Technical University of Denmark

02452 Machine Learning
DTU Compute
Fall 2025

Machine Learning Project 1

Data: Feature extraction, and visualization

September 30, 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	40%	0%	60%
Section 2	30%	50%	20%
Section 3	100%	0%	0%
LaTeX	95%	5%	0%

(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

1 The *Glass Identification* dataset

The Glass Identification dataset comes from forensic science research in the 1980s, originally compiled at the Institute of Forensic Medicine, with support from the UCI Machine Learning Repository [2]. It was created to help develop methods for identifying types of glass found at crime scenes, such as window glass, containers, or headlamps, based on their chemical makeup.

For each of the 214 observations, the measured attributes are the glass `type`, the refractive index (`RI`) and 8 different oxides, all of which can be seen in Table 2.

1.1 Previous analysis of the data

Several studies have analysed the dataset to improve multi-class classification performance. Zhou et al. applied machine learning techniques in 2023, focusing on optimising Support Vector Machines (SVM) through grid search and Bayesian methods [3]. Their optimised SVM model achieved a remarkably high accuracy of 99.25%, outperforming baseline models such as logistic regression, and demonstrated strong predictive ability when only chemical composition was available.

In contrast, Bhowmick and Saha concentrated on addressing class imbalance and noisy data, also in 2023 [4]. They removed outliers using interquartile range filtering, applied ReliefF feature ranking, and balanced classes with SMOTE before training an inverse-distance-weighted k-nearest-neighbor (kNN) classifier. Their improved pipeline reached 78.9% accuracy with an F-measure of 0.791, showing how preprocessing can boost kNN's effectiveness on this imbalanced dataset.

1.2 Goals and *main machine learning aim*

The objective is to build models for both regression and classification. For the *regression task*, the target variable will be the Refractive Index (`RI`), predicted using eight oxides that were measured for each observation. For the *classification task*, the target variable is the glass type (`type`), which is divided into seven distinct classes. This label will be predicted based on the other attributes, namely the refractive index of the glass and its measured oxide composition

Since the dataset is inherently structured around glass type classification, it is clear that the *main machine learning aim* of this project is classification, while regression serves as a secondary task.

2 A close look at the different attributes

2.1 Feature Types

The glass type is, of course, a categorical, *nominal* variable, since no glass type is considered 'better' or 'worse' than another.

The oxide components are measured in terms of weight percentage in the corresponding oxide, and are classified as *continuous*, *ratio* variables: a value of zero (e.g., 0.0% Ba) signifies a true absence of that component, making multiplicative comparisons (ratios) meaningful. The refractive index (`RI`), though also continuous, is an *interval* variable. This is because it is calculated as $\texttt{RI} = \frac{c}{v}$ where c is the speed of light in vacuum and v is the speed of light in the medium. The existence of a true zero would imply infinite light speed so `RI` cannot be considered ratio-scaled.

It should be noted that the oxide components collectively form *compositional data*, meaning their values are statistically interdependent because they represent parts of a whole (i.e., they sum to approximately 100% for each glass sample). This inherent interdependence can influence statistical modeling.

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: The 10 studied attributes, along with the observation ID, which was excluded from the analysis.

	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: The different glass types in the dataset.

2.2 Need for scaling

Despite the compositional nature, we should definitely scale the oxide features, primarily due to the significant disparity in their numerical ranges. This can be seen from the summary statistics in Table 4a, as well as on the unscaled boxplots in Figure 1a. For example, Si (Silicon) often has values around 72%, whereas Fe (Iron) and Ba (Barium) frequently register values close to 0.0%. If these features were used in their raw form, distance-based machine learning algorithms (such as k-Nearest Neighbors or k-Means Clustering) would be unfairly dominated by features with the largest magnitudes, such as Si or Ca.

2.3 Distributions and potential outliers

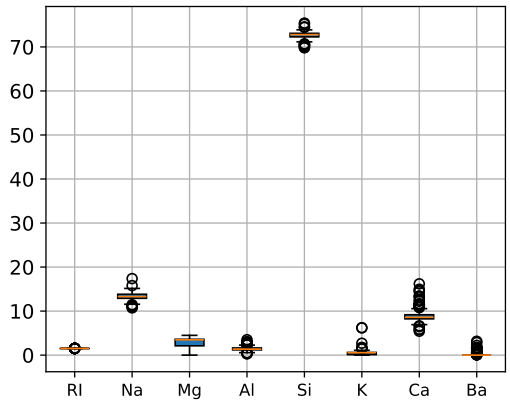
Figures 1b and 2 reveal that K, Ba, and Fe are highly skewed with occasional extreme values. Mg displays a bimodal pattern, suggesting differences between glass types. Ca shows wide spread but is roughly symmetric, whereas RI, Na, Al, and Si appear closer to normal. Standardisation reduces scale differences and partially mitigates skewness, but highly skewed or multimodal variables may require additional preprocessing such as log-transformations or class-sensitive handling.

