

**CENG 480 MACHINE LEARNING**

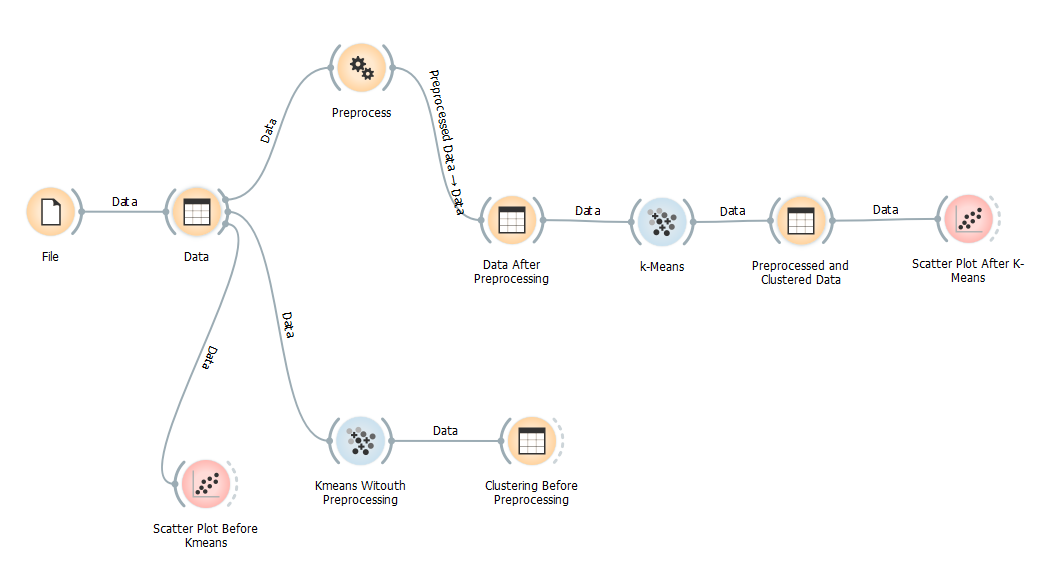
**FINAL EXAM**

**CLUSTERING ANALYSIS**

**Name/Surname:** Nurseli BAL

**Student ID:** 201611656

In this experiment, a dataset will be clustered by eliminating class label and predicting the class labels using clustering K-means method. In figure 1, we will see the flow chart of the experiment.



