Dimensionality Reduction

by Juan Manuel Alonso

Installation

```
$  pip3 install -r requirements.txt
```

—————————

Running

Custom hyperparameters in a textfile i.e. *“./configs/config.txt”*.

```
$  python3 experiments.py ./configs/config.txt
```

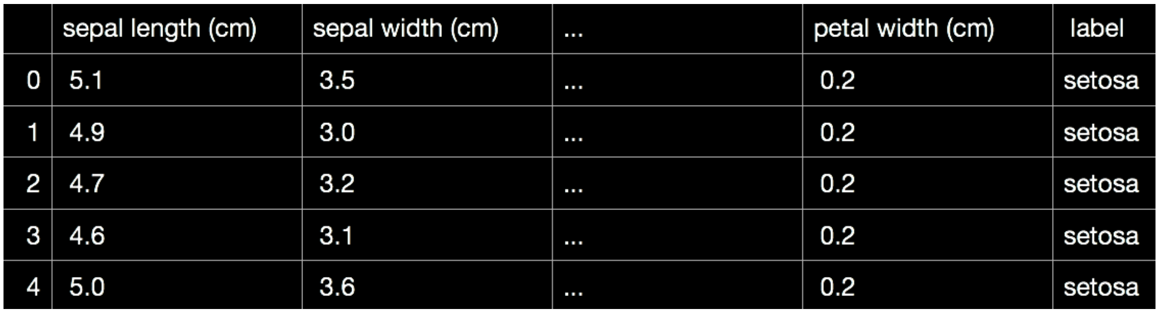
A *results* folder will contain a timestamp directory with the latest results.

—————————

Datasets

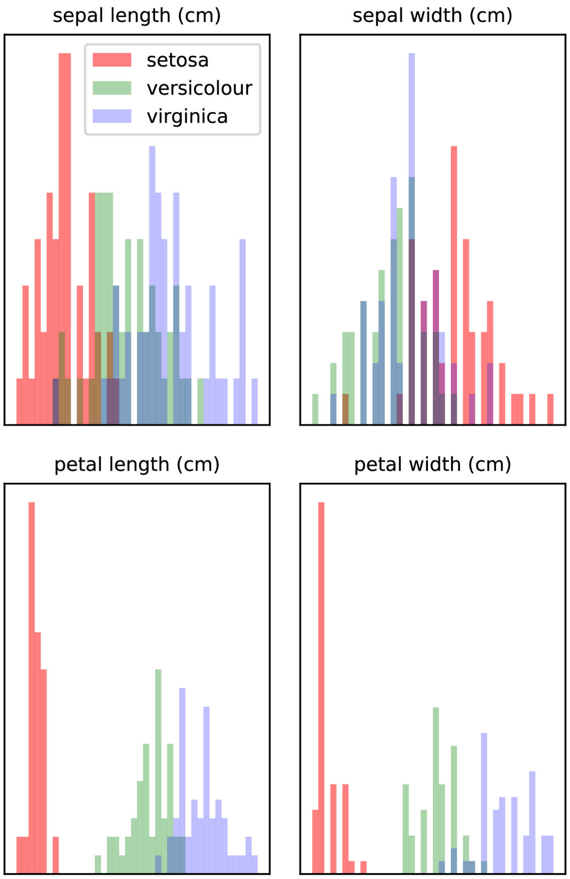
- Iris (http://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html)

The first rows of this dataset’s dataframe are shown below:



	sepal length (cm)	sepal width (cm)	...	petal width (cm)	label
0	5.1	3.5	...	0.2	setosa
1	4.9	3.0	...	0.2	setosa
2	4.7	3.2	...	0.2	setosa
3	4.6	3.1	...	0.2	setosa
4	5.0	3.6	...	0.2	setosa

