Title: Glass classification model for criminological investigation

Name and email (contact info): Minh Khoi Nguyen

Date of Report: 26/5/2024

Table of Contents

Abstract

Knowledge extracted from criminal evidence databases is crucial for conducting effective criminological investigations. Implementing data modelling techniques helps to reveal clear and logical patterns extracted from those databases, providing essential knowledge for correctly identifying evidence at the scene of the crime. This criminological investigation determines the study of the classification of glass type, where the classification model's predictive power was considered comparable with other rule-based systems identifying the types of glass. (German, 1987)

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are widely used by the machine learning community as a primary source of machine learning datasets (UCI Machine Learning Repository, 2024)

Introduction

The Glass Identification dataset from the UCI Machine Learning Repository is sourced from the USA Forensic Science Service and contains information on 6 types of glass. (German, 1987)

The attributes of glass types include: “1: building_windows_float_processed”, “2: building_windows_non_float_processed”, “3: vehicle_windows_float_processed”, “5: containers”, “6: tableware”, and “7: headlamps”. (German, 1987). It is important to note that the glass attribute type “4: vehicle_windows_non_float_processed” does not have any evidence to be recorded on this dataset, so it will be excluded from the whole process.

Each attribute is categorised based on its “RI: refractive index” and oxide element weight percentages of “Na: Sodium”, “Mg: Magnesium”, “Al: Aluminum”, ”Si: Silicon”, “K: Potassium”, “Ca: Calcium”, “Ba: Barium” and “Fe: Iron”. (German, 1987)

This report aims to compare the results of machine learning models trained by different classification techniques. Hence, identify the technique that gives the best result on the same dataset, which would potentially used in the context of criminal investigations.

The process will implement both k-nearest Neighbours (kNN) and Decision Tree (DT) classification techniques on the same dataset and then compare the performance of models generated by the two. The evaluation metric mainly considers the accuracy score. Feature selection analysis and basic hyperparameter tuning will also be conducted to optimise the model's performance.

The rest of the report will be structured as follows: (Methodology) describes how the dataset is retrieved, explored, and modelled. (Results) shows the relevant results regarding the performance of each model. (Discussion) compares the performance of the two models and discusses the impact of feature importance analysis and hyperparameter tuning on model performance.

Methodology

About the data package

The archived package contains four data files, one of which is “glass.data” contains the main dataset that we are going to use. The other contains descriptive information about the dataset.

Data Retrieving

As the raw data of “glass.data” is stored in a comma-separated format, I simply changed its extension to “glass.csv.” This would facilitate easier data extraction using the Python pandas library.

Using the pandas.read_csv() method, I retrieved a data frame, as shown in Fig. 1, with 10 features and 214 instances named according to the dataset description. See Introduction for more information.

Figure 1: Glass Identification data frame.

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Type_of_glass
1	1.52101	13.64	4.49	1.10	71.78	0.06	8.75	0.00	0.0	1
2	1.51761	13.89	3.60	1.36	72.73	0.48	7.83	0.00	0.0	1
3	1.51618	13.53	3.55	1.54	72.99	0.39	7.78	0.00	0.0	1
4	1.51766	13.21	3.69	1.29	72.61	0.57	8.22	0.00	0.0	1
5	1.51742	13.27	3.62	1.24	73.08	0.55	8.07	0.00	0.0	1
...	...	...	...	...	...	...	...	...	...	...
210	1.51623	14.14	0.00	2.88	72.61	0.08	9.18	1.06	0.0	7
211	1.51685	14.92	0.00	1.99	73.06	0.00	8.40	1.59	0.0	7
212	1.52065	14.36	0.00	2.02	73.42	0.00	8.44	1.64	0.0	7
213	1.51651	14.38	0.00	1.94	73.61	0.00	8.48	1.57	0.0	7
214	1.51711	14.23	0.00	2.08	73.36	0.00	8.62	1.67	0.0	7

214 rows × 10 columns

Data Preparation

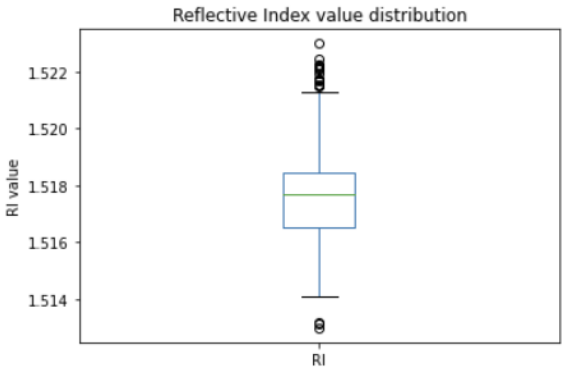
The glass data set is relatively clean as it has no missing or unexpected datatype values. However, there is an inspected duplicate pattern, so I removed it, trimming the data records to 213 instances in total. As the field of chemical substances and experiments varies, there is no justifiable reason for identifying such a record as an outlier without performing further research. I will use a statistical approach to identify those outliers in each feature and replace them with the average value of their feature. This approach, as described by (Saul Mcleod, 2023), uses the box plot components to identify the distribution range of values and then treat any values that are not numerically lying under this range as outliers.

Data Exploration

Feature exploration

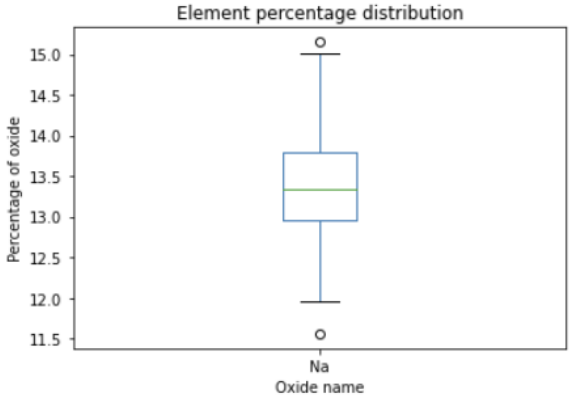
Apart from features 1 which is the index value and specified as index for the data frame (Fig.1), the other features are RI (Fig.2), Na (Fig.3), Mg (Fig.4), Al (Fig.5), Si (Fig.6), K (Fig.7), Ca (Fig.8), Ba (Fig.9), Fe(Fig.10) and Type_of_glass(Fig.11). As the data is numerical, I will mainly plot a box graph for each feature to see how their values are distributed and possibly identify patterns in the value sets. For some features, there might be additional plots to discover the patterns of distribution. It is also important to note that the data used in each plot has already been processed in the Data Preparation step.

Figure 2



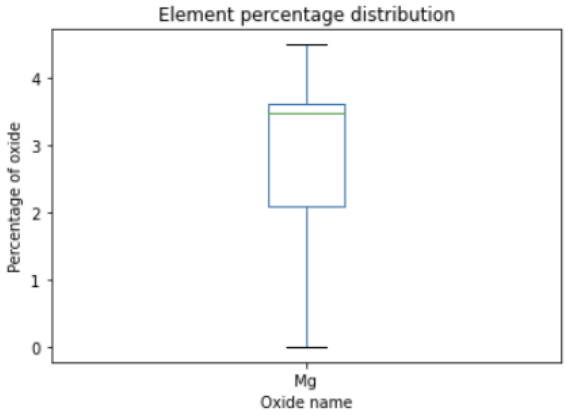
The majority of the data points are clustered in the range of 1.516 to 1.519. The data distribution is relatively symmetrical, with a few outlier values remaining, but not extreme, which might demonstrate some samples with special RI values.

Figure 3



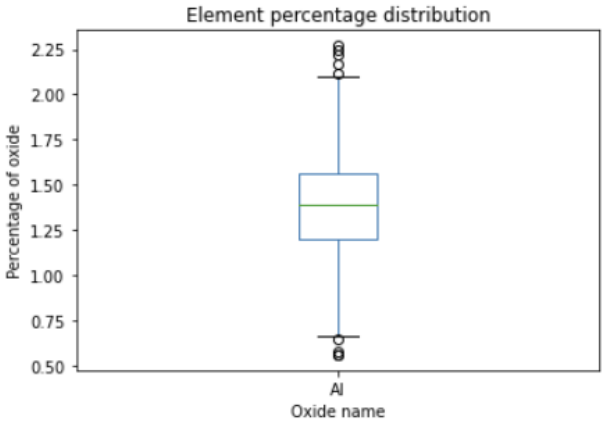
Na oxide distribution is fairly symmetrical and wide, with majority fall around the range of 13-14%, which make it one of the main component types of oxide found in given samples.

Figure 4



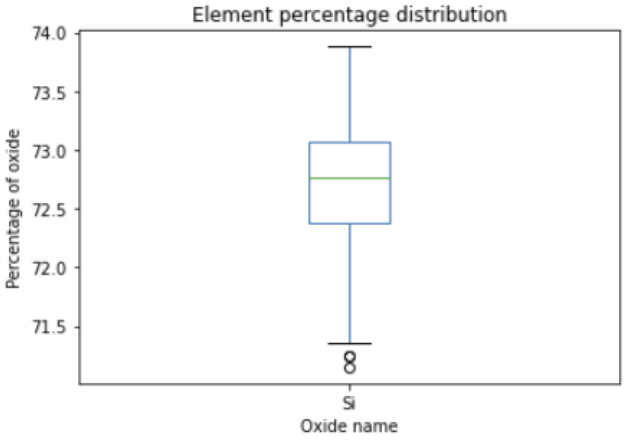
The percentage of Mg oxide contributed mostly around 2-3.5% to sample components, whereas some special ones could have a lower portion of this oxide type.

Figure 5



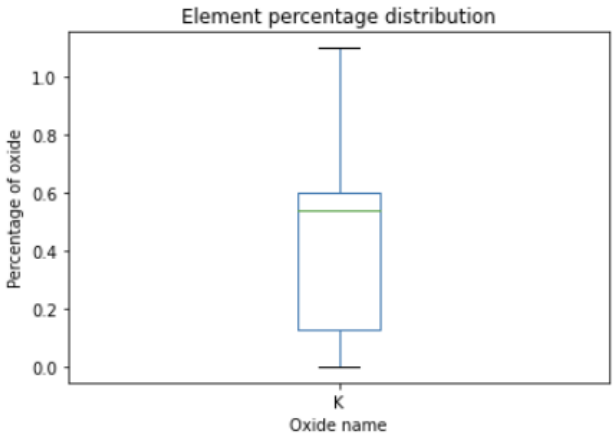
Al oxide percentage consistently contributed to all data sample, majority of records ranging around 1.25-1.50% which is not huge in variance. Although there are still some sample with special amount of Al oxide.

Figure 6



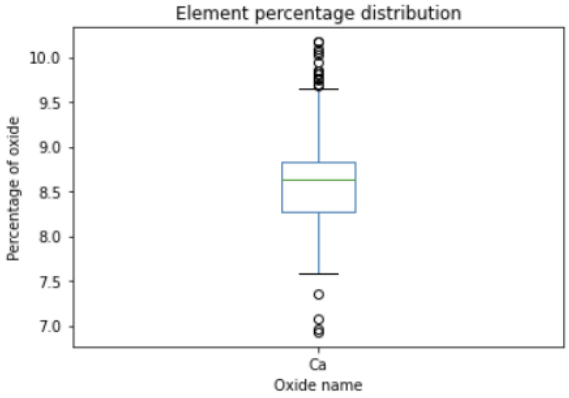
A huge portion of Si oxide is recorded. Populated around 72 and 73%, the plot also indicates a symmetrical distribution amongst all data points.

Figure 7



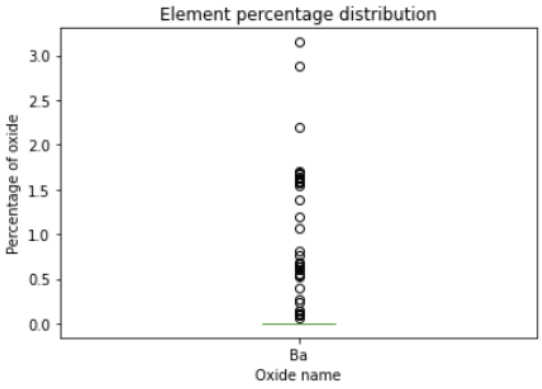
K oxide mainly contributed less than 1% of all samples. However, most of the records are skewed towards the lower amount, which is under 0.6%.

Figure 8



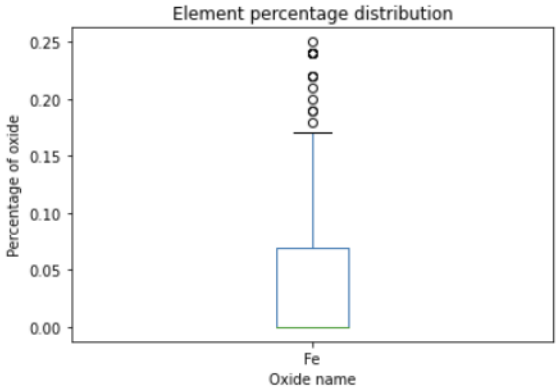
Ca oxide consistently contributes around 8-9% across all samples. It is also considered one of the most popular oxide components in these records. Also, there are noticeably high records of Ca oxide indicating special samples.

Figure 9



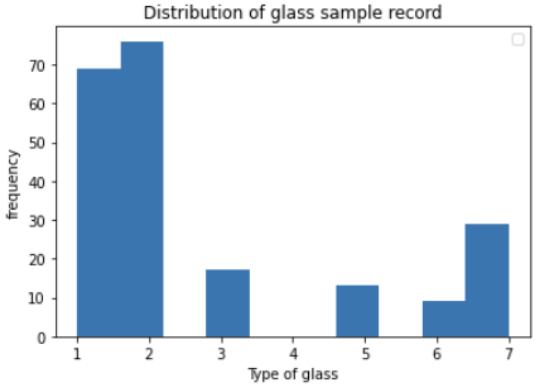
The distribution of Ba oxide in the recorded samples is separated with no noticeable pattern; however, most of the records fall under 2% in percentage, which indicates a low component of Ba oxide in general.

Figure 10



Fe oxide is also considered the least component to be found in those sample, with just less than 0.1%. however, record also indicated some special record with noticeably high portion of Fe oxide.

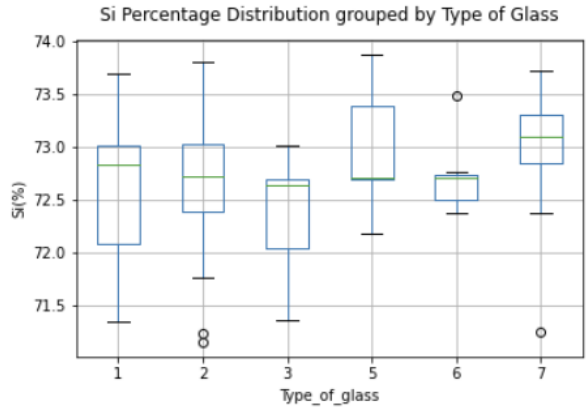
Figure 11



The majority of glass samples in this data set belong to either type 1 or 2, which are building window float and non-float glass, respectively. It also shows that the sample 4 record is absent from the set with 0 samples. Additionally, samples 7, 3, 5, and 6 are recorded with relatively fewer samples than the other two, with type6-tableware having the fewest samples.

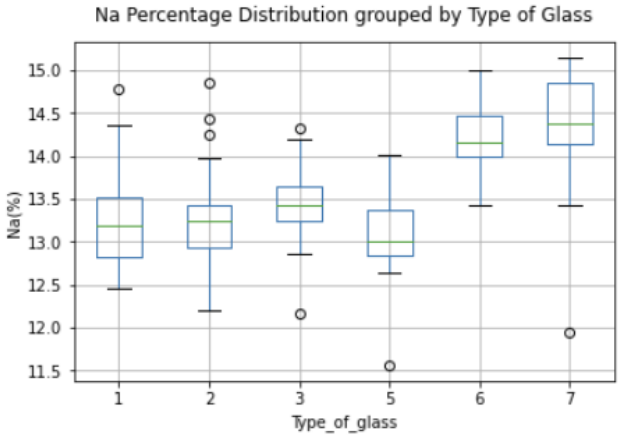
Hypothesis drawn from Feature pairs.

Regarding the percentage of components, Si oxide is considered the most popular item in each type of glass.



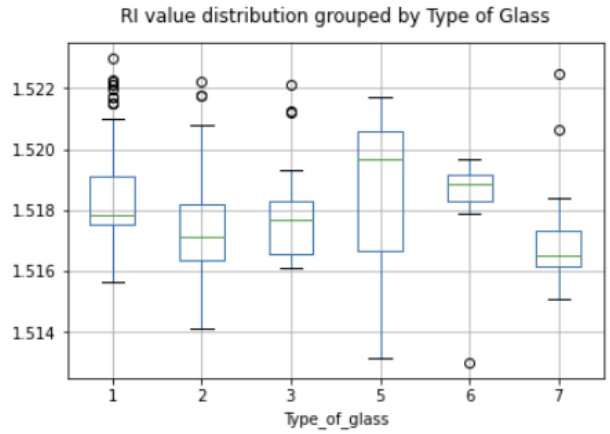
This is obvious as records shows a consistently proportion of Si oxide (72-73%) is found in each type of glass, indicating it high popularity.

Due to its noticeable percentage, all types of glasses should contain a consolidated amount of Na oxide.



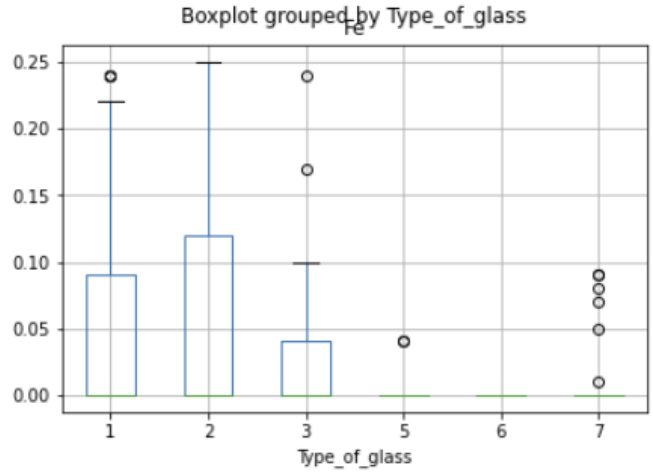
This is accepted since each type of glass contains more than 12% of Na, especially types 6 and 7.

RI could be used to differentiate the type of glass.



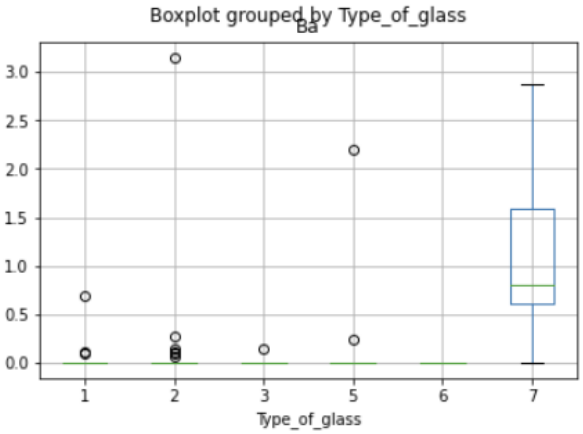
This hypothesis is not valid. Considering the RI, as shown in the plot, although there are some variations across the types, there are also noticeable overlaps indicating the similarity of value distribution, clarifying that RI could not be used as a definite distinction among glass types.

Fe oxide components are rarely used in all glass types due to their generally low proportions.



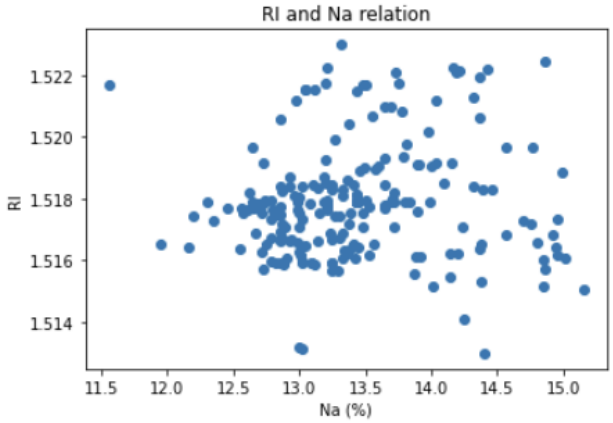
Rejected, the plot of Fe distributions across glass types shows that although Fe oxide is not used in types 5 and 6, it is found in high proportions in types 1, 2, and 3.

Ba component is rarely being used in all types of glass.



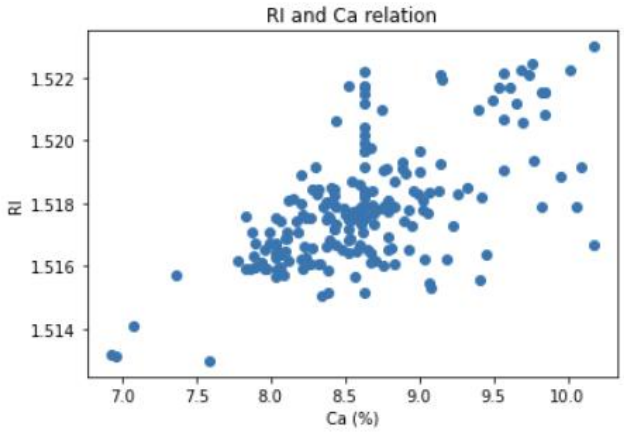
This could be considered true. Most of the glass types have very little to no Ba oxide. Especially across the popular samples like type 1 and 2(see fig11). However, a solid amount of this oxide is found in type 7; this might indicate the special formula of this glass type, not the majority.

Higher sodium content may affect the refractive index of the glass.



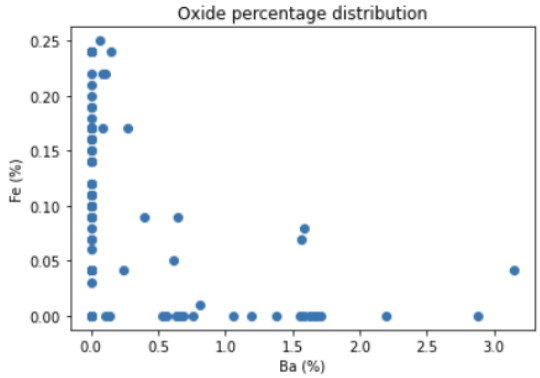
This is not correct as there is no clear pattern in this plot indicate the relationship between the twos.

Higher Calcium content may affect the refractive index of the glass.



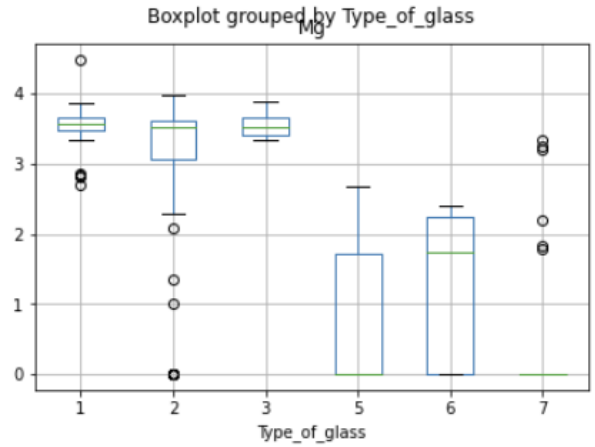
There is a noticeable positive correlation between the two variables. As the percentage of calcium (Ca) increases, the refractive index (RI) also tends to increase. Hence this hypothesis is accepted.

Both having low distribution in general, Fe oxide and Ba oxide will be correlated in distribution pattern.



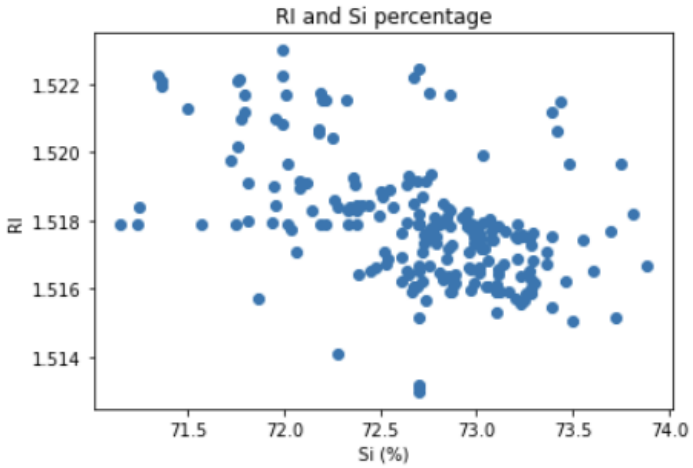
This is clearly not provable; in fact, as shown in the plot, the higher the percentage of Ba oxide, the less (if not 0) the percentage of Fe oxide.

Mg oxide proportion is distinguished amongst different types of glass.



The distribution of MgO proportions for each glass type shown in the boxplot clearly differs, considering the distribution range and outliers. Hence the hypothesis is observed.

Being used at a high portion in all glass types, Si should not affect the RI value at all.



This is correct, as expected. No linear relationship is found in this plot.

Data modelling

The validation technique is important to model training as it will evaluate the predictive ability of a model against the ground truth that it has not seen. For this purpose, I will split the dataset into two unequal parts, with the majority of data goes into training (80%-20%).

The two common classification techniques, k-Nearest Neighbour and Decision Tree will both be implemented in this step, with hyperparameter tuning to improve the models' accuracy.

k-Nearest Neighbour considers the k nearest instances from an instance and assumes that the most frequent class is the class of that instance. As there are still many outliers in the dataset, considering the metric weight by distance would be a more robust choice since it will ignore the distant relatives (which are potential outliers) as less relevant. I would keep the default Minkowski distance with hyperparameter p=2 as it is giving optimal results. Below is the accuracy score using different k values with metrics as described.

Basic tuning kNeighborClassifier			
n_neighbors	weights	Accuracy Score	p
1	distance	0.7907	2
3	distance	0.7907	2
5	distance	0.7441	2
7	distance	0.7674	2
9	distance	0.7674	2

With a small neighbour value, the model tends to be overfitted. Since it does not reduce the accuracy score, I will pick 3 as the better k-value choice.

Decision trees are another powerful technique used in classification tasks. However, their performance highly relies on those hyperparameters. For the purpose of this report, to find a comparable accuracy score to the rival k-Nearest Neighbour technique, I will hyper-tune only some of those parameters, while the remaining will be kept as default; as there is no need to change, or tuning them reduces the accuracy score of the model (not suitable to work with the data set). The tuned parameter is described as follows:

min_samples_split=6 defines the minimal number of samples that are needed to split a node. This is set to 6 to ensure a node must have at least 6 samples before splitting.

min_samples_leaf=3: the required minimal number of samples to become a leaf node. This is set to 3, which ensures each leaf node in the decision tree must contain at least 4 samples and prevents further splitting.

min_weight_fraction_leaf=0.1: tuned to address class imbalance issue, ensure each leaf node in the decision tree must hold at least 10% of the sum of sample weights.

Results

Comparing the performance of the classification models made using both techniques, we could state that out of the two classification algorithms, the Decision Tree could generate a better model than its viral k-nearest Neighbour in terms of accuracy score measurement. Which appropriate hyperparameter tuning, they respectively give 0.8372 and 0.7907 for their accuracy scores.

Discussion

The hill-climbing technique reflects on the impact of features that help improve the classifier's performance.

```
Score with 1 selected features: 0.46511627906976744
Score with 2 selected features: 0.5116279069767442
Score with 3 selected features: 0.5813953488372093
Score with 4 selected features: 0.7209302325581395
Score with 5 selected features: 0.7209302325581395
Score with 6 selected features: 0.7209302325581395
Score with 7 selected features: 0.7441860465116279
Score with 8 selected features: 0.7906976744186046
Score with 9 selected features: 0.7906976744186046
```

The result shows that the fourth and eighth features substantially improve accuracy, whereas features 5, 6, and 9 have no impact on the model's predictive power.

Conclusion

Although classification models' predictive power is comparable to rule-based technologies in criminal investigations, it is important to carefully choose the right classification techniques, algorithms, and hyperparameters. These choices can significantly impact the predictive performance of the models, especially their accuracy, which can vary based on the dataset they are built upon.

References

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

Saul Mcleod, P. (2023). Box Plot Explained: Interpretation, Examples, & Comparison. *simplypsychology*. Retrieved 5 22, 2024, from https://www.simplypsychology.org/boxplots.html#

UCI Machine Learning Repository. (2024). *About.* Retrieved from https://archive.ics.uci.edu/about