The distribution of the glass type - as can be see in Table 4b - is imbalanced: 70 float-processed building windows, 76 processed building windows, 17 vehicle windows, 13 containers, 9 tableware, and 29 headlamps. Class 4 (VW-NFP) has no samples.

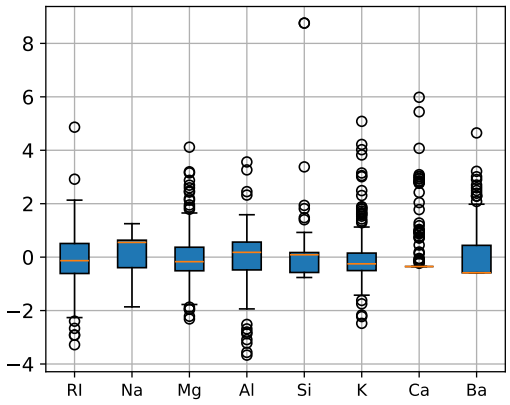
2.4 Correlation between attributes

The correlation structure of the nine numerical attributes is summarised by the correlation-matrix heatmap (Figure 3a). Several clear patterns are visible in the *linear* correlations:

- **Strong positive correlation between refractive index and calcium:** RI and Ca have a large positive linear association (Pearson $r \approx +0.81$). This indicates that samples with higher calcium



(a) Unscaled boxplots of the refractive index (`RI`) and the 8 oxides.



(b) Scaled boxplots of the refractive index (`RI`) and the 8 oxides.