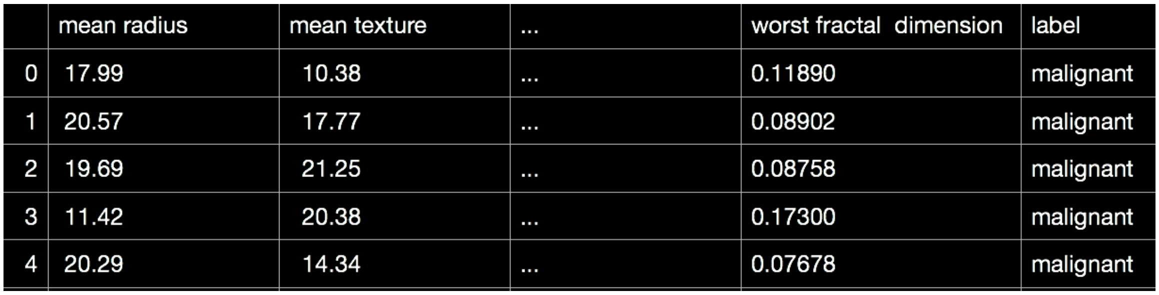
This dataset is often used as a benchmark for multiple machine learning fields.
It is composed of 4 features of flower petals (dimensionality = 4). There are three possible classes of petals,
with 50 samples per class.



From the image above one can see at a glance a histogram of each feature with respect to the three available classes.
Some features distinguish the labels more than others, for instance, the setosa type is more easily identified when
conditioning on petal width.

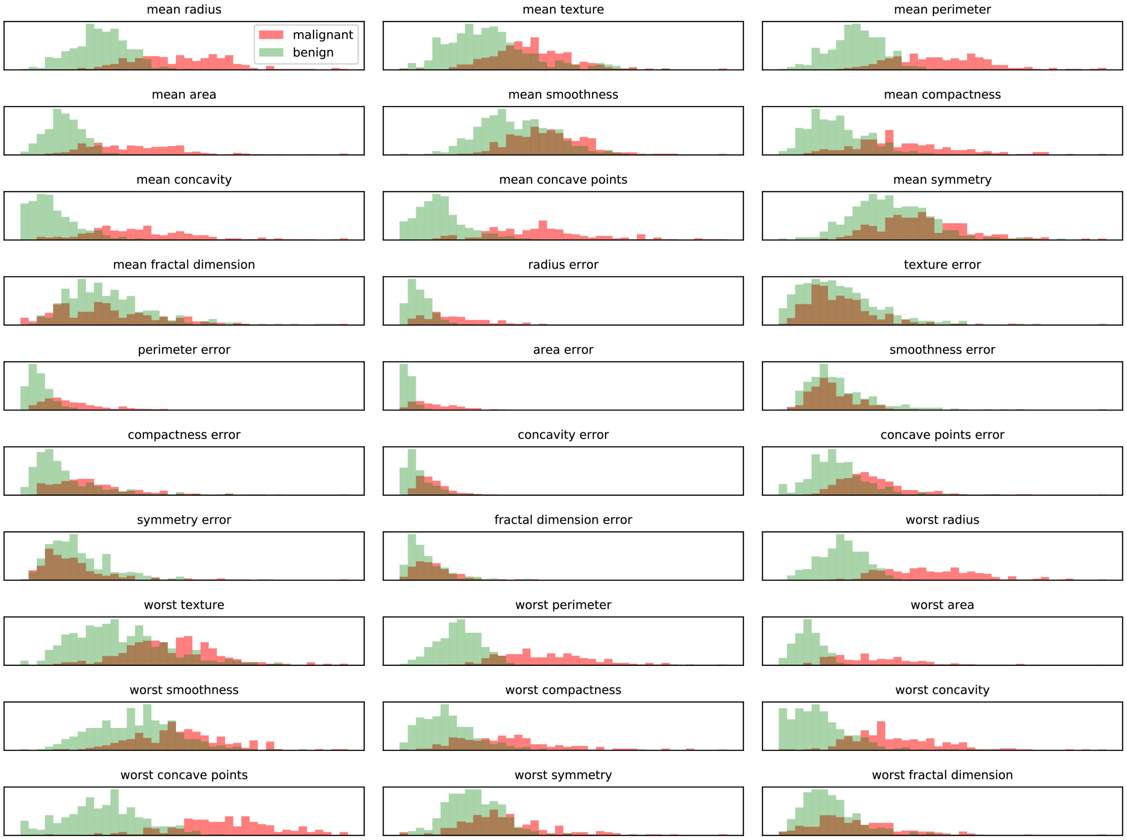
- Breast Cancer (http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic))