**Fig.1.** Workflow chart

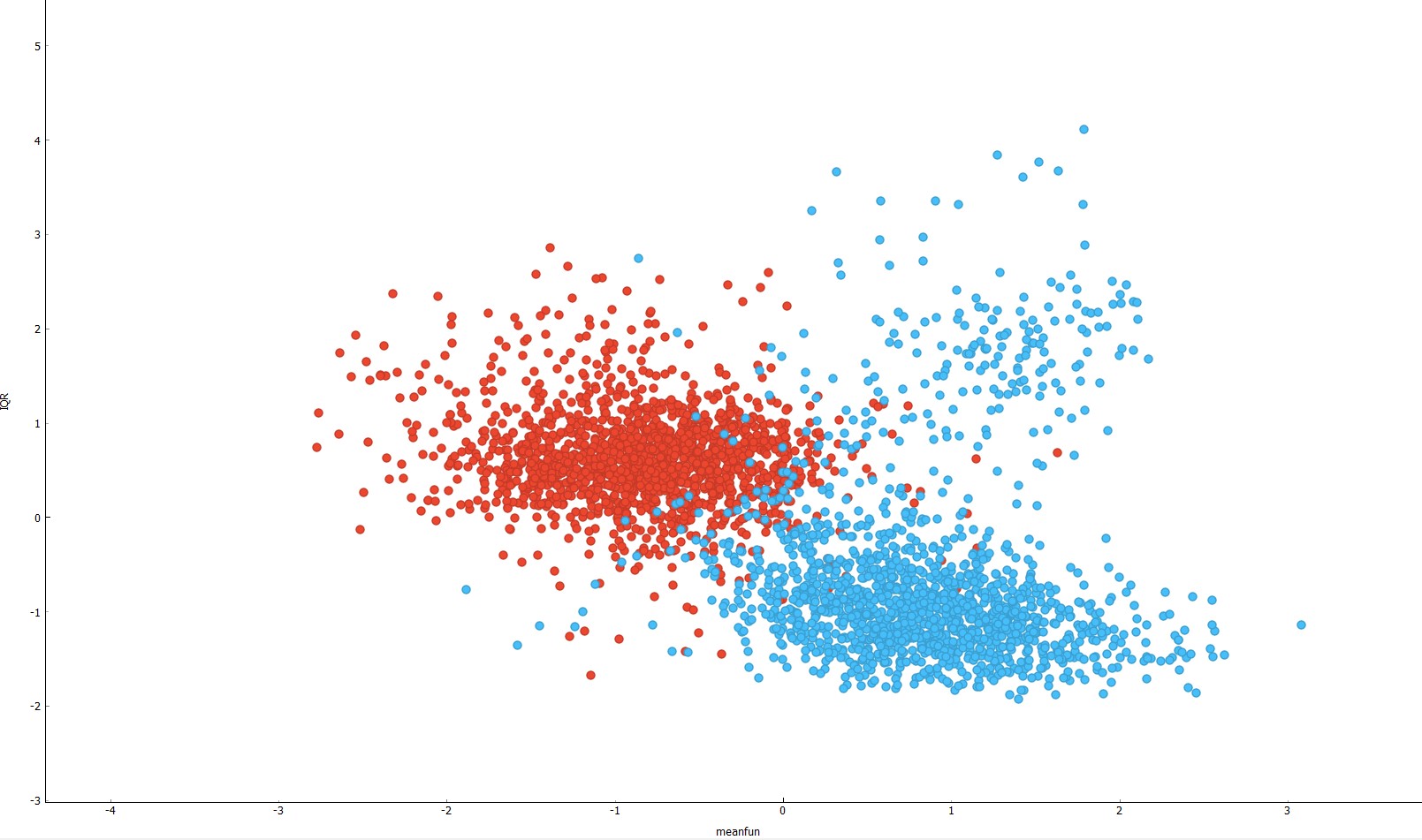
In the workflow above, File widget is used to upload the data. Then we can see our data by using Data table widget. The widget that has been named as Data is used to see our data. After the Data widget, we see 3 different ways from Data widget. One is Preprocess, the other one is Kmeans Without Preprocessing, and the last one is Scatter Plot Before Kmeans. As the name suggests, Scatter Plot Before Kmeans widget represents the scatter plot before kmeans clustering. I have used that widget to visualize my data before the clustering. We can see the plot in Figure 2.



**Fig. 2.** Scatter Plot before Kmeans Clustering

And also Kmeans without preprocessing name suggests that the widget is for kmeans clustering before the preprocessing. In this step, Since I have 2 class labels, I have used 2 clusters as C1 and C2. In this step, I saw that the clustering is not seperated well. Most of the data went to the C1 cluster. What it means is we need to do a step called Preprocessing step which I already used in Preprcess widget. In this step, I select only 2 important features to find the clusters. And it surely did. I applied the Kmeans algorithm after preprocessing in the widget k-Means. Then I saw my clustered and prepocessed data in the step Preprocessed and Clustered Data. I saw that the data seperated better than it did in the step Clustering before Preprocessing widget.

