Technical University of Denmark

02452 Machine Learning
DTU Compute
Fall 2025

Machine Learning Project 1

Data: Feature extraction, and visualization

September 23, 2025 Group 94

Name	Student number
Vincent Van Schependom	s251739
Diego Armando Mijares Ledezma	s251777
Albert Joe Jensen	s204601

(a) Group members.

Task	Vincent	Diego	Albert
Section 1	30%	30%	40%
Section 2	40%	30%	30%
Section 3	30%	40%	30%
Section 4	30%	30%	40%
LaTeX	90%	5%	5%

(b) Contributions & responsabilities table.

Table 1: Group information & work distribution.

Introduction

The objective of this report is to apply the methods that were discussed during the first section of the course *Machine Learning* [1] to a chosen dataset. The aim is to get a basic understanding of the data prior to the further analysis (project report 2).

The particular dataset that is being investigated is the *Glass Identification* dataset from 1987 by B. German [2]. Table 1a lists our full names and student numbers, while Table 1b shows an overview of the contribution of each team member.

Contents

Attribute	Description	Type of variable
ID	Observation ID (excluded from analysis)	Numeric (discrete)
RI	Refractive Index	Continuous
Na	Sodium oxide (Na_2O)	Continuous
Mg	Magnesium oxide (MgO)	Continuous
Al	Aluminum oxide (Al_2O_3)	Continuous
Si	Silicon oxide (SiO_2)	Continuous
K	Potassium oxide (K_2O)	Continuous
Ca	Calcium oxide (CaO)	Continuous
Ba	Barium oxide (BaO)	Continuous
Fe	Iron oxide (Fe_2O_3)	Continuous
type	Type of glass	Nominal

Table 2: [TODO: Fill in.]

	Abbreviation in dataset	Description
1	BW-FP	Building Window, Float Processed
2	BW-NFP	Building Window, Non Float Processed
3	VW-FP	Vehicle Window, Float Processed
4	VW-NFP	Vehicle Window, Non Float Processed
5	containers	Containers
6	tableware	Tableware (e.g. ...)
7	headlamps	Headlamps (e.g. ...)

Table 3: [TODO: Fill in.]

1 The *Glass Identification* dataset

[TODO: The introduction to the data set.]

2 A close look at the different attributes

[TODO: Detailed explaination of the attributes of the data.]

3 Descriptive analysis of the dataset

[TODO: ...]

3.1 Extreme values and outliers

Note that the IQR 'rule' is not a ground truth, but rather a way to detect *possible* outliers.

[TODO: ...]

3.2 Distribution of the attributes

Note that there are no non-float processed vehicle windows! (#VW-NFP=0)

[TODO: ...]

3.3 Correlation between attributes

	Mean	Std.	Min	Max	Skew	Kurtosis
RI						
Na						
Mg						
Al						
Si						
K						
Ca						
Ba						
Fe						

(a) Summary statistics.

	type	Cnt.	Freq.
1	`BW-FP`		
2	`BW-NFP`		
3	`VW-FP`		
4	`VW-NFP`		
5	`containers`		
6	`tableware`		
7	`headlamps`		

(b) Absolute and relative frequencies of `type`.

Table 4: [TODO: Fill in.]

[TODO: ...]

4 Principal Component Analysis

4.1 Need for standardisation

Since the attributes are expressed on different numerical scales (see Section 3.2), the variables are standardised by subtracting the mean and dividing by the standard deviation. This ensures that no single attribute dominates the analysis merely due to its scale.

4.2 Dimension reduction

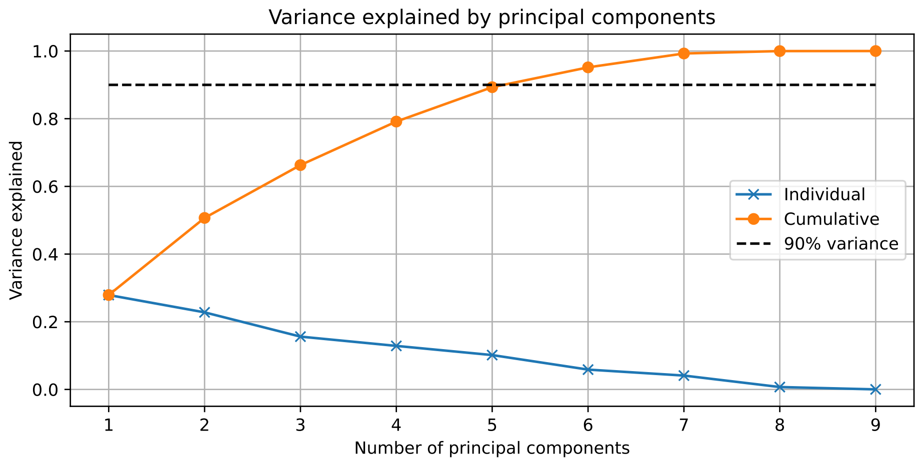


Figure 1: Explained (cumulative) variances for the 9 principal components $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_8$

.