The first rows of this dataset's dataframe are shown below:



	mean radius	mean texture	...	worst fractal dimension	label
0	17.99	10.38	...	0.11890	malignant
1	20.57	17.77	...	0.08902	malignant
2	19.69	21.25	...	0.08758	malignant
3	11.42	20.38	...	0.17300	malignant
4	20.29	14.34	...	0.07678	malignant

This dataset represents characteristics of cell nuclei from images of breast mass. It is composed of 569 samples of 30 features for each cell nuclei (dimensionality = 30). There are 2 possible classes for a breast mass, ‘malignant’ or ‘benign’. Some characteristics are: texture, concavity, symmetry.



As in the previous database, from the image above one can see a histogram of each feature with respect to the two possible outcomes. In the same manner, the difference in overlaps along the features can suggest a certain covariance between the classes. These last histograms were generated based on algorithms from https://towardsdatascience.com/dive-into-pca-principal-component-analysis-with-python-43ded13ead21.

—————————

Techniques

- PCA
- t-SNE
- Multi Dimensional Scaling (MDS)

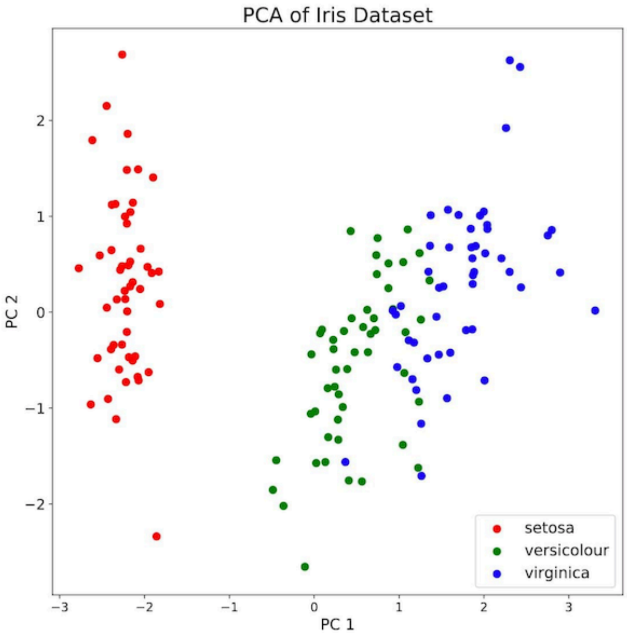
—————————

Results

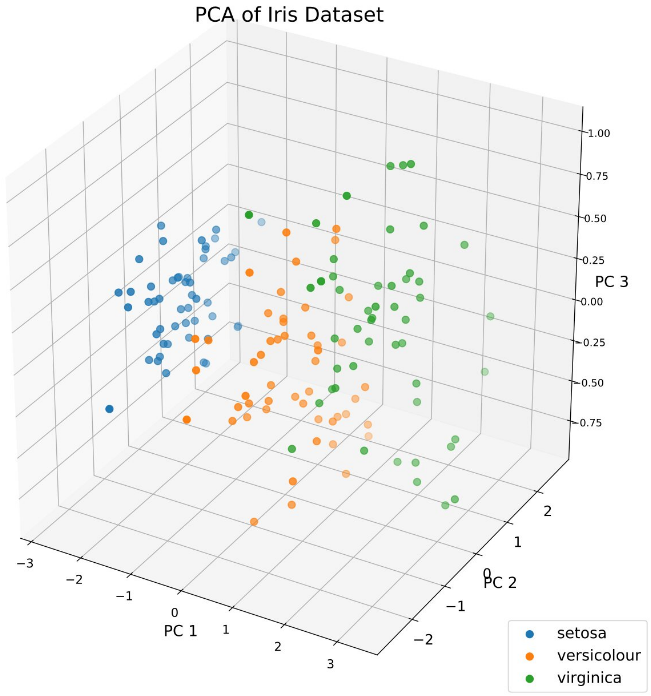
Description of the results from the visualizations

