Technical University of Denmark

02452 Machine Learning
DTU Compute
Fall 2025

Machine Learning Project 1

Data: Feature extraction, and visualization

September 29, 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	80%	10%	10%
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

1 The *Glass Identification* dataset

The Glass Identification dataset [2] comes from forensic science research in the 1980s, originally compiled at the Institute of Forensic Medicine, with support from the UCI Machine Learning Repository. 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. The dataset has been widely shared since its inclusion in the UCI Repository in 1987, and it remains a popular benchmark for testing classification methods in research and teaching.

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 Goal

[TODO: Explain, in the context of your problem of interest, what you hope to accom- plish/learn from the data using these techniques?]

[TODO: Explain which attribute you wish to predict in the regression based on which other attributes?]

[TODO: Which class label will you predict based on which other attributes in the classi- fication task?]

→ (**V:**) The target variable is the type of glass (`type`), divided into 7 distinct classes. [TODO: Complete this and tell that this will be the classification objective blabla.]

[TODO: Describe **main machine learning aim**: One of these tasks is likely more relevant than the rest and will be denoted the main machine learning aim in the following.]

→ (**V:**) Mention that it's pretty obvious (and previously mentioned) that the main objective is **classification**.

1.3 Data transformations

[TODO: Explain if you need to transform individual attribues in order to carry out these tasks (e.g. centering, standardization, discretization, log transform, etc.) and how you plan to do this.]

→ (**V:**) I would just mention overall standardisation is necessary, since it's all on different scales, a bit like in the PCA section (maybe move the text from there over here, and afterwards say something like "As we already mentioned in Section ..., standardisation is necessary" in the PCA section.). Also refer to the boxplots (unscaled and scaled).

2 A close look at the different attributes

"Assignment: *Describe if the attributes are discrete/continuous and whether they are nominal/or- dinal/interval/ratio.*"

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.]

The data set consists of 214 samples of glass, each characterised by glass type, along with 9 numeric features. There are *no missing values*. The attributes are all presented in Table 2 and the `type` attribute is unfolded in Table 3.

The refractive index (`RI`), and concentrations of eight different oxides, including sodium (`Na`), magnesium (`Mg`), aluminum (`Al`), silicon (`Si`), potassium (`K`), calcium (`Ca`), barium (`Ba`), and iron (`Fe`) all are *continuous* and [TODO: ratio or interval].

3 Descriptive analysis of the dataset

3.1 Summary statistics

[TODO: The assignment is as follows:]

"Include relevant summary statistics of the attributes. Reflect on the values. If your data set contains many similar attributes, you may restrict yourself to describing a few representative features (apply common sense). You can place additional results in the appendix if needed."

[TODO: Reshuffle the text such that it follows the structure of this LaTeXdocument.]

[TODO: Put boxplots, histograms.]

3.2 Extreme values and outliers

[TODO: Reshuffle the text such that it follows the structure of this LaTeXdocument.]

[TODO: Mention boxplots.]

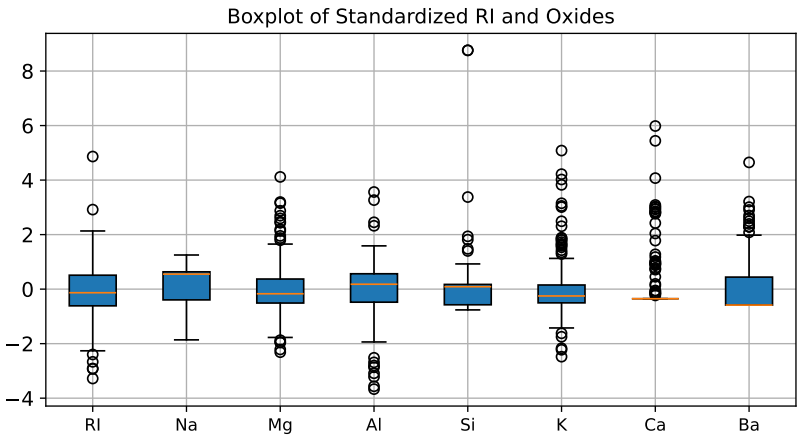


Figure 1: Boxplots of the refractive index (`RI`) and the 8 oxides.

