02450 Introduction to machine learning - Report 1

Edison Von Matt(223534), Georgios Kapakoglou(223001), Matija Šipek (222736)

October 3, 2022

Contents

1 Description of the data set

1.1 Overall problem of interest

For this report we have used a data set for glass identification. It was used by the USA forensic science service to distinguish between seven different types of glass based on their oxide content and refractive index. The data set is multivariate and contains 9 attributes (RI, Na, Mg, Al, Si, K, Ca, Ba, Fe) and the subject area is physical. The overall problem of interest is to correctly classify glass types from a crime scene based on the above described attributes in order to use them as evidence.

1.2 Reference

This data set was retrieved from the UC Irvine Machine Learning Repository
You can find the data on the next link Glass Identification

1.3 Previous analysis of the data

1) Object classification using support vector machines with Kernel-Based Data pre-processing

The goal of this paper was to prove if kernel-based data pre-processing can improve the performance of Support Vector Machines (SVM). Glass data set was one of the data sets used in for testing the correctness of this method. This is done by conducting appropriate data transformation, in this case feature extraction, prior to the classification step. For this they used 2 non-linear methods : kernel Principal Component Analysis (kPCA) and Supervise kernel Principal Component Analysis (SkPCA).

Namely, when using SVM on raw data the performance drops between 4 % and 8 %, while when SkPCA is used the drop stays within 3 %. If no fine-tuning is introduced, the classification drops by 20 % for a nonlinear sigmoid kernel. Lastly the paper was assessing the performance of evaluation of k-NN algorithm derived using kPCA and SkPCA, which shows one can achieve results comparable to SVM, but only in some cases. Finally, it was followed by SVM or k-NN classification. The conclusion varies, as the classification results improved only for some pre-processing methods.

[1] Adamiak, Krzysztof Duch, Piotr Slot, Krzysztof. (2016). Object Classification Using Support Vector Machines with Kernel-based Data Preprocessing. Image Processing & Communications. 21. 10.1515/ipc-2016-0015.

2) The Mixture of Neural Networks as Ensemble Combiner

This paper is referred to the proposition of two new ensemble combiners based on the Mixture of Neural Networks model. They did it by applying on multiple data sets (Glass Identification included) two different network architectures on the methods: The Basic Network (BN) and the Multilayer Feed-forward Network (MF). Furthermore, in order to compare the combiners, they used ensembles of different networks previously trained with Simple Ensemble and then they compared the Mixture Combiners with the Mixture Models by building the Mixture models with different experts and a single gating network. At last, they calculated the mean Increase of Performance and the mean Percentage of Error Reduction with respect to a single MF network.

Lastly, the results showed that the mixture combiners proposed are the best way to build Multi-net systems. Moreover, we can also conclude that the accuracy of an ensemble of Multilayer feed-forward network can be improved by applying the gating network of the Mixture of Neural Networks as an ensemble combiner.

[2] Fernández-Redondo, M., Torres-Sospedra, J., Hernández-Espinosa, C. (2008). The Mixture of Neural Networks as Ensemble Combiner. In: Prevost, L., Marinai, S., Schwenker, F. (eds) Artificial Neural Networks in Pattern Recognition. ANNPR 2008. Lecture Notes in Computer Science(), vol 5064. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-69939-2_17

1.4 Report goals

1.4.1 Classification

As indicated earlier, the main problem of interest is the identification of glass types based on the attribute *refractive index* (RI) and the oxide content described in the chemical attributes (Na, Mg, Al, Si, K, Ca, Ba, Fe). The class attribute *Type of glass* stores the different class labels. There are seven distinct classes in our data set described by the labels: *building_windows_float_processed*, *building_windows_non_float_processed*, *vehicle_windows_float_processed*, *vehicle_windows_non_float_processed, containers, tableware* and *headlamps*. However, it is to note that there is no record with the label *vehicle_windows_non_float_processed* present in this data set.

1.4.2 Regression

The main purpose of this data set is classification based (described above). However, it could also be interesting to find out how the *refractive index* (RI) is dependent on the different oxide levels described by the attributes (Na, Mg, Al, Si, K, Ca, Ba, Fe) as it is continuous and thus suitable for regression. We chose the refractive index as the dependent variable as it measures a ray of light as it travels from one medium to another which makes it very distinguishable from the other attributes that simply summarize the chemical composition of the glass. The data providers already applied *One-out-of-n encoding* to transform the class labels to discrete numbers in the class attribute. Apart, from this no additional data transformation was necessary in this data set.

2 A detailed explanation of the attributes of the data

2.1 Attribute description

No.	Description	Type
x_1	Id number: 1 to 214	Discrete, Nominal
x_2	RI: Refractive index	Continuous, Interval
x_3	Na: Sodium	Continuous, Ratio
x_4	Mg: Magnesium	Continuous, Ratio
x_5	Al: Aluminum	Continuous, Ratio
x_6	Si: Silicon	Continuous, Ratio
x_7	K: Potassium	Continuous, Ratio
x_8	Ca: Calcium	Continuous, Ratio
x_9	Ba: Barium	Continuous, Ratio
x_{10}	Fe: Iron	Continuous, Ratio
x_{11}	Type of glass	Discrete, Nominal

Table 1: Attribute information

In Table 2.1 we summarized the attributes along with a short description and whether they are discrete, continuous or binary and whether their scale is nominal, ordinal, interval or ratio.

2.2 Data Issues

As noted above, there is no data record storing the label called *vehicle_windows_non_float_processed* in the class attribute x_{11} meaning that this class is not represented in the data set. Apart from that, we checked whether there are missing values or if any of these were corrupted but we could not detect any issues which was also described by the data provider. Thus, we can confidently say that there are no data issues present and continue with our analysis.

2.3 Summary statistics

Below, in Table 2.3 we included summary statistics of the 9 relevant attributes. While the unit measurement is weight percent for all chemical attributes (Na, Mg, Al, Si, K, Ca, Ba, Fe), we can see that some of them seem to have very different scales (i.e. by comparing the mean of 72.651% for Si and 0.175% for Ba). We can also see that the refractive index (RI), for example, has a comparatively low standard deviation indicating that most values tend to be close to the mean, while the comparatively high standard deviation for Mg and Ca indicates a greater spread of values among these attributes.

Measure	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
n	214	214	214	214	214	214	214	214	214
mean	1.518	13.408	2.685	1.445	72.651	0.497	8.957	0.175	0.057
std	0.003	0.817	1.442	0.499	0.775	0.652	1.423	0.497	0.097
min	1.511	10.730	0	0.290	69.810	0.000	5.430	0	0
max	1.517	12.908	2.115	1.190	72.280	0.122	8.240	0	0
25%	1.518	13.300	3.480	1.360	72.790	0.555	8.600	0	0
50% (median)	1.519	13.825	3.600	1.630	73.088	0.610	9.172	0	0.100
75%	1.534	17.380	4.490	3.500	75.410	6.210	16.190	3.150	0.510

Table 2: Summary statistics of the attributes

3 Data Visualization and PCA

3.1 Exploratory data analysis

3.1.1 Boxplot visualization

Looking at figure 3.1.1 below, it becomes evident that outliers are present in mostly all of the attributes apart from *Mg*. However, after taking a closer look we argue that all of the present outliers still represent realistic values for each of the attributes and could thus contain valuable information. For example, it is not unrealistic for the refractive index of a glass to fluctuate between 1.51 and 1.54 (*) and additionally, we argue that the weight percentage of the different chemical components in glass is likely to fluctuate between the min and max values that are evident in the boxplots below for each of the related attributes.

For example, the Ba boxplot at the bottom of the figure 3.1.1 shows a very skewed distribution with a high amount of outliers. Here, the whole interquartile range is tightly centered around 0% while the outliers can go up to 3% which is still realistic as many glass types might have been produced without the related chemical component (leading to a IQR squeezed around 0), while this component is only used for specific glass types (leading to some outliers up to 3%).

J.A. Lambert, FORENSIC SCIENCES — Glass, Elsevier, 2005, Pages 423-430, https://doi.org/10.1016/B0-12-369397-7/00202-

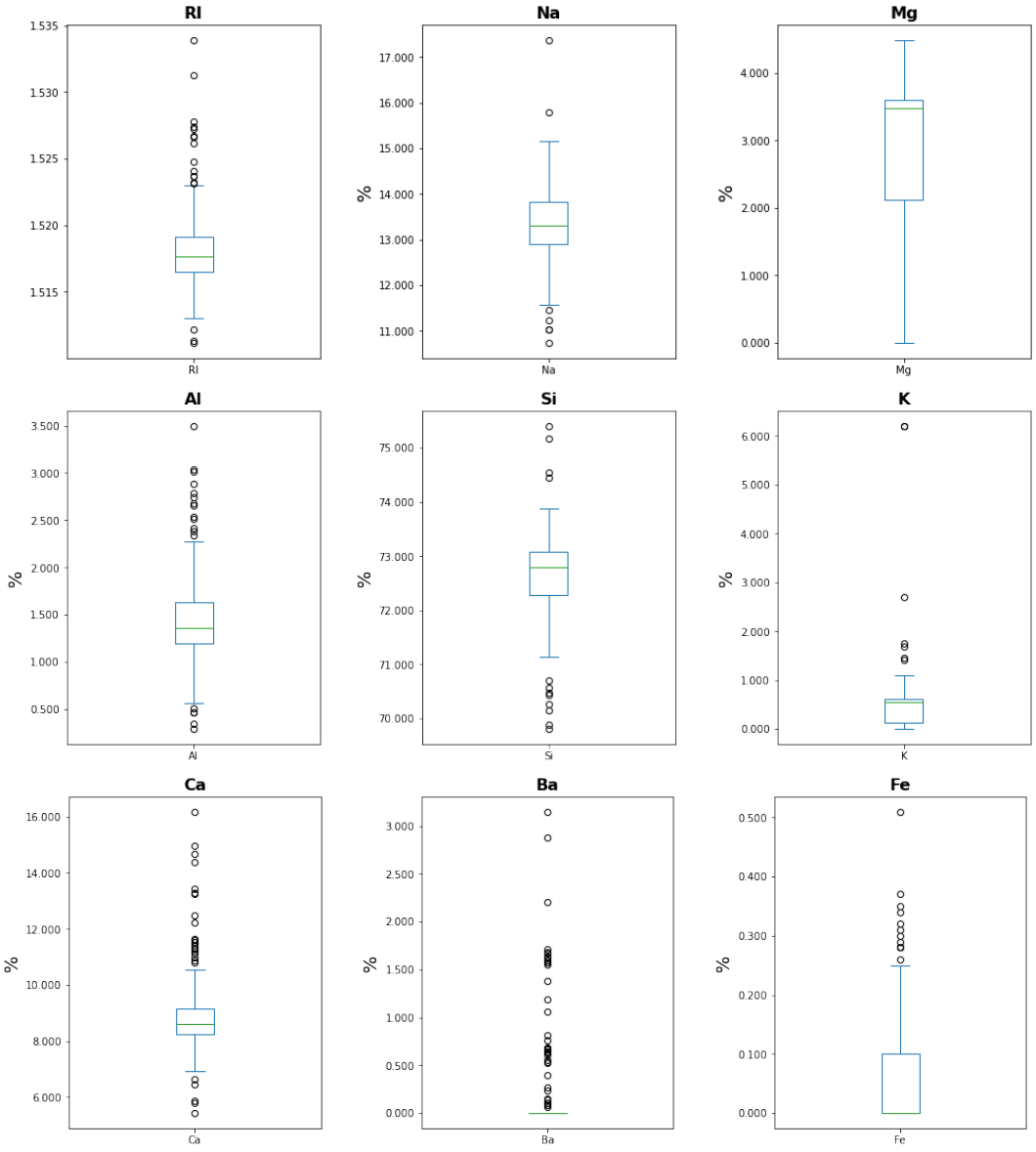

Figure 1: Boxplot representation of spread and skewness groups of data

3.1.2 Attribute distributions

Looking at 3.1.2 we can see that there is significant difference in the distribution of attributes. We plotted the histograms with their probability density so that the area under the histogram integrates to 1. Additionally, we plotted the red line which reflects the theoretical normal distribution based on the mean and standard deviation of each attribute. By comparing the real distribution of the data with the red line, we can see that the data for Na, Si, Al, Ri and maybe even Ca appear to be normally distributed, while K, Ba and Fe appear to follow a left-skewed distribution, while Mg seems to follow a bimodal distribution.

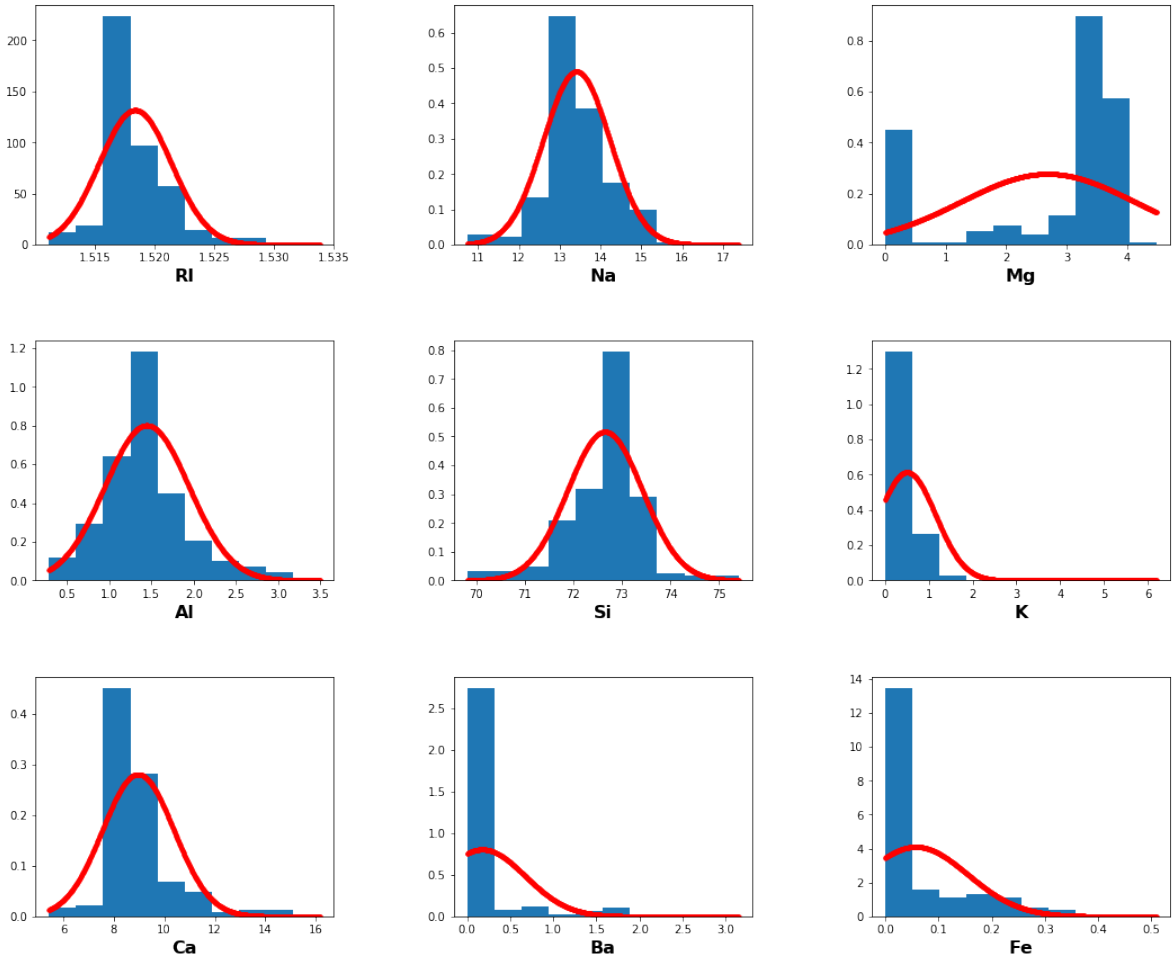

Figure 2: Histograms showing probability density for each attribute

3.1.3 Correlation coefficients

To determine whether a PCA analysis would be beneficial, we assessed the pair-wise correlation coefficients between all of our attributes. Looking at 3.1.3, we see that there is a strong positive correlation (0.81) between the attributes RI and Ca, while there is also a fairly strong negative correlation (-0.54) between RI and Si. Furthermore we can see a moderate correlation between RI and Al (-0.41), Ba and Na (0.33), Al and Mg (-0.48), Ca and Mg (-0.44), Ba and Mg (-0.49), K and Al (0.33) and finally Ba and Al (0.48). To summarize, we can see that there are numerous mutual relationships between the attributes providing us with an incentive to perform the PCA analysis and summarize these relationships inside of principal components to reduce the dimensions of the data.

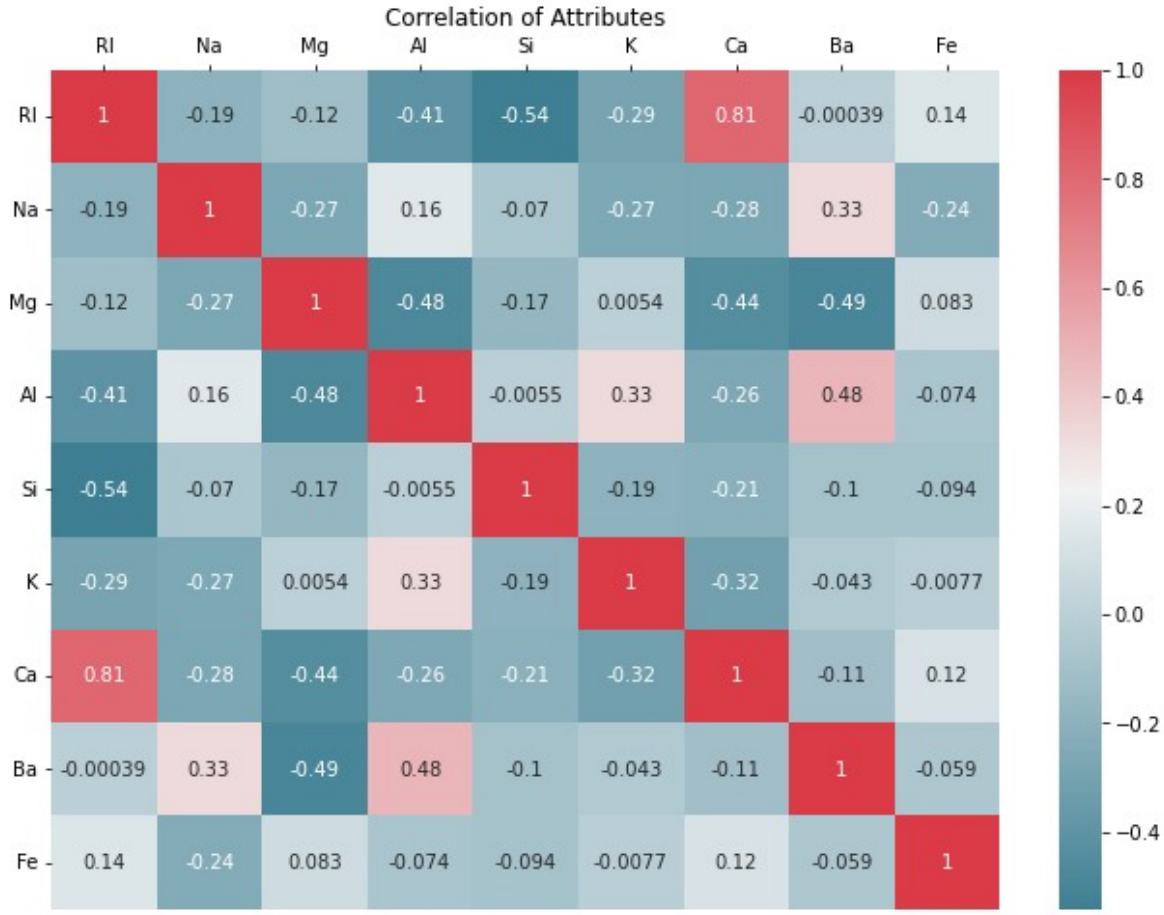

Figure 3: Pairwise correlation coefficients accross attributes

3.2 Principal Component Analysis

From Table 3.1.3 above we can interpret that our attributes (RI, Na, Mg, Al, Si, K, Ca, Ba, Fe) had different scales, so in order to carry out the PCA analysis we not only subtracted the mean, but also divided the data by the standard deviation to normalize the data by ensuring a mean of 0 and a standard deviation of 1 for each attribute column. We then carried out the PCA by applying the *scipy.linalg.svd()* function on our normalized data matrix.

3.2.1 How much variation in a data set can be attributed to each of the principal components?

We plotted the explained variation in the data set of each computed principal component below in 3.2.1. From the graph we can see that the difference in the explained variance among the first seven principle components does not seem to be particularly high as the percentages range from (27,9% to 4,1%). Thus, six principal components are needed to explain more than 90% of the variance in the data set. Thus, the PCA analysis was not as effective as we hoped for in reducing the dimensions of the data set, as we would only discard three principle components so we can still explain enough variance ($\approx 90\%$) with the rest of the principle components.

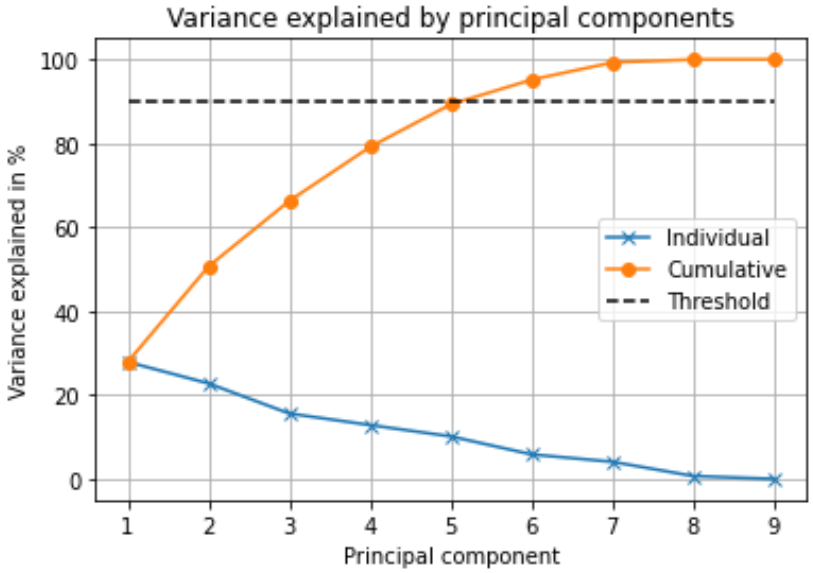

Figure 4: Percentage of variation in the glass dataset explained by each of its principal components

3.2.2 Principal directions of the relevant PCA components

In figure 3.2.2, we plotted the different coefficients of the six principle components with the highest eigenvalues as these explain more than 90% of the variation in the data set below. Focusing on the first three principle components, we can see that the PC1 mainly captures the variation of the attributes RI, Na and Ba, while the PC2 mainly captures the variance of RI, Na, Mg, Al, K and Fe. Finally, PC3 mainly captures the variance of Na, Al, K, and Fe. Additionally, it becomes visible that the first three principal components seem to capture the variance of the some of the same attributes.

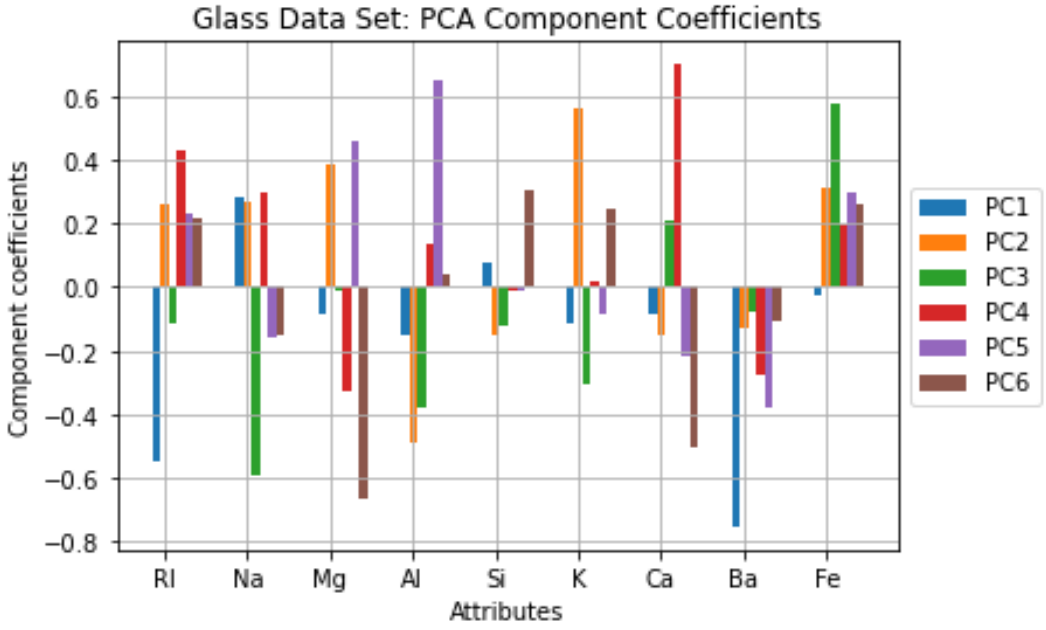

Figure 5: PCA Component Coefficients

3.2.3 The data projected onto the considered principal components

We plotted the data projected onto the computed PC1 and PC2 below in fig 3.2.3. It becomes evident that both principle components PC1 and PC2 are not able to separate the data points into distinctive clusters in respect to their classes. A reason for this could be that, the PC1 and PC2 together only explain up to 50% of the variance in the data (see 3.2.1). Also, the spread of data points among the X and Y axis seems to be similar as both principle components do not differ greatly in their explained variance of the data set (27.9% and 22.8%). We can also see that some classes experience a higher spread among both axes compared to other classes. For example, the class vehicle_windows_float_class (green) seems to have a greater variation among both principle components axes while the data points of the class building_windows_non_float_processed (orange) are centered around a smaller projection area. This could mean, that both principle components do capture more variance of attributes that are relevant to the (green) class than compared to the (orange) class.