Iris Dataset

PCA

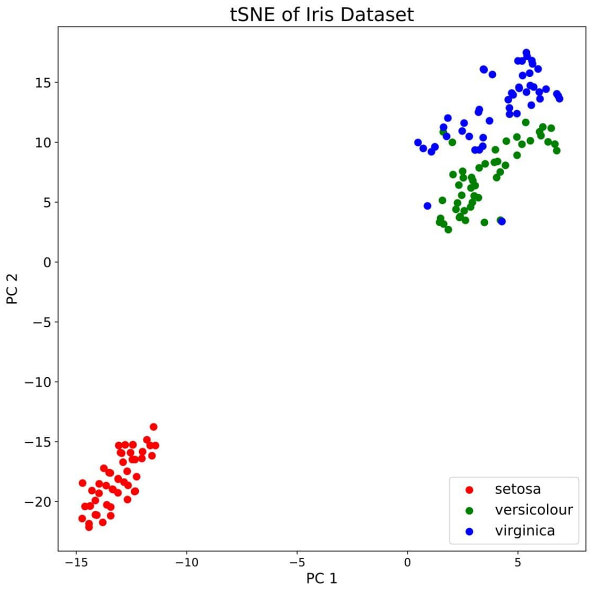


The cumulative explained variation for 2 principal components for this 2D visualization was 0.958. The versicolour and virginica classes are not as identifiable as the setosa variation.

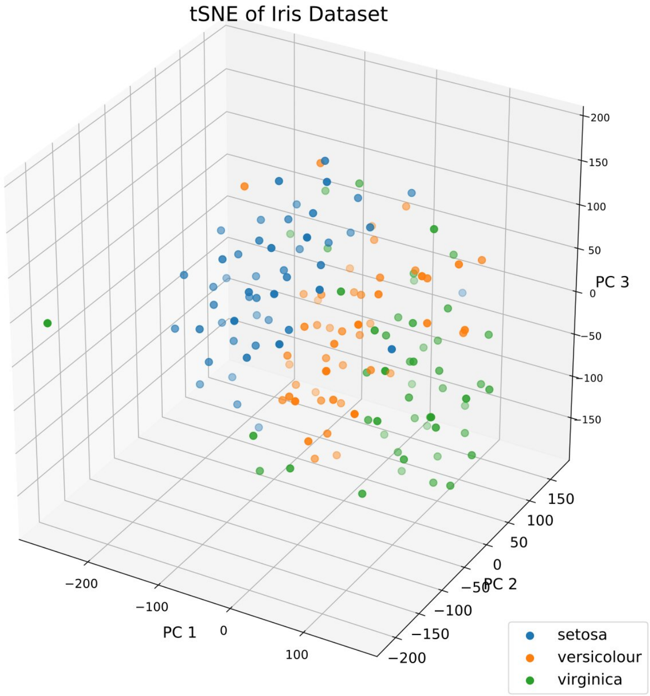


The cumulative explained variation for 3 principal components for this 3D visualization was 0.995, a 4% improvement. In other words, the loss of information is 0.5%. In this case, the classes appear to be more separated, but the 3D visualization is not as clear as its 2D version.

t-SNE

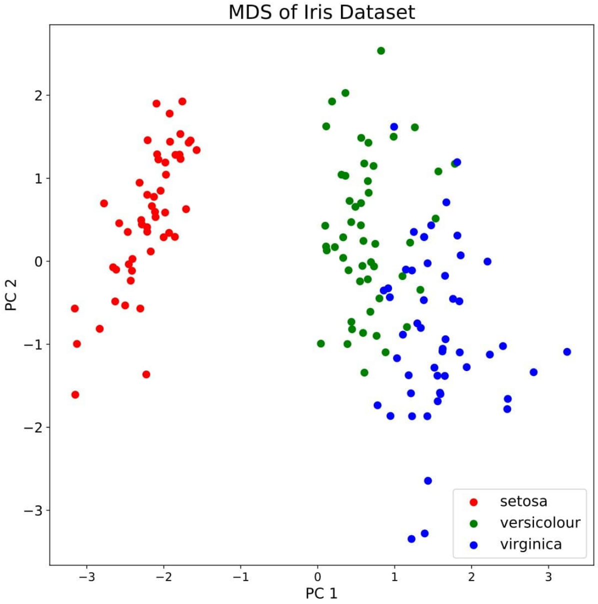


The time taken to calculate this t-SNE reduction was 0.678 seconds. Similar to the last 2D visualization, the setosa class is more identifiable. Also, there are some outliers from the remaining classes in each others areas.

