Data mining
Analysis of adult dataset part 2

Katarzyna Gunia, Krystian Walewski

February 2024

1 Problem description

In the second part of our project, we will compare different methods of clustering as well as dimensionality reduction. To validate results of clustering, multiple indices will be used. We will also check how these results change after dimensionality reduction. In the last step, we will compare performance of classifiers on data after dimensionality reduction with results from the previous report.

2 Clustering

We compared the following methods of clustering:

- k-means,
- PAM — Partitioning Around Medoids,
- AGNES — Agglomerative Nesting with different linkage methods.

The first method was used only on the numerical features, which were standardized.

The rest of algorithms were applied to the dissimilarity matrix, calculated using the Gower's distance. Due to large number of observations and computational complexity, we used it on 50% of our data. Obviously, the label column (income) wasn't used during the clustering procedure.

2.1 Internal indices

The first internal index that we were looking at was average silhouette (1). K-means performs the best from all models, and it has the highest average silhouette for 3 clusters, and then it rapidly decreases. PAM and AGNES for single and average linkage achieve the best score for 2 clusters. AGNES with complete linkage has the highest index for $k = 4$. However, even these best scores are not very close to ideal value 1, which indicates that clusters are not very compact and are not well separated.

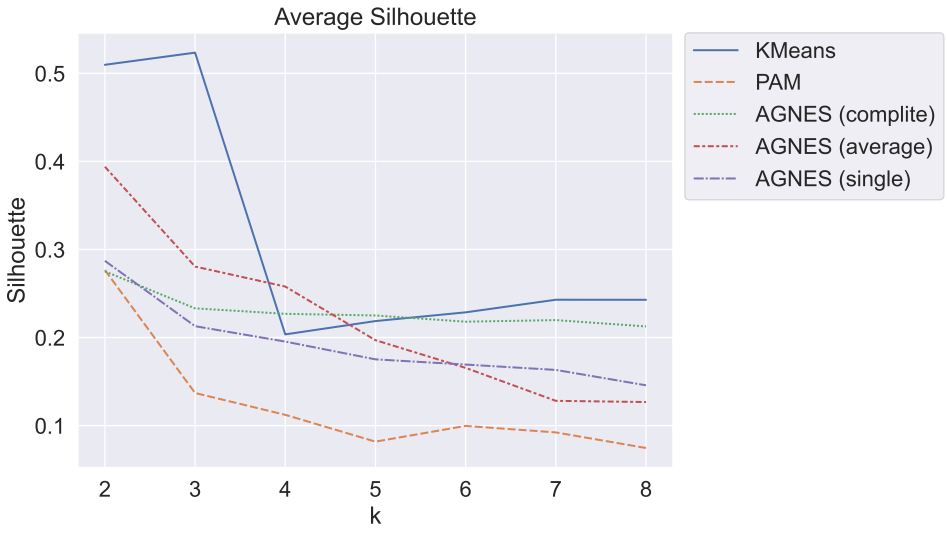

Figure 1: Comparison of average silhouette index for different number of clusters.

We checked how the silhouette plot looked like for the best case, which was k-means algorithm with 3 clusters (2). We see that the algorithm distinguished one large cluster and two very small ones. It behaves similar for $k = 2$. The next silhouette plot (3) contains the results for AGNES with single linkage and $k = 2$, which had the second-best average silhouette. We can see that this time the algorithm distinguished a cluster containing only one observation, it probably suffers from chain effect.

Comparing silhouette plots for different models and different k, we concluded that all clustering methods works rather bad on our data. Surplisingly, results for AGNESS with complete linkage (figure 4) looks better, than results of other linkage methods, even though complete linkage had worse average silhouette value.

Looking at the Dunn index (5), which is another internal one, we see a totally different order of best performing methods. However, still the values on the plot are not very high. This time AGNES with single linkage works the best, second place takes AGNES with average linkage, for both $k = 2$ and $k = 3$. Surprisingly, PAM has almost 0 for all number of clusters and this time k-means perform very bad.

2.2 External indices

To calculate external indices for our models, we had to assume the number of clusters equal to 2, as we have '$\leq$ 50k' and '$>$ 50k' labels. Table (1) contains values of different external metrics. Starting with the rand index, we see that

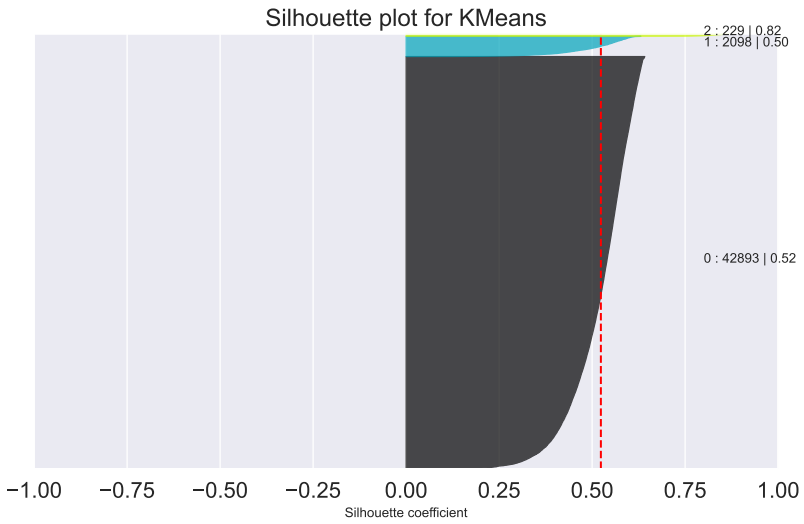

Figure 2: Silhouette plot for k-means algorithm and 3 clusters. Vertical line corresponds to an average on the whole sample. Next to the cluster label are its size and average Silhouette for this cluster.

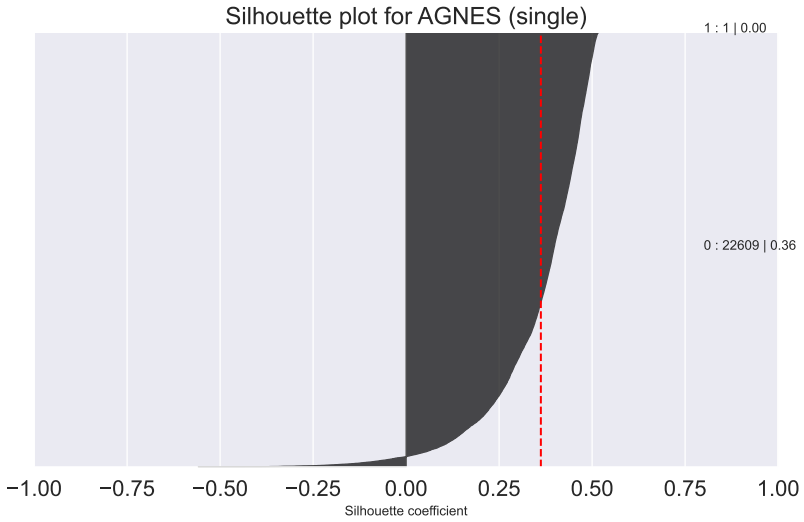

Figure 3: Silhouette plot for AGNESS algorithm with single linkage and 2 clusters. Vertical line corresponds to an average on the whole sample. Next to the cluster label are its size and average Silhouette for this cluster.

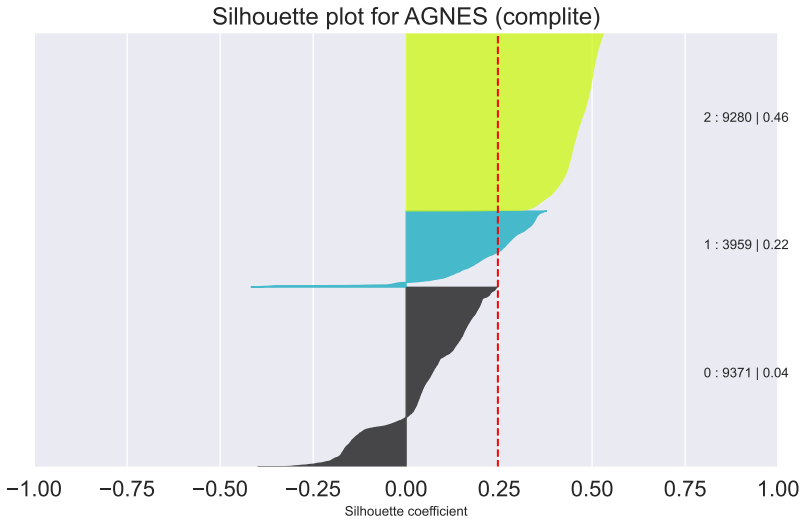

Figure 4: Silhouette plot for AGNESS algorithm with single linkage and 2 clusters. Vertical line corresponds to an average on the whole sample. Next to the cluster label are its size and average Silhouette for this cluster.

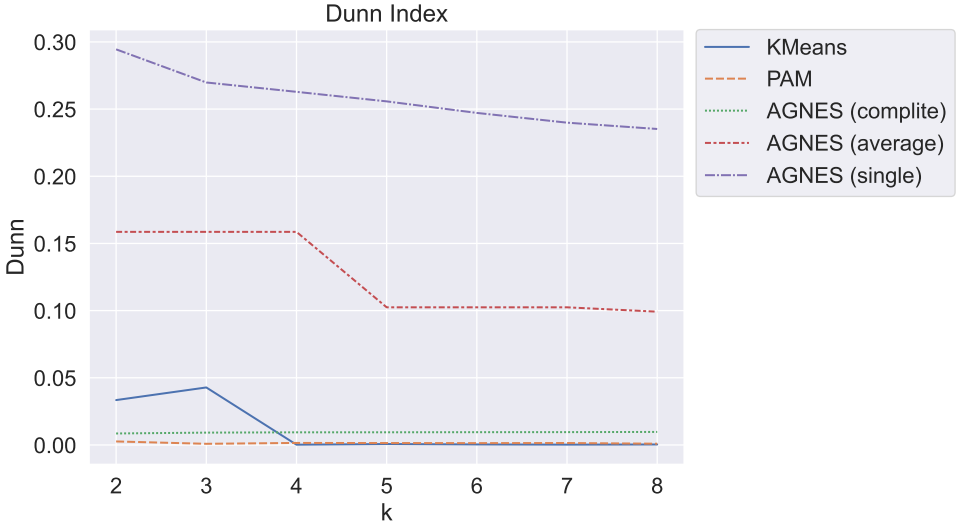

Figure 5: Comparison of dunn index for different number of clusters.

model\index	KMeans	PAM	AGNES (C)	AGNES (A)	AGNES (S)
Rand	0.629	0.545	0.514	0.623	0.623
Adjusted Rand	0.062	0.090	0.021	-0.000	-0.000
Homogeneity	0.015	0.144	0.084	0.000	0.000
Completeness	0.045	0.117	0.070	0.026	0.026
V-Measure	0.023	0.129	0.077	0.000	0.000
Fowlkes-Mallows	0.772	0.600	0.576	0.790	0.790
Jaccard	0.585	0.257	0.309	0.560	0.560

Table 1: Comparison of external metrics for clustering models. Where C, A and S denotes complete, average and single linkage respectively.

k-means performs the best, but it's not much better than other models. Values for adjusted rand are very close to 0, however PAM has the highest value. Values for homogeneity, completeness and V-measure are also rather poor, close to 0, however again PAM has the best one. On the other hand, AGNES with average and single linkage performs best for Fowlkes-Mallows metric and this time we are not so far away from the optimal value 1. Lastly, the maximal Jaccard is for k-means, but complete and single AGNES is close. Summarizing, it's hard to decide which model works the best on our dataset.

Summarizing, it is very hard to decide which model works the best on our dataset, since the results are not consistent. In the next part of our analysis, we decided to focus on hierarchical clustering with complete linkage.

2.3 Visualization of clustering

Looking at the dendrogram (Figure 6) of AGNES with complete linkage, there are clearly visible 2 similar clusters and the third larger one. We can see the same pattern on the plot that visualizes our data in 2 dimensions using first 2 components calculated by UMAP — Uniform Manifold Approximation and Projection, method of dimension reduction (Figure 7). There are 2 clusters that are close to each other and quite compact, while the third one is definitely further from them and less compact.

To better understand the differences between clusters, we compared their distributions of different features. For numerical variables we will present box-plots, and for categorical — barplots, with percent of all observations in that cluster having a specific characteristic.

We can observe that in cluster number 1, almost all people are from the USA, while in the two remaining ones about 10% of people are not from the USA (figure 8).

On the plot 9, we can see that the last cluster consist exclusively of people married to civil spouse. While the remaining clusters include mainly never-married and divorced persons.

Looking at the occupation feature (plot 10), we see that cluster 0 has the

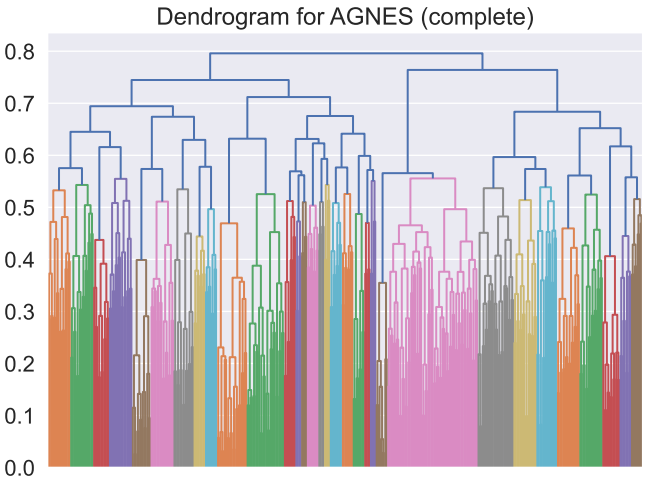

Figure 6: Dendrogram for AGNES with complete linkage.

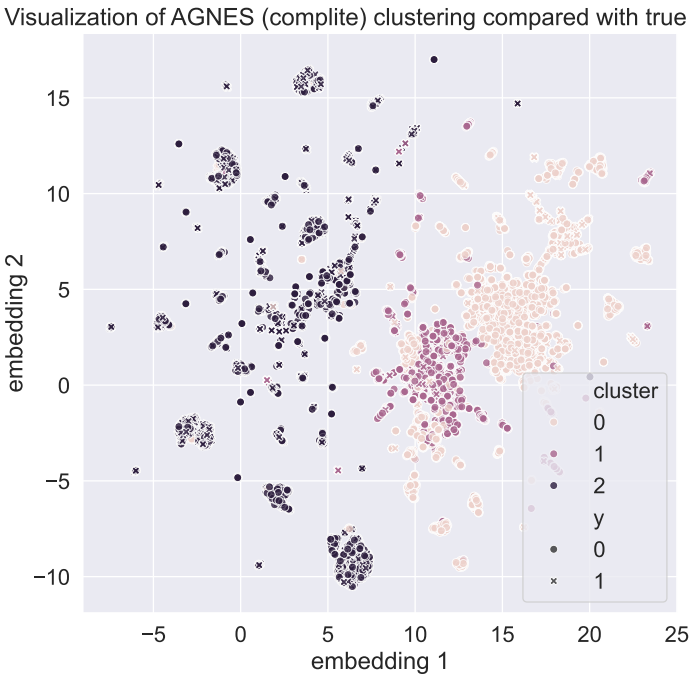

Figure 7: Visualization of clusters using first 2 components of UMAP.

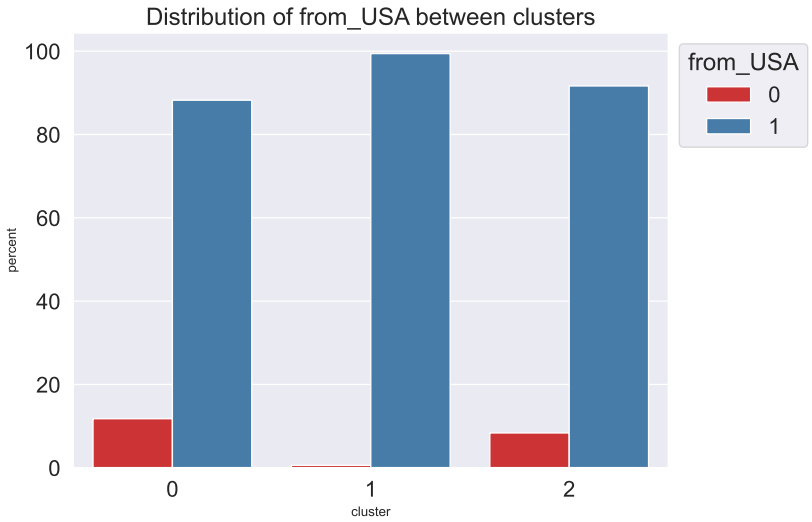

Figure 8: Distribution of from_USA in clusters returned by AGNES with complete linkage.

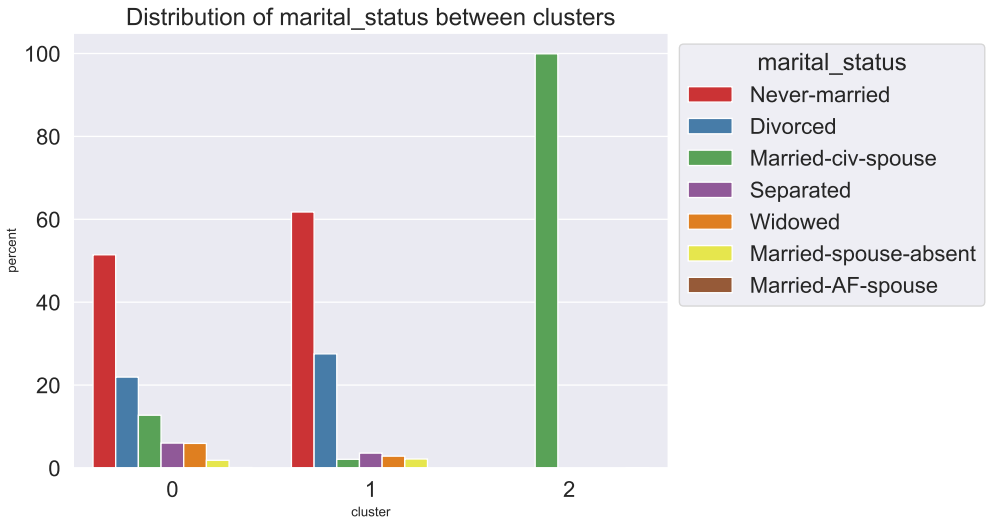

Figure 9: Distribution of marital_status in clusters returned by AGNES with complete linkage.

highest percent of administrative-clerical and other-service workers. While in clusters 1 and 2 the most numerous is a group of craftsmen (Craft-repair), the last cluster has also the highest present of executive managers.

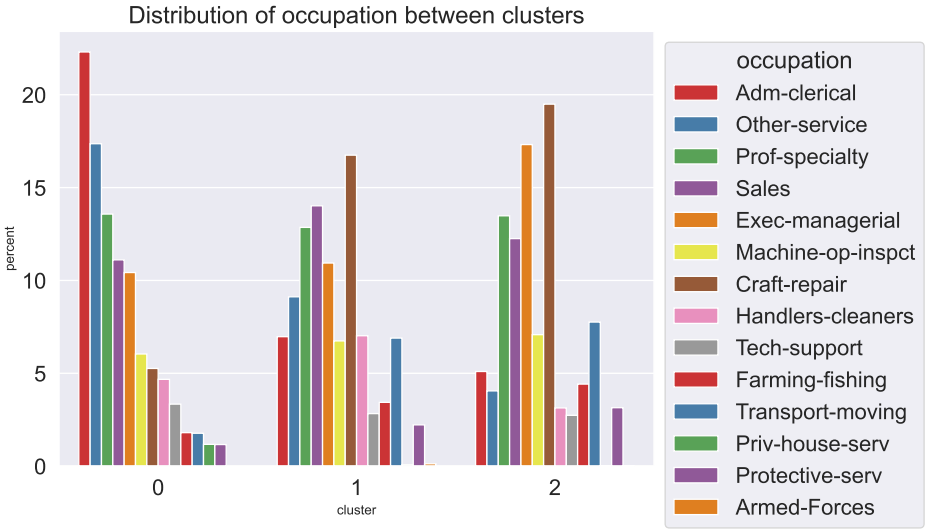

Figure 10: Distribution of occupation in clusters returned by AGNES with complete linkage.

When it comes to a race variable, clustering algorithm haven't really based on that one, as the distribution of races in all the clusters is similar (figure 12). The same we observed for workclass feature.

It might be surprising that almost all husbands from our dataset felt into cluster 2. Furthermore, there are no other observations in that cluster. In cluster 1 more than 70% of people are not in a family, which is clearly visible majority (13).

Considering sex feature, in cluster 2 we have only men. In cluster 1, the majority of people are also men, but in cluster 0 — women (figure 14).

Looking at the boxplots of age, we can say that the distribution is similar in cluster 0 and 1, while in cluster 3 are older people (15).

For education number variable, medians in all clusters are the same, but the IQR in cluster number 0 is significantly smaller (figure 16).

We can also observe, that people in cluster 0 tend to work less than the others (17).

Boxplots for the other numerical features don't bring us any other valuable information, so we decided to not include them in the report.

Lastly, let's look at the distribution of the most interesting feature — income level, which of course wasn't used during clustering. On the figure 19 we can see that clusters 0 and 1 contains in majority people earning less than $50K, in the last cluster the percentage of people earning less and more than $50K is on

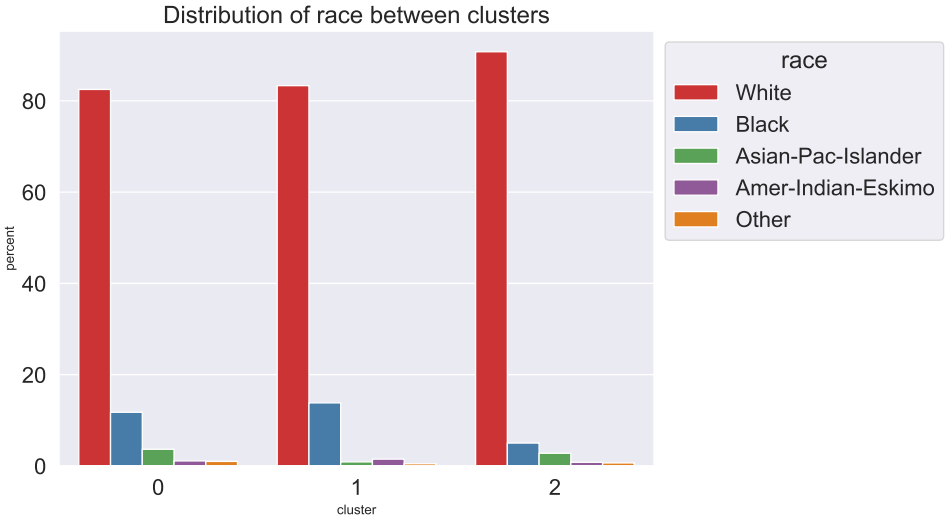

Figure 11: Values distribution for race feature.

Figure 12: Distribution of race in clusters returned by AGNES with complete linkage.

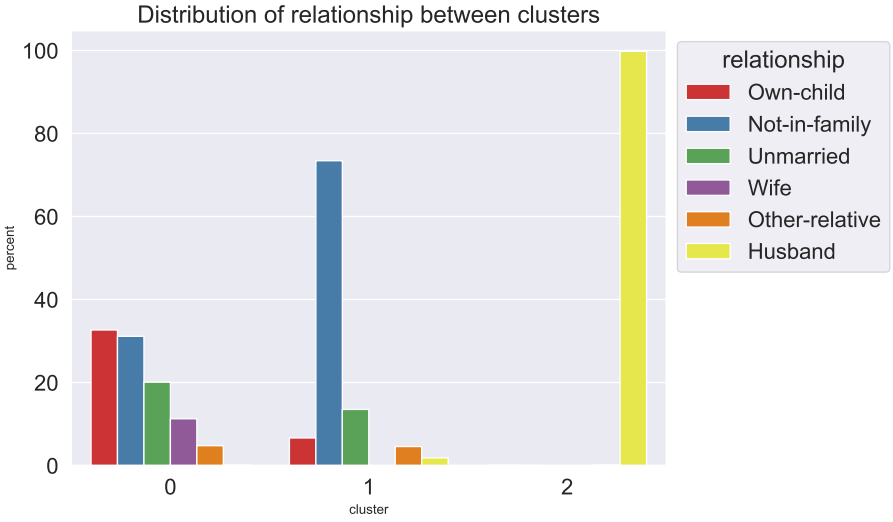

Figure 13: Distribution of relationships in clusters returned by AGNES with complete linkage.

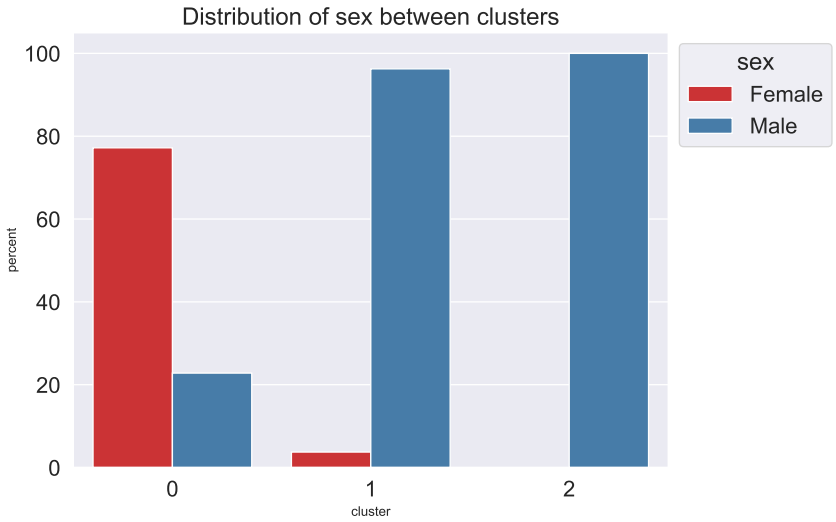

Figure 14: Distribution of sex in clusters returned by AGNES with complete linkage.

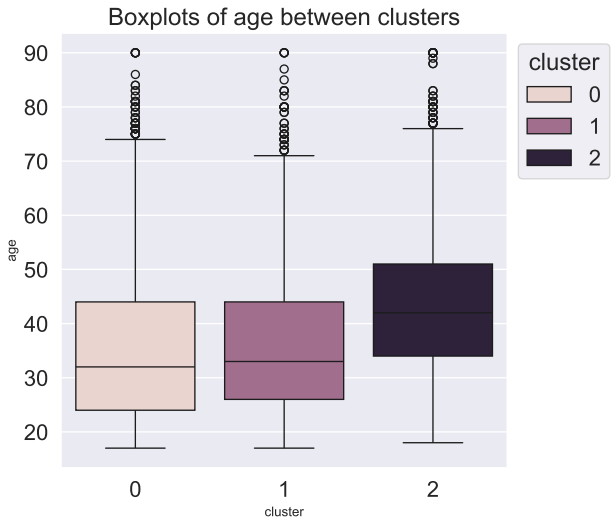

Figure 15: Distribution of age in clusters returned by AGNES with complete linkage.

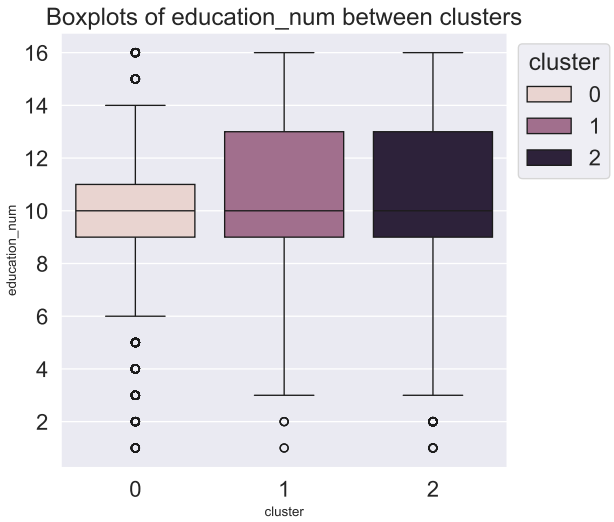

Figure 16: Distribution of education_number in clusters returned by AGNES with complete linkage.

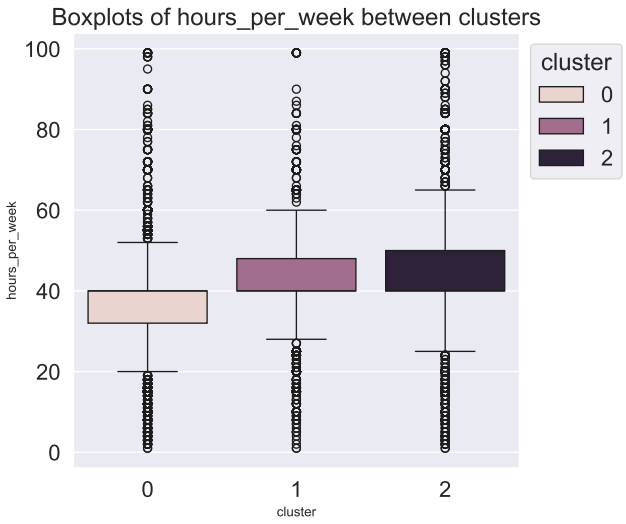

Figure 17: Distribution of hours_per_week feature in clusters returned by AGNES with complete linkage.

the similar level. In general, the cluster with mostly married men, usually older and working as an executive manager or craftsman, has the highest proportion of larger income.

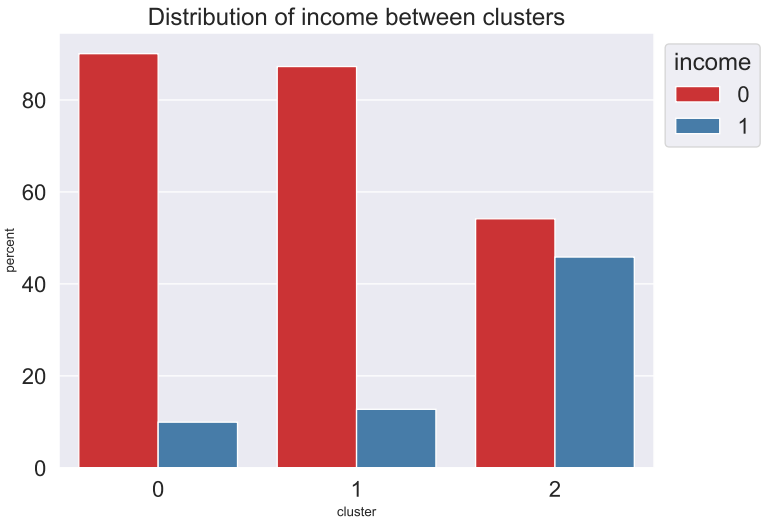

Figure 18: Values distribution for hours_per_week feature.

Figure 19: Distribution of income in clusters returned by AGNES with complete linkage.

3 Dimensionality reduction

For the dimensionality reduction part of our project, we decided to test two methods:

- PCA - Principal Component Analysis,
- UMAP - Uniform Manifold Approximation and Projection.

The first method works only on numerical data, which we decided to standardize, so that each feature has the same impact. On the other hand, UMAP can take any precomputed distance matrix, so as in clustering we inputted matrix of Gower's distance computed on 50% of data.

3.1 PCA

On the plot (20) we see the visualization of our data for the first two principal components. We can observe 3 clusters, 2 of them are really close to each other, while the third one is better separated and consists only of people earning more

than $50k. In total, these two principal components explain almost 40 % of variance. The next two plots present the variance explained by each component and the cumulative variance (figure 21). In this case it is hard to reduce dimension as we need 5 components to achieve more than 80% of explained variability.

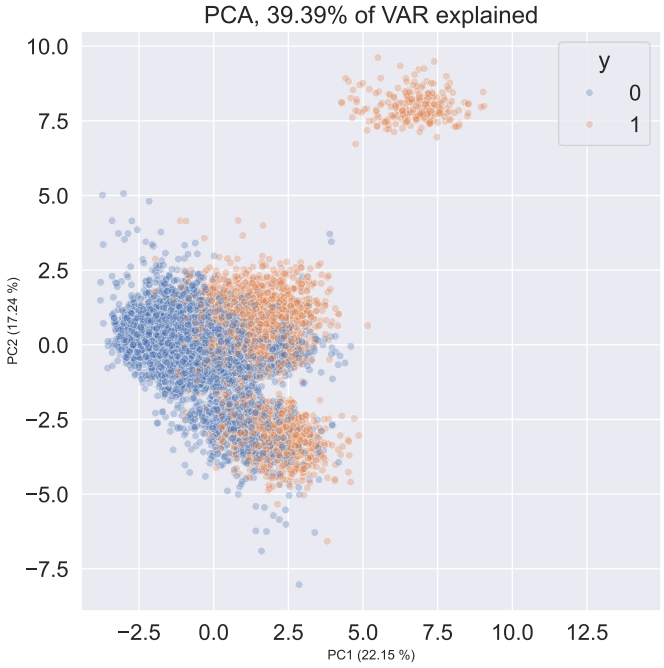

Figure 20: Visualization of data for the first two principal components with class label.

In the table (2) we can find the vector of loadings for each principal component. From that, we can conclude that the first component corresponds mostly to the education and number of working hours per week. The second one corresponds mostly to capital loss and gain.

3.2 UMAP

The plot (22) presents the visualization of 50% of our dataset after the projection into the first two dimensions, labels represent the income level. We can see there are 2 quite compact clusters with mostly 0 label and the points without any visible structure containing both labels.

To evaluate the performance of UMAP we used the trustworthiness, which indicates to what extent the local structure is retained, it is within 0 and 1. The next plot shows this core as a function of number of components in the UMAP (figure 23). The score is almost 1 for three components.

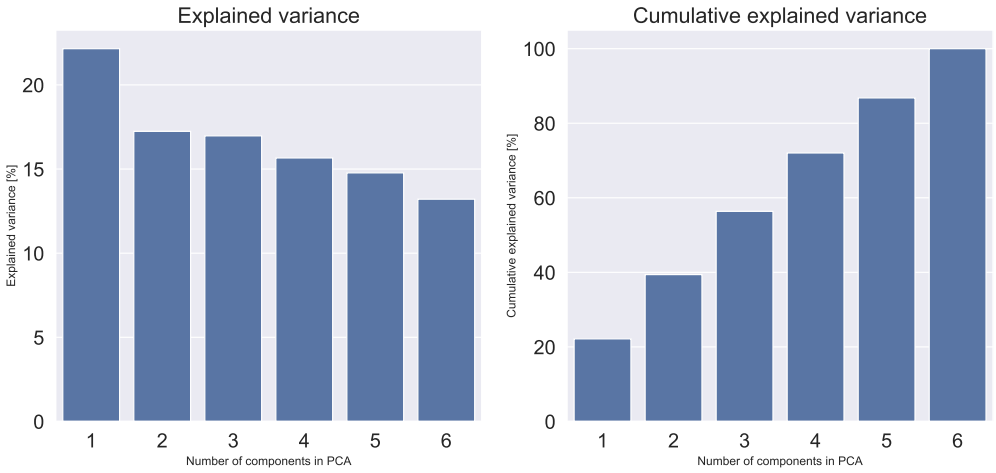

Figure 21: Scree plot for PCA, showing percent of explained variance by each principal component.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
age	0.415	-0.264	-0.373	0.672	0.083	-0.401
fnlwgt	-0.201	0.389	0.692	0.528	-0.002	-0.224
education_num	0.528	0.142	0.259	-0.502	0.090	-0.611
capital_gain	0.415	0.593	-0.143	0.094	0.509	0.434
capital_loss	0.254	-0.636	0.517	0.027	0.395	0.327
hours_per_week	0.522	0.054	0.163	0.090	-0.755	0.347
Explained VAR [%]	22.152	17.239	16.971	15.658	14.771	13.208

Table 2: Vectors of loadings for principal components.

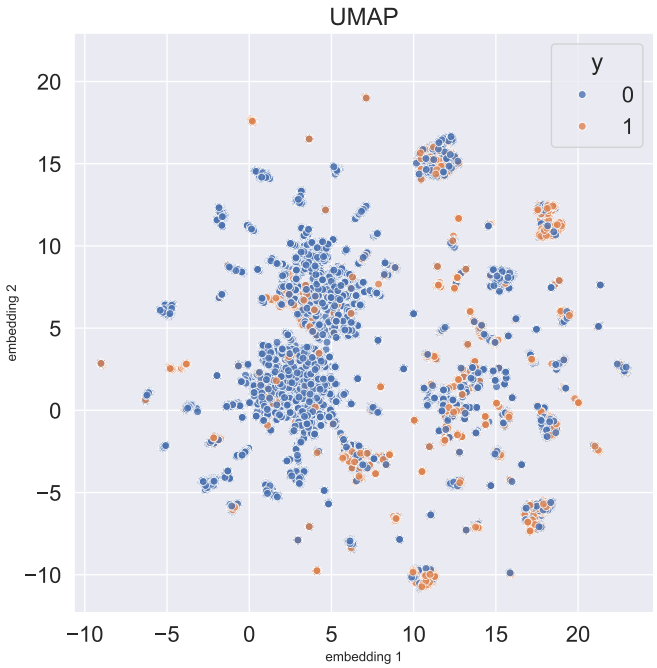

Figure 22: Visualization of data for the first two embeddings of UMAP with class label.

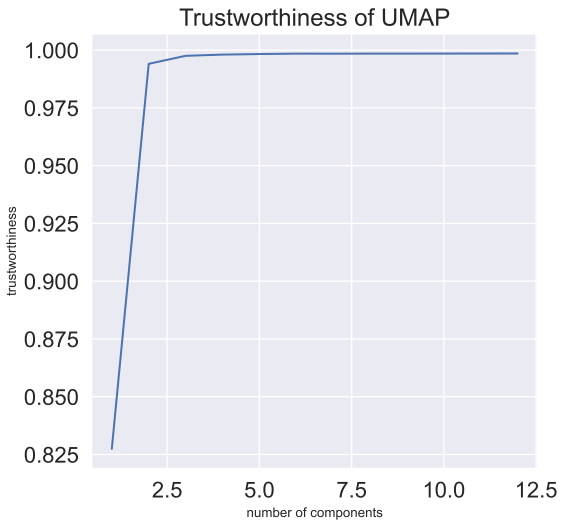

Figure 23: The trustworthiness score of UMAP for different number of components.

4 Clustering for reduced dimension

Being so satisfied with UMAP results, we decided to perform clusterization for reduced dimension of our data by taking the first 3 components. We can see the results for average silhouette score on the plot (24). Comparing it to the previous results for all dimensions, we can conclude that for AGNES with single linkage method we obtained much worse outcome. KMeans for 2 and 3 clusters also achieved worse results. We can also observe that all models without AGNES (single) yield now similar results. However, this time, it is hard to choose the best number of clusters as the silhouette doesn't really depend on it.

Looking at the Dunn index, we can also observe other results than before (figure 25). AGNES with complete linkage performs better than before, so as KMeans for 2 clusters. The rest of methods yield similar or worse values. However, for all of them, the best number of clusters is equal to 2.

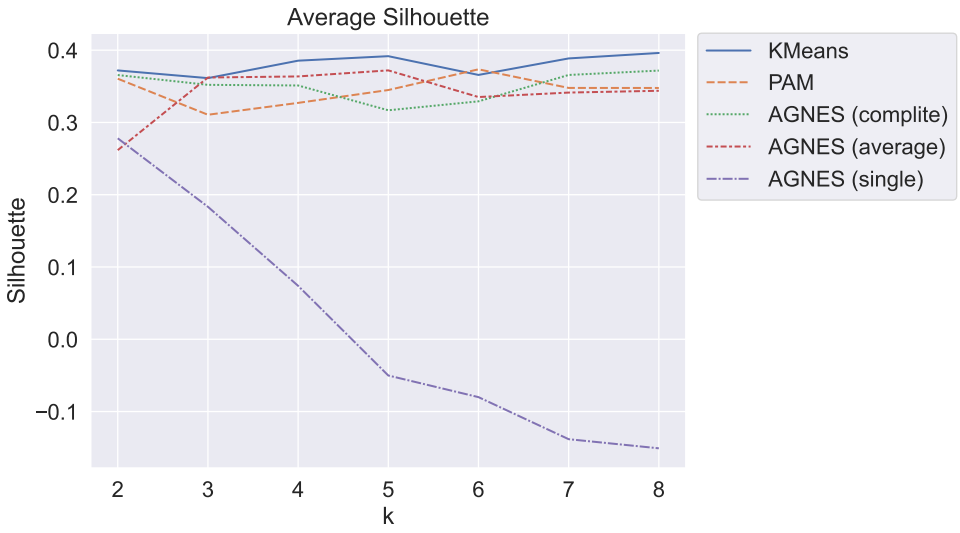

Figure 24: Comparison of average silhouette index for different number of clusters (after dimensionality reduction).

5 Classification for reduced dimension

To check the performance of classification models after dimensionality reduction, we once again took first 3 components calculated by UMAP. On the figure (26) we can see the boxplots of accuracy calculated using cross validation on training data. The results are really satisfying as 3 best models (KNN, Random Forest and XGBoost) are only a little worse than before. QDA gives now even better and more stable results. However, the accuracy for LDA and linear regression is significantly worse than before.

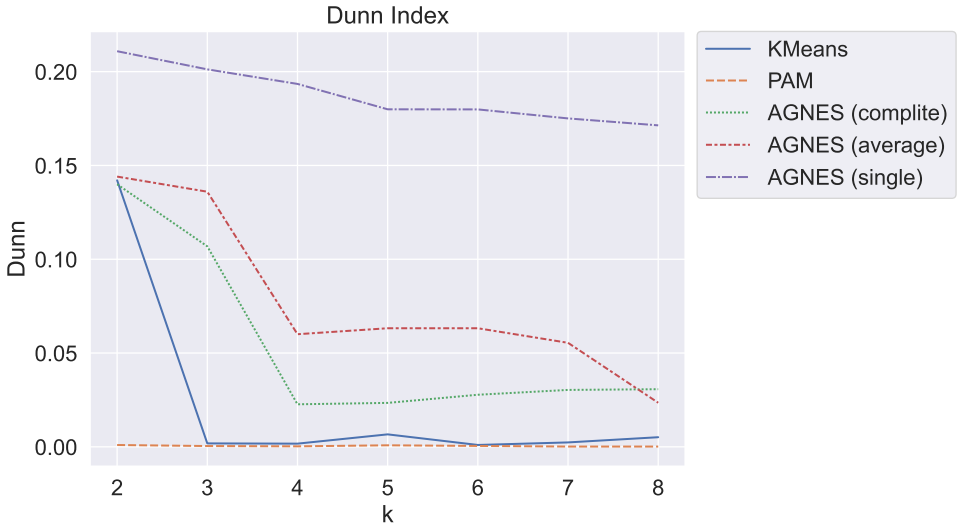

Figure 25: Comparison of dunn index for different number of clusters (after dimensionality reduction).

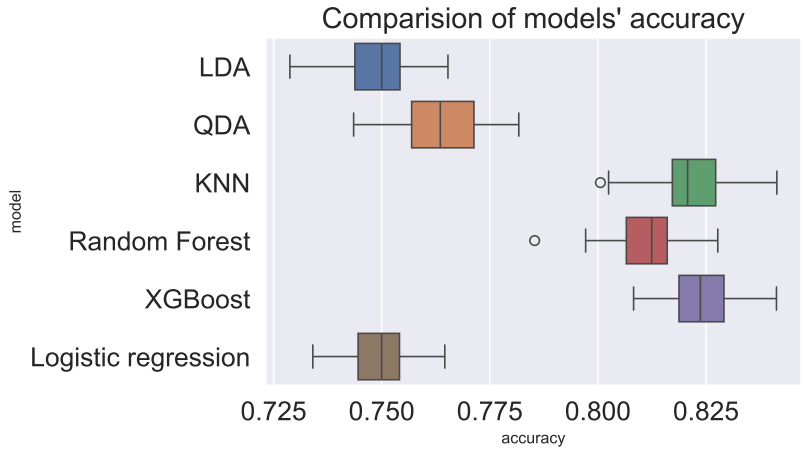

Figure 26: Comparison of classification models' accuracy (after dimensionality reduction).

model	accuracy before UMAP	accuracy after UMAP
LDA	0.835	0.738
QDA	0.821	0.765
KNN	0.846	0.820
Logistic Regression	0.849	0.739
Random Forest	0.851	0.805
XGBoost	0.871	0.825

Table 3: Comparing the accuracy before and after dimensionality reduction.

We can better compare the results looking at the final accuracy score for the test dataset (table 3). The best model turns out to be XGBoost (the same as before) with accuracy 0.825 (about 5 percentage points less). The second-best algorithm — KNN, lost only 2 percentage points. The worst models, which are LDA and logistic regression yield accuracy about 10 percentage points smaller than before. For sure, these algorithms were more affected by the new set of data.

6 Summary

We compared different clustering algorithms, however it was hard to decide which one performed the best, since the different internal and external metrics returned different conclusions. Nevertheless, we get an interesting characteristic of the clusters. The cluster with the highest proportion of income above \$50k — approx. 50% of people, consists mostly of married men working as an executive managers or craftsmen.

Later, we compared different methods of dimensionality reduction. We tested PCA and UMAP. We also applied MDS and IsoMap algorithms, but they yield similar or worse results, so we didn't include them in this report.

We used UMAP to visualize our dataset, however the clusters were not always clearly visible, well separated and compact. Emdebings returned by UMAP were also used as an input into the clustering and classification algorithms. Clustering for data with reduced dimension gave rather worse results. It was also much harder to choose the optimal number of clusters. When it comes to the classification, reducing the dimension of the data decreased the accuracy of models, however in some cases it wasn't a big drop. On the other hand, it significantly reduced the time of model training.