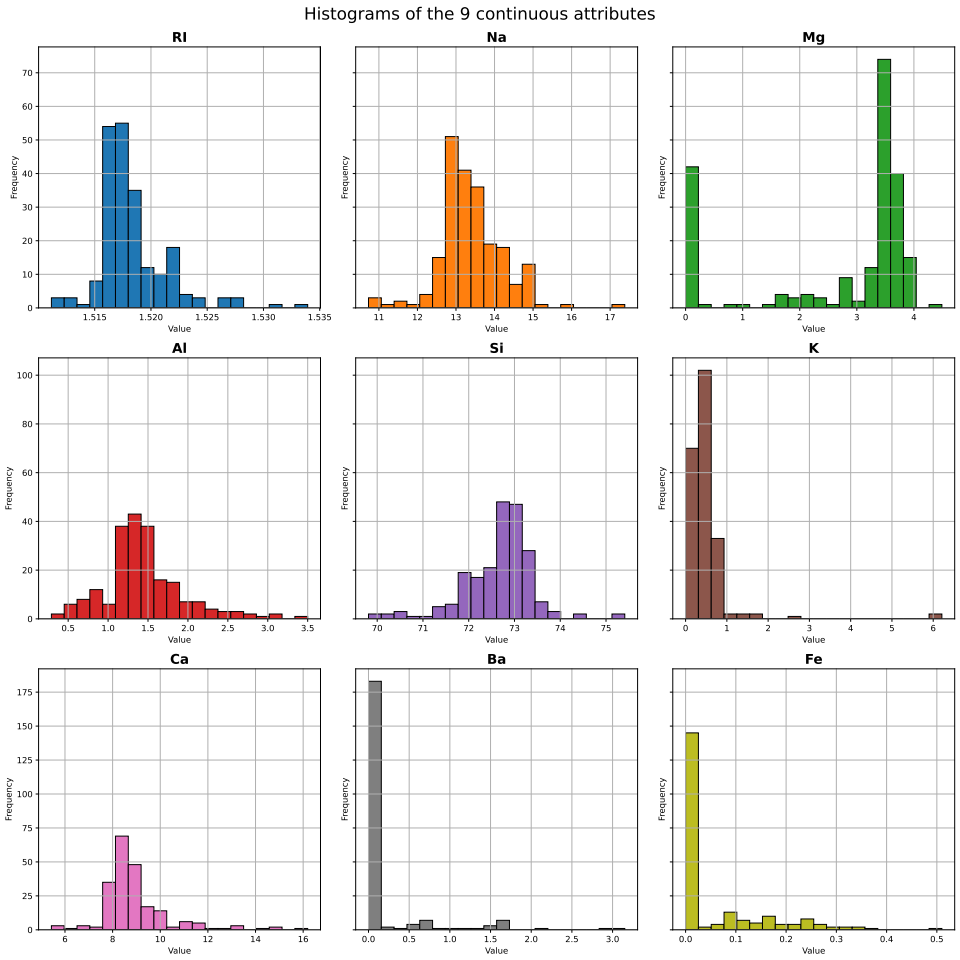


Figure 2: Relative frequency histograms for the nine numerical attributes

	Mean	Std.	Min	Max	Skew	Kurtosis
RI	1.518	0.003	1.511	1.534	1.625	4.932
Na	13.408	0.817	10.730	17.380	0.454	3.052
Mg	2.685	1.442	0.000	4.490	-1.153	-0.410
Al	1.445	0.499	0.290	3.500	0.907	2.061
Si	72.651	0.775	69.810	75.410	-0.730	2.968
K	0.497	0.652	0.000	6.210	6.552	54.690
Ca	8.957	1.423	5.430	16.190	2.047	6.682
Ba	0.175	0.497	0.000	3.150	3.416	12.541
Fe	0.057	0.097	0.000	0.510	1.754	2.662

(a) Summary statistics.

	type	Cnt.	Freq.
1	BW-FP	70	0.327
2	BW-NFP	76	0.355
3	VW-FP	17	0.079
4	VW-NFP	0	0.000
5	containers	13	0.061
6	tableware	9	0.042
7	headlamps	29	0.136

(b) Absolute and relative frequencies of type.

Table 4: Summary statistics of continuous variables and frequency distribution of glass type.

content tend to have larger refractive indices in this dataset.

- **Notable negative correlation between RI and Si:** RI and Si are negatively correlated ($r \approx -0.54$), so silica-rich samples tend to have somewhat lower RI values.
- **Moderate correlations involving Mg, Al and Ba:** Mg is negatively correlated with both Ba and Al (e.g. Mg–Ba $r \approx -0.49$, Mg–Al $r \approx -0.48$), while Al and Ba show a moderate positive correlation (Al–Ba $r \approx +0.48$). Na and K show only weak-to-moderate associations with other oxides (e.g. Na–Ba $r \approx +0.33$, Al–K $r \approx +0.33$).
- **Generally weak correlations for Fe:** iron (Fe) has only small correlations with the other variables (all $|r| \ll 0.5$).

These patterns have two important consequences for downstream analysis. First, several pairs/groups of variables show moderate-to-strong linear dependence (e.g. RI/Ca, Mg/Ba, Al/Mg). Such *collinearity* can inflate variance of coefficient estimates in linear methods and makes interpretation of individual coefficients difficult. Second, the oxides are *compositional* (the percentages sum to one), which can induce spurious correlations; standard multivariate techniques (PCA, regularised regression, or compositional transforms such as log-ratios) should be considered if the analysis requires interpretable linear relationships.

For visualization and dimensionality reduction we used PCA (Section 3). The correlation structure above is consistent with the PCA loadings: PC1 places large positive weight on RI and Ca, and negative weight on Al and Si; PC2 emphasises Mg and Ba. See Figure 3a for the full matrix and the scatter example (RI vs Ca) that illustrates the strongest linear relationship identified above.

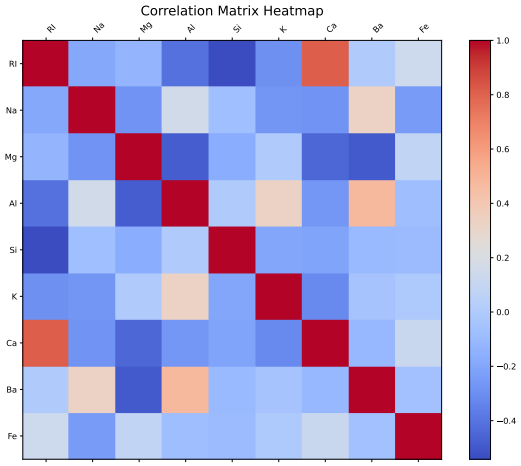
3 Principal Component Analysis

3.1 Need for standardisation

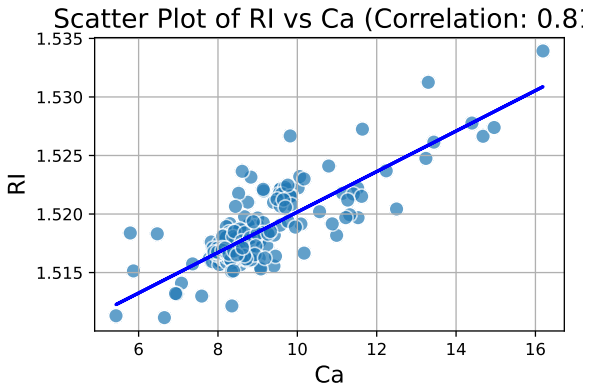
Since the attributes are expressed on different numerical scales (see Section 2.2), the variables are standardised by subtracting the mean and dividing by the standard deviation. This ensures that no single attribute dominates the analysis merely due to its scale.

3.2 Dimension reduction

The goal is to reduce the original 9-dimensional dataset to an M-dimensional representation with $M < 9$, using the first M principal components $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_M$. As shown in Figure 4, 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.



(a) Correlation matrix for the nine numerical attributes.



(b) Individual correlation between `RI` and `Ca`

Figure 3: Whole correlation matrix and individual correlation between `RI` and `Ca`.

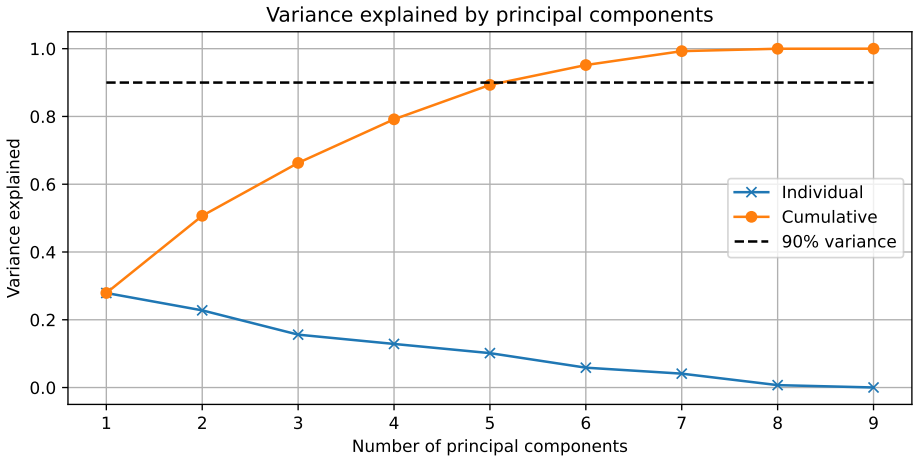


Figure 4: Explained (cumulative) variances for the 9 principal components $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_8$

.

Variable	$\mathbf{PC}_1$	$\mathbf{PC}_2$	$\mathbf{PC}_3$	$\mathbf{PC}_4$	$\mathbf{PC}_5$
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 $\text{PC}_i = \boldsymbol{v}_i$ in the rotation matrix $\boldsymbol{V}_M$. Larger absolute values indicate stronger influence of a variable on a given component.

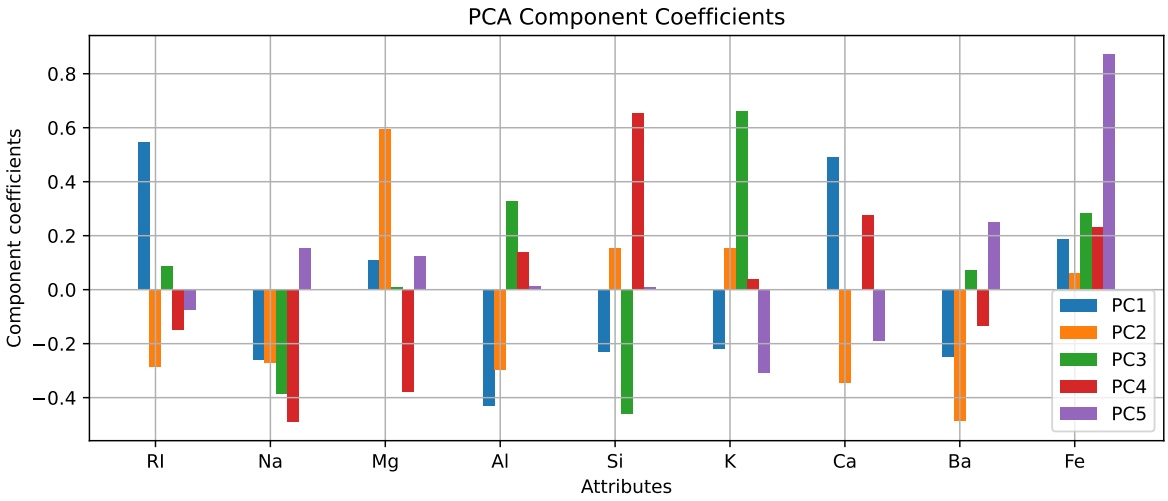


Figure 5: Loadings of the original variables on the first five principal components. Positive and negativ values indicate the *direction* of influence, while the magnitude reflects the *strength* of the contribution.

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

3.3 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 $\tilde{\boldsymbol{X}}$ to obtain the projected representation $\boldsymbol{B} = \boldsymbol{V}_M\tilde{\boldsymbol{X}}$. The new coordinates $\boldsymbol{B}$ capture the structure of the data in fewer dimensions while emphasising the directions of greatest variance. Based on the loadings in Table 5 and the corresponding visualisation in Figure 5, we interpret the first 5 principal components:

- The first principal component (PC_1) is most strongly influenced by `RI` and `Ca`, which carry positive loadings, and negatively by `Al` and `Si`. This suggests that PC_1 captures a trade-off between

refractive index and calcium content versus aluminium and silicon.

- The second principal component (PC_2) assigns a large positive loading to `Mg`, while strongly down-weighting `Ba` and, to a lesser degree, `Na`. This indicates that PC_2 primarily reflects a contrast between magnesium concentration and the presence of barium and sodium.
- The third principal component (PC_3) 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_4) 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_5) is characterised by a dominant positive loading for `Fe`, while moderately influenced by `Ba` and negatively by `K`. This indicates that PC_5 primarily isolates variation in iron concentration, with some contribution from the balance between barium and potassium.

3.4 Projected data

3.4.1 Biplots

The projection of the standardized dataset onto the principal component axes produces a lower-dimensional representation that can be visualized 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.

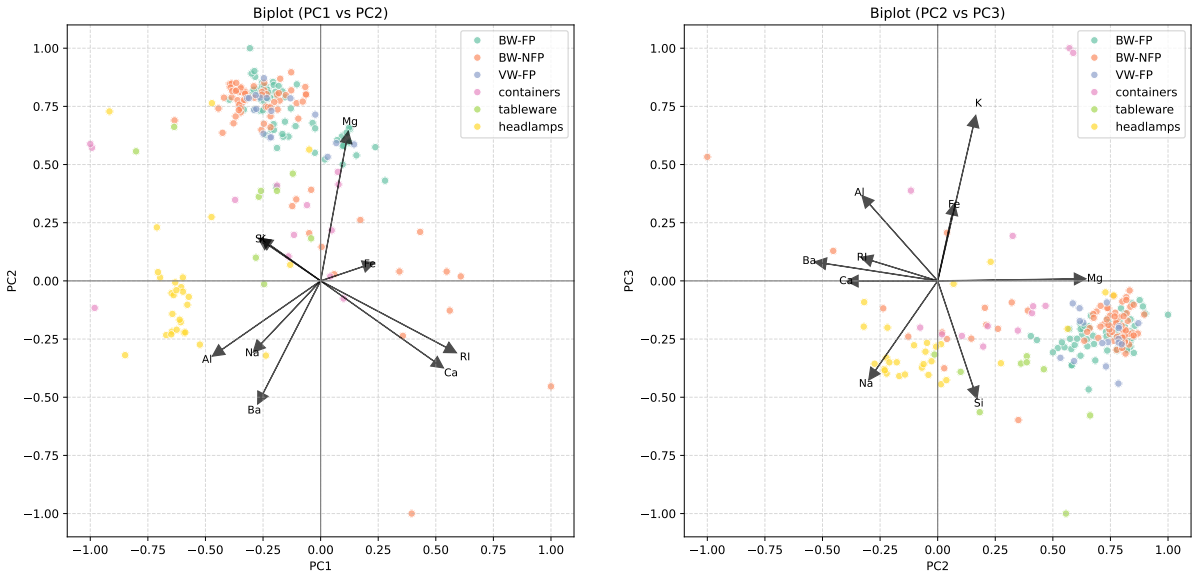
3.4.2 Interpretation

Figure 6 shows two biplot visualizations. 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. We can, for example, see a clear separation of headlamps, who on average score low on PC1 and have about a zero-loading for PC2. The building and vehicle windows, on the other hand, show significant overlap and can thus not be distinguished from each other by just looking at the first two principal components.

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.

4 Discussion

Summary of main findings. From the descriptive analysis and visualisations we draw the following principal conclusions. The dataset contains $N = 214$ samples described by nine numerical attributes (`RI` + eight oxides). There are no missing values. Many variables are skewed (notably `K` and `Ba`) and some have occasional extremes (especially `Ba` and `Ca`); `Mg` shows a bimodal-like pattern across the dataset. The class distribution is imbalanced: building-window classes dominate (70 and 76 samples), headlamps



(a) PC1 vs PC2

(b) PC2 vs PC3

Figure 6: Biplots of the glass dataset showing both projected samples (colored by class) and variable loadings.

are moderately represented (29), while container/tableware classes are small (13 and 9 samples) and one of the UCI classes is absent in our download. The correlation analysis (Section 2.4) reveals substantial linear dependence (e.g. `RI`–`Ca` $r \approx +0.81$, `RI`–`Si` $r \approx -0.54$, `Mg`–`Ba` $r \approx -0.49$), which motivated the PCA and dimensionality-reduction work in Section 2.2.

Feasibility of the classification aim. Our baseline supervised experiments (Random Forest and logistic regression on standardised oxide-only features) show that the multi-class classification task is feasible but non-trivial. The Random Forest achieved a test accuracy of approximately 0.674 on the held-out set and a 5-fold cross-validated mean accuracy of around (0.748 ± 0.029). The PCA biplots indicate that some classes (notably the two building-window classes) substantially overlap in the leading principal-component space and are therefore inherently harder to separate; by contrast, the headlamps class appears relatively well separated along PC1/PC2. Small sample sizes for containers and tableware make reliable classification for those classes challenging without further data or targeted rebalancing.

Actionable next steps and points of attention. Based on the visual and numerical diagnostics we consider the following steps:

1. **Address skewness and extremes:** apply robust transformations (e.g. log or Box–Cox for strictly positive oxides), and consider robust estimators or tree-based models that handle outliers naturally.
2. **Reduce redundancy:** use PCA (or regularised models) to control multicollinearity in linear approaches; alternatively, select a compact subset of chemically interpretable predictors.
3. **Compositional caution:** because oxide measurements are parts of a composition, compositional data methods (e.g. log-ratio transforms) could be considered.
4. **Class imbalance:** the class imbalance should be handled with care.

Taken together, these steps should increase both predictive performance and interpretability. The baseline results and the PCA visualisations indicate that the classification objective is realistic, but substantial care is needed for small classes and for correct preprocessing of compositional and skewed measurements.

Limitations. The main limitations at this stage are (i) class imbalance and small sample counts for some classes, (ii) compositional structure that can bias naive correlation-based interpretations, and (iii) sensitivity to outliers for some attributes (Ba, Ca, K). These caveats should guide model selection and evaluation in the next project phase.

Use of GenAI

For the sake of transparency, the project used generative-AI tools in a limited and documented manner: GenAI assisted primarily with *work-organisation suggestions* and language polishing when going from draft to final report. All quantitative analyses, figures, and numerical results reported in this document were produced by the code in the repository and the authors' own computations were *not* fabricated by GenAI. It should be noted that some team members make use of the VSCode Copilot code completion to speed up their programming work. Where GenAI suggestions were used, the group reviewed and validated them against the code outputs and figures before inclusion.

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.

[3] Jiasheng Zhou, Zhihong Fan, and Wenxin Zhang. Research on glass classification and identification based on machine learning and monte carlo algorithm. *Highlights in Science, Engineering and Technology*, 39:863–871, 04 2023.

[4] Sandipan Bhowmick and Ashim Saha. Enhancing the performance of knn for glass identification dataset using inverse distance weight, relieff ranking and smote. *AIP Conference Proceedings*, 2754(1):020021, 09 2023.

Appendix

A Repository and supplementary materials

The full notebook, scripts, and generated figures for this project are available in the project repository:

`https://github.com/schependom/DTU_machine-learning-projects/tree/main`

This repository contains the data-loading and analysis code that produced the tables and figures cited above (see the `figures/` folder for the PDF outputs referenced in the report).