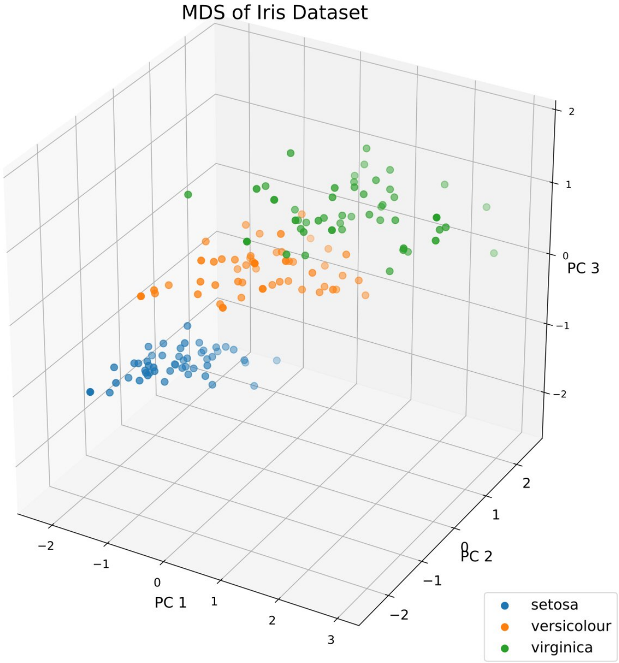


The time taken to calculate this t-SNE reduction with an extra dimension 1.080 seconds, a 60% increase. As in the previous case, the visualization may result more confusing than the one with 1 dimension less.

Multi Dimensional Scaling (MDS)



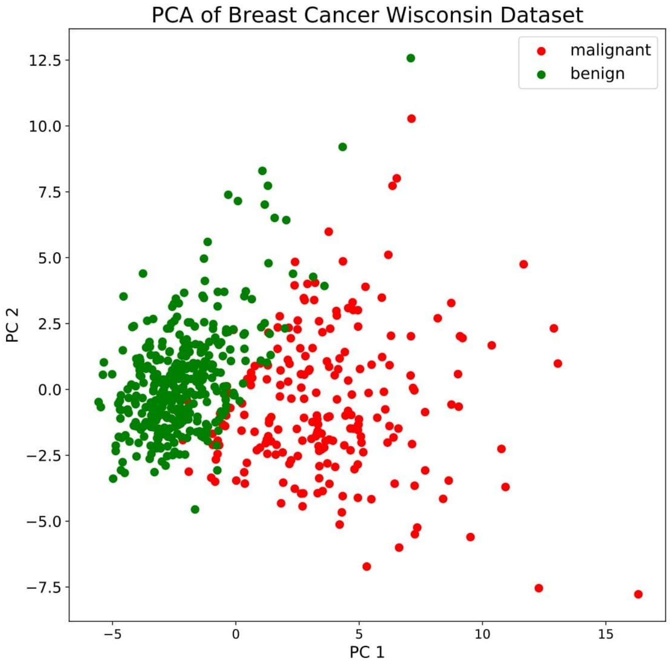
This visualization resembles its PCA variation, though it seems to expand more area, compared to the t-SNE visualization which clearly places setosa far apart form the others.



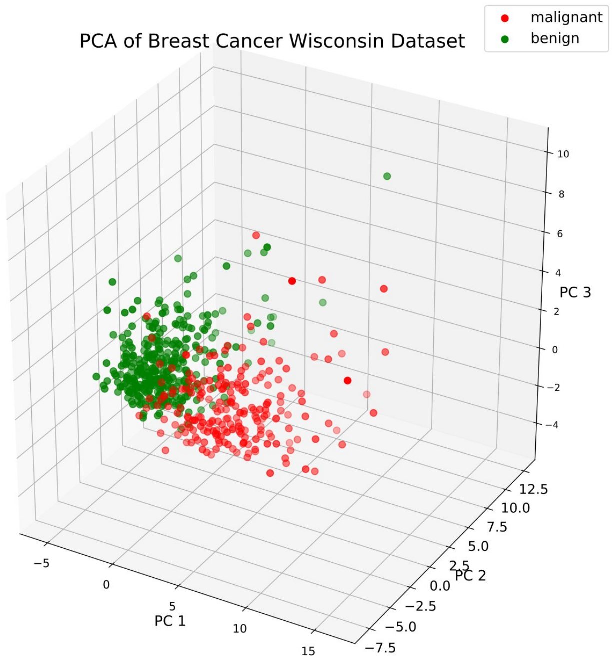
This visualization with an extra dimension seems to be a better representation of the data, with the datapoints arranged as in ‘layers’ in 3D space.

Breast Cancer Dataset

PCA

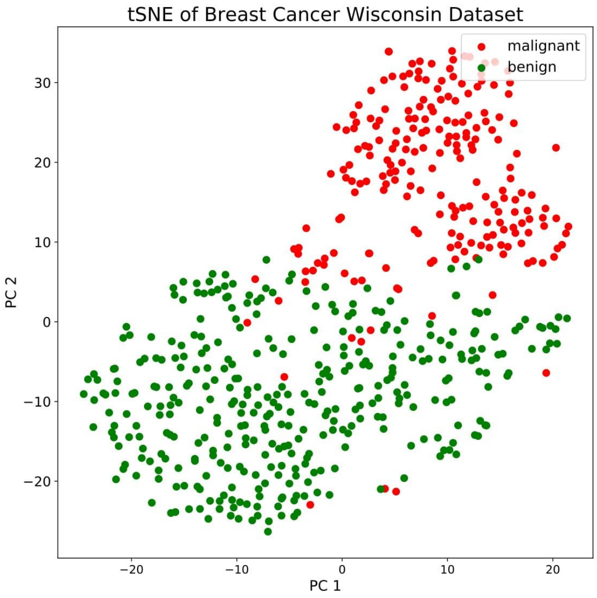


The cumulative explained variation for 2 principal components for this 2D visualization was 0.632. The malignant and benign data points are dispersed around distinct areas, with a certain section that borders both.

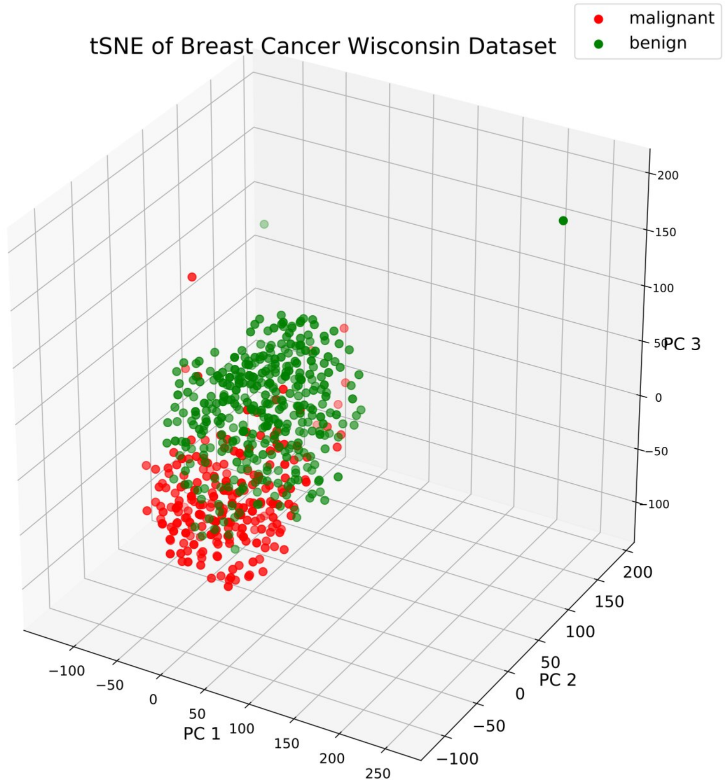


The cumulative explained variation for 3 principal components for this 3D visualization was 0.726, a 15% improvement. In other words, the loss of information is about 30%. In this case, outliers seem to be more clear, but the perspective does not make the visualization more clear than its 2D version.

t-SNE

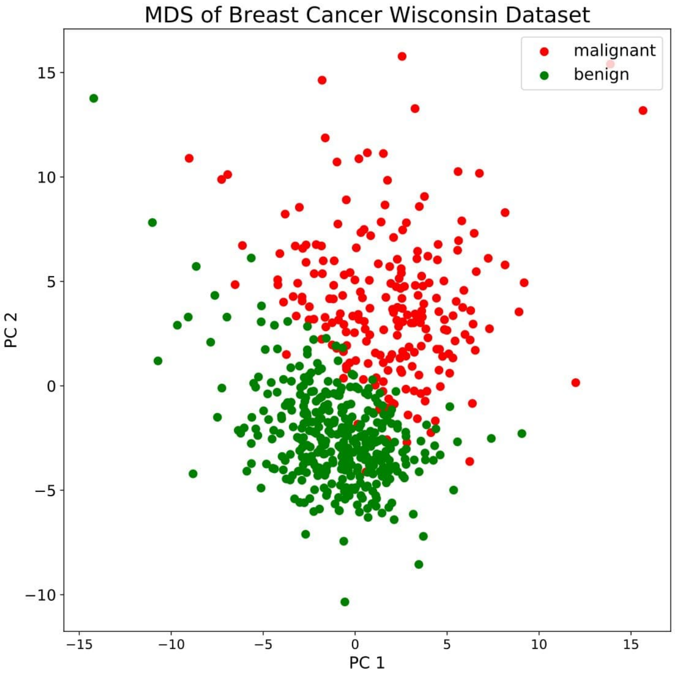


The time taken to calculate this t-SNE reduction was 2.711 seconds. This 2D visualization places the classes further apart in general terms, though some malignant samples are around benign points.

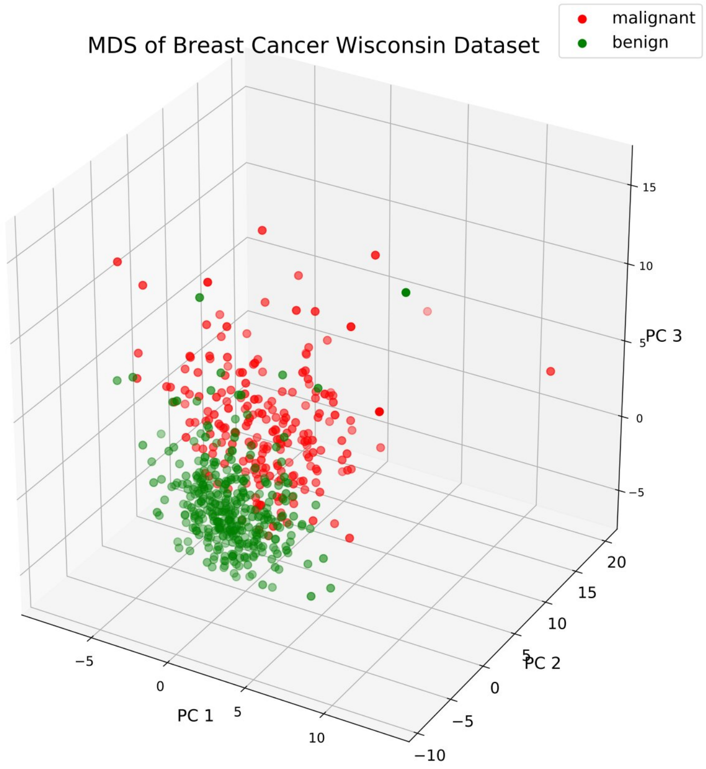


This 3D variation displays clusters for each label, and took 5.932 seconds to calculate, more than double the effort. Though not clear from the current perspective, the extra component seems to separate the classes on the z-index.

Multi Dimensional Scaling (MDS)



This MDS visualization splits the data in two general areas, with data points spreading from each cluster, which seem to be the best representations of the characteristics in the histograms from above.



This 3D visualization adds another component to the previous visualization, still portraying a cluster-like separation, but the sparsity of most datapoints does not provide a clearer picture than the 2D alternative.