Title: Practical Data Science With Python – Assignment 2 – Glass Identification

Student ID: s4081442

Student Name and email (contact info): Peter Ljubisic s4081442@student.rmit.edu.au

Affiliations: RMIT University.

Date of Report: 19/05/2024

I certify that this is all my own original work. If I took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in my submission. I will show I agree to this honor code by typing "Yes": *Yes*.

Table of Contents

Abstract

The k-Nearest Neighbours and Decision Tree classification models were used and compared for the classification of glass samples based on their properties. The training and testing dataset was gathered from the UC Irvine Machine Learning Repository, and contains data on the refractive index of each sample, the proportions of various oxides formed on each sample, along with the classification type of each sample. The models were trained to classify the glass (Windows (Normally Processed), Windows (Float Processed), Vehicles (Normally Processed), Containers, Tableware, and Headlamps) by processing oxide content and refractive index data. The most optimal kNN model showed an accuracy of 0.84 with parameters of k = 1 and p = 1. The most optimal decision tree had an accuracy of 0.80 with parameters *max_depth = 7, max_leaf_nodes = 12, min_samples_split = 2, min_samples_leaf = 1, and max_features = 9*. The kNN model was shown to be superior due to its capabilities of handling continuous variables. These results show promise in the acute identification of glass, which is valuable in the discipline of forensic science to interpret the scene of a crime.

Introduction

Forensic science is a crucial area of expertise in the world. It is critical that investigators are able to research a crime scene to uncover clues, and this means being able to conduct analyses on the various objects of relevance to the crime. Glassware is an extremely common but fragile substance: meaning it is often shattered at a crime scene. In the event that pieces of multiple glassware are found scattered over the same location, or found in a different but related location, it is vital to sort which pieces belong to which object to understand the events of the crime. Depending on the function of the glassware, they would have undergone different methods of smelting with different materials and chemical impurities. Most materials are easy to analyse to extract this information (courtesy of advancements in material sciences and chemistry), but glass remains a

challenge. The property of glass transmitting light rather than absorbing it (and it's innate chemical impurity) makes spectroscopy and other techniques very challenging to employ to determine its composition and thus its associated object at the crime, but fortunately glassware contains measurable properties. The goal of this report is to outline the application and effectiveness of classification models on forensic glass data to classify samples into distinct categories of their associated glassware object.

Literature Review

Glass is not a pure substance but rather a mixture: a mix of many types of elements and molecules without a proper overarching chemical structure. Muddy water is also a mixture: there are many H_2O molecules present in it, but also lots of extraneous materials found in the mud that don't chemically bond to the water molecules. Different types of mud with different molecules can be used to make muddy water, but it is still muddy water. Diamonds in contrast are a compound: they consist of carbon atoms that join together in a very specific shape, and a pure diamond will only consist of carbon atoms joined in the specific structure shown in Figure 1. Glass is like muddy water, and certain elements commonly found in glass form chemical bonds with the oxygen in the air to form oxides (stains) on the surface. These oxides are visible and can be analysed to determine the proportions and types of chemical impurities in the glass. Different types of glass undergo different processes to be made depending on their function, creating different levels of impurities. Analysis of oxides on glass can therefore allow investigators to trace which object a shard of glass originally belonged to.

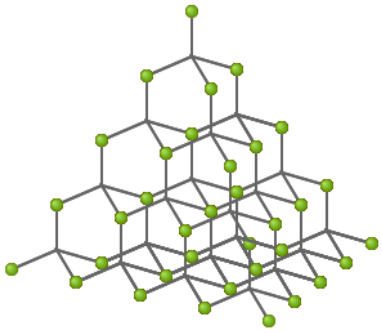

Figure 1: A diamond lattice: each carbon atom bonds with four neighbouring carbon atoms repeatedly in a specific shape. https://www.physics-in-a-nutshell.com/article/13/diamond-structure

Glass is made by smelting sand rich with silicon crystals at high temperatures of roughly 1700°C. Depending on the elements and molecules included in the sand, the properties of the resulting glass (such as the oxides that form on the surface) can change. As the heat turns the sand into glass, it can also be reshaped to an intended object, and made permanent during cooling. Additional chemicals can be added to reinforce the durability of the glass or to reduce the cost of manufacture. Soda ash (composed of sodium) is often mixed to reduce the melting temperature of the sand, while limestone (composed of calcium) adds structural integrity to the glass. This is the basic process by which glass is made, and a particular method that is often used today to create high quality glass is the float process: molten glass is poured over molten metal (generally tin) to reshape itself evenly to the desired shape. These explanations of the creation of glass are relevant to understand some of the categories that glass can be grouped into, and why certain impurities exist within it that can be used for analysis. There is another property of glass however that needs to be explained to understand the measures used to classify samples: the ability to refract light.

Glass is a material that is primarily known to allow for the transmission of light. Objects that transmit light have an interesting property where they can bend light rays depending on the angle of incidence of the light to the surface. This is shown in Figure 2: A light ray hits the material at an angle Θ_1, and bends to a new direction Θ_2 inside the material. This is a property known as refraction: light will bend in a new direction when it enters a new medium (material or vacuum) to travel in. The new trajectory of a light ray can be calculated using Snell's Law:

$$\mathbf{n_1 \sin(\theta_1) = n_2 \sin(\theta_2)}$$

The variable n called the refractive index of the medium. The refractive index is a unitless measurement that describes how much the material refracts light. A perfect vacuum (pure emptiness: no air) has a refractive index of exactly 1. Any material will therefore have a refractive index greater than 1. Glass may not be a pure substance, but it too has a refractive index that can be measured.

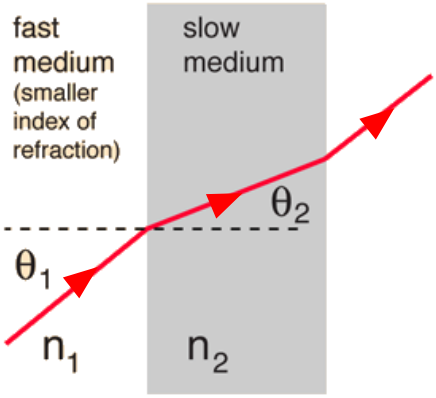

Figure 2: An incident ray of light travelling from one medium to another. The angles are measured with respect to the direction normal to the surface of the mediums. http://hyperphysics.phy-astr.gsu.edu/hbase/geoopt/refr.html

Methodology

This report is entirely centred around the Glass Identification dataset supplied by the UC Irving Machine Learning Repository. The data was collected by the USA Forensic Science Service, and was donated on the 31st of August 1987 by Vina Speihler. It features 214 samples of glass belonging to the following classifications under the *Type of Glass* variable: *Windows*, *Windows (Float Processed)*, *Vehicles*, *Vehicles (Float Processed)*, *Containers*, *Tableware*, and finally *Headlamps*. Each sample features a unique ID as well as measurements of their refractive index and percentages of weights of oxides formed on their surfaces. The refractive index and oxide percentages were treated as features to be visually analysed and used for modelling while the type of glass was treated as the target variable.

	ID	Refractive Index	Na %	Mg %	Al %	Si %	K %	Ca %	Ba %	Fe %	Type of Glass
0	1	1.52101	13.64	4.49	1.10	71.78	0.06	8.75	0.0	0.0	1
1	2	1.51761	13.89	3.60	1.36	72.73	0.48	7.83	0.0	0.0	1
2	3	1.51618	13.53	3.55	1.54	72.99	0.39	7.78	0.0	0.0	1
3	4	1.51766	13.21	3.69	1.29	72.61	0.57	8.22	0.0	0.0	1
4	5	1.51742	13.27	3.62	1.24	73.08	0.55	8.07	0.0	0.0	1

Figure 3: The first five rows of data for the glass identification dataset.

The dataset was first transformed into a csv file, and then imported into Python using the read_csv() function from the Pandas library. The variables were scanned for missing values, outliers, and logically impossible values (refractive indexes had to be above 1.00, all percentages needed to be non-negative to a maximum of 100%, and the sum of percentages of each sample needed to be 100%) as a sanity check. Then the data exploration stage was initiated by analysing the distribution of values per variable. The relationships of the features to each other and the target variable were analysed using visualisations to better understand some of the characteristics of the data, especially in relation to the type of glass sample. With sufficient understanding, the data was split into training and testing proportions by a ratio of 60:40 with stratification between glass types. Two models were trained initially under the assumption that performance was maximised with all features used simultaneously. The models were the k-Nearest Neighbours and Decision Tree algorithms supplied by the sklearn library. The models were evaluated in terms of performance and accuracy using confusion matrices and classification reports detailing the precision, recall and F1-scores of each model. Then, the parameters of each model were adjusted using loops to find an accuracy-optimised version of each. The most promising kNN model underwent hill-climbing procedures to find the best set of features that maximised the accuracy of the model, while the Decision Trees were equivalently streamlined by adjusting the *max_features* parameter. In reaching these two final models, they were compared to each other in terms of accuracy with classification reports and confusion matrices.

Results

Feature Distributions

As specified by the Methodology section, the distributions of the features were plotted to begin the Data Exploration stage of the research. This was accomplished with the use of histograms. The histogram for the refractive index of samples is shown in Figure 4 to the right. The refractive index of glass samples have a mean of 1.5184 and a standard deviation of 0.0030, but the distribution is heavily skewed to the left with a minimum of 1.5111 and a maximum of 1.5339. The distribution contains several distant outliers to the right.

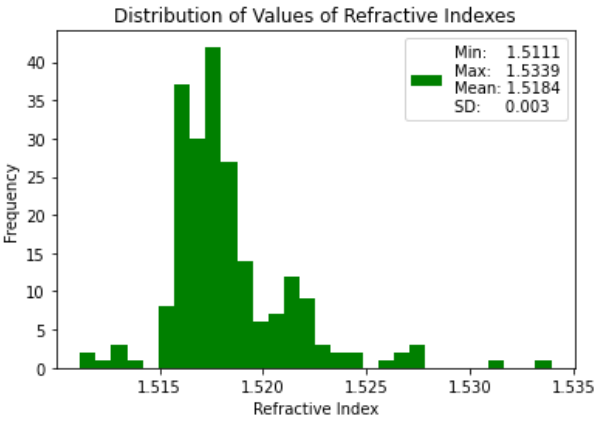

Figure 4: A histogram describing the distribution of values of the refractive index variable with summary statistics.

Figure 5 in the next page provides the distributions for the oxide features of the data. Summarising the results of each graph in sequence:

- Na_2O: Features a mean of 13.41% and standard deviation of 0.82%. It has a minimum value of 10.73%, a maximum value of 17.38% and is slightly skewed to the left.
- MgO: Features a mean of 2.68% and a standard deviation of 1.44%: the highest of any variable. It has a minimum value of 0.00% comprised of over 40 samples, and a maximum value of 4.49%. The distribution is U-shaped: narrow peaks both ends.

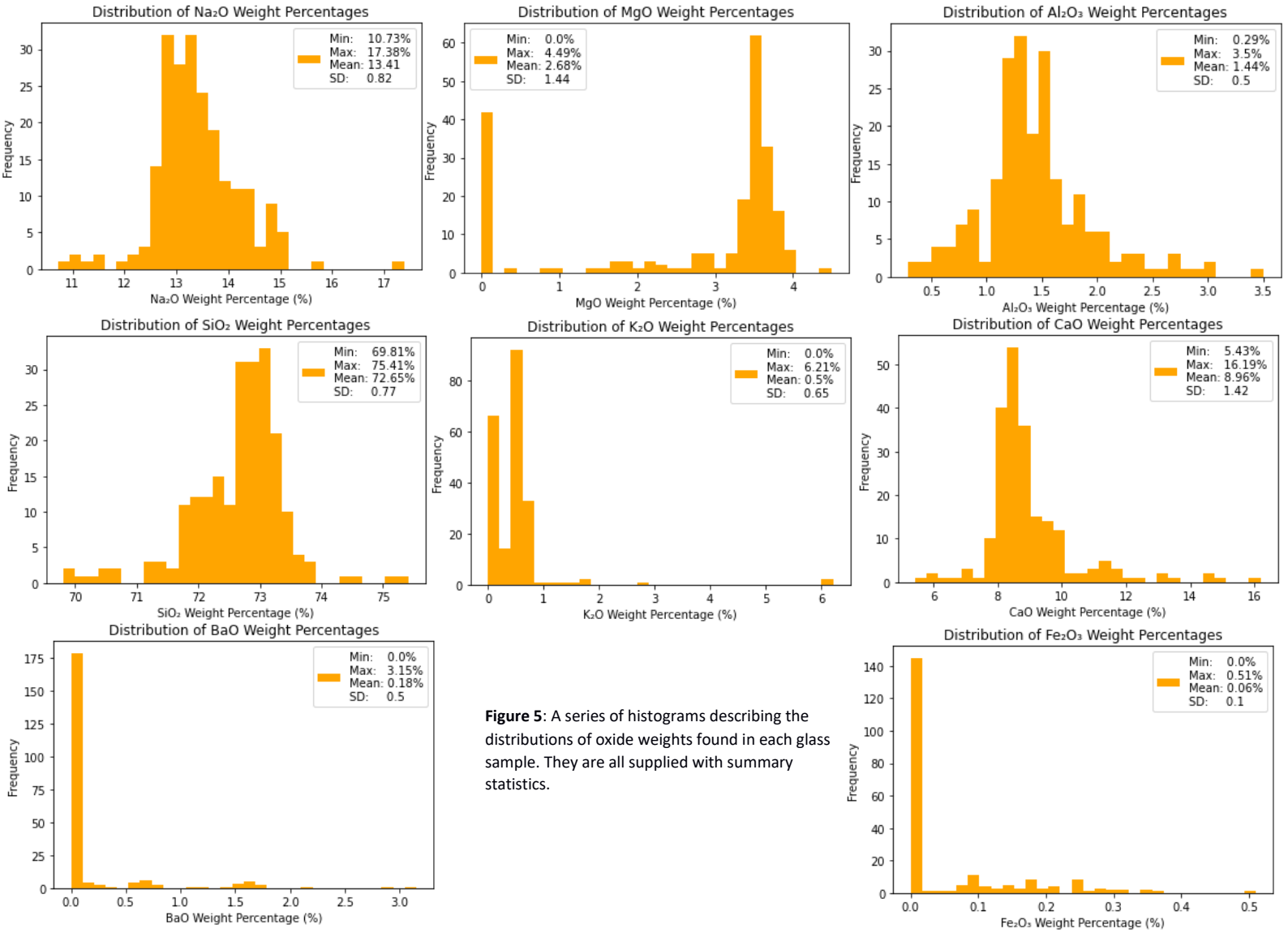

Figure 5: A series of histograms describing the distributions of oxide weights found in each glass sample. They are all supplied with summary statistics.

- Al_2O_3: Features a mean of 1.44% and a standard deviation of 0.50%. It has a minimum value of 0.29%, and a maximum value of 3.50%. It resembles a normal distribution the most out of all the histograms, with an extended tail to the right.
- SiO_2: Features a mean of 72.65% and a standard deviation of 0.77%. It has a minimum value of 69.81%, and a maximum value of 75.41%. It is a centralised distribution skewed to the right, with a raised plateau followed by an abrupt peak.
- K_2O: Features a mean of 0.50% and a standard deviation of 0.65%. It has a minimum of 0.00% comprised of over 60 samples, and a maximum of 6.21% courtesy of extremely large outliers. It has a second peak at the mean with over 80 samples.
- CaO: Features a mean of 8.96% and a standard deviation of 1.42%. It has a minimum value of 5.43%, and a maximum value of 16.19%. The distribution is centralised, but features long tails especially in the rightwards direction that create large variances.
- BaO: Features a mean of 0.18% and a standard deviation of 0.50%. It has a minimum value of 0.00% comprised of over 170 samples, and a maximum value of 3.15%. The non-zero values of this variable proportionally behave as outliers.
- Fe_2O_3: Features a mean of 0.06% and a standard deviation of 0.10%. It has a minimum of 0.00% comprised of over 140 samples, and a maximum of 0.51%. Unlike the BaO distribution, there is a sparse cluster of entries centralised around 0.20%.

There is finally the target variable of this data set: the type of glass of the sample. This is a categorical variable with seven possible values. A pie chart is shown to the right that details the proportions of each sample type. Most notably: there are no samples of vehicle glass constructed using the float process. Furthermore, window glass makes up roughly 2/3 of the entire data set, leaving the other classifications with scarce amounts of data. The worst in this regard is the tableware class, which only features 9 samples.

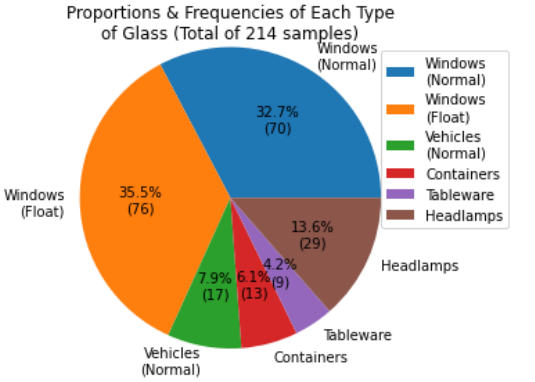

Figure 6: A pie chart showing the quantities and proportions of classifications of glass for the dataset.

Feature-Pair Exploration

After having analysed each variable individually, it is constructive to plot them against each other. Doing so would check if each variable carries distinctive behaviour to the target variable, and ascertain whether each feature is independent of one another or related in some fashion. Figure 7 in the next page contains a series of box plots. They exist to respond to the null hypothesis that each feature is independent of the target variable. If the null hypothesis were true, then the distribution of values for every variable would not change with respect to the glass type. It is therefore sufficient to show that at least two different glass types carry differing distributions for a shared variable that cannot be a result of random error in sampling to reject the null hypothesis for that variable.

- The Na_2O box plots show changes across different glass types. The windows, vehicles and containers samples do not vary much between each other, but the values for tableware and headlamps samples on the other hand are noticeably greater, such that it is reasonable to assume that sodium oxide levels in glass do vary with respect to the type of glass. The null hypothesis is rejected for the sodium oxide variable.
- The MgO box plots are more obvious in rejecting the null hypothesis. Windows and vehicle glassware contain much greater traces of MgO than glass for containers, tableware and headlamps.
- The Al_2O_3 box plots again vary significantly across different classifications. Al_2O_3 levels are notably weaker in normally processed windows compared to container and headlamp glassware: rejecting the null hypothesis.
- SiO_2 is the least likely variable to reject the null hypothesis as the distributions look somewhat similar across glass types. The saving grace of the variable is the vehicle and tableware distributions: the former contains generally lower SiO_2 levels than possible in the latter, and the latter contains higher values than possible for the former distribution. The null hypothesis is therefore rejected.
- K_2O levels are generally highest among the containers and (with outliers) headlamps samples, while tableware glass is notably absent of the oxide. The null hypothesis is therefore rejected.
- CaO levels contain a significant number of outliers in float-processed window glass compared to the other distributions. They stretch to higher values than shown in the other glass types: rejecting the null hypothesis.
- BaO presence is extremely likely in Headlamp glassware (even if in trace amounts), while other types almost always have zero BaO levels: rejecting the null hypothesis.
- Fe_2O_3 has almost no traces found in container, tableware and headlamp glassware compared to window and vehicle glass which generally do contain the oxide: rejecting the null hypothesis.

The refractive index variable rejects the null hypothesis on the same grounds as the CaO variable. In fact, the distributions of these two variables look remarkably similar, even to the presence of outliers across different glass types. This insight sparks the creation of a new null hypothesis to be addressed: the refractive index of a glass sample is independent of the oxide contents of the glass. The null hypothesis can be reasonably rejected if a correlation can be identified between the two features using a scatter plot (such as in Figure 8).

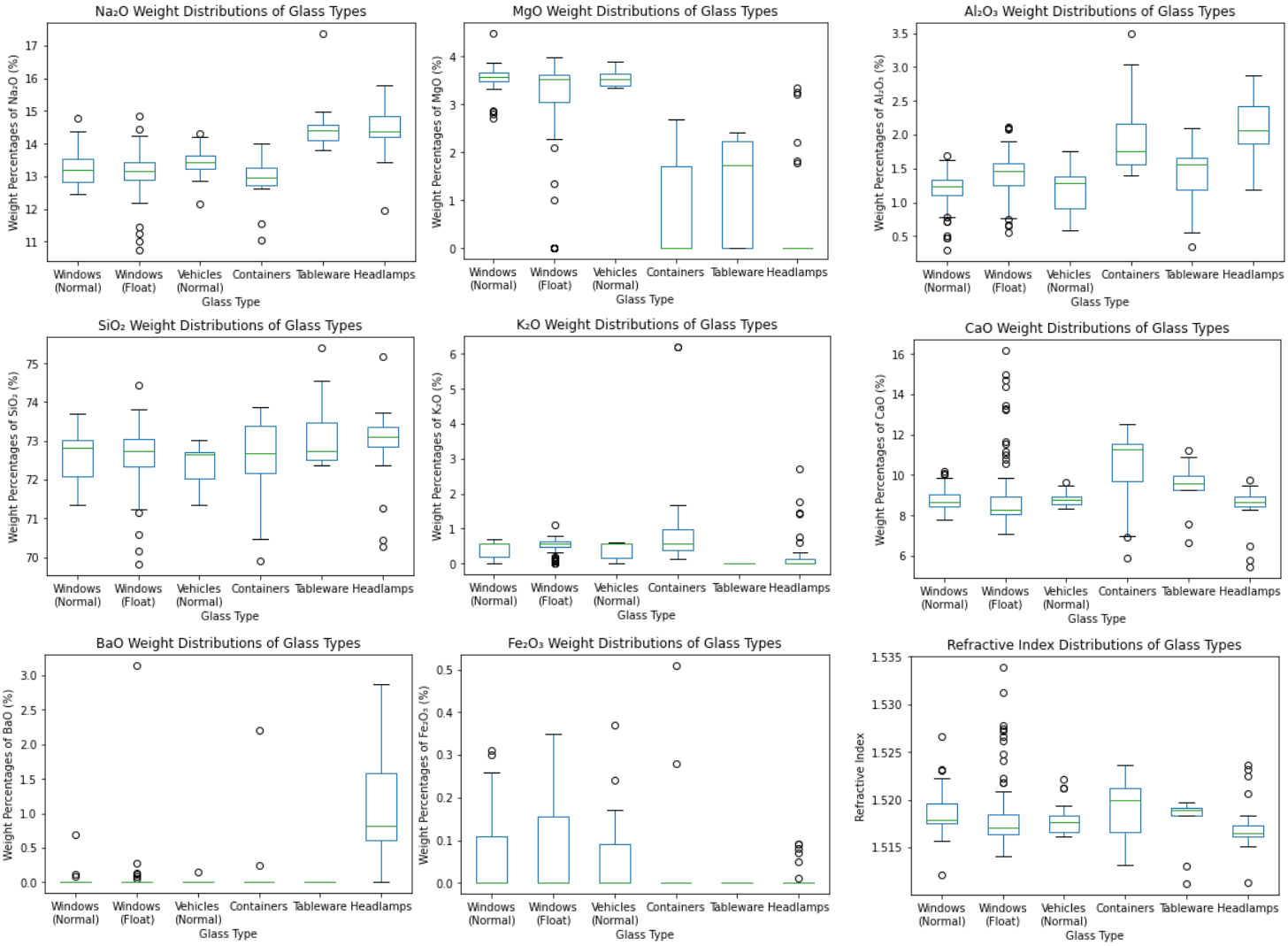

Figure 7: A series of box plots showing the distribution of data points for each feature of the model based on the type of glass. These were created to better understand how each feature can be used to classify the sample, and if any potential features show signs of being uncorrelated to the target variable.

- The Na_2O vs Refractive Index data points are centralised into a triangular shape. The coordinates outside this triangle may hint of decreases in refractive index as Na_2O percentages rise, but the awkward distribution of points cannot conclusively reject the null hypothesis.
- The scattered nature of the data points for the MgO graph suggests that the refractive index of glass is independent of the presence of MgO in the glass: accepting the null hypothesis.
- The Al_2O_3 plot shows a loose downwards trend of the refractive index with respect to increases in Al_2O_3 percentages, but the noise and the lack of slope for the dense cluster create uncertainty over whether the correlation exists. The null hypothesis cannot be rejected between these two variables.
- The SiO_2 plot seems to contain significant noise as well but the central cluster also supports the notion of a decrease in refractive index as SiO_2 percentages increase: rejecting the null hypothesis.
- The K_2O plot is clustered around low values of K_2O such that a trend cannot be observed as there is not enough data for higher percentages: failing to reject the null hypothesis.
- The CaO plot clearly shows that a linear increase in CaO creates a linear increase in the refractive index: rejecting the null hypothesis.
- The BaO plot fails to reject the null hypothesis as the refractive index does not seem to vary with respect to BaO percentages.
- Fe_2O_3 similarly fails to reject the hypothesis, but the presence of outliers is much higher above the gradient than below it.

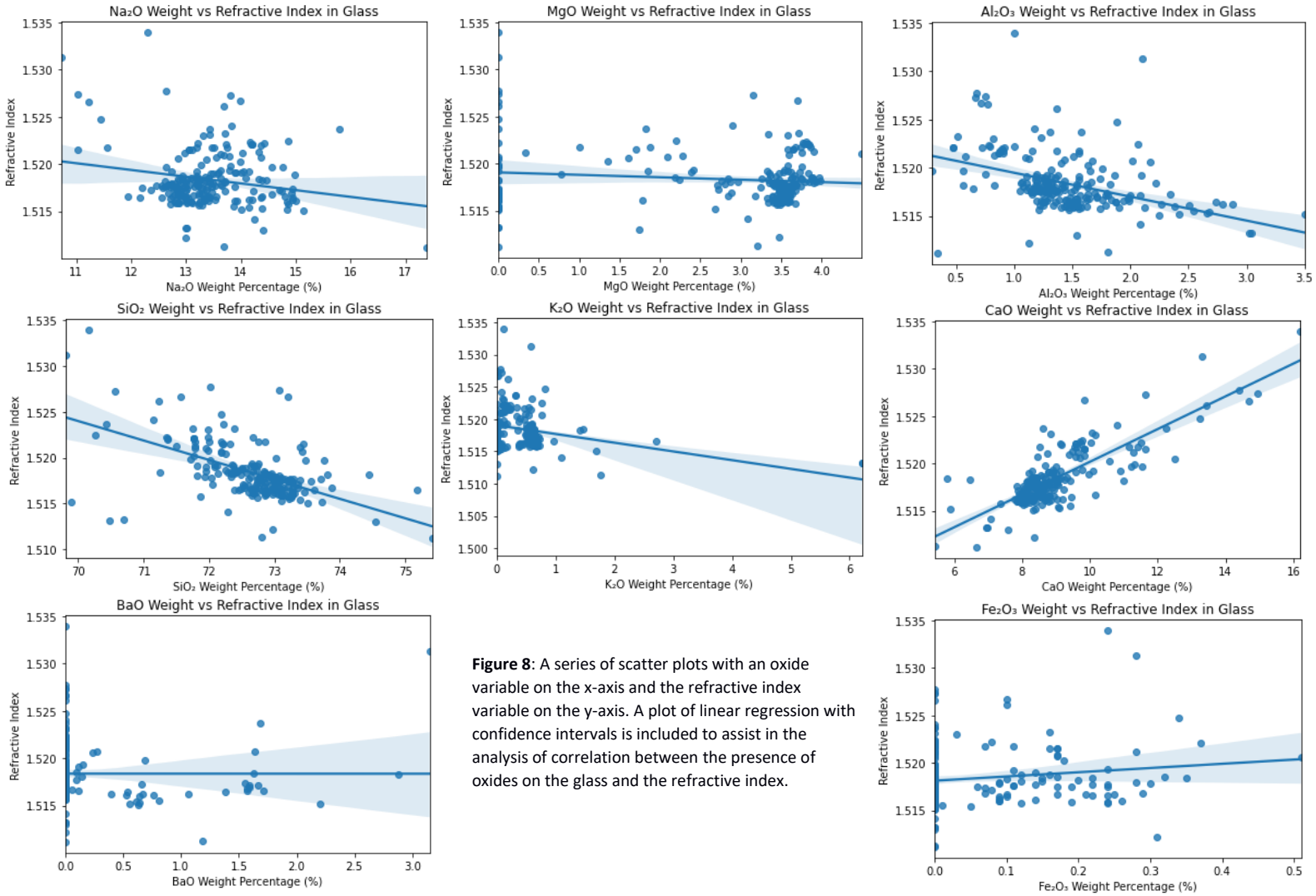

Figure 8: A series of scatter plots with an oxide variable on the x-axis and the refractive index variable on the y-axis. A plot of linear regression with confidence intervals is included to assist in the analysis of correlation between the presence of oxides on the glass and the refractive index.

Based on these visualisations, SiO_2 seems to be loosely correlated to the refractive index, while CaO displays a strong correlation to the refractive index. The refractive index of a glass sample is shown to be independent of the presence of other oxides. A final null hypothesis can be generated: the oxide percentages are independent of each other. The null hypothesis can be rejected if a trend can be observed between them. Figure 9 in the next page shows a scatter matrix of all oxide features of the data against each other. There are 28 unique combinations, so only those that are considered to reject the null hypothesis will be listed:

- The plots indicate that a linear increase in BaO (x-axis) leads to a logarithmic increase in Na_2O (y-axis) plateauing at levels of 15%. The null hypothesis is therefore rejected.
- Increases in Na_2O of glass seem to be correlated to a linear decrease in CaO and vice-versa. There is significant noise in the data, but there is enough evidence to reject the null hypothesis.
- Linear increases in MgO seem to offer a linear decrease in CaO and vice-versa, rejecting the null hypothesis between these variables.
- Linear increases in CaO (x-axis) seem to create exponential decreases to K_2O (y-axis) plateauing to 0%: rejecting the null hypothesis between these variables.
- The presence of BaO often means there is no Fe_2O_3 in the glass sample and vice-versa. This may simply be because both oxides are very uncommon in glass, so the chance of both appearing in the same sample is low. This does not necessarily reject the null hypothesis, but it is an interesting observation.

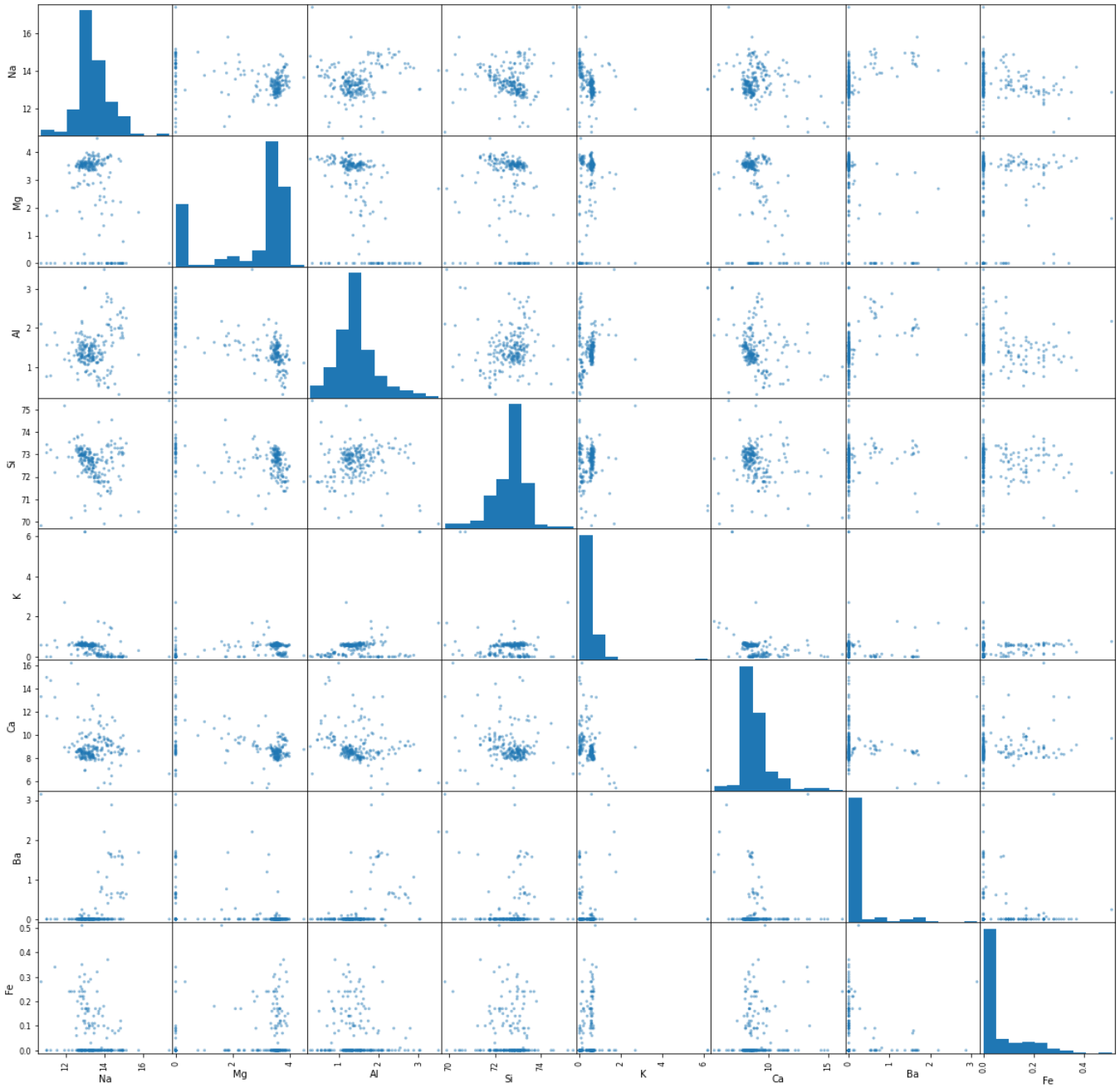

Figure 9: A scatter matrix with each oxide on the x and y-axes. The main diagonal consists of a linear density graph of the distribution of values for each oxide.

k-Nearest Neighbours Model Performance

After all these visualisations, the distributions and relations between the variables of the data were better understood. The data was then used to train a kNN classifier with a stratified 60:40 train/test split using initially all features. The kNN parameters *k*, *p*, and *weights* were adjusted to different values to test how the performance of the model would vary: the results of which can be found in Figure 10 in the next page. The best performing model had parameters *k = 1* and *p = 1* with an accuracy score of 0.81. Models with *p = 1, weights = distance* and *k = 1 & 2* are equivalent in function to this model so they also had the same accuracy scores. The best performing models for higher levels of k were those with distance-based weighting, with *p = 1* variants outperforming p = 2 variants by at least 5%. In general however, the *p = 1* models were supreme to the *p = 2* variants, and universally the *k = 1* versions were also the most effective across both p values. Application of distance-based weighting greatly increased model performance for higher values of k, but failed to exceed the results of models with *k = 1*.

Accuracy scores of kNN model based on different parameters

		Values of k parameter									
P	Weight	1	2	3	4	5	6	7	8	9	10
2	None	0.78	0.74	0.70	0.70	0.70	0.66	0.63	0.63	0.64	0.63
2	Dist	0.78	0.78	0.72	0.70	0.71	0.71	0.70	0.70	0.67	0.67
1	None	0.81	0.71	0.74	0.69	0.72	0.64	0.65	0.66	0.66	0.67
1	Dist	0.81	0.81	0.74	0.73	0.73	0.71	0.70	0.73	0.74	0.73

Figure 10: A table of accuracy scores for kNN models using the stratified data based on adjustments to their parameter values.

Confusion Matrix and Performance Scores of kNN model for k = 1, p = 1

		Predicted classifications								
		Windows (Normal)	Windows (Float)	Vehicles (Normal)	Containers	Tableware	Headlamps	Support	Recall	F1 Score
Actual classifications	Windows (Normal)	28	0	0	0	0	0	28	1.00	0.84
	Windows (Float)	6	21	2	0	1	0	30	0.70	0.81
	Vehicles (Normal)	4	0	3	0	0	0	7	0.43	0.50
	Containers	0	0	0	5	0	0	5	1.00	1.00
	Tableware	0	1	0	0	3	0	4	0.75	0.67
	Headlamps	1	0	0	0	1	10	12	0.83	0.91
	Precision	0.72	0.95	0.60	1.00	0.60	1.00			

Macro Precision	0.81	Macro Recall	0.79	Accuracy	0.81
Weighted Precision	0.84	Weighted Recall	0.81		

Figure 11: A combination of a confusion matrix and the classification report information of the kNN model with parameters k = 1, p = 1.

Figure 11 shows the confusion matrix and performance scores collected for the most successful model out of the ones tested so far. It has the following characteristics:

- Model attained an overall precision, recall and accuracy of 0.84, 0.81 and 0.81 respectively.
- Windows (Normal) samples were always correctly classified, but vehicle glass and float-processed window glass would also occasionally be classified as such.
- Samples that were classified as float-processed window glass were almost certain to be correctly classified, but actual samples would get classified into normal window glass and vehicle glassware.
- Vehicle glassware was heavily mistaken for window glassware.
- Containers glassware would always be perfectly predicted, but only 5 samples were tested.
- Tableware glass would generally be correctly classified, but would also be confused with float-processed window glass.
- Samples that were classified as headlamp glass were certain to be correctly classified, and most actual samples were placed into the correct category.

Confusion Matrix and Performance Scores of feature-optimised kNN model for k = 1, p = 1

		Predicted classifications								
		Windows (Normal)	Windows (Float)	Vehicles (Normal)	Containers	Tableware	Headlamps	Support	Recall	F1 Score
Actual classifications	Windows (Normal)	26	0	2	0	0	0	28	0.93	0.85
	Windows (Float)	3	25	1	0	1	0	30	0.83	0.88
	Vehicles (Normal)	3	0	4	0	0	0	7	0.57	0.57
	Containers	0	1	0	4	0	0	5	0.80	0.89
	Tableware	0	1	0	0	3	0	4	0.75	0.67
	Headlamps	1	0	0	0	1	10	12	0.83	0.91
	Precision	0.79	0.93	0.57	1.00	0.60	1.00			

Macro Precision	0.81	Macro Recall	0.79	Accuracy	0.84
Weighted Precision	0.85	Weighted Recall	0.84		

Figure 12: A combination of a confusion matrix and the classification report information of the kNN model with parameters k = 1, p = 1, and optimised feature selection.

Hill-climbing algorithms were applied to optimise this model by finding the key set of features to include and exclude. Model accuracy with the same parameters was improved to 0.84 by removing the Sodium and Barium Oxide features. The improved results are shown in Figure 12. The notable changes are a 7% improvement in precision for normal window glass at the expense of precision in float processed window glass and vehicle glass by 2% and 3% respectively. The recall of float-processed window glass and vehicle glass increased by 13% and 14% respectively at the expense of a 7% and 20% decrease for normal window glass and container glass respectively. Model precision and recall was improved by 1% and 3% respectively.

Decision Tree Model Performance

Decision trees feature a larger number of parameters to adjust compared to kNN models, thus it was important to have a baseline measurement to compare further models against. Using all default parameters and access to all features, Figure 13 shows the decision tree, and Figure 14 shows the test metrics. Other models needed an accuracy above 0.74 to replace this variant.

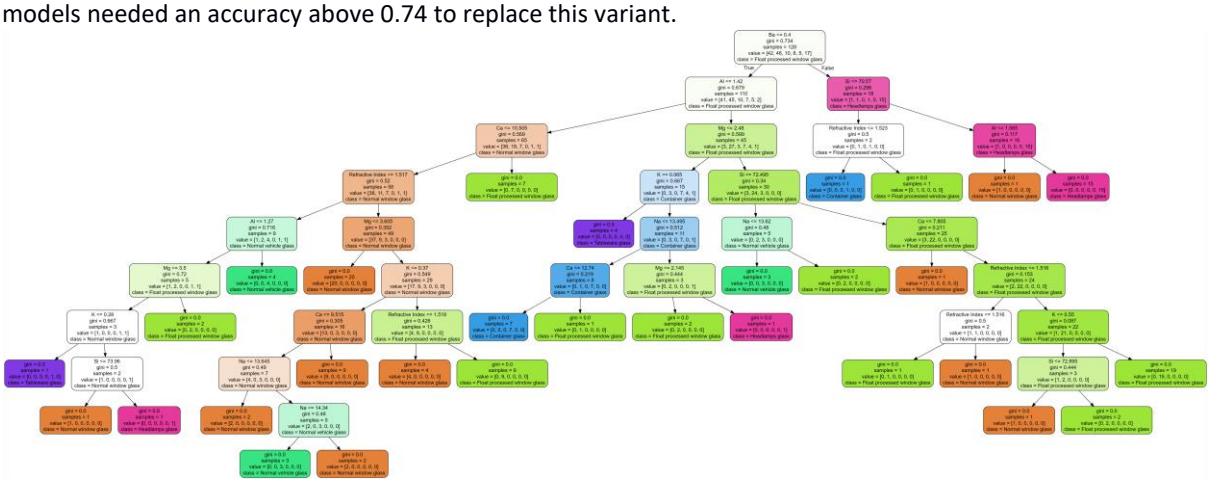

Figure 13: A decision tree for glass identification with all parameters set to default values while using all possible features for the model.

Confusion Matrix and Performance Scores of Decision Tree with default parameters

		Predicted classifications								
		Windows (Normal)	Windows (Float)	Vehicles (Normal)	Containers	Tableware	Headlamps	Support	Recall	F1 Score
Actual classifications	Windows (Normal)	24	2	2	0	0	0	28	0.86	0.76
	Windows (Float)	4	22	3	0	0	1	30	0.73	0.77
	Vehicles (Normal)	5	0	2	0	0	0	7	0.29	0.29
	Containers	0	1	0	4	0	0	5	0.80	0.89
	Tableware	1	1	0	0	2	0	4	0.50	0.67
	Headlamps	1	1	0	0	0	10	12	0.83	0.87
	Precision	0.69	0.81	0.29	1.00	1.00	0.91			
	Macro Precision	0.78	Macro Recall	0.67	Accuracy	0.74				
	Weighted Precision	0.76	Weighted Recall	0.74						

Figure 14: A combination of a confusion matrix and the classification report information on the decision tree with default parameters.

The final optimised decision tree is shown in Figure 16 in the next page, and its accuracy metrics are also shown in Figure 15. The optimised decision tree was constructed with the following parameters using nested loops: *max_depth = 7, max_leaf_nodes = 12, min_samples_split = 2, min_samples_leaf = 1, and max_features = 9*. In developing the optimised decision tree, the most noteworthy statistical developments are:

- Precision, recall and accuracy all increase to 0.80 (previously it was 0.76, 0.74 and 0.74 respectively).
- Normal window glass classification sits 4% below the default version in recall, but carries 8% more precision. Commonly, samples are mistaken for float-processed window and vehicle glassware.
- Float-processed window samples gain 20% in recall but lose 5% in precision. Like for normal window glass, samples are often mixed up with normal window and vehicle glass.
- Vehicle glass gained 21% more precision but lost 15% recall. Most samples are mistaken for window glass.

- Container glassware retains 100% precision but loses 20% in recall.
- Tableware samples retain a recall of 0.50 and a precision of 100%.
- Headlamp samples increase in precision by 9% to perfection, recall increases by 9%.
- The decision tree was optimised if all features were available for splitting nodes.

Confusion Matrix and Performance Scores of Decision Tree with optimised parameters

Actual classifications	Predicted classifications: Windows (Normal)	Windows (Float)	Vehicles (Normal)	Containers	Tableware	Headlamps	Support	Recall	F1 Score
Windows (Normal)	23	4	1	0	0	0	28	0.82	0.79
Windows (Float)	2	28	0	0	0	0	30	0.93	0.84
Vehicles (Normal)	3	3	1	0	0	0	7	0.14	0.22
Containers	0	1	0	4	0	0	5	0.80	0.89
Tableware	1	1	0	0	2	0	4	0.50	0.67
Headlamps	1	0	0	0	0	11	12	0.92	0.96
Precision	0.77	0.76	0.50	1.00	1.00	1.00			
Macro Precision	0.84	Macro Recall	0.69	Accuracy	0.80				
Weighted Precision	0.80	Weighted Recall	0.80						

Figure 15: A combination of a confusion matrix and the classification report information on the decision tree with optimised parameters.

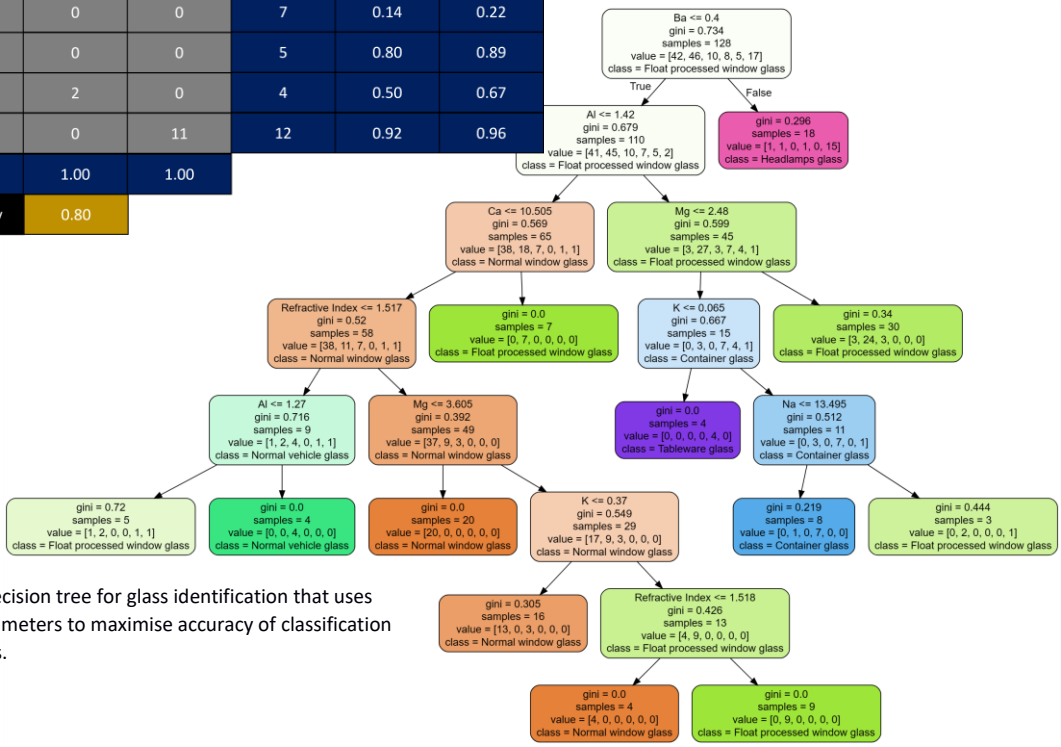

Figure 16: A decision tree for glass identification that uses optimised parameters to maximise accuracy of classification of test samples.

Discussion

There are a significant number of results to address. Regarding the distributions of features, the oxides of the samples wildly vary in presence. SiO_2 for example has a mean of 72.65% (far more than any other oxide) while BaO has a mean of 0.18% with many samples containing no BaO at all. Some distributions are unexpected, such as MgO's U-shaped distribution with two peaks. The reasons for these discrepancies can be explained by the molecular properties of glass and the creation process that was established in the literature review. Sand is a chemical mixture, but contains large amounts of silicon crystals within. Even after the smelting process, the silicon remains and is one of the most frequent elements found in glass, hence the especially large presence of silicon oxide. Limestone and soda ash are frequently used as agents to give glass better commercial value, and they contain large amounts of Sodium and Calcium respectively. This is why they are the second and third most prevalent oxides formed respectively, as they are copiously used for all types of glass. Other oxides exist from: impurities in the sand, the presence of agents to reinforce the function of the glass, or the smelting process leaving traces of metals in the glass. These reasons explain why they are significantly less common.

Regarding feature relations, this reasoning also extends to why the distributions of oxides differ for different glass types. Different processes or brands create different types of glassware for different purposes, hence some metals like Barium may be found in some glassware but not others. This may also be why some oxides show some relationship to others: they may be involved in the same smelting process of glass but fulfill different material purposes. The influence of oxides on the refractive index of glass is a natural outcome in material science: different materials change in how well they refract light and so if the chemical composition of

the glass changes, so can the refractive index. Calcium seems to be the strongest agent that affects the refractive index of glass, while silicon seems to also show influence in this.999

Regarding the performance of the models, it was expected that the kNN classifier worked best for low values of k. For one, the distributions of the features were remarkably similar between (normal and float) window glass and vehicle glass. This meant that introducing even a moderately high value for k would likely correspond to introducing neighbours of differing classes: creating confusion in results. There was also the issue that there was scarcely any container and tableware samples to use in comparison, so higher values of k could easily introduce other classifications simply because statistically there was not enough data for those categories. Meanwhile, the parameter p was also expected to optimise performance if set to 1. This is because the kNN model operated on higher-dimensional vectors, and as a rule of thumb p = 1 is the optimal choice for those.

The performance of the decision tree classifier however was underwhelming, but expected. While decision trees are useful and easy to visualise, they are known to struggle with continuous variables as they only split nodes by categorising a certain range or value into one thing or another. They are also prone to overfitting, which was minimised by cycling through a number of parameters until the optimal set were found. In fact, the poorer accuracy of the initial tree can be attributed to the large number of nodes and branches present, and thus the overfitting of the data. The similarity in data for the windows and vehicle classifications mentioned earlier meant that a significant amount of error was generated in attempting to distinguish those samples using continuous features, and it was especially apparent for vehicle glass with a horrendous recall of 14%.

In comparing the two competing models, the kNN classifier reigned supreme. It had an accuracy of 0.84 to the decision tree's 0.80, a precision of 0.85 to 0.80, and a recall of 0.84 to 0.80. This can be attributed to the fact that kNN is designed to work with continuous variables, while decision trees handle categorical variables. Other notable strengths included higher precision for window and vehicle classes as well as better recall for normal window, vehicle, and tableware classes. The decision tree had a much higher recall for float-processed window glass, but the significantly lower precision meant that samples were just more likely to be classified into that class compared to the normal window and vehicle classes. It also notably outperformed the kNN classifier in recall for headlamp glass, and in precision for tableware glass. But overall, it is clear that the kNN classifier for parameters $k = 1$ & $p = 1$ was the best option to model the data.

Conclusion

In conclusion, glass samples were shown to be effectively distinguished between each other using data models. The k-Nearest Neighbours classifier with parameters $k = 1$ & $p = 1$, and exclusion of the sodium and barium oxide features was the most effective for this purpose, with an accuracy of 0.84 that surpassed all decision trees. This model shows promise for wider use in the identification of glass samples in forensic investigations, but there is still room for improvement. In reconstructing the events of the crime and identifying evidence, an accuracy of 0.84 is still too low. Fortunately, this can be improved by supplying more data points for the model in all categories, especially for vehicle, container, and tableware samples which had hardly any. The model can then be expanded to consider other oxides and properties of glass, and to explore further glassware classifications such as float-processed vehicle glass.

References

German, B. (1987). *Glass Identification*. UCI Machine Learning Repository. https://doi.org/10.24432/C5WW2P.

Corning. (n.d.). *How Glass is Made*. https://www.corning.com/worldwide/en/innovation/materials-science/glass/how-glass-made.html

Levy, D. (2016). *What Is The Float Glass Process?* Glass.com. https://info.glass.com/what-is-the-float-glass-process/

Physics in a Nutshell. (n.d.). *Diamond Structure*. https://www.physics-in-a-nutshell.com/article/13/diamond-structure

Hyperphysics. (n.d.). *Refraction*. http://hyperphysics.phy-astr.gsu.edu/hbase/geoopt/refr.html