BBM469

Data Intensive Applications Laboratory

Report of Assignment-2

Topic: Clustering and Classification with Python
Group Name: Budgeregiar- Group 6

Name of Student-1: Yusuf Efe Kalaycı

ID of Student-1: 21827517
Name of Student-2: Mert Tazeoğlu

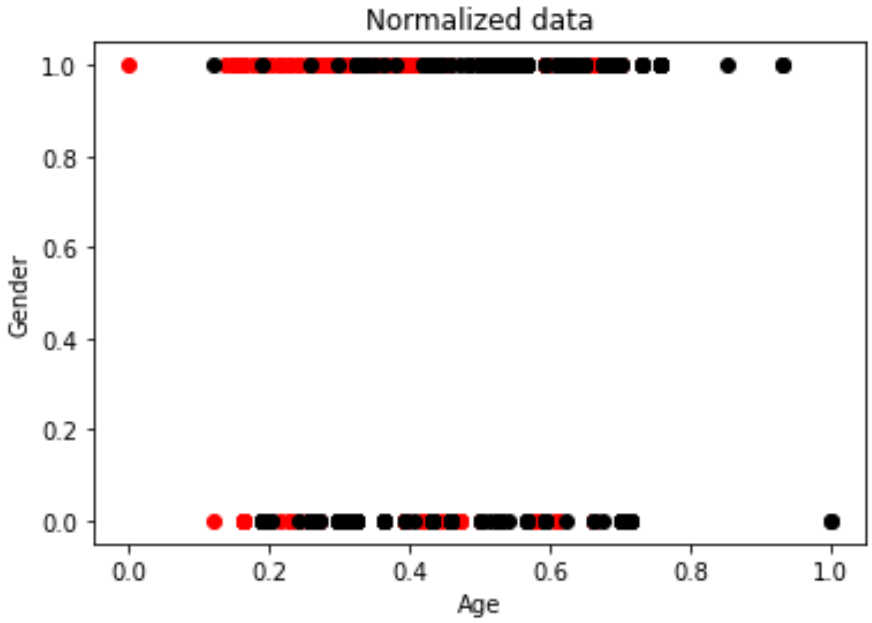
ID of Student-2: 21946606

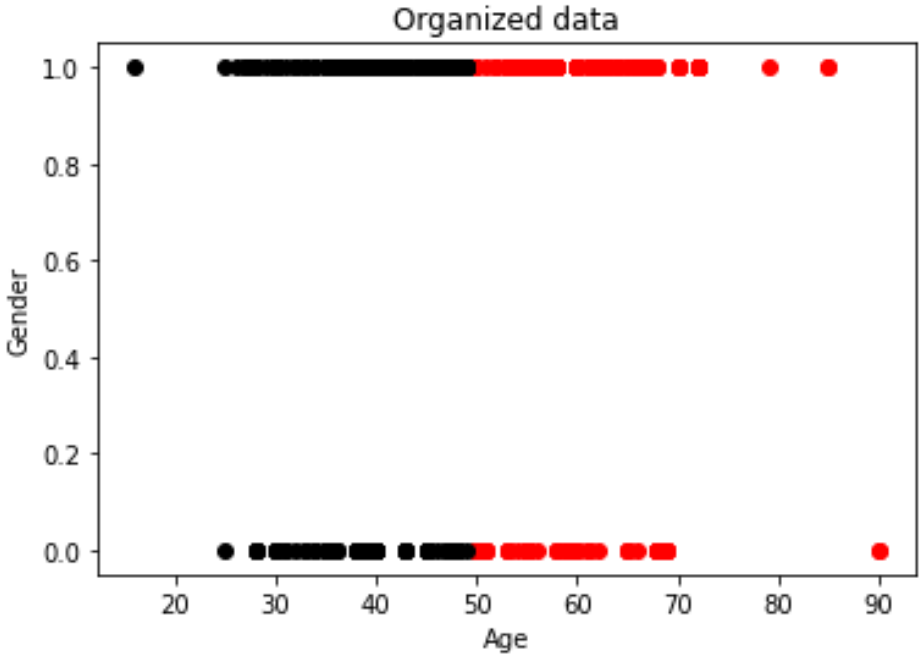
Instructors:

- Fuat Akal
- Tuğba Erdoğan

We have chose Kmeans for clustering and Gaussian naïve bayes for classification.We thought classification with using probabilities will fit best for this dataset. It was hard to decide for clustering because of the discrete values.

For the organized data clustering rand score is 0.49 since we have two classes it is almost random. And for the normalized data clustering rand score is 0.63 . This is better than random and the organized data. Since most of the features are discrete clusterization is not working well on this dataset.





From scatter plots aboe we can see that in organized data ‘age’ feature has more impact to clustering. Since the organized datas rand score is worse than normalized data. This may be bad decision for algorithm.

We used min-max normalization on ‘age’ feature in [0-1] range.

Graph in the below is unfortunately true for the both organized and normalized data. We think that’s because normalization on age is close to just divide to 100 because of the old person’s age is close to 100 and babies are close to 0. To avoid this we can use 0.25 and 0.75 quartiles. There also can be naïve bayeses problem with continous features.

	mean accuracy	recall	precision	f1score	confusion matrix
k=3	0.8788	0.8512	0.83	0.8404	166 34 29 291
k=4	0.8807	0.8556	0.83	0.8426	166 34 28 292
k=5	0.8634	0.8412	0.795	0.89	159 41 30 290
k=7	0.8826	0.8527	0.84	0.8463	168 32 29 291

In this graph we can see that we have best values in mean and precision with k=7. K=5 is the best for f1 score but it is also worst in the other metrics

K=4 has the best recall value.

Best feature to consider in this dataset is polyuria. Since only this feature shows with %82 accuracy value.

References:

https://www.askpython.com/python/examples/plot-k-means-clusters-python

https://scikit-learn.org/stable/modules/clustering.html#k-means

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMeans.predict

https://scikit-learn.org/stable/modules/naive_bayes.html

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html