Title: Classification for Glass Indetification dataset

Student ID: s4040536

Student Name and email (contact info): Sowmiya Kumar and sowmikumar20@gamil.com

Affiliations: RMIT University.

Date of Report: 29/05/2024

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:

Data modeling and presentation simplify complex data, extract insights, and support decision-making, helping to find hidden patterns, identify trends, highlight correlations, and provide a full and informative view of the data. They enable understanding, communication, and collaboration among stakeholders, facilitating strategic planning and problem-solving. By visualizing trends and patterns, they empower organizations to drive performance and make informed decisions. This study focuses on utilizing a glass dataset for classification purposes, leveraging machine learning algorithms to perform effective data modeling. The study begins with meticulous data preparation, ensuring cleanliness and effectiveness for analysis. Exploratory data analysis is conducted to reveal relationships and correlations among features, providing insights into the underlying structure of the data. Standardization is applied to normalize the data, followed by modeling using K-Nearest Neighbors (KNN) and Decision Tree classifiers. Evaluation techniques such as train-test split and k-fold cross-validation are employed to robustly assess model performance. Parameter tuning and feature selection are then performed to optimize classification algorithms and enhance predictive accuracy. This comprehensive approach underscores the significance of effective data modeling and its application in glass classification, particularly in scenarios like criminal investigations where precise identification is crucial.

Introduction:

The motivation for this study arises from the field of criminal investigation. Glass fragments found at crime scenes can serve as crucial evidence if accurately identified and classified. These glass particles are often extremely small, making it essential to employ precise methods to identify and compare them. Proper identification of these glass fragments can significantly aid forensic experts in drawing connections between the crime scene and potential suspects. Each type of glass is composed of different elements, including Sodium (Na), Calcium (Ca), Magnesium (Mg), Barium (Ba), Silicon (Si), Aluminium (Al), Iron (Fe), and Potassium (K), along with varying Refractive Index (RI) values. These properties are crucial in distinguishing between different types of glass, such as building windows float glass, building windows non-float glass, vehicle windows float glass, containers, tableware, and headlamps. Float glass, for instance, is produced by floating molten glass on a bed of liquid metal, resulting in high-quality glass with no finishing required.

The goal of this research is to develop a machine learning model to accurately classify the type of glass based on its chemical composition and refractive index. The study aims to explore the features of the glass classification dataset using relevant data visualization techniques, build classification models using K-Nearest Neighbors (KNN) and Decision Tree algorithms with appropriate parameters and features, and compare the performance of these two classification algorithms to recommend the most suitable model.

K-Nearest Neighbors (KNN) is a simple, yet powerful, classification algorithm that assigns a class to a sample based on the majority class of its nearest neighbors. It is highly intuitive and effective for small datasets but can become computationally expensive with large datasets. Decision Tree classifiers, on the other hand, split the data into subsets based on the value of input features, creating a tree-like model of decisions. They are easy to interpret and can handle both numerical and categorical data, but they can also be prone to overfitting.

Methodology:

Retrieving and Preparing the Data:

Data retrieving and Data Preparation are part of the Data Curation process. Data retrieving involves collecting the dataset from various sources and ensuring the dataset is in the right format for further processing. Data Preparation focuses on cleaning and transforming the data suitable for analysis.

The Glass Identification dataset is taken from the UCI Machine Learning Repository. Once the data is retrieved it is necessary to inspect the dataset for better understanding. The given data is structured and stored in text format, with each value separated by a comma. The text format dataset is manually converted to CSV format using MS spreadsheet to facilitate easier data manipulation and integration with various data analysis techniques. The dataset consists of 214 instances and 10 numerical attributes (including Id_number, which is an incremental value) and a class attribute "Type_of_glass"(target variable) which indicates the type of glass (numbers from 1-7 representing 7 types of glass). The columns/attributes along with their datatype and information are listed Table 1 and Table 2.

Attributes	Datatype	Information
Id_number	int	Continuous number representing an instance
RI: Refractive Index	float	measure of how much light bends when passing from one medium into another. For glass, RI ~ 1.5
Na: Sodium	float	Element used in glass manufacturing. Measured using the weight percent in the corresponding oxide.
Mg: Magnesium	float	Element used in glass manufacturing. Measured using the weight percent in the corresponding oxide.
Al: Aluminium	float	Element used in glass manufacturing. Measured using the weight percent in the corresponding oxide.
Si: Silicon	float	Element used in glass manufacturing. Measured using the weight percent in the corresponding oxide.
K: Potassium	float	Element used in glass manufacturing. Measured using the weight percent in the corresponding oxide.
Ca: Calcium	float	Element used in glass manufacturing. Measured using the weight percent in the corresponding oxide.
Ba: Barium	float	Element used in glass manