Project 1

Glass Identification Data set

Albert Frisch Møller s214610 and Mark Andrawes s214654

	Section 1	Section 2	Section 3	Section 4	Exam questions
s214610	40%	40%	60%	40%	60%
s214654	60%	60%	40%	60%	40%

1 Description

The dataset which will be used in this project is entitled 'Glass Identification'[1].It contains forensic information, such as the weight percentage of magnesium in magnesium oxides found in the glass and refractive index, about different types of glass, such as headlamps or building windows. The study of glass identification was motivated by criminal investigations, so that evidence in the form of glass could be correctly identified and used to help solve crimes.

After applying classification and regression to our data, we hope to learn more about the relations between attributes and the dominant attributes for each glass type. In the context of our problem of interest, we wish to be able to successfully classify a piece of glass, as well as predict particular attributes based on others. More specifically, we would like to predict the class labels 'headlamps', 'tableware', 'containers', 'vehicle windows' (2 types), and 'building windows' (2 types). We wish to be able to successfully classify a piece of glass given it's element composition. We also wish to be able to predict the refractive index based on the values of other attributes related to the chemical content. After reading about the different attributes, it has become apparent that due to the fact that refractive index is measured on a different scale than the other attributes, the data will have to be standardized.

Despite there being very limited available literature about its past usage, it was found that the dataset was used to investigate rule induction in forensic science. The general result of this investigation was that the nearest neighbor was able to hold its own with respect to the rule-based system.

2 Explanation of attributes

The dataset contains 11 attributes, namely Refractive Index, as well as Sodium, Magnesium, Aluminium, Silicon, Potassium, Calcium, Barium and Iron weight percentages in their respective oxides. Note that the unit measurements for the elements is in weight percent of the corresponding elements oxide. Therefore, they are expected to lie between 0% and 100%.The refractive index attribute is continuous and interval, as it cannot have a value less than 1. The 'Type' attribute is discrete and nominal. The rest of the attributes describe the amount of a particular substance that the sample in question contains, and are therefore all continuous and ratio. There are no missing values or corrupted data in the dataset, therefore there are no data issues.

[1]B.German Central Research Establishment, Glass Identification Data set, `https://archive.ics.uci.edu/ml/datasets/glass+identification`

To obtain a clearer understanding of the attributes, the table below contains the summary statistics of each continuous attribute.

Attribute	**Min**	**Max**	**Median**	**Mean**	**Standard Deviation**
RI	1.511	1.534	1.518	1.5184	0.00304
Na	10.730	17.380	13.300	13.41	0.817
Mg	0.000	4.490	3.480	2.685	1.442
Al	0.290	3.500	1.360	1.445	0.499
Si	69.810	75.410	72.790	72.651	0.775
K	0.000	6.210	0.555	0.497	0.652
Ca	5.430	16.190	8.600	8.957	1.423
Ba	0.000	3.150	0.000	0.175	0.497
Fe	0.000	0.510	0.000	0.0570	0.0974

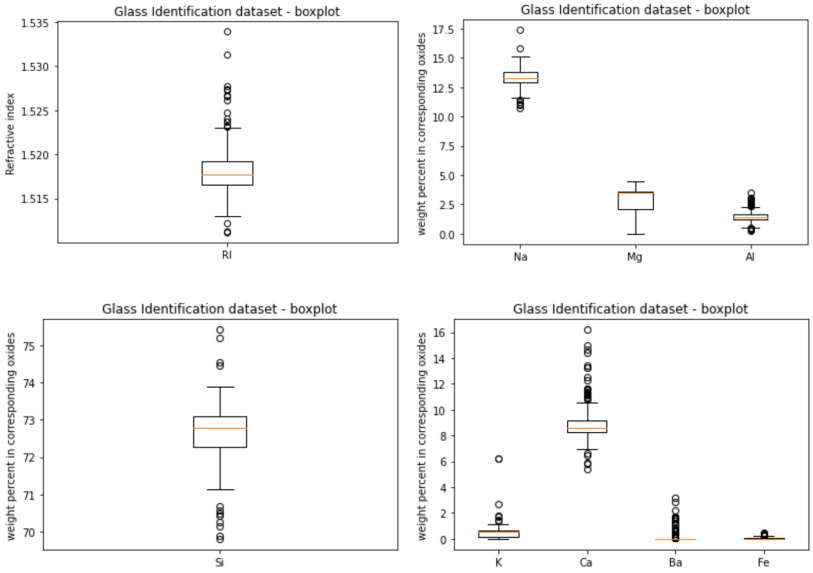
Table 1: Summary statistics for all continuous attributes of the dataset.

