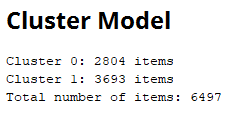
Nick Iudiciani

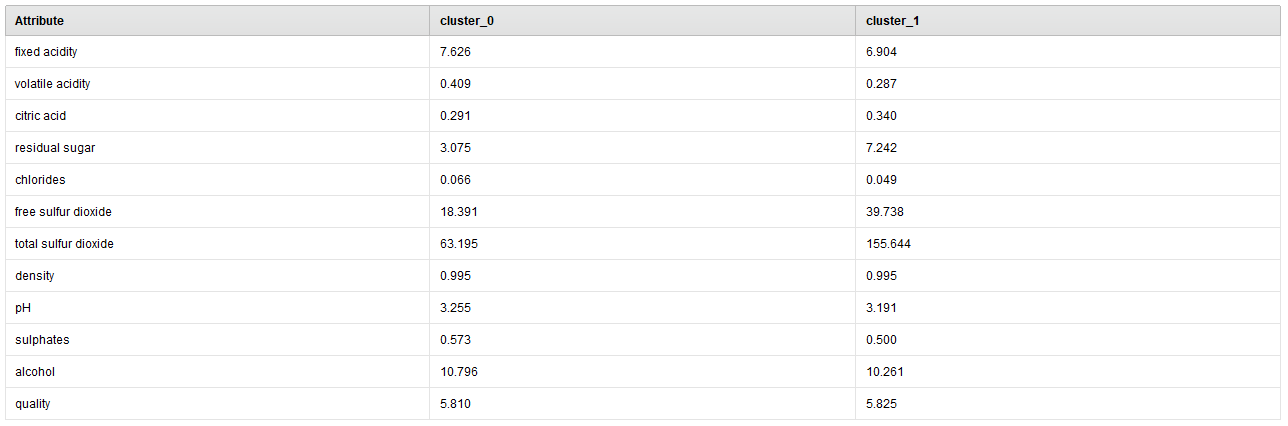
Prof. Amit Deokar

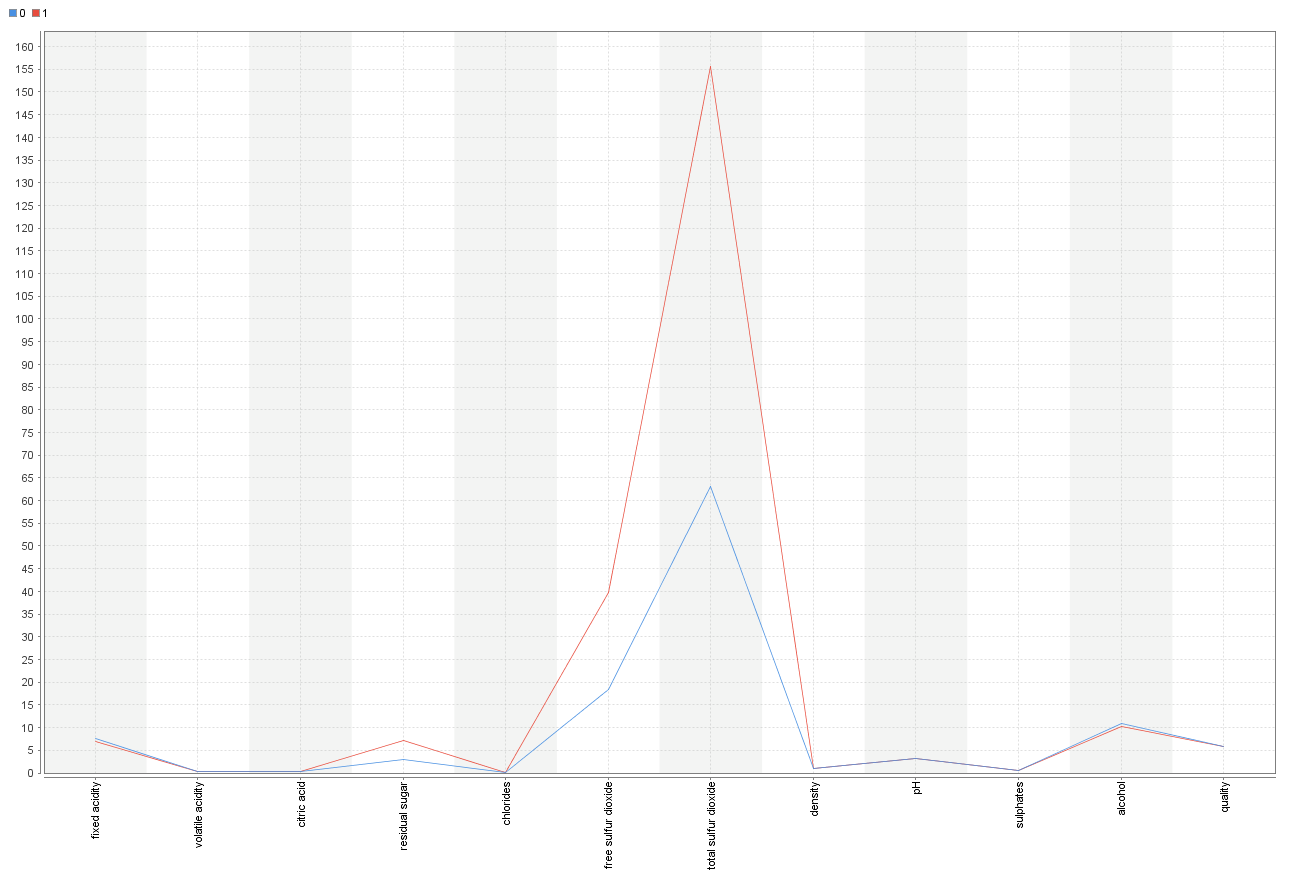
MIST4060 Data Mining for Business Intel

Assignment 5: SOM and k-means Clustering

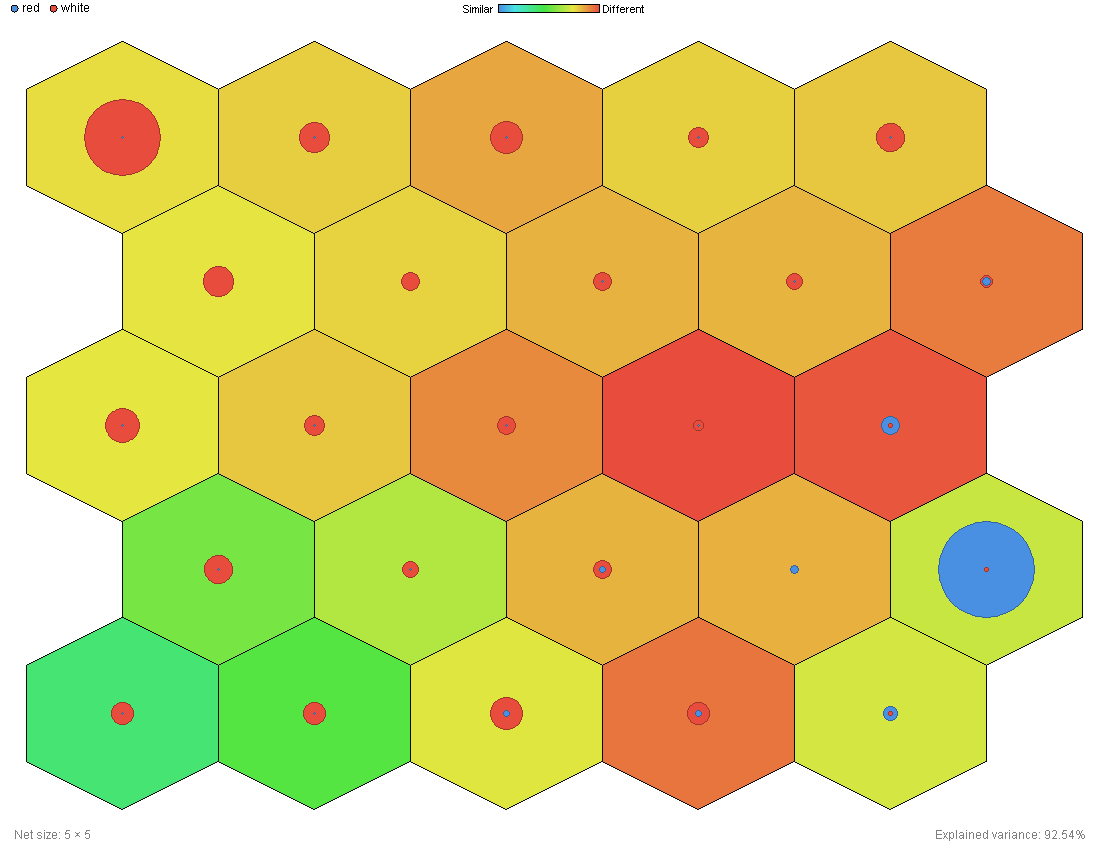


1.

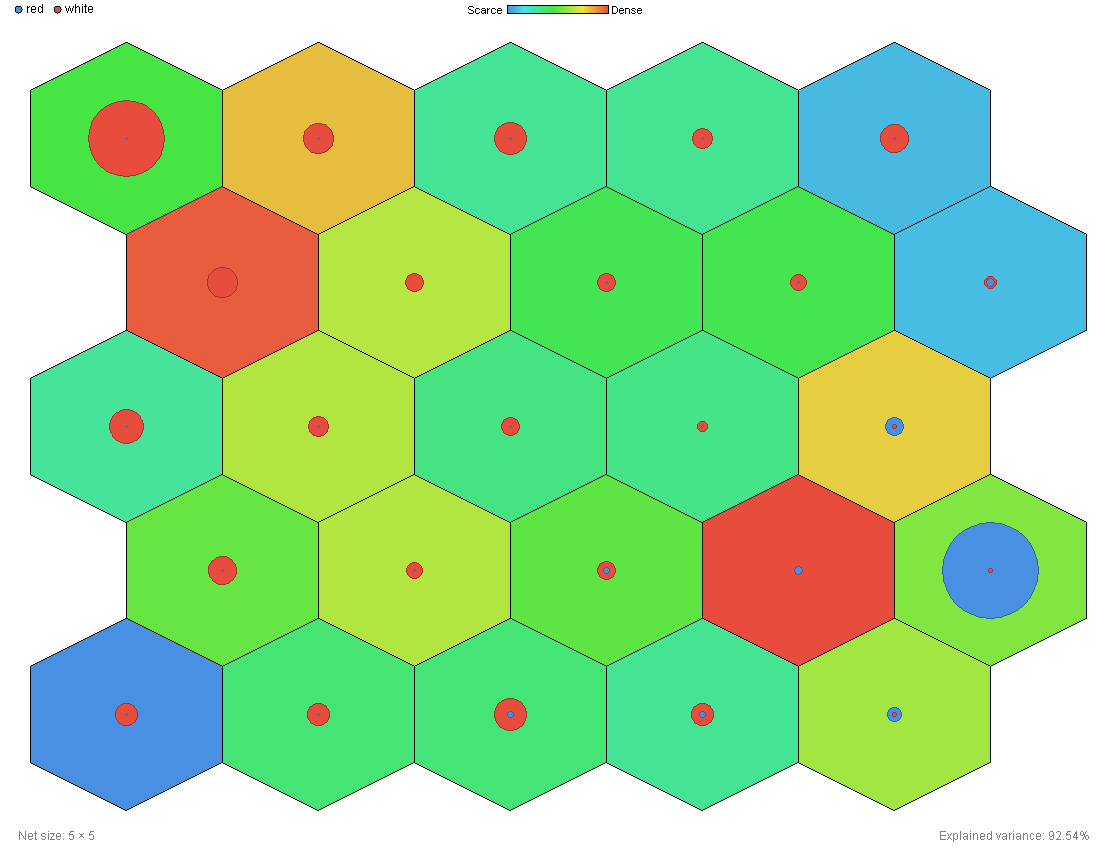
2. 



3. The Centroid table and the parallel coordinates plot show the same concept, just in different manners. Each cluster (whatever the number is for the “k” value) has a prototype centroid value for each attribute of the wine example set. This centroid value will be used for the clustering, as it is an imaginary data point that represents the characteristics of the cluster’s data points. The values are derived from the data point closest to each groups of data within the data plot. The parallel coordinates plot shows a line graph for both clusters and their respective centroid values on the y-axis, and the appropriate attribute on the x-axis. While the lines show the different data points, the main thing to note here is the distance between the lines at those points. We want a large difference in space because we want these clusters to be well-separated to divide the data set.

4. 

In a U Matrix the color of the hexagon shows the distance between each centroid within a cluster. The lighter the color hexagon the similar (close) each centroid is to its neighbors. The circles within each hexagon are based on a histogram. This means that each hexagon has its own distribution of numerical data. In this case the red circle shows the distribution of wine with the white quality and the blue circle shows the distribution of red wine.

5. 

In a P matrix the hexagon’s colors show the number of example data points within each cluster. A darker color shows a dense population, while a lighter color shows a scarce population within each cluster. Similar to number 4, the wine qualities are represented in a histogram so the circles represent the numerical distribution of the wine qualities within each cluster. Looking at the top left hexagon and the 20th hexagon (with the big blue circle in it), you can see that red and white wines are vastly dominant in those hexagons, but the number of example points is very small given by the P matrix. So, for example, if there’s 100 examples of the dominant wine type, the hexagon only has about 105 examples, out of an overall 6497 items.