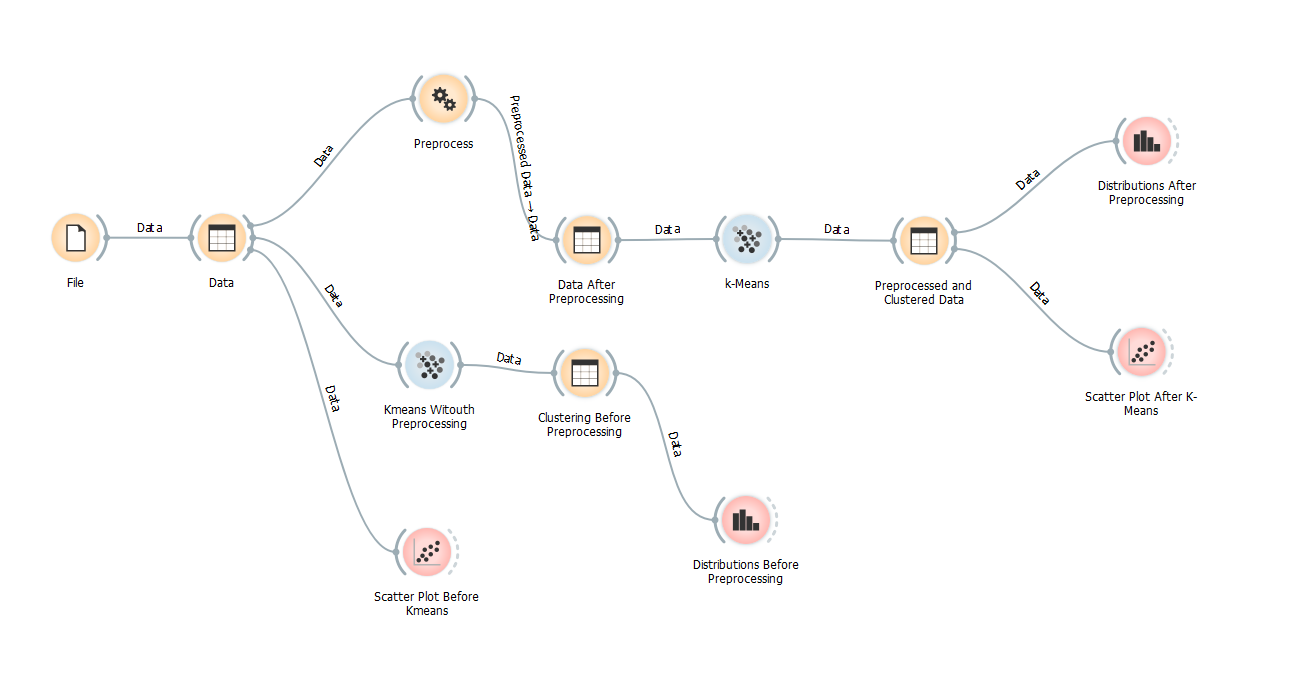
Finally, at the end of the workflow, in the step Scatter Plot After Kmeans, I have put a scatter plot to see how good is the Kmeans clustering. Figure 3 shows the Scatter plot after kmeans. We easily say that it seperated better.