As shown in Table 1, the refractive index has the smallest standard deviation, meaning that the samples in the dataset have a very similar refractive index value. Additionally, calcium has the greatest range as well as the greatest standard deviation, meaning that its value in the sample varies the most out of all the attributes.

3 Data visualization & Principal Component Analysis

3.1 Investigating outliers using boxplot

Before we are able to address if there are any outliers in the dataset, we need to create a boxplot for the different attributes. The boxplots are given below:

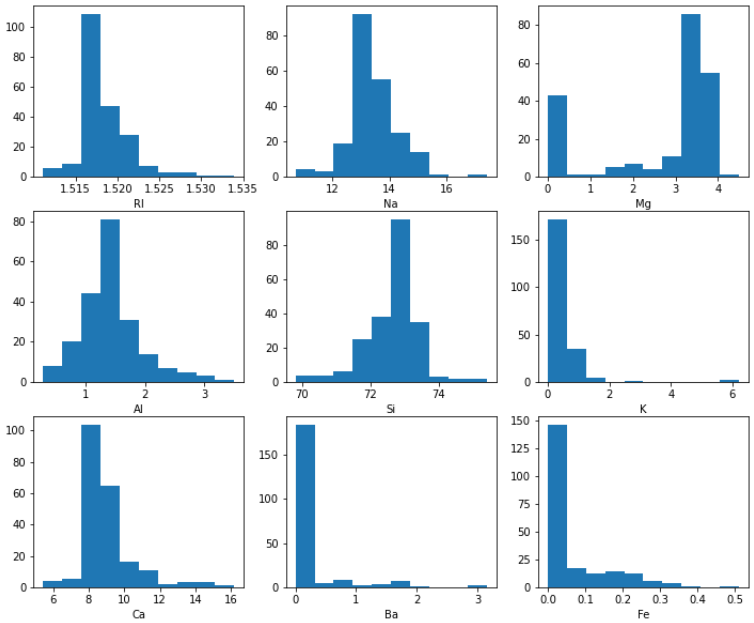


Figur 1: Boxplots for all continuous attributes of the dataset.

From the above plot, we notice that the spread of the boxplots for Al, K, Ca, Ba and Fe are very small. Whereas the spread of the boxplot for RI and Si is significantly greater. There appears to be no values greater than 100% or smaller than 0%. As indicated by the boxplots, the attributes seem to have outliers that lie beyond the upper and lower bounds. Approximately 10% of the data points are outliers, which is quite significant. However, if we were to remove them, then our dataset would be significantly reduced which could lead to inaccurate results.

3.2 Distribution of attributes

We now wish to investigate the distribution of the attributes in the dataset, particularly to see if they are normally distributed. This can be explored by creating histograms for each attribute and assessing whether or not the data is centered around the mean, as well as the shape of the plot. The respective histograms are plotted below.

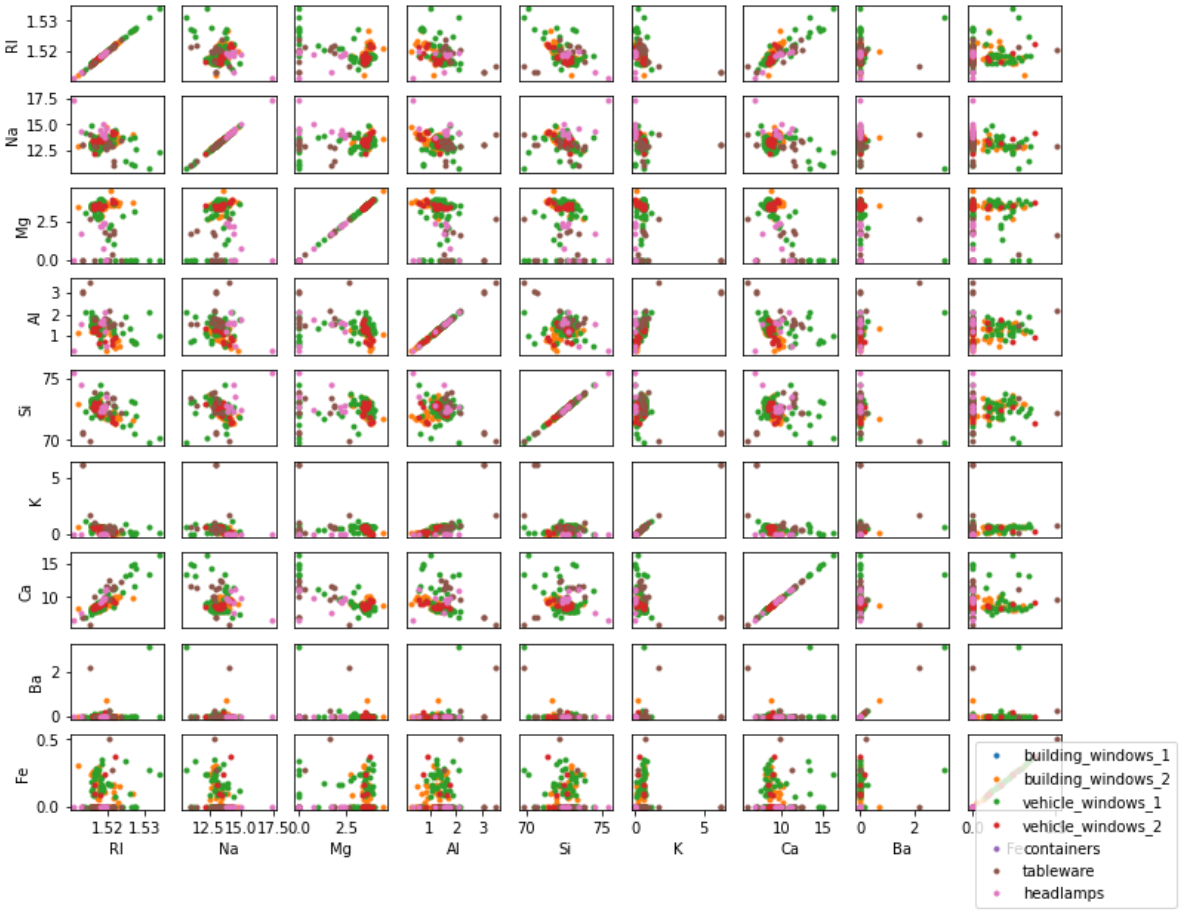


Figur 2: Histograms for all continuous attributes of the dataset.

In order to build a good model, we hope to have attributes in the dataset which resemble a normal distribution. By way of inspection, we see that the attributes Na and Si best resemble a normally distributed attribute. Moreover, the attributes Al, Ca, and RI all resemble slightly right-skewed normal distributions. On the other hand, the attributes Ba, Fe, Mg and K are not normally distributed.

3.3 Correlation

Before we are able to state anything about the correlation of attributes we need to plot a matrix of scatter plots. This is given below:



From the above plot, we can see that RI and Ca are strongly positively correlated. For the other attributes they are weakly correlated. We notice there are some outliers, for instance in the scatter plot between RI and Ba. We therefore need to look at the correlation values between attributes. Given below is a correlation matrix:

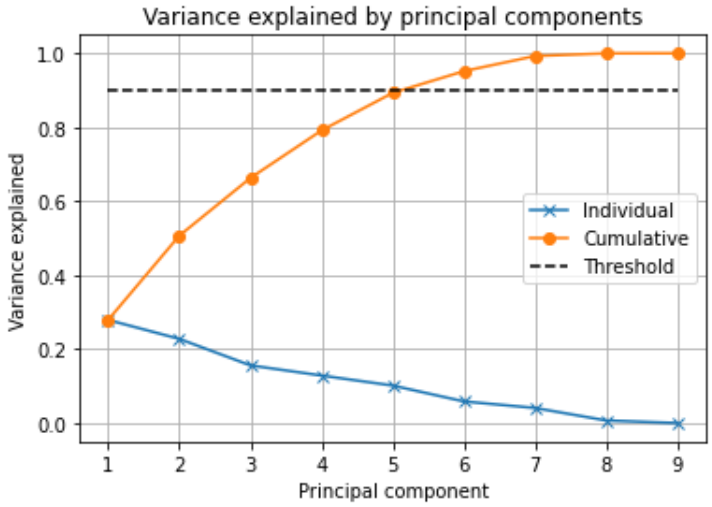
	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
RI	1	-0.19	-0.12	-0.41	-0.54	-0.29	0.81	-0.00039	0.14
Na	-0.19	1	-0.27	0.16	-0.070	-0.27	-0.28	0.33	-0.24
Mg	-0.12	-0.27	1	-0.48	-0.17	0.0054	-0.44	-0.49	0.083
Al	-0.41	0.16	-0.48	1	-0.0055	0.33	-0.26	-0.48	-0.074
Si	-0.54	-0.070	-0.17	-0.0055	1	-0.19	-0.21	-0.10	-0.094
K	-0.29	-0.27	0.0054	0.33	-0.19	1	-0.32	-0.043	-0.0077
Ca	0.81	-0.28	-0.44	-0.26	-0.21	-0.32	1	-0.11	0.12
Ba	-0.00039	0.33	-0.49	0.48	-0.10	-0.043	-0.11	1	-0.059
Fe	0.14	-0.24	0.083	-0.074	-0.094	-0.0077	0.12	-0.059	1

Table 2: Correlation matrix between all continuous attributes of the dataset.

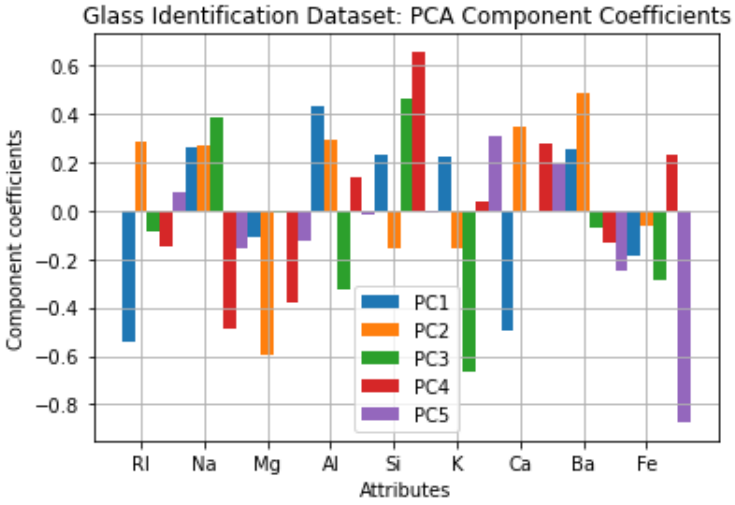
From the above matrix, we can clearly see that there is a strong positive correlation between RI and Ca. From this in combination with the rest of the matrix, it becomes clear that regression is feasible.

3.4 Principal Component Analysis

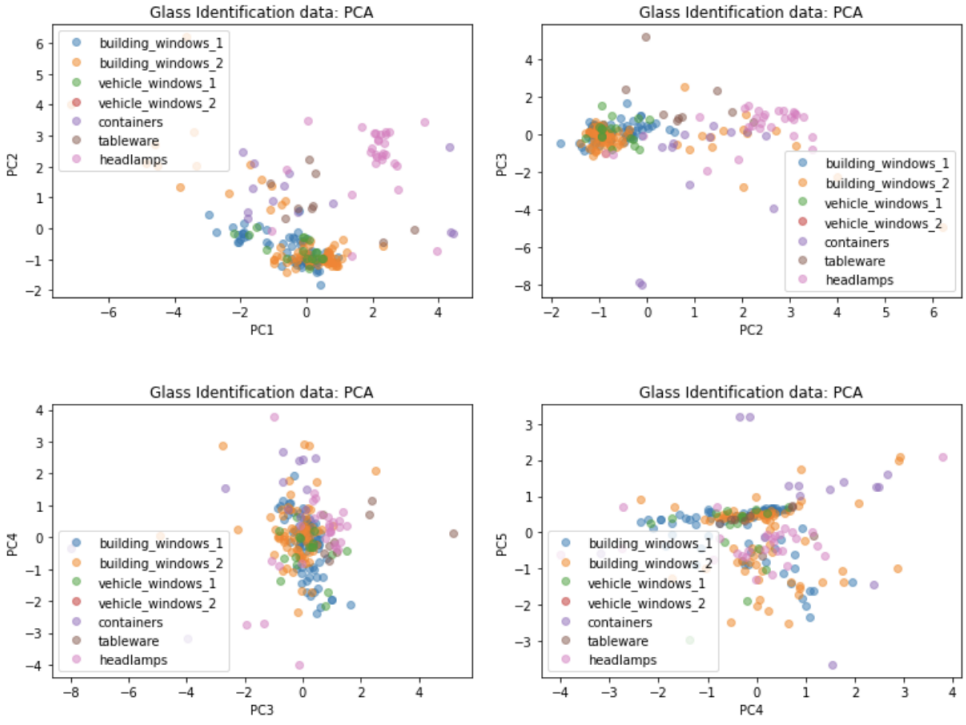
From our investigation of the dataset, we noticed that the RI was on a different scale compared to the other attributes. Hence, we had to standardize the data by subtracting the mean and dividing by the standard deviation. This would ensure that the principal components would not be bias towards attributes with greater variation. Having conducted PCA on the dataset, it was important to analyse the amount of variance that was explained with the inclusion of each principal component. A plot was made to demonstrate the number of principal components necessary to reach an explained variance of around 90%. This plot is displayed below, and shows that 5 principal components were needed.



