Project

Pattern Recognition

Jonasz Kulpinski EP1418

**Step 1 – Choosing data**

I chose Haberman’s Survival Data Set.

Features:

1. Age of patient at time of operation (numerical)

2. Patient's year of operation (year - 1900, numerical)

3. Number of positive axillary nodes detected (numerical)

Classes:

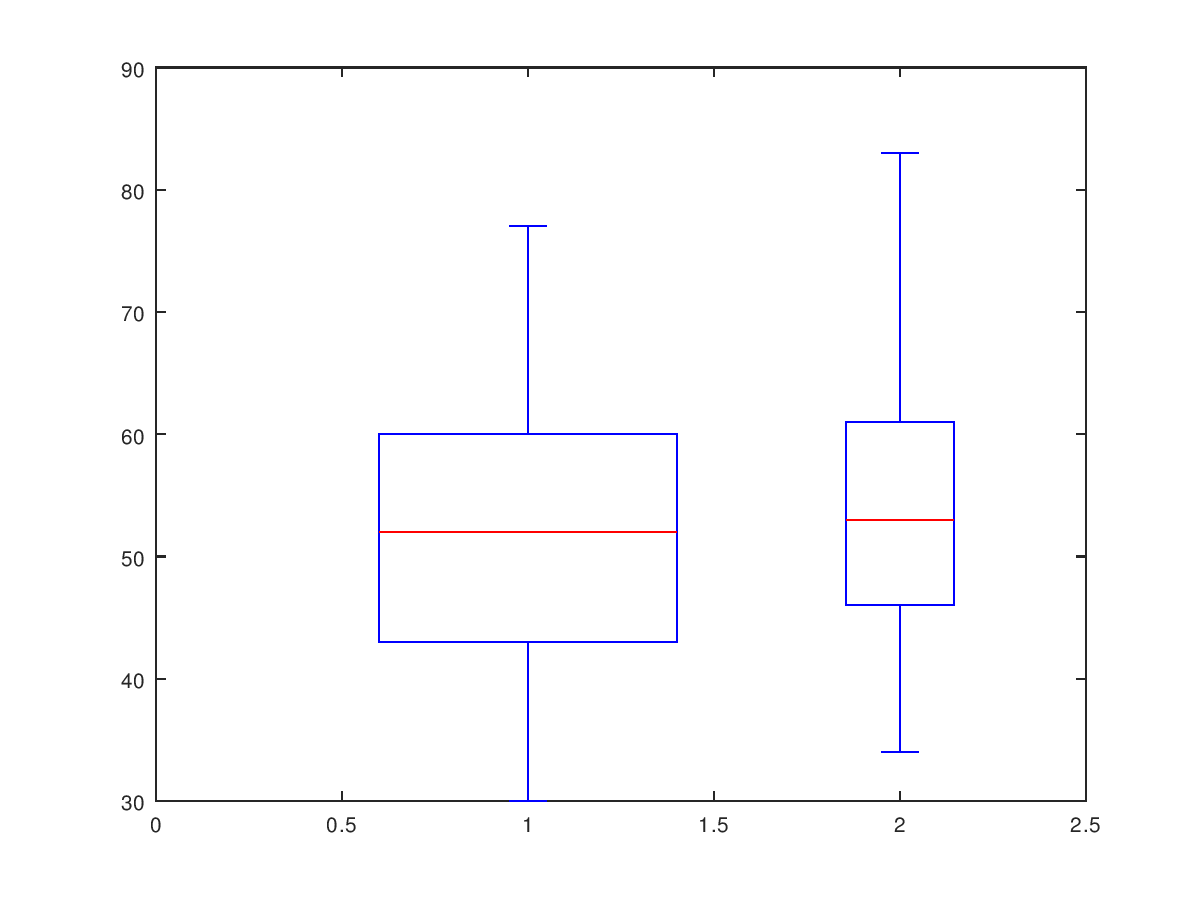
1 = the patient survived 5 years or longer

2 = the patient died within 5 year

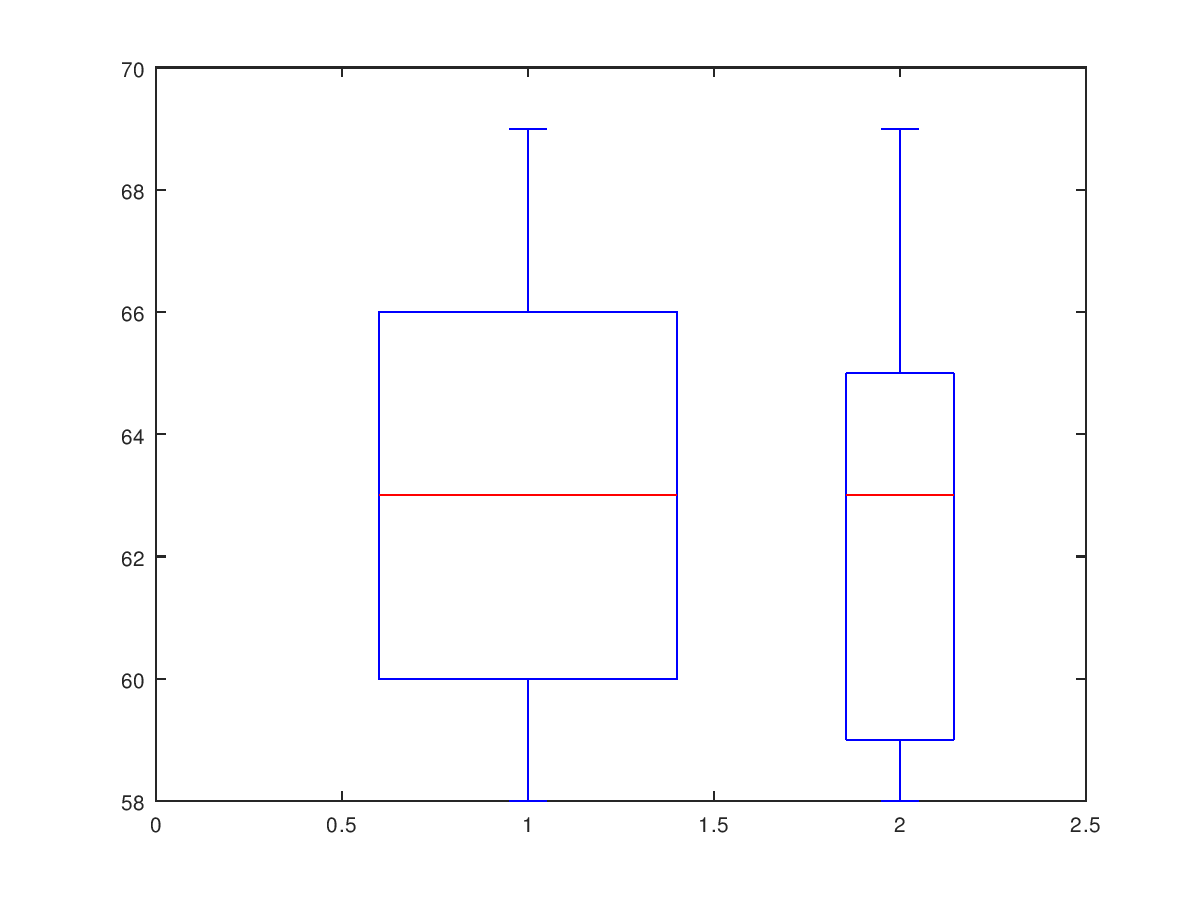
**Step 2 - Statistical Analysis**

After loaded dataset I sorted data by class. There are 225 instances of class (1) and 81 - class (2).

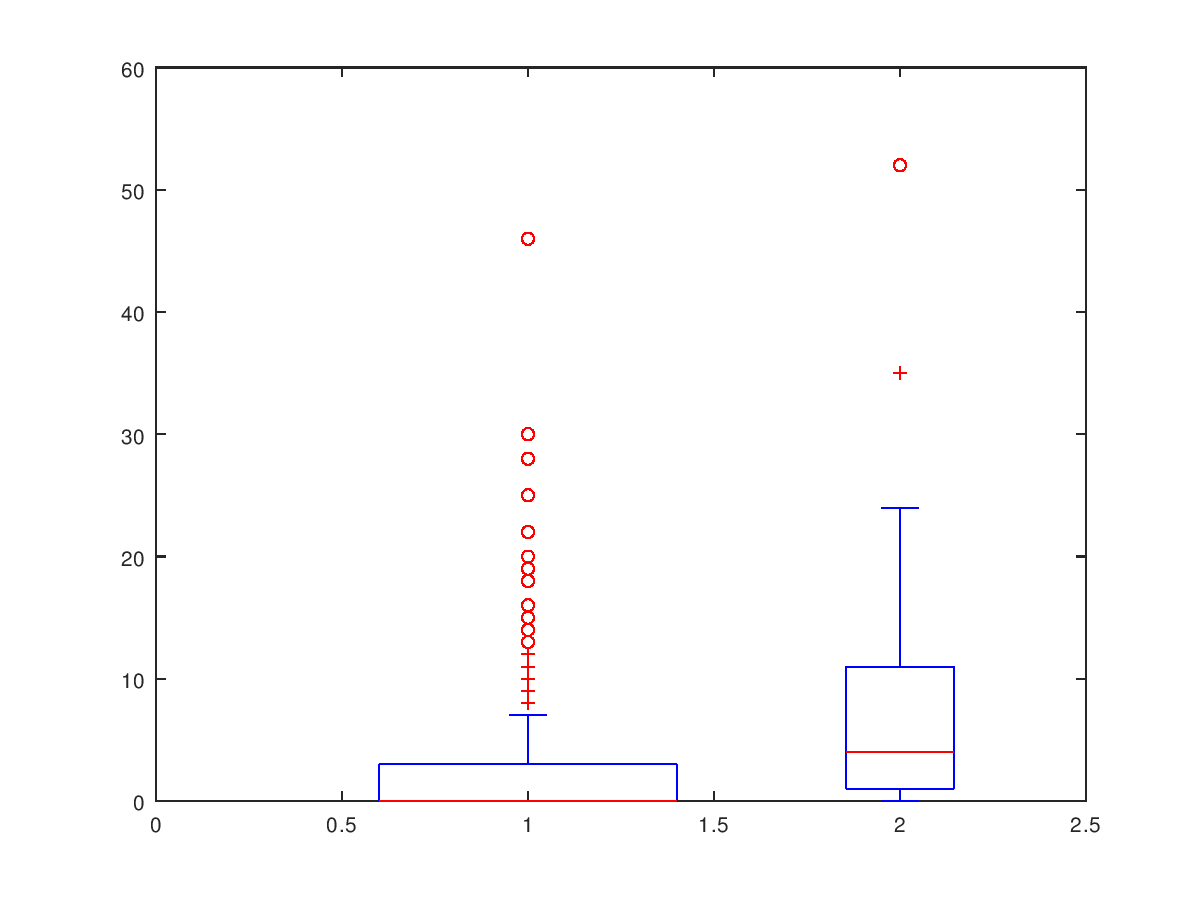
Later, I made boxplots every of the 3 features:



1. Age of patient at time of operation



2. Patient's year of operation



Number of positive axillary nodes detected

**Conclusions:**

People who have lived for more than 5 years (1st class) are 2,7 times more than those from the 2nd class. The first boxplot shows that the median age of people who underwent surgery is similar for both classes (about 52). However, people representing class 2 were a little older. The year of operation for both classes has the same median (1963). The difference can be seen in the first and third quartiles, they have lower values for the 2nd class. Differences occurred on the boxplot of number of positive axillary nodes detected. The median for 1st class is 0, although there have been cases of even tens nodes for individual results. For patients who lived less than 5 years after surgery, the average number of nodes is definitely higher. The conclusion is the higher result for this feature is characterized by class 2.

The best feature is number of nodes (feature 3) because we can see clearly the difference between the results for classes. In feature 1 and 2 there are also differences, however smaller than in feature 3.

**Step 3- Preprocessing data**

There’s no missing values in haberman’s dataset.

I made matrix with informations about instances class and deleted class column from dataset.

Later, found 5 outliers, removed them and transformed data into [0 1].

Then split dataset into training and validation set : 80-20%, 70-30%, 90-10%, the best accuracy was for 80-20% so I used this split percentage in next measurements. The biggest error was for 70-30%.

I made gscatters for **features: patients age, number of nodes**. These two features will be use in next learning methods too.

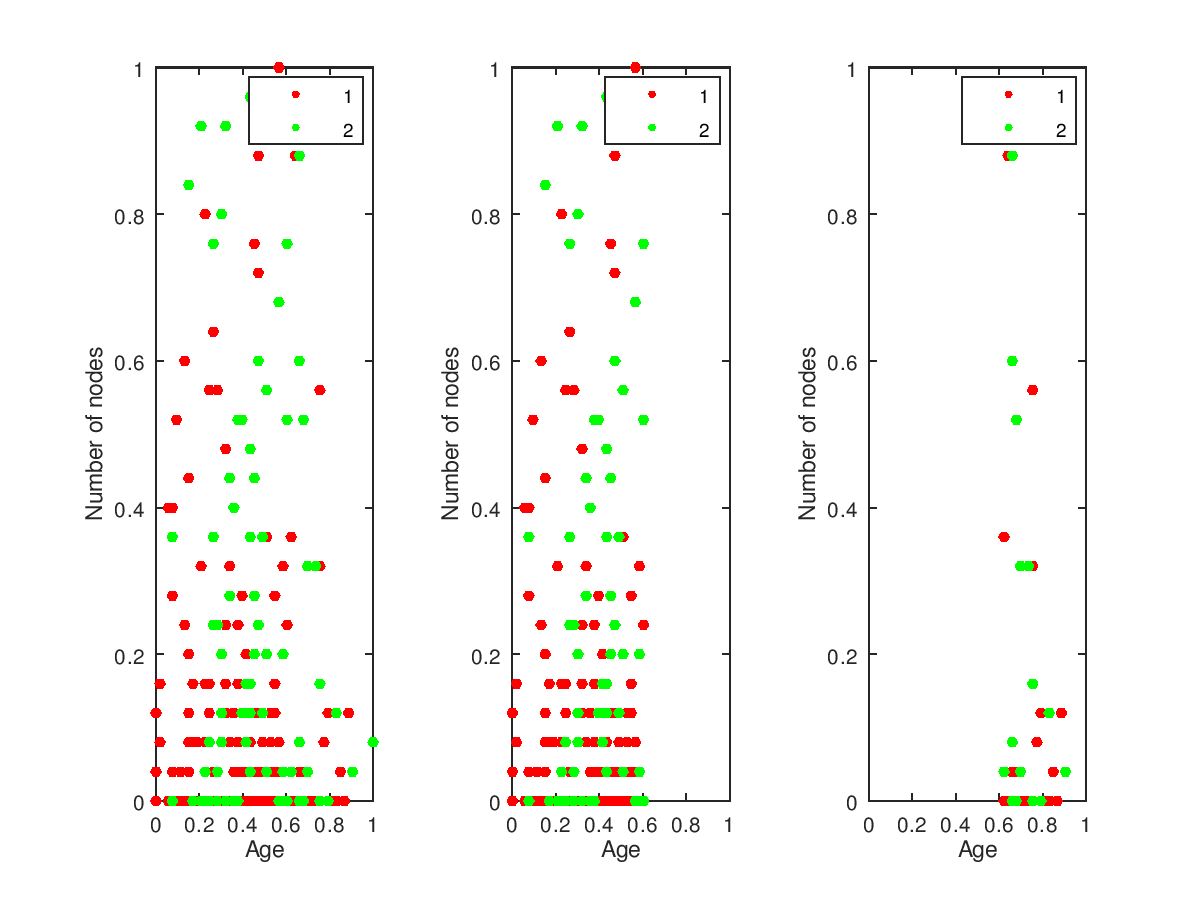
Three subplots in gscatter:

1 subplot: instances in dataset

2: training set

3. validation set

**Legend: 1 – survived more than 5 years , 2 – died within 5 years**



**Conclusions:**

On plots we can see larger number of instances belonging to class1 of the dataset than to class2. A larger number of nodes for class 2 is also visible.

For the training set, instances with a lower value of the age feature were selected and to the validation set – larger.

**Step 4 – Classifiers**

I implemeted linear and quadratic classifier.

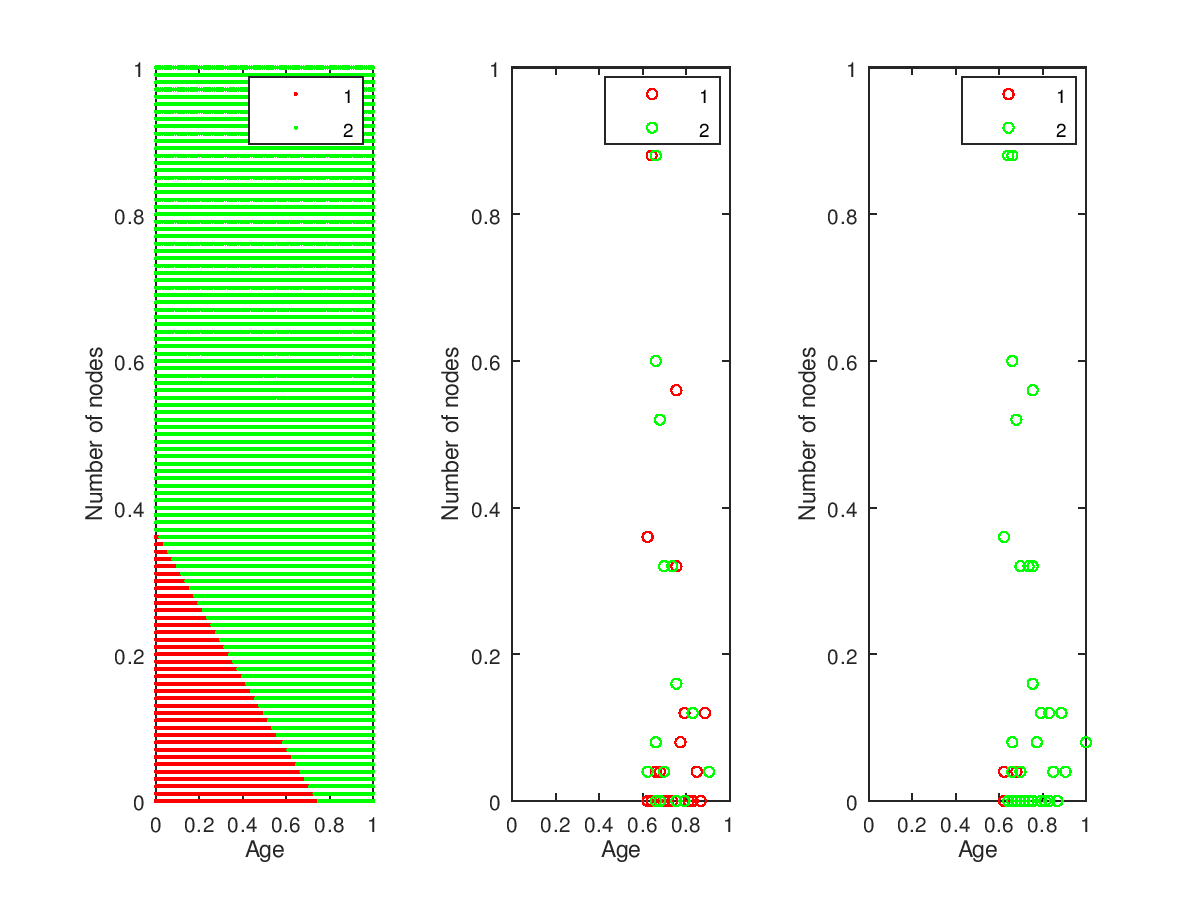
Error (errn):

**Linear: 0,25**

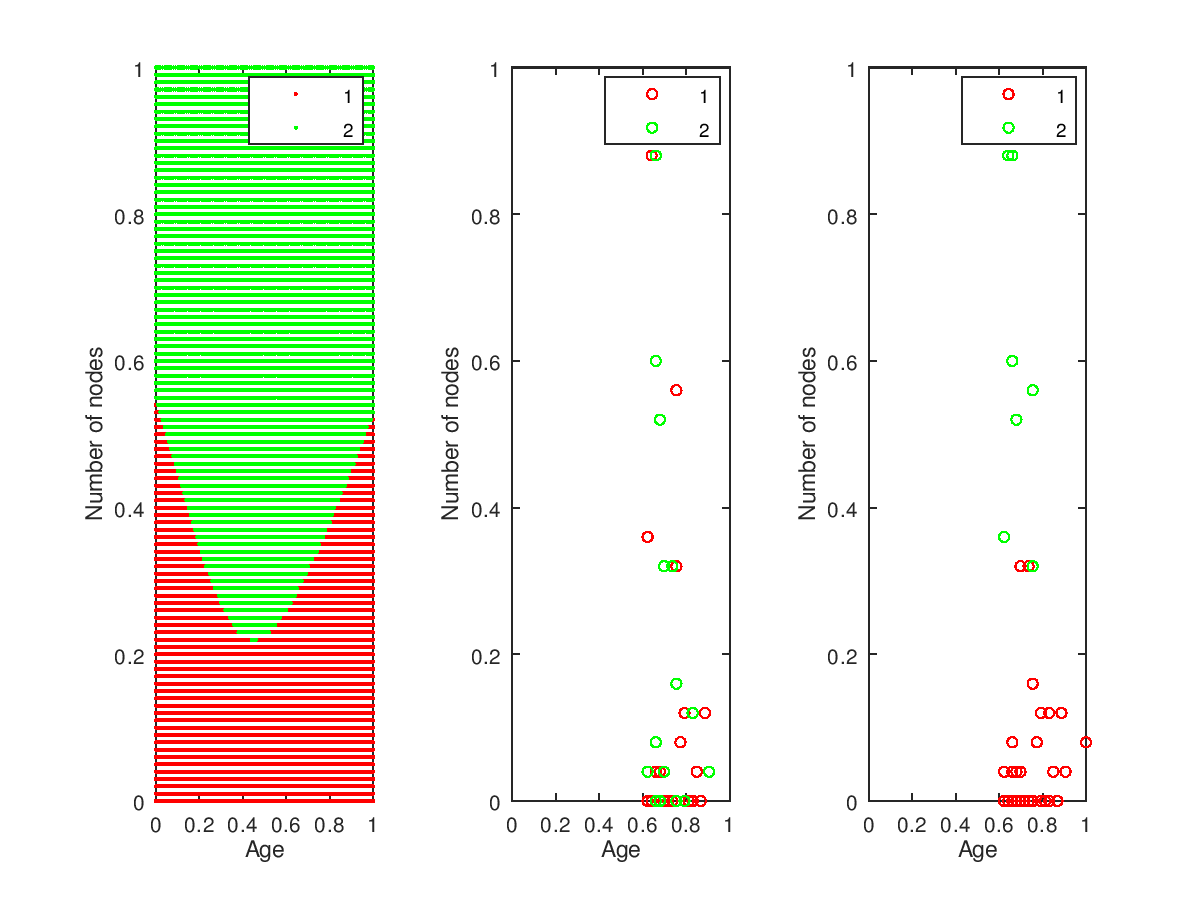
**Quadratic: 0,2375**

Visualization classification results and Validation Set before and after classification:

**Linear classifier:**



**Quadriatic Classifier:**



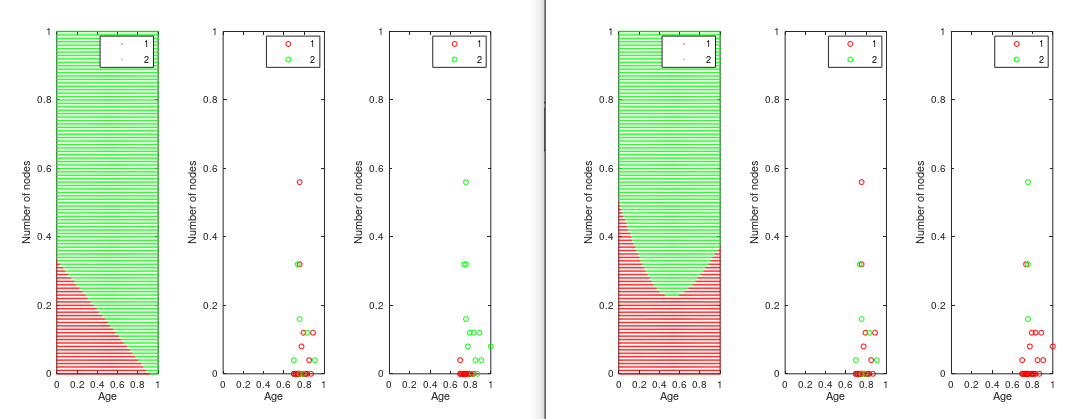
**Conclusions:**

There were differences between the results of classifiers. For the linear, only a few instances, with a small number of nodes Validation Set after classification belongs to class 1. Most is in Class 2, unlike Quadratic classifier.

Quadratic classifier has lower error and the result is better for this classifier as we can see on plots. To class 1 belong instances with a small number of nodes, and to class 2 those with a larger number, what agrees with gscatters.

I used both classify methods with another percentage splits of dataset. The results were similar but more instances belongs to class 1 in validation set after using linear classifier:

90-10%, linear and quadratic classifier gscatters:

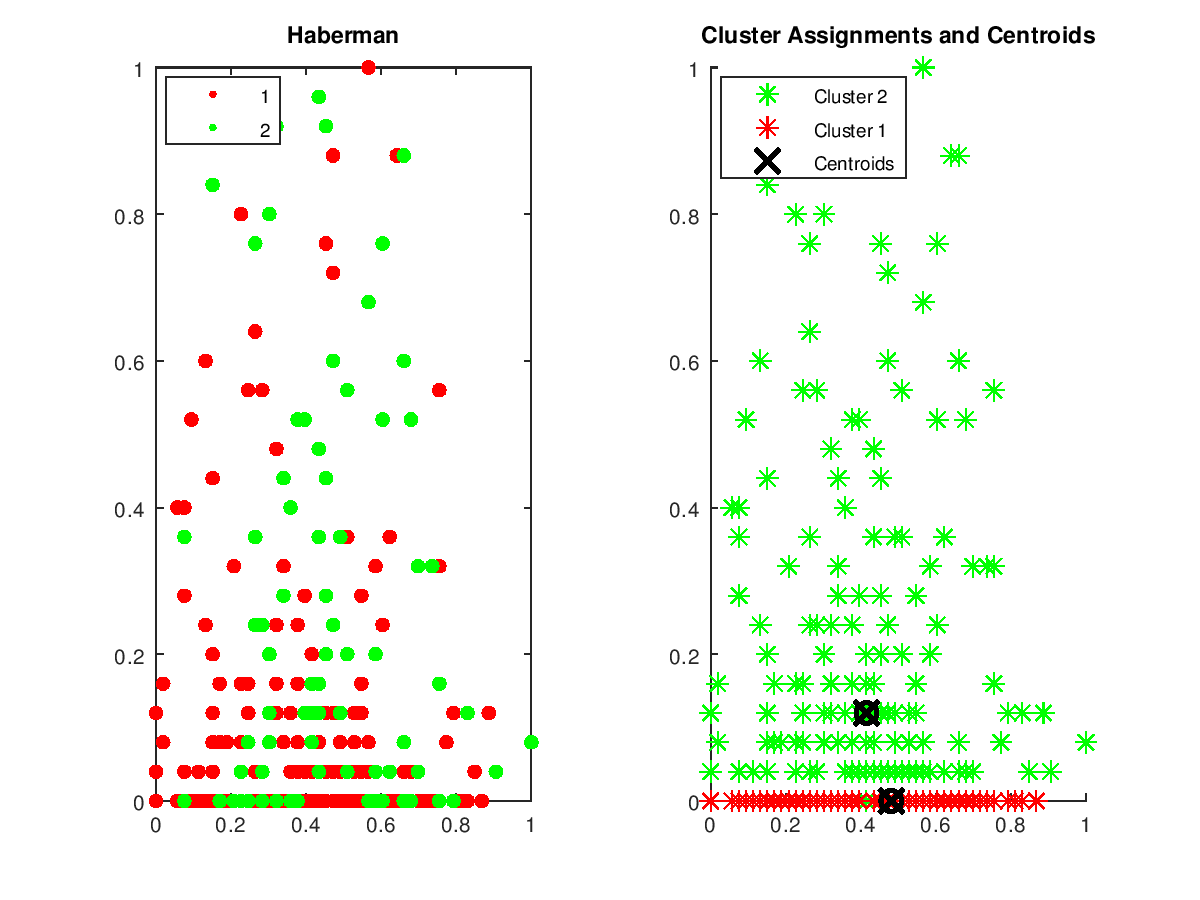


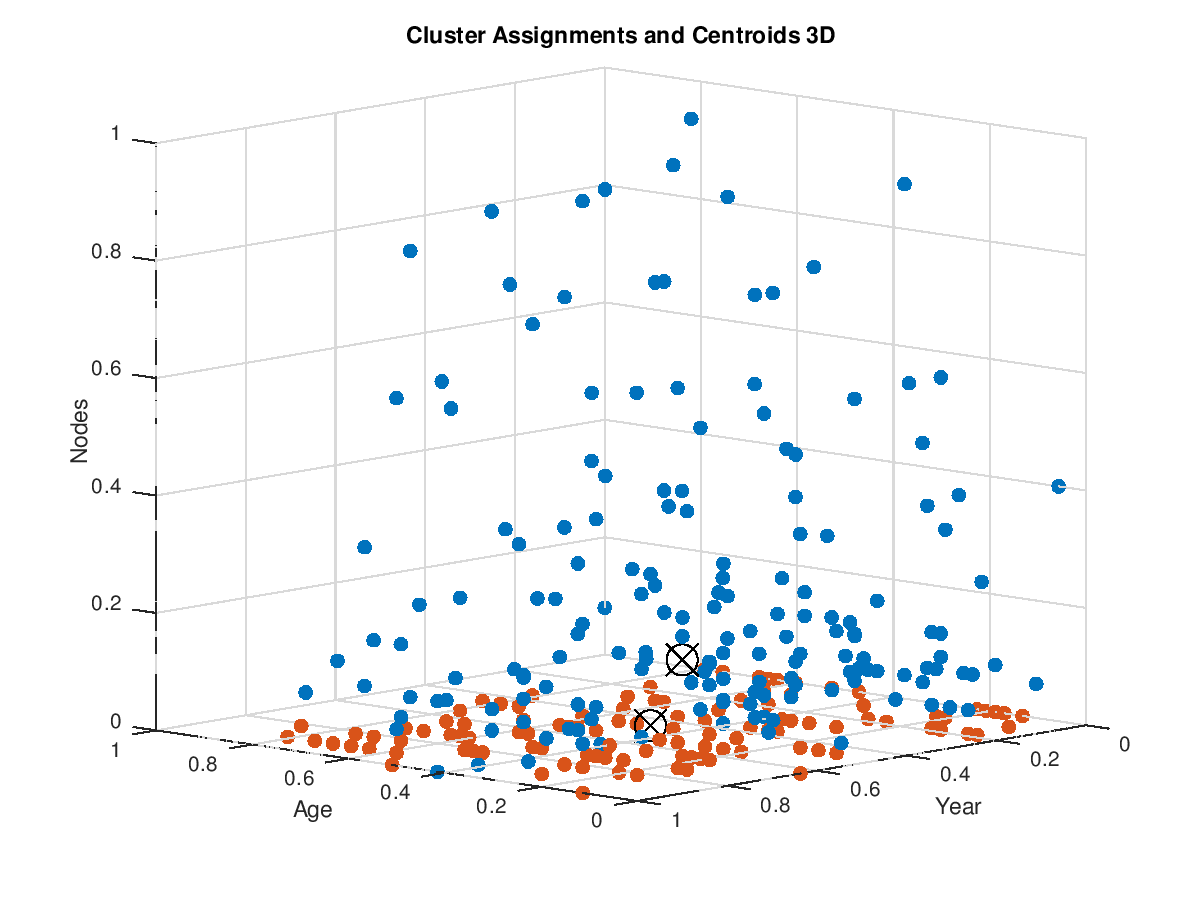
**Step 5 -Kmeans**

I used K-means alogrithm to cluster data.

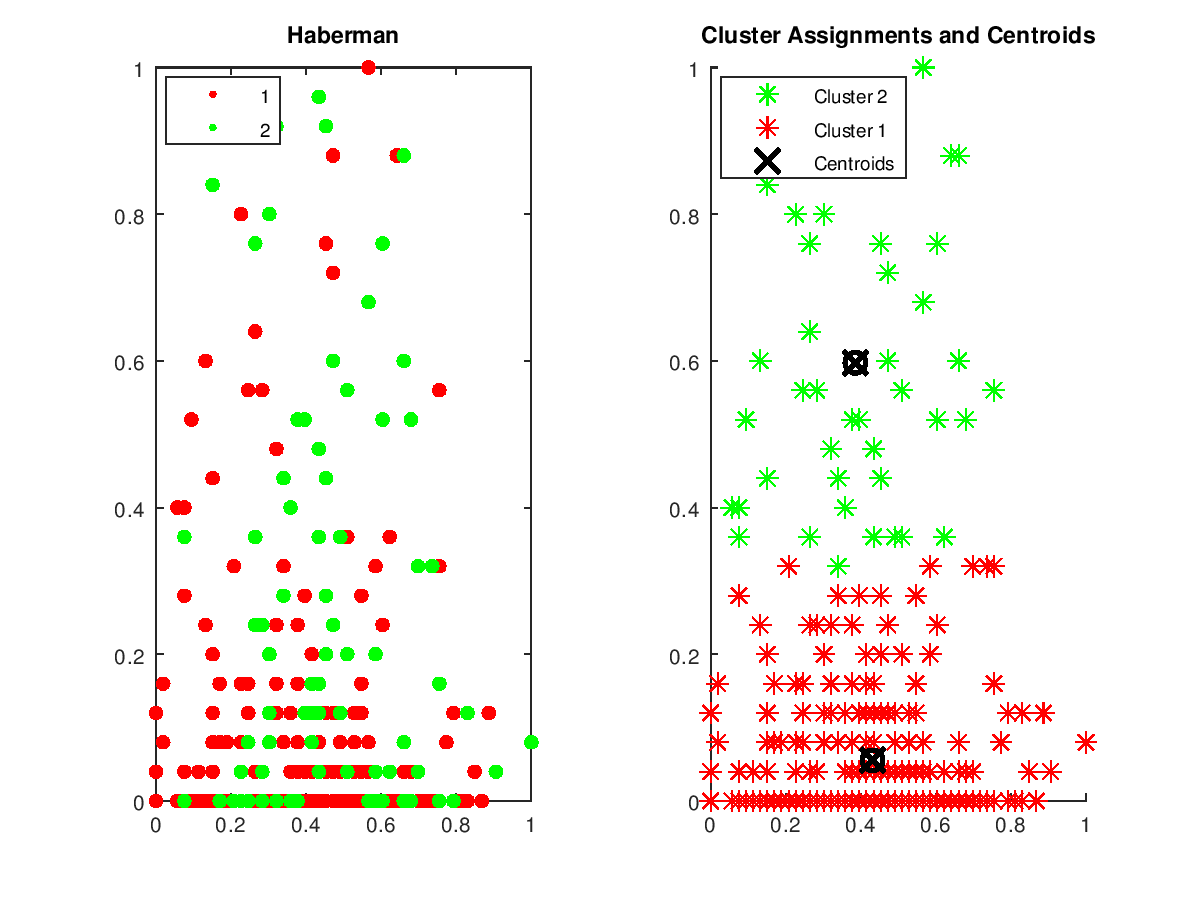
Firstly I changed number of **clusters to 2** and distance measure setting to **‘hamming’.**

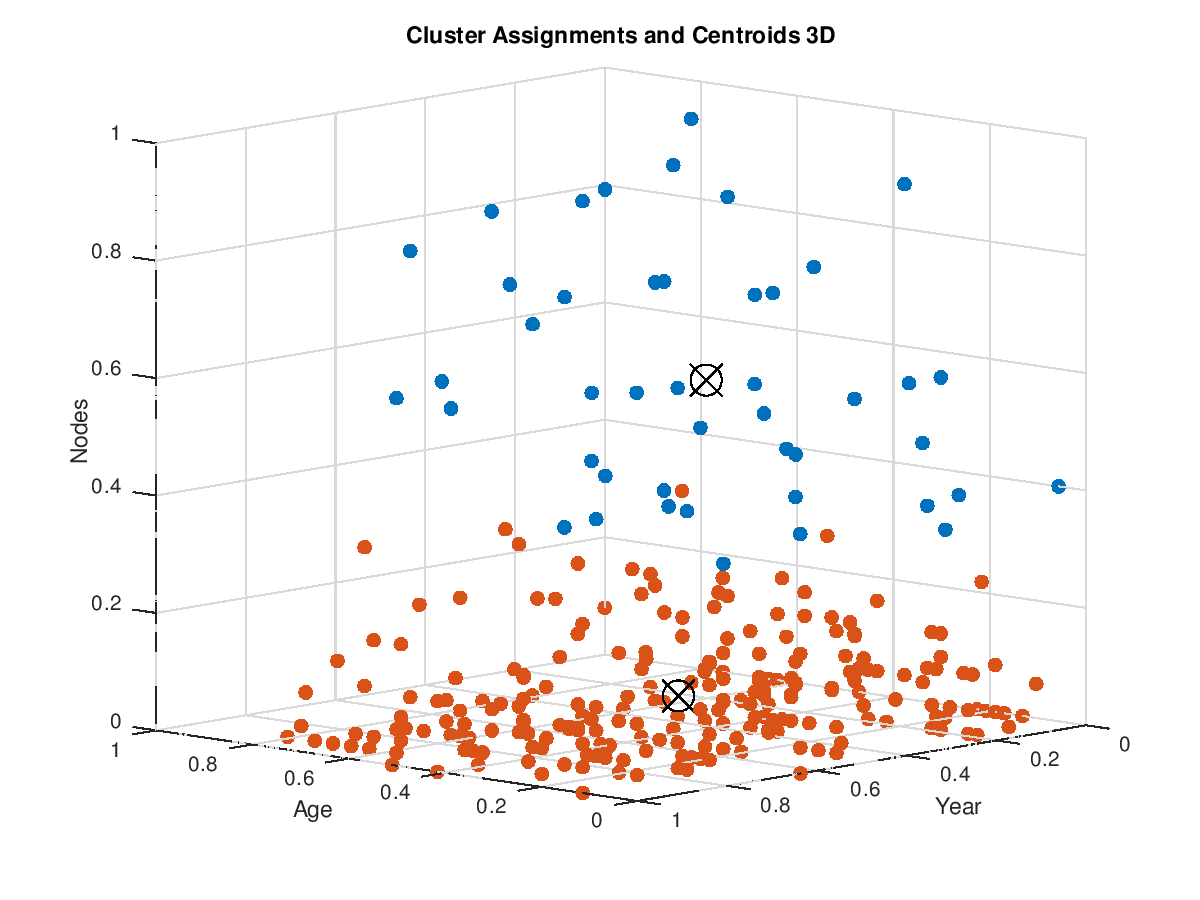
**Legend: 1-class1 (live more than 5 years), 2- died in 5 years.**



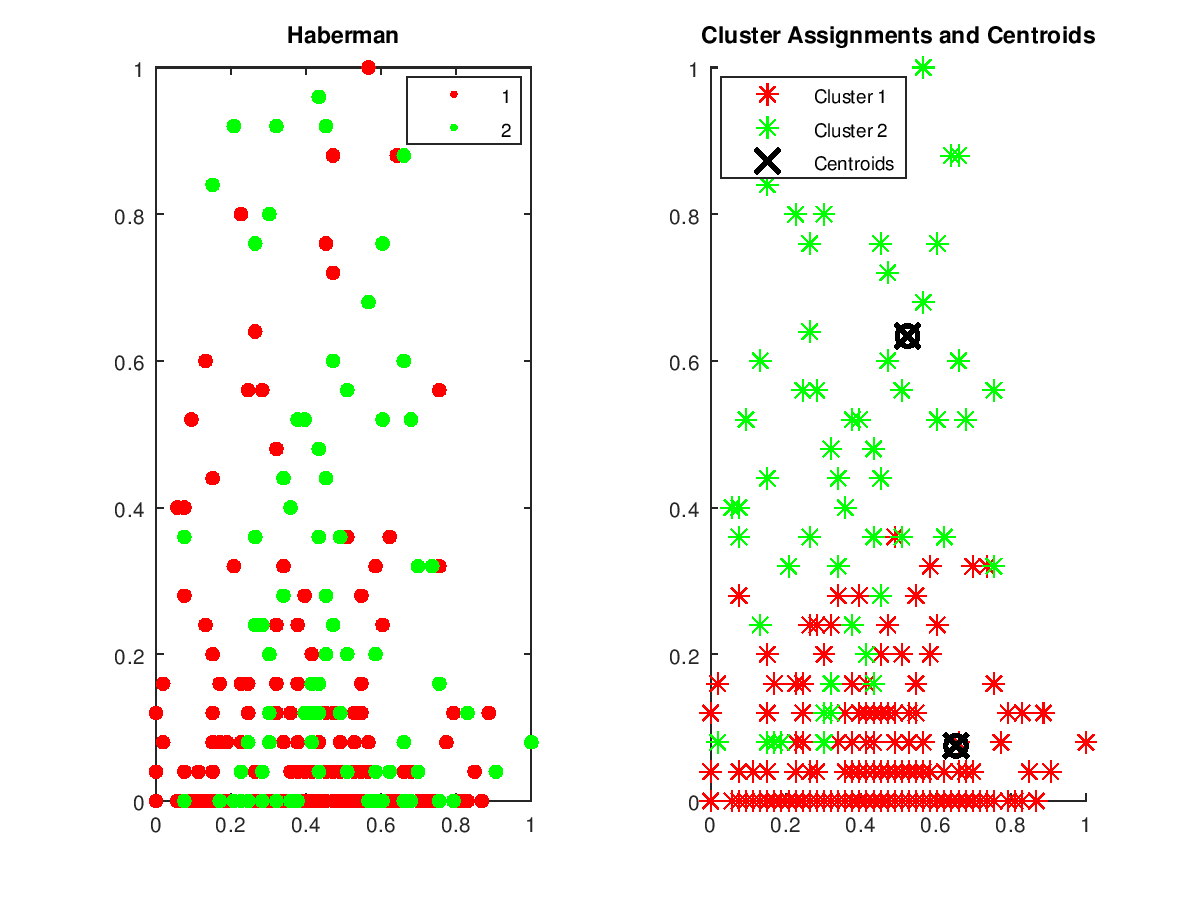


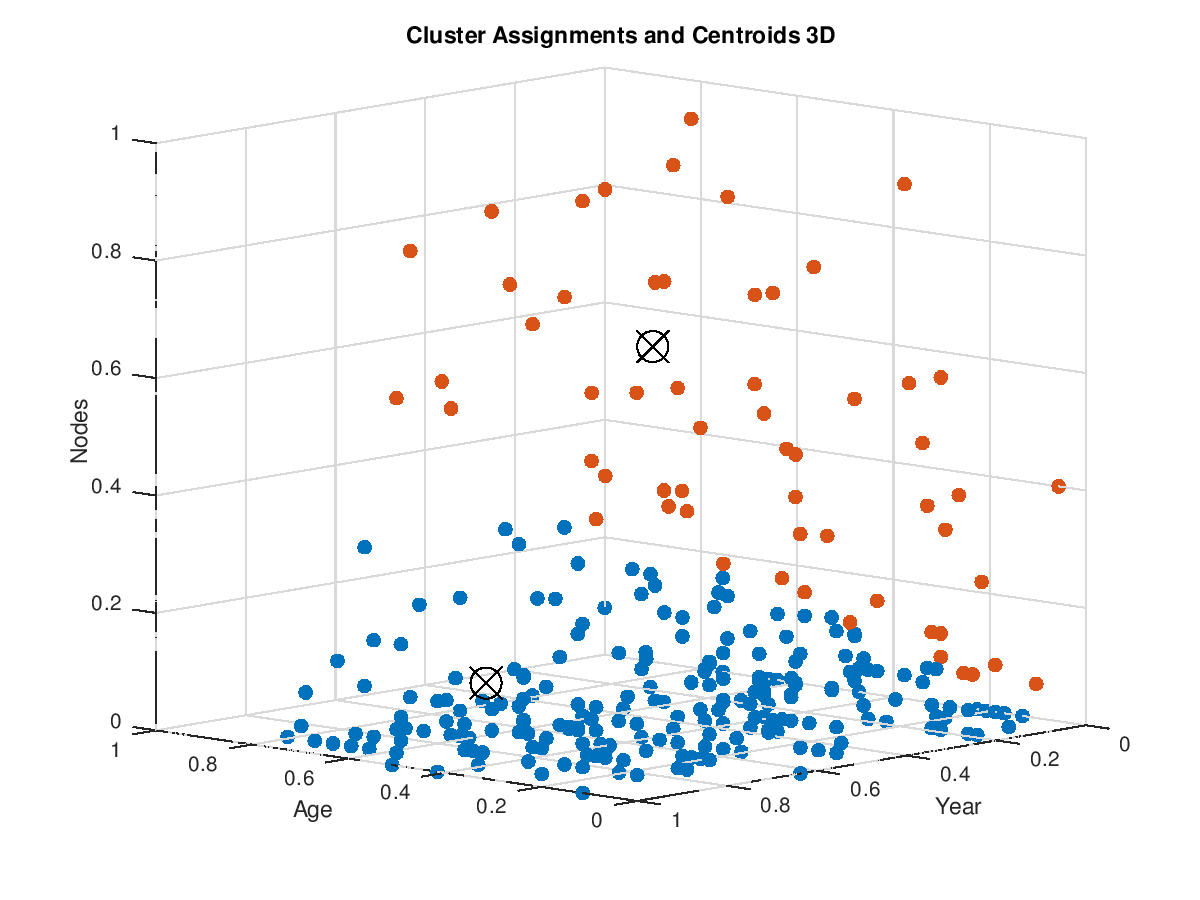
**Clusters: 2. Distance: ‘sqeuclidean’**



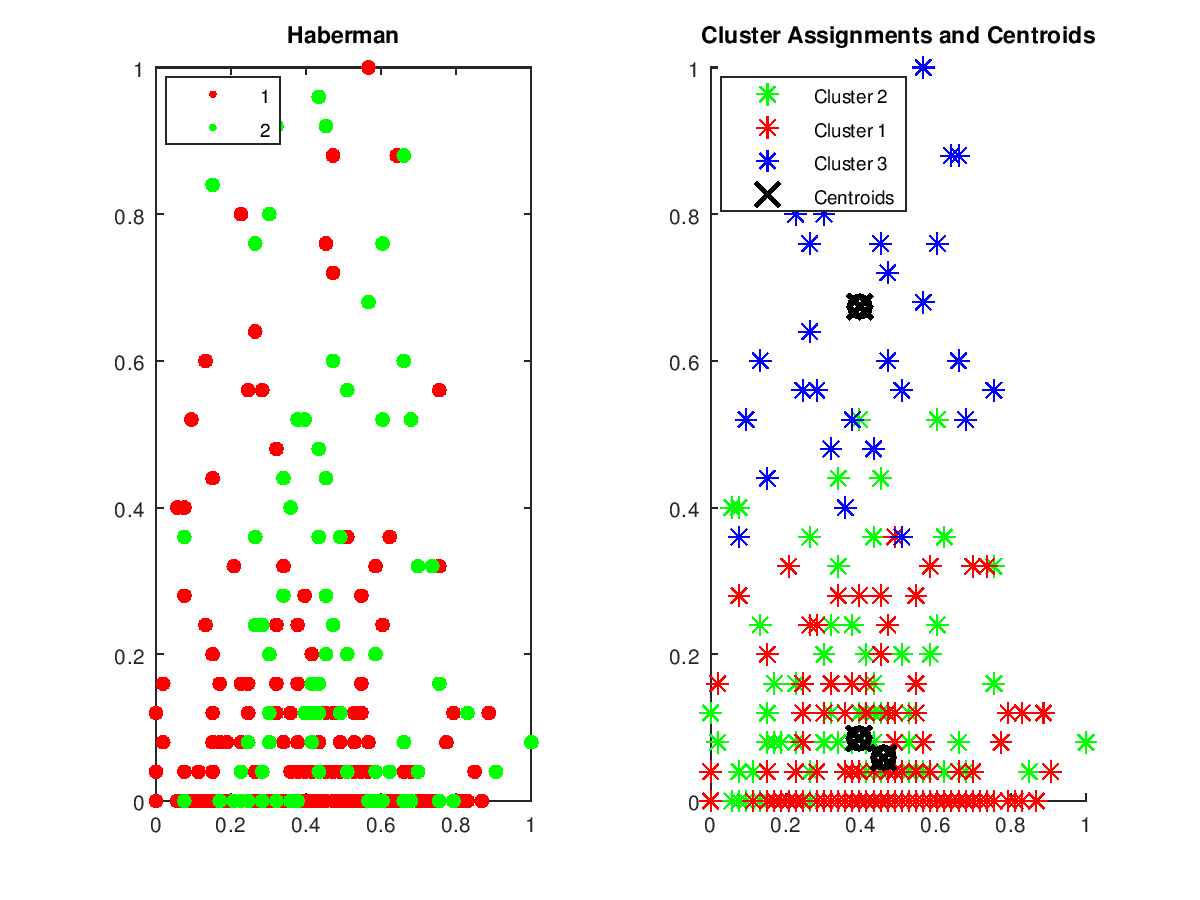


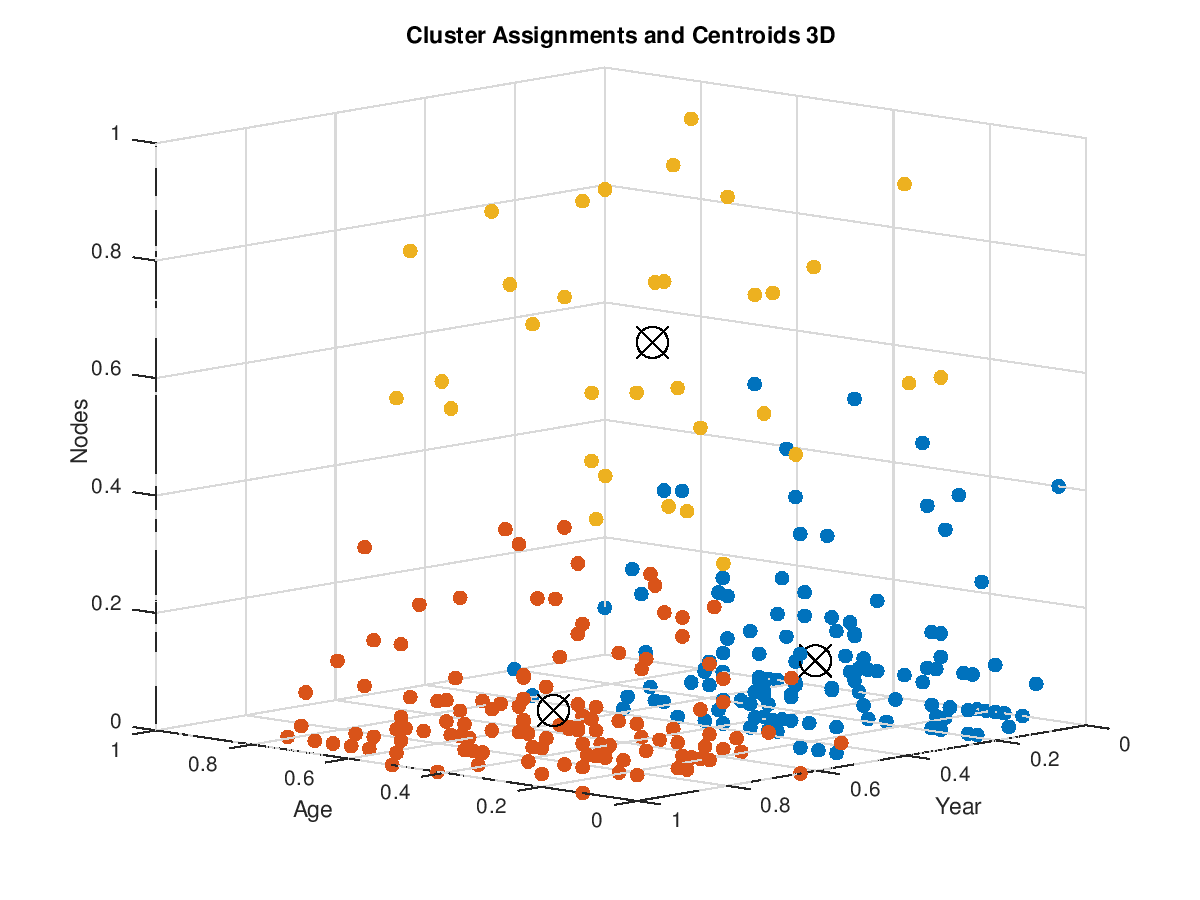
**Clusters 2, method: ‘cosine’**





**Custers: 3. Distance: ‘sqeuclidean’**





**Conclusions**:

In kmeans using the Hamming distance with creation of 2 clasters, you can see that the class 1 cluster was created close to zero nodes because many of the class 1 has very low values of these nodes, and class 2 above, because their values in this class are higher.

After changing the Euclidean distance, cluster 1 includes instances with a higher value of nodes feature than before, but it does not seem bad because up to 0.4 on the vertical axis there are still many instances of class 1.

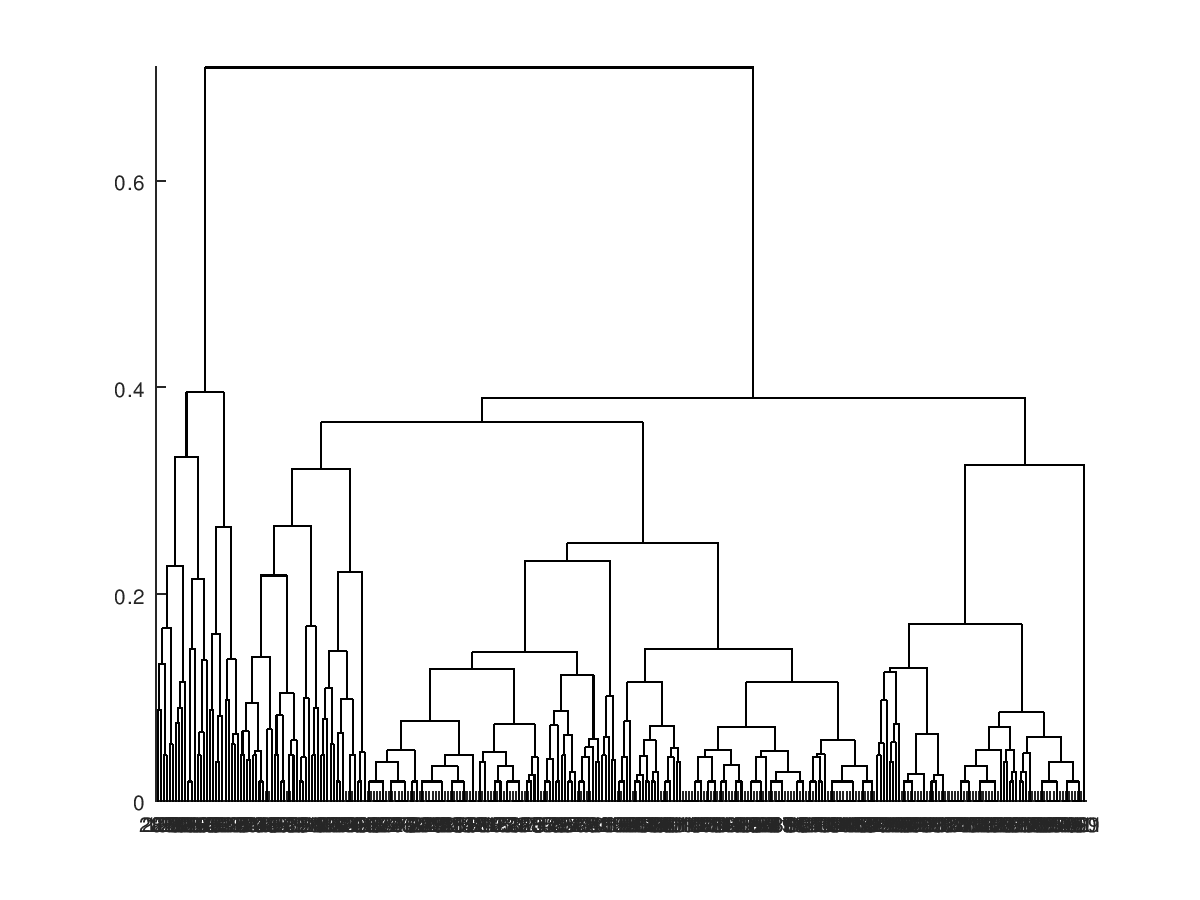
In the experiment with the creation of 3 clusters, cluster 1 with a small node value, was divided into 2 clusters. The division is on the axis of the year of operation.

**Step 6 – Hierarchical Grouping**

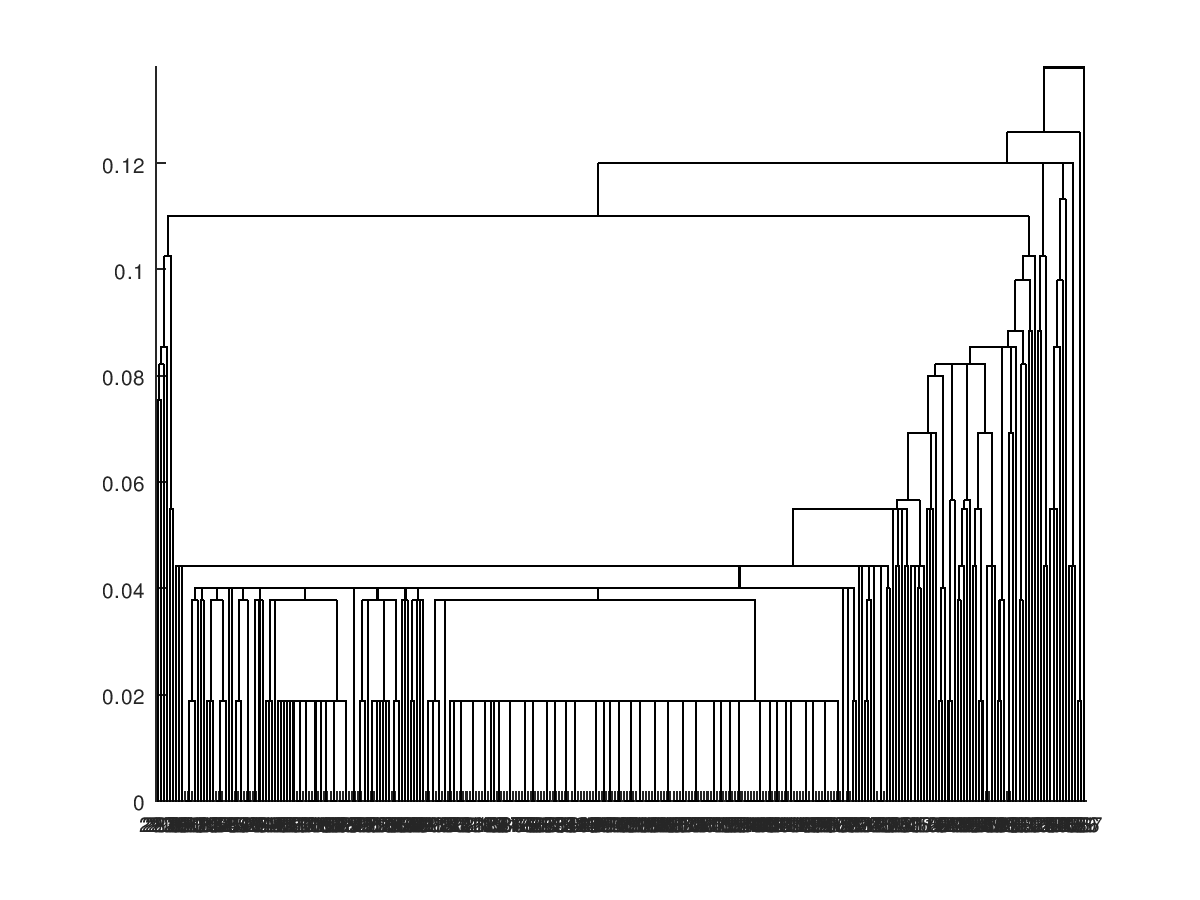
I used linkage alghoritm to create clusters and build the dendrogram for **features age and number of positiv nodes.**

Dendrograms:

Distance method: Euclidean ,linkage method: ‘average’



Distance: Euclidean , method : default



**Conclusions:**

The largest node in first dendrogram (method: average) tree is much higher than the other nodes, so there are two large groups. The first of the 2 major nodes which is on the left contains many instances in a small range of x axis values. Second major node which is on the right contains instances with a large range of values on the x-axis. As in the previous methods, there are 2 groups, one of which contains a lot of instances in a small area like class 1, the other one in a larger area like class 2.

On the second dendrogram (method:default) near to the beginning of coordinate system is one narrow high node. Then nodes high is growing with value on the x-axis.

**Step 7 – Discussion**

Both supervised and unsupervised learning methods gave satisfied results of classification collections in the data set. The best results were given by quadratic classifier and kmeans. Both recognized a large number of instances with a small number of nodes as class 1, and those with a large number as class 2. Linkage dendrogram confirming also what was visible before: there are two large, separate groups. Linear classifier after classifying the majority of validation set as class 2, which was a less correct result compared to Quadratic.

Distinct differences occurred between results using similar methods. Linear and quadratic differed in results, similar to K-means (Euclid and Hamming methods).

In general, based on the results of learning algorithms, it can be concluded that the number of axillary nodes can affect the life expectancy of patients after breast cancer surgery. Those who survive longer have a small number of nodes in relation to those living shorter.