**Fig. 3.** A figure that shows the scatter plot visualization after the clustering.

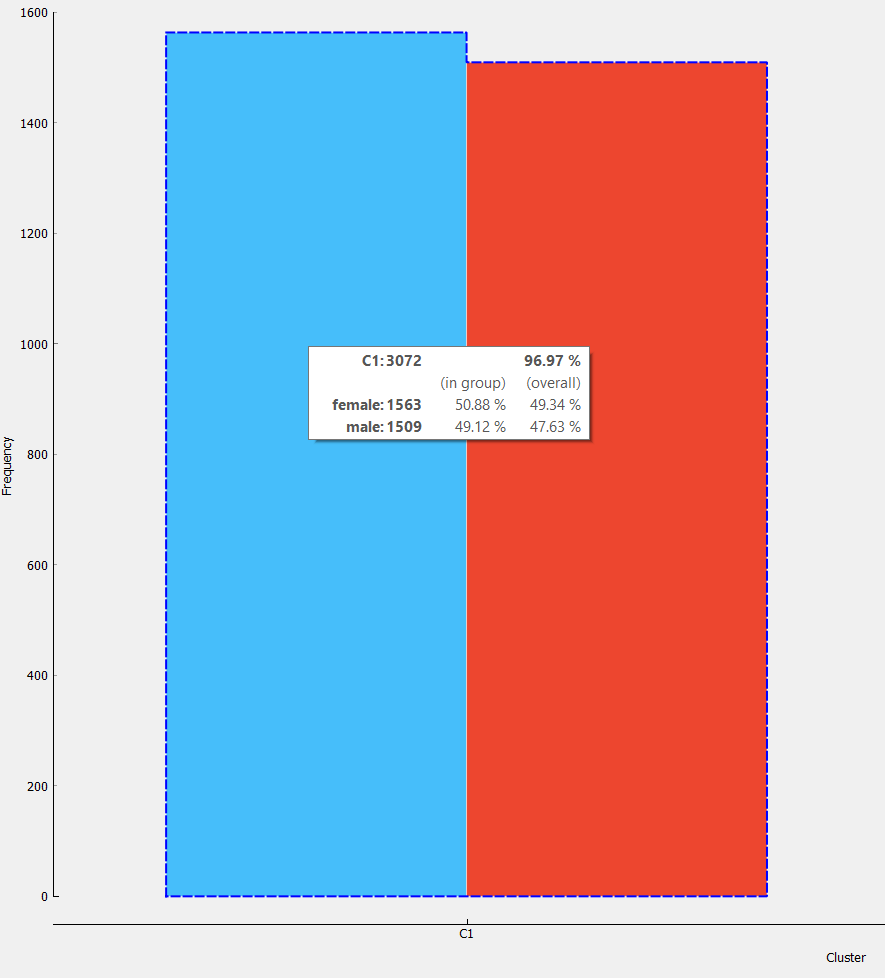
**EVALUATION**

We need to evalute the performance of the methods after the clustering. The reason is that we need to find how good job we have done. There might be a lot of methods here. I have used the visualization method to compare the results. Figure 4 shows the workflow of evaluation.

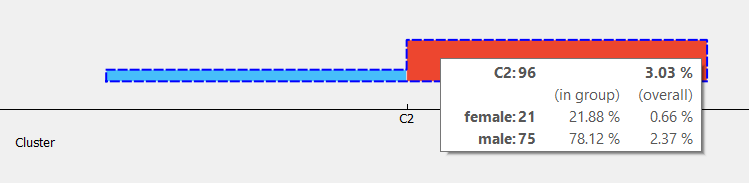


**Fig. 4.** Workflow of evaluation

After we have done the clustering steps, as you can see in the above image, I have used the widgets named Distribution After Preprocessing and Distribution Before Preprocessing. The results are given in the Figure 5 and Figure 6, represents the Distribution Before Preprocessing. Red ones represents the males and blue ones represents the females.