3.3 Distribution of the attributes

[TODO: Reshuffle the text such that it follows the structure of this LaTeXdocument.]

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

[TODO: Mention the histograms and maybe the boxplots, too.](Figure `\ref{fig:histograms}`)

	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`.

3.4 Correlation between attributes

[TODO: Fill out this section.]

[TODO: Reference the matrix correlation plot.](Figure `\ref{fig:correlation}`)

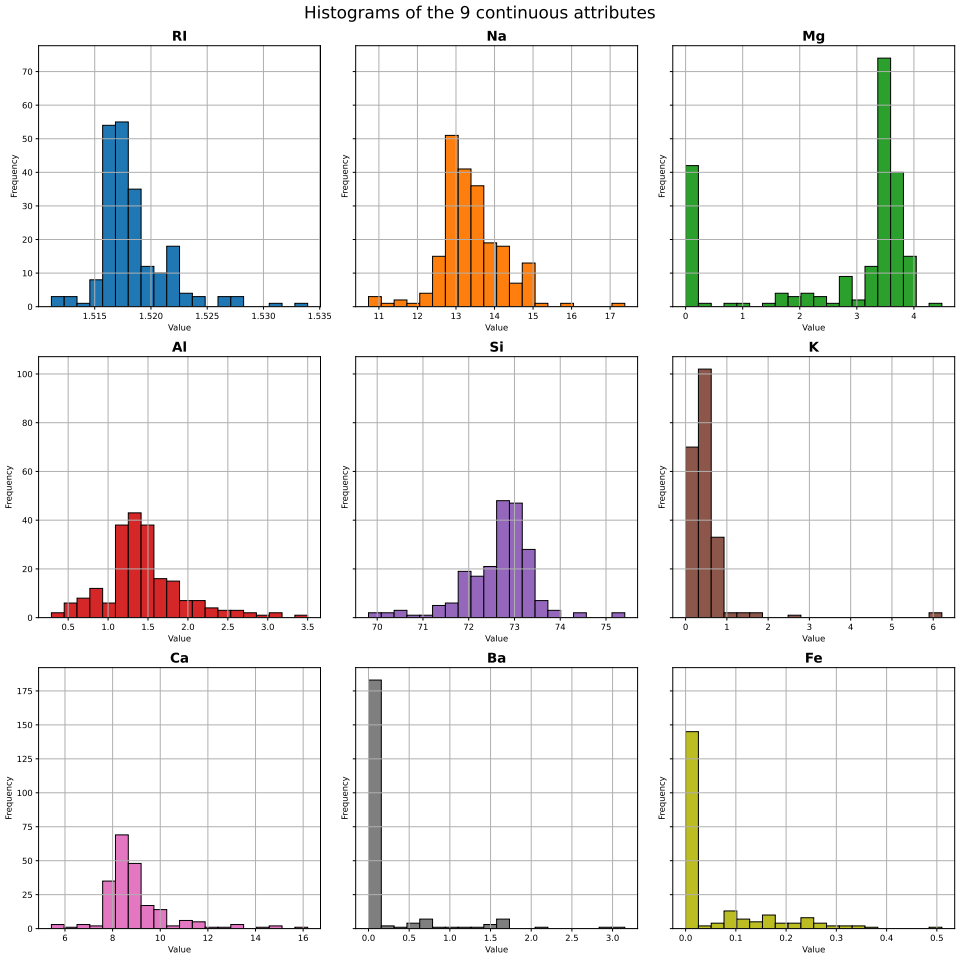
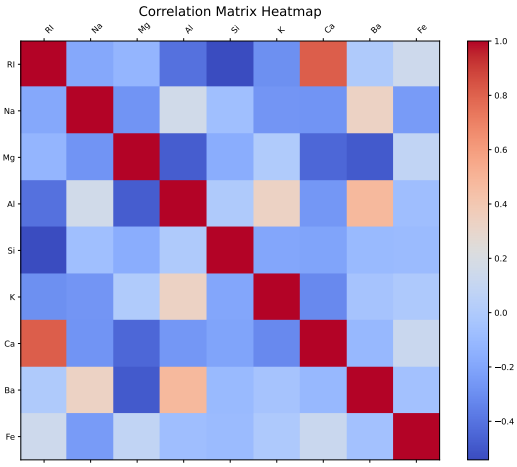
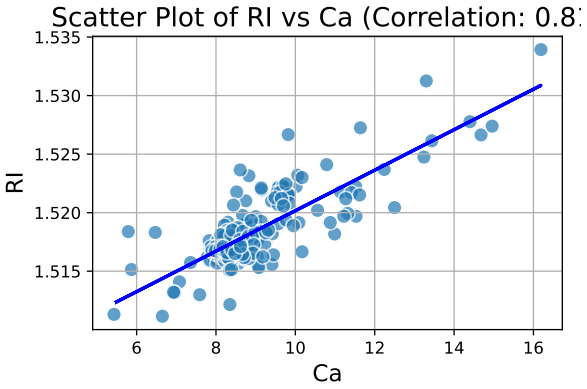