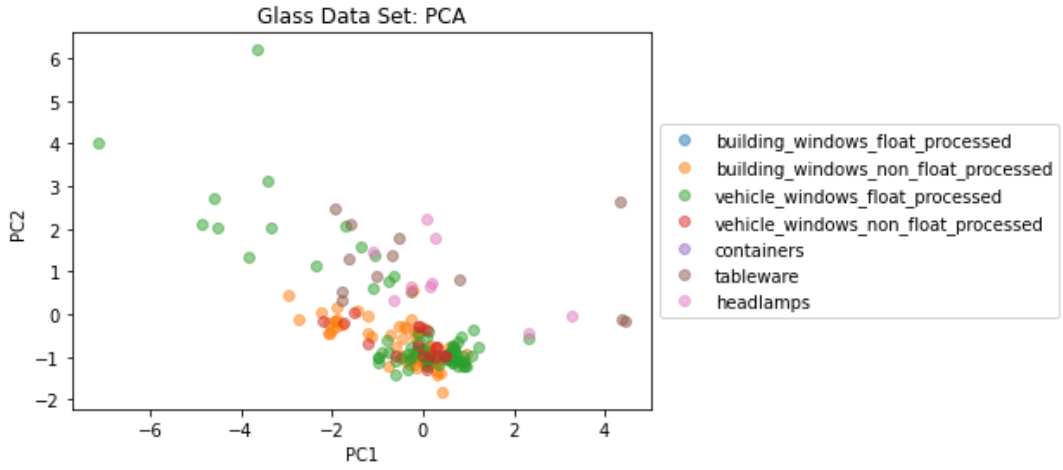

Figure 6: Scatterplot: Data projected onto PC1 and PC2

In the 3D-Scatterplot below in 3.2.3 we plotted the first 3 principle components to see if the class clusters will be more distinguishable. However, this seems to be not the case as many data points of different classes are still sitting on top of each other. It seems to be that we need more principle components to project the data in detail. Again, a reason for this could be that the cumulative variance of these principle component only adds up to around 66% of the data set.

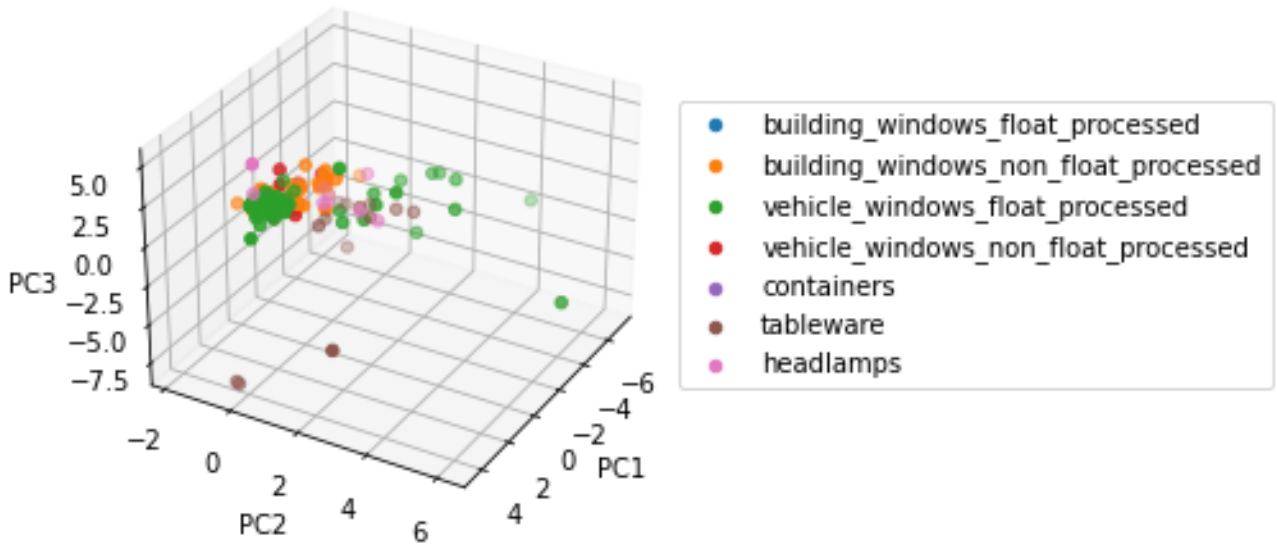

Figure 7: Scatterplot: Data projected onto PC1, PC2 and PC3

4 Summary

In summary, we have 214 records, each belonging to a specific type of glass, of which there are six different ones in the data set (excl. *vehicle_windows_non_float_processed*). Apart from the class, each of these records consists of nine relevant attributes storing information about the refractive index of the glass (RI) and its chemical composition (Na, Mg, Al, Si, K, Ca, Ba, Fe). Our data set included no missing or corrupted values and we argue that the outliers being present in some of the attributes are realistic and should be taken into account. Additionally, five attributes (Na, Si, Al, Ri and Ca) seemingly follow a normal distribution and the rest of the attributes seem to be of mixed distributions. There is a strong correlation among the attributes RI and Ca as well as RI and Si, while seven other attribute combinations have a medium correlation. After performing the PCA analysis, we saw that the first six principle components are able to summarize more than 90% of the variance in the data set. Focusing on the first three principle components, PC1 mainly captures the variance of RI, Na and Ba, while PC2 captures the variance of Ri, Na, Mg, Al, K and Fe. Lastly, PC3 mainly captures the variance of Na, Al, K, and Fe. When projecting the data onto the first two and first three principle components in a 2-D and 3-D scatter plot, we can see that many data points of different classes are sitting on top of each other. This means that the first two/three PCs are not able to clearly distinguish between the classes, which indicates that we need more principle components and thus, a higher dimension for the main problem of interest which is the classification of different glass types.

5 Exam Problems

5.1 Question 1

Answer: D

The scale of x_1 (Time of the day) is interval, as there is a clear distance in time between two different values given by a unit scale of 30 minutes. x_6 (Traffic lights) and x_7 (Running over) is ratio, as there is a clear distance among values and a value of 0 means the absence of broken traffic lights and run over accidents. x_7 (congestion level) is ordinal as it is ranked from low to high, but without a clear distance among these rankings.

5.2 Question 2

Answer: A

The p-norm distance of ∞ is given by:

$dp(x, y) = max|x - y|$, which equals 7 for the vectors x_{14} and x_{18}.

5.3 Question 3

Answer: A

The variance explained by a number of principal components is given by:

$$\frac{\Sigma_{j=1}^{m}\sigma_i^2}{\Sigma_{i=1}^{n}\sigma_j^2}$$

where m is the number of assessed principal components and n is the total number of principal components. So, the variance explained by the first four principal components is 86.67% , which is greater than 80%.

5.4 Question 4

Answer: D

$$V_{PC2} = \begin{bmatrix} -0.5 \\ 0.23 \\ 0.23 \\ 0.09 \\ 0.8 \end{bmatrix}$$

The directions for the second principle component are given in the second column of the V matrix. Here we can see a negative coefficient of -0.5 for the variable *Time of the day*, a positive coefficient of 0.23 for the variables *Broken truck* and *Accident victim*, and a positive coefficient of 0.8 for the variable *Defects*. This means that when an observation has a low *Time of the day* value, and high values in the rest of the above noted variables, the projection of this observation onto this principle component should typically be positive.

5.5 Question 5

Answer: D

The Jaccard similarity is given by the following formula:

$$J = \frac{F_{11}}{K - F_{00}}$$

, where $F_{00} = K - F_{11} - F_{10} - F_{01}$, $K = 2000$,
$F_{11} = 2$, $F_{10} = 8 - 2$ and $F_{01} = 7 - 2$

We can also apply the following formula:

$$J = \frac{|A \cap B|}{|A \cup B|}$$

, where $|A \cap B| = 2$ describes the number of shared words between both documents and $|A \cup B| = 13$ describes the total number of distinct words in both documents.

5.6 Question 6

Answer: B

$$P(x_2 = 0|y = 2) = P(x_2 = 0, x_7 = 0|y = 2) + P(x_2 = 0, x_7 = 1|y = 2) = 0.81 + 0.03 = 0.84$$

6 Distribution of workload

	S223534	S223001	S222736
Part1	40%	30%	30%
Part2	30%	30 %	40 %
Part3	30%	40 %	30 %
Part4	33.33%	33.33%	33.33%
Questions	33.33%	33.33%	33.33%
Grand total	33.33%	33.33%	33.33%