The goal is to reduce the original 9-dimensional dataset to an M-dimensional representation with $M < 9$, using the first M principal components $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_{M-1}$. As shown in Figure 1, selecting $M = 5$ principal components explains 89.31 % of the total variance. By increasing the dimensionality to $M = 7$, as much as 99.27 % of the variance can be retained, essentially preserving almost all of the original information.

Choosing $M = 5$ provides a balance between dimensionality reduction and information retention. While $M = 7$ captures nearly all of the variance, the marginal gain in explained variance beyond the first five

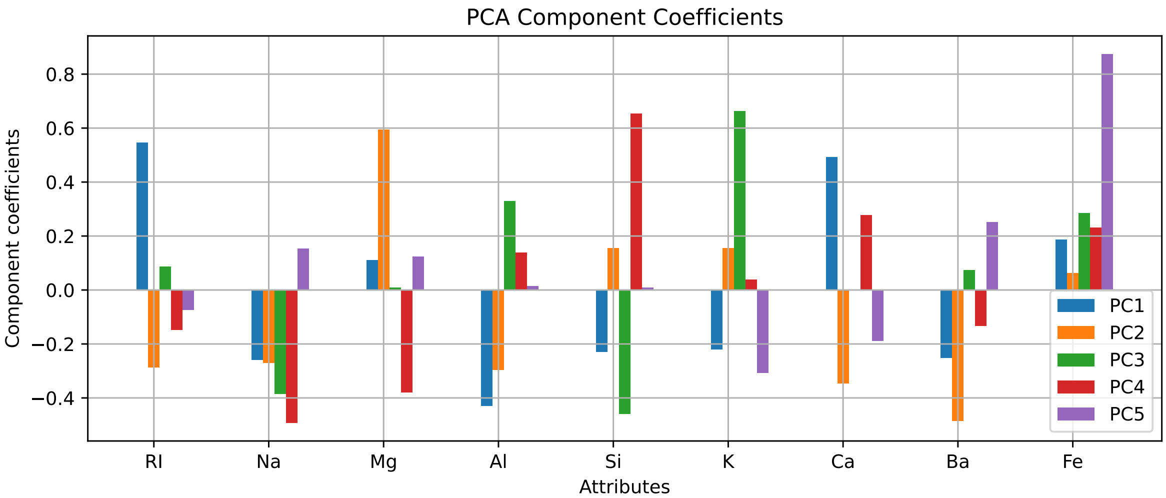


Figure 2: Loadings of the original variables on the first five principal components. Positive and negativ values indicate the *direction* of influence, while the magnitude reflects the *strength* of the contribution.

components is relatively small compared to the added complexity. Retaining five dimensions reduces computational cost, simplifies subsequent analysis, and removes noise while still preserving the majority of the data's variability.

4.3 Principal directions

The *principal directions* of the first M principal components are defined by the eigenvectors $\boldsymbol{v}_i$ that span the subspace of reduced dimensionality. These vectors form the transformation matrix $\boldsymbol{V}_M$, which is applied to the standardised data $\boldsymbol{X}$ to obtain the projected representation $\boldsymbol{B} = \boldsymbol{V}_M\boldsymbol{X}$. The new coordinates $\boldsymbol{B}$ capture the structure of the data in fewer dimensions while emphasising the directions of greatest variance.

Variable	**PC_0**	**PC_1**	**PC_2**	**PC_3**	**PC_4**
`RI`	0.545	-0.286	0.087	-0.147	-0.074
`Na`	-0.258	-0.270	-0.385	-0.491	0.154
`Mg`	0.111	0.594	0.008	-0.379	0.124
`Al`	-0.429	-0.295	0.329	0.138	0.014
`Si`	-0.229	0.155	-0.459	0.653	0.009
`K`	-0.219	0.154	0.663	0.039	-0.307
`Ca`	0.492	-0.345	-0.001	0.276	-0.188
`Ba`	-0.250	-0.485	0.074	-0.133	0.251
`Fe`	0.186	0.062	0.284	0.230	0.873

Table 5: The principal directions (a.k.a. the *loadings*) of the first $M = 5$ principal components $PC_i = \boldsymbol{v}_i$ in the rotation matrix $\boldsymbol{V}_M$. Larger absolute values indicate stronger influence of a variable on a given component.

Based on the loadings in Table 5 and the corresponding visualisation in Figure 2, the first principal component (PC_0) is most strongly influenced by `RI` and `Ca`, which carry positive loadings, and negatively by `Al` and `Si`. This suggests that PC_0 captures a trade-off between refractive index and calcium content versus aluminium and silicon.

The second principal component (PC_1) assigns a large positive loading to `Mg`, while strongly downweighting `Ba` and, to a lesser degree, `Na`. This indicates that PC_1 primarily reflects a contrast between magnesium concentration and the presence of barium and sodium.

The third principal component (PC_2) is dominated by a large positive loading for `K`, balanced by strong negative contributions from `Si` and `Na`. This points to a dimension that separates potassium-rich compositions from those with higher silica and sodium content.

The fourth principal component (PC_3) shows high positive contributions from `Si`, `Ca`, and `Fe`, with negative influence from `Mg` and `Na`. It therefore captures variation where higher levels of silicon, calcium, and iron occur together in opposition to magnesium and sodium.

Finally, the fifth principal component (PC_4) is characterised by a dominant positive loading for `Fe`, while moderately influenced by `Ba` and negatively by `K`. This indicates that PC_4 primarily isolates variation in iron concentration, with some contribution from the balance between barium and potassium.

4.4 Projected data

4.4.1 Biplots

The projection of the standardised dataset onto the principal component axes produces a lower-dimensional representation that can be visualised in two-dimensional *biplots*. In these plots:

- Each point corresponds to a *score*, representing the coordinates of an original observation in the space spanned by the selected principal components. The scores have been *rescaled* to the interval $[-1, 1]$ to facilitate comparison with the arrows, representing variable loadings.
- The arrows correspond to the *loadings*, i.e., the contributions of the original variables to the principal components. The direction of an arrow indicates how the variable influences the component axes, and its length reflects the magnitude of this influence.

Together, scores and loadings allow for simultaneous interpretation of both the observations and the variables.

4.4.2 Interpretation based on glass types

Figure 3 shows two biplot visualisations. In the first biplot (PC1 vs. PC2), which captures 50 % of the variance, the dominant trends in the data are visible, and some separation between glass types can be observed [TODO: explain]. For the second biplot, [TODO: fill out.]

The projected scores reveal that some classes, such as `BW-FP` and `R`, cluster distinctly in the reduced space, whereas others, such as `LW-FP` and `GL`, exhibit more overlap. The loadings indicate that the separation of classes along the axes is largely driven by specific chemical compositions; for example, PC1 contrasts samples with high `RI` and `Ca` against those with higher `Al` and `Si`, explaining why some classes separate along this component. [TODO: Explain the concrete projections of the different classes further.]

4.4.3 Extreme scores

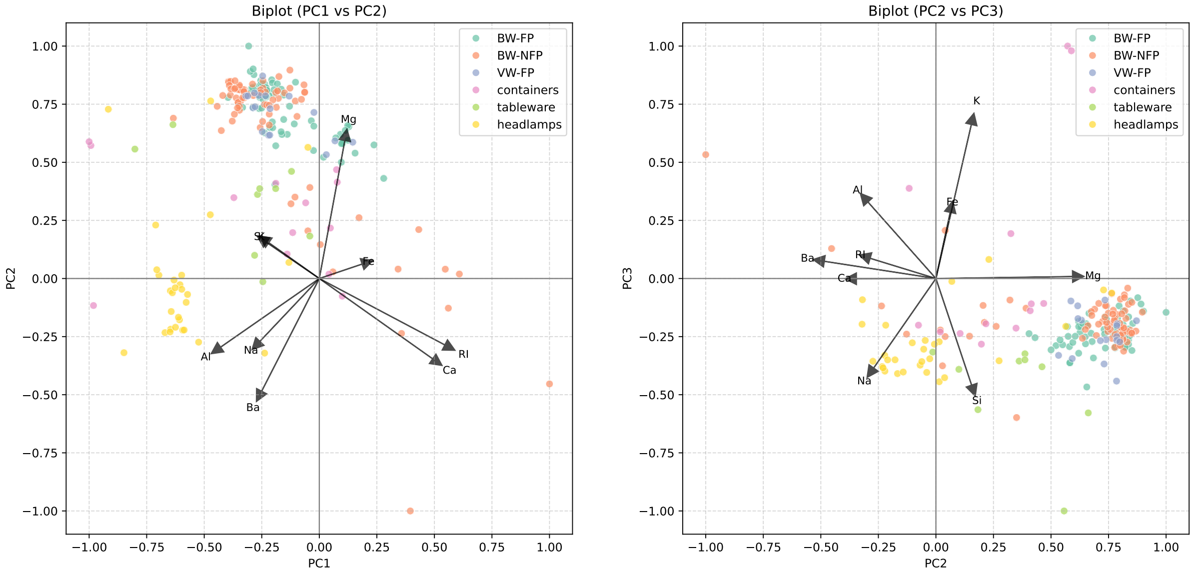
[TODO: Explain the observations that score extremely high/low in certain principal directions.]

5 Use of GenAI

[TODO: Explain what kind of GenAI you used.]

Summary

[TODO: A short summary of what we discussed in the whole paper.]



(a) PC1 vs PC2

(b) PC2 vs PC3

Figure 3: Biplots of the glass dataset showing both projected samples (colored by class) and variable loadings.

References

[1] Tue Herlau, Mikkel N. Schmidt, and Morten Mørup. *Introduction to Machine Learning and Data Mining*. Technical University of Denmark (DTU), Lyngby, Denmark, 2023. Lecture notes, Fall 2023, version 1.0. This document may not be redistributed. All rights belong to the authors and DTU.

[2] B. German. Glass Identification. UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C5WW2P.

Appendix

A Test