Figure 2: Relative frequency histograms for the nine numerical attributes. [TODO: Describe the general distributions.]



(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`.

4 Principal Component Analysis

4.1 Need for standardisation

Since the attributes are expressed on different numerical scales (see Section 3.3), 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

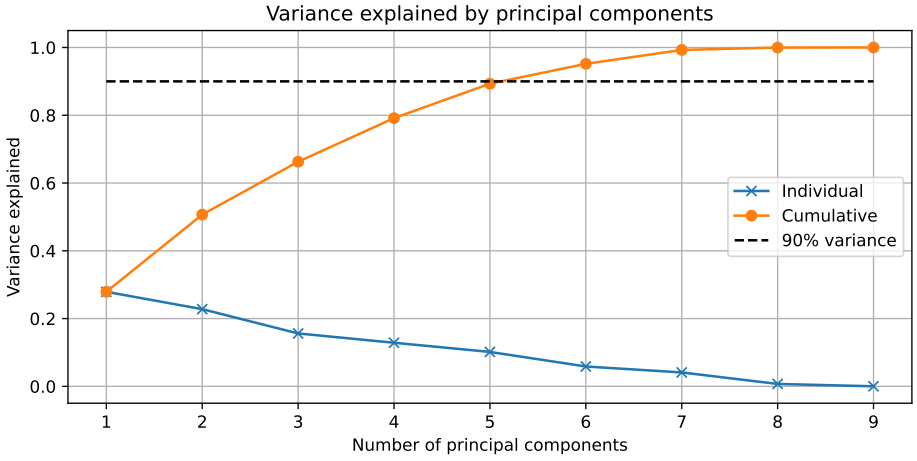


Figure 4: Explained (cumulative) variances for the 9 principal components $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_8$

.

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 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.

Choosing $M = 5$ provides a balance between dimensionality reduction and information retention. While