We see from this figure that the first principal component explains the most variance in our dataset. We also see that the individual variance explained decreases as we include more principal components. To get a better understanding of how the attributes and principal components are linked, we need to look at the principal component coefficients.



From the above plot, we can for example see that PC5 has the greatest negative effect on the attribute Fe. We can also observe the attributes that contribute the most to the different principal components. For example, PC1 has a significant effect on RI, Al and Ca. Hence, in order for the projection to be negative, then the coefficient of RI and Ca must be large, and the coefficient for Al should be small. Given below is the projection of the data set onto considered principal components, namely PC1 and PC2, PC2 and PC3, PC3 and PC4, and PC4 and PC5.



It is important to note that the dataset does not include any samples from the class vehicle windows 2. There are therefore no red points in the above plots. After looking at the above principal component projection plots, we can see clusters of pink, clusters of orange, and clusters of green. There is therefore some variance and some distinct separation between the different classes. Hence, classification is feasible.

4 Discussion & Summary

In conclusion, we have uncovered that some of the attributes, namely Na and Si, appear to be normally distributed. This will be very beneficial for model building. We have also discovered that there are strong correlations between attributes. After performing principal component analysis and projecting the data set onto the considered principal components, it was observed that there was some distinct separation between classes. In addition, we saw that clusters began to form. We were also able to investigate the relationship between particular principal components and the attributes, which provided us with more insight on the behavior of the dataset. All of the above observations and discoveries ultimately signify that our primary machine learning aim appears to be feasible. We are therefore able to apply classification and regression, in the hopes of learning more about the attributes and how they are related to the different classes in our dataset.

5 Exam Problems

5.1 Question 1

Option A. We solve this by seeing that y is ordinal, as it can be ranked. We also see that x_1 is nominal. Hence the answer is A.

5.2 Question 2

Option A. We solve by using the formula for p-norm. When $n = \infty$, then $\max(\ x_{14} - x_{18}) = 26.0 - 19.0 = 7$

5.3 Question 3

Option A. We solve this problem by using the **S** matrix diagonal. The explained variance of the first four principal components is found by $\frac{13.9^2+12.47^2+11.48^2+10.03^2}{13.9^2+12.47^2+11.48^2+10.03^2+9.45^2} = 0.867$
The calculations were also made for the other three options which were all found to be incorrect.

$$\frac{(13.9^2 + 12.47^2 + 11.48^2 + 10.03^2)}{(13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2)}$$

0.8667931475

$$\frac{11.48^2 + 10.03^2 + 9.45^2}{(13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2)}$$

0.4798501562

$$\frac{(13.9^2 + 12.47^2)}{(13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2)}$$

0.5201498438

$$\frac{(13.9^2 + 12.47^2 + 11.48^2)}{(13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2)}$$

0.7167331912

5.4 Question 4

Option D. We solve this by looking at the principal components's coeffecients. For a projection to be positive, it needs a high value in the positive coeffecients, and a low value in the negative coeffecients. Hence the answer is D.

5.5 Question 5

Option A. s_1 and s_2 have 2 words in common, so $f_{11} = 2$. s_1 has 6 unique words from s_2, so $f_{10} = 6$. Similarly, s_2 has 5 unique words from s_1, so $f_{01} = 5$. Therefore $J(s_1, s_2) = \frac{2}{5+6+2} = \frac{2}{13}$. So the answer is A.

5.6 Question 6

Option C. We solve this by: $p(x_2 = 0|y = 2) = p(x_2 = 0, x_7 = 0|y = 2) + p(x_2 = 0, x_7 = 1)$ $= 0.81 + 0.1 = 0.91$

6 References

UCI Machine Learning Repository, Glass Identification Data Set, Data donated in 1987, `https://archive.ics.uci.edu/ml/datasets/glass+identification`