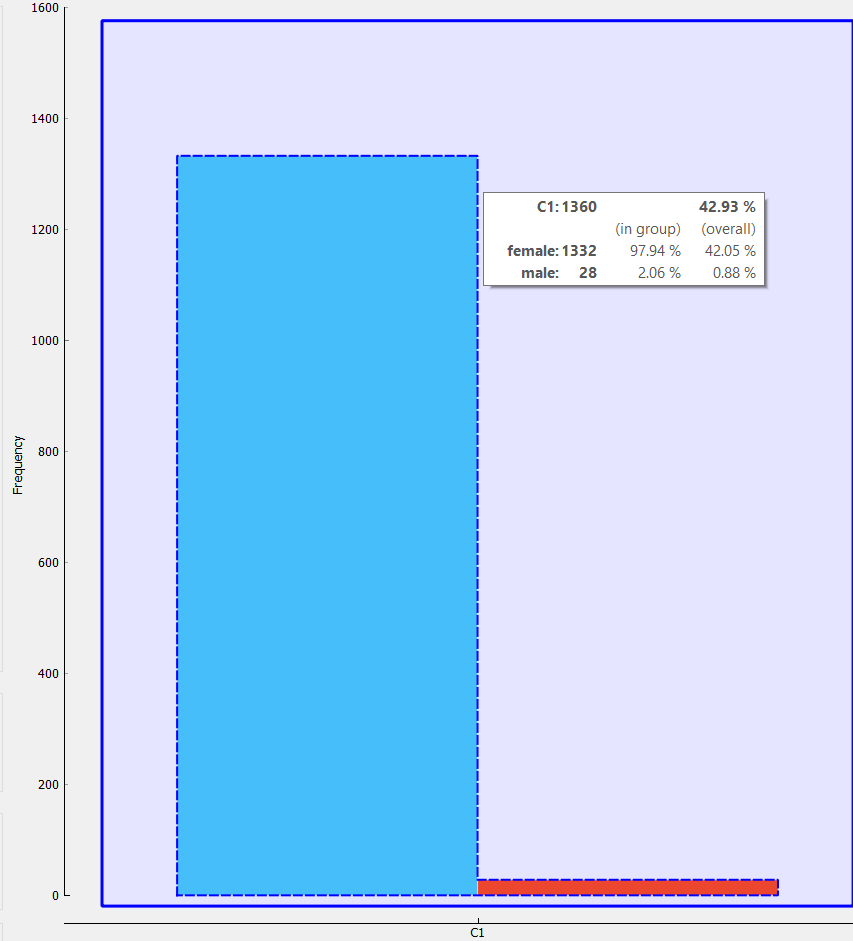


**Fig. 5.** Shows the distribution of female and male in C1 cluster before preprocessing

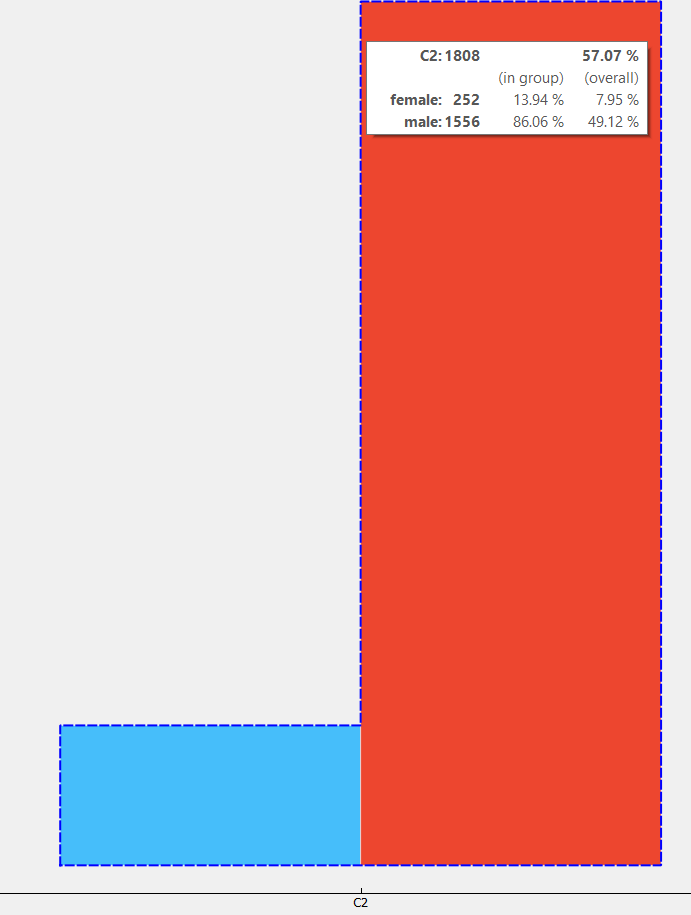


**Fig. 6.** Shows the distribution of female and male in C2 cluster before preprocessing

Now we can see the results after the preprocessing in the Figure 7 and Figure 8.



**Fig. 7.** Shows the distribution of female and male in C1 cluster after preprocessing



**Fig. 8.** Shows the distribution of female and male in C2 cluster after preprocessing