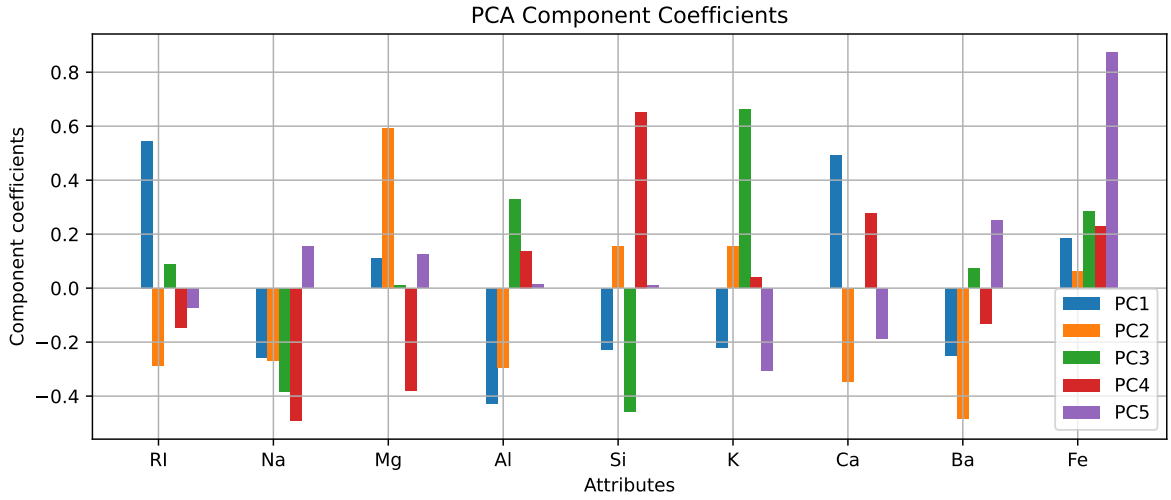


Figure 5: Loadings of the original variables on the first five principal components. Positive and negativ values indicate the *direction* of influence, while the magnitude reflects the *strength* of the contribution.

$M = 7$ 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.

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	$\mathbf{PC}_0$	$\mathbf{PC}_1$	$\mathbf{PC}_2$	$\mathbf{PC}_3$	$\mathbf{PC}_4$
`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 $\mathrm{PC}_i = \boldsymbol{v}_i$ 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 5, 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.

[TODO: Review all the stuff that is below and check for errors.]

The second principal component (PC_1) assigns a large positive loading to `Mg`, while strongly down-weighting `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 6 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.]

[TODO: Check the paragraph below and fill out further.]

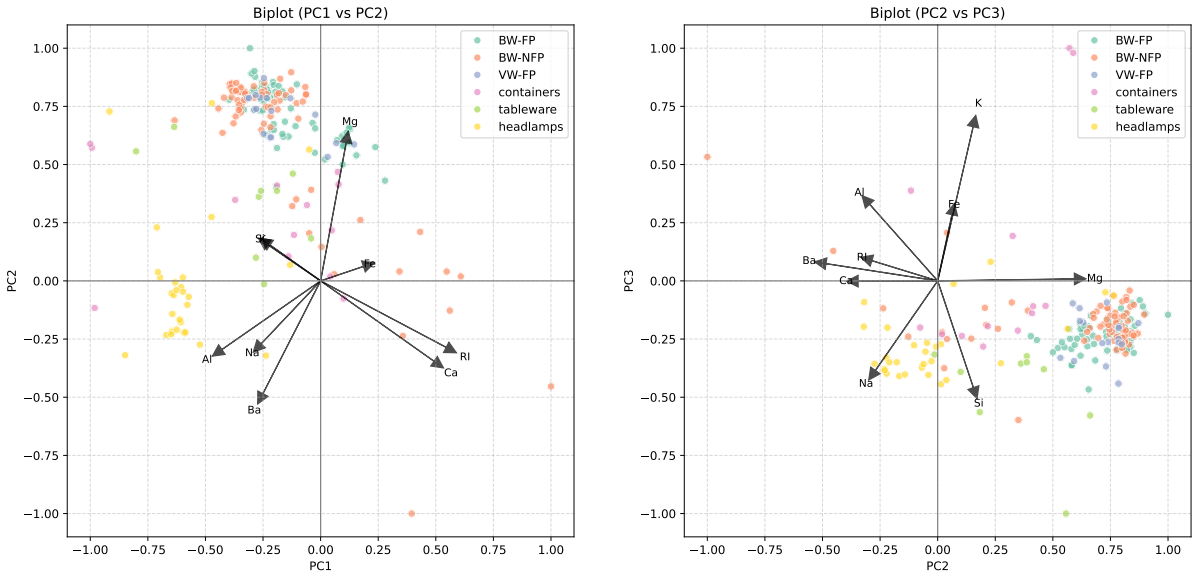
The projected scores reveal that some classes, such as `BW-FP` and `R`, cluster distinctly in the reduced space, whereas others, such as `LW-FP` and `GL`, exhibit more overlap. The loadings indicate that the separation of classes along the axes is largely driven by specific chemical compositions; for 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

Keep in the back of your mind for what you used GenAI and add it at the end.



(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.

Summary

[TODO: Assignment: "A discussion explaining what you have learned about the data. Summarize the most important things you have learned about the data and give your thoughts on whether your primary machine learning aim appears to be feasible based on your visualization."]

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.