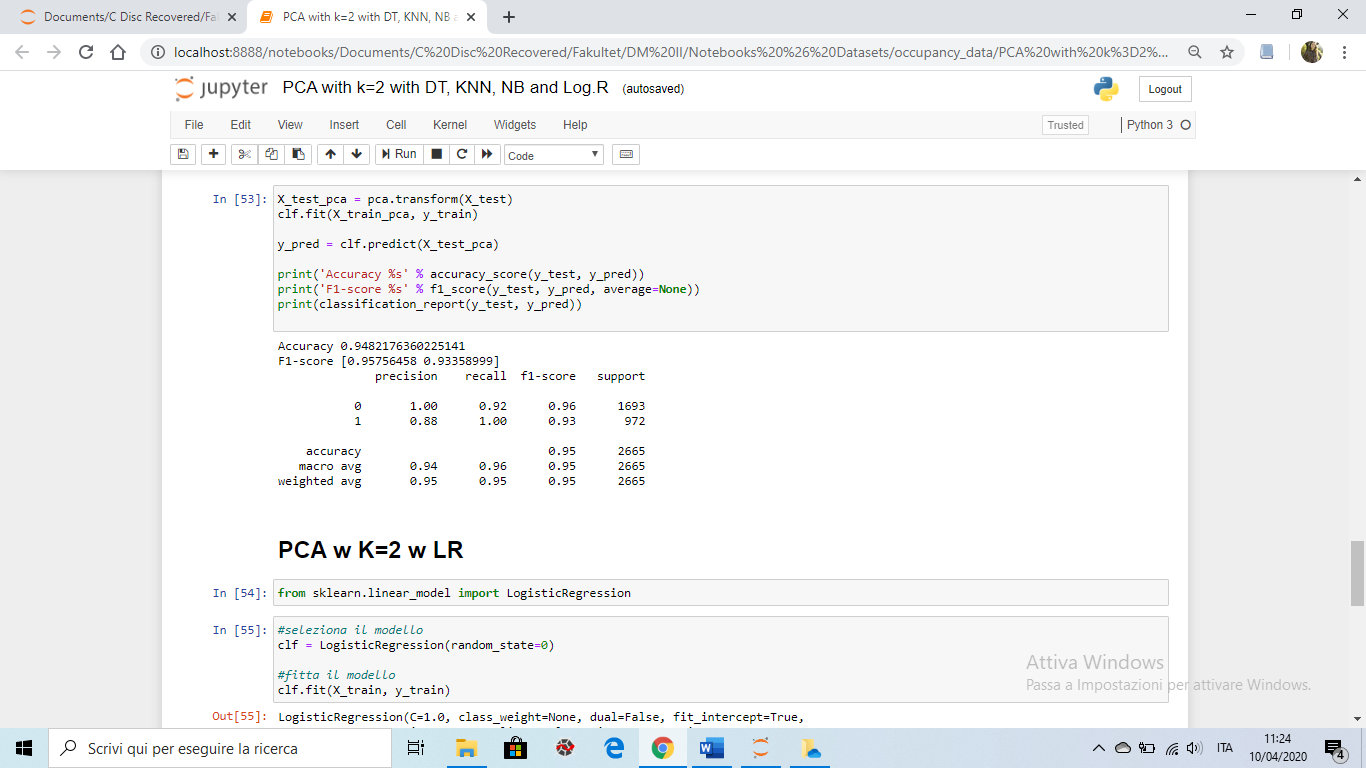


Figure 5

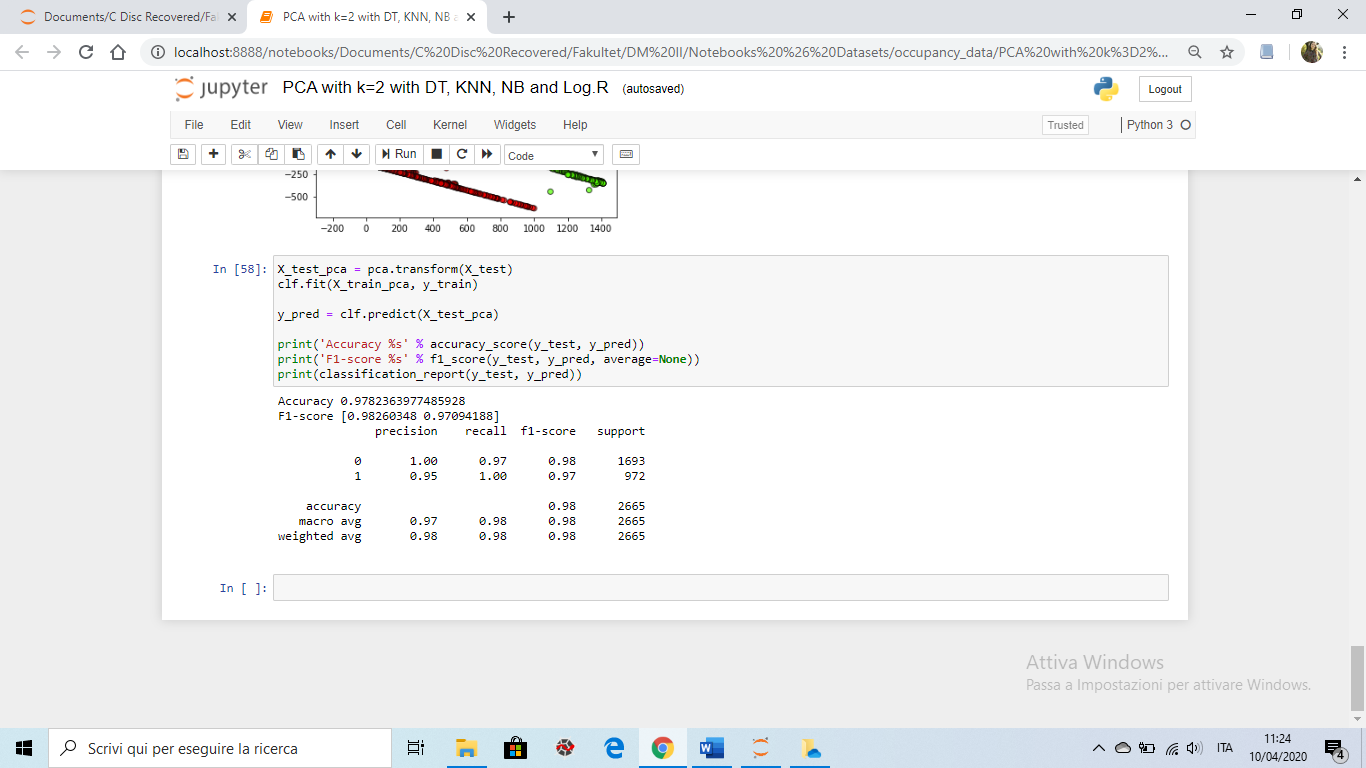


Figure 6

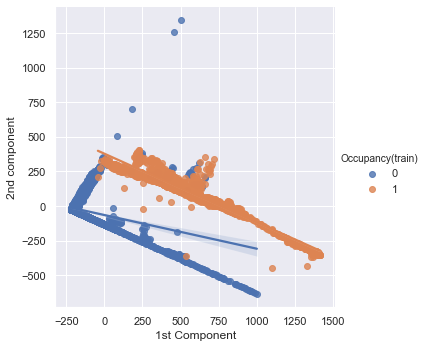


Figure 7

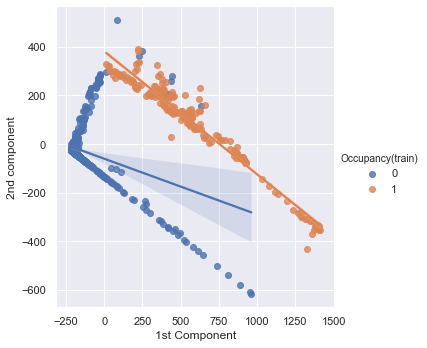
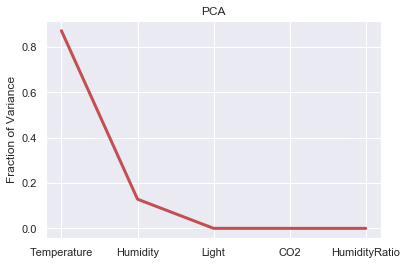


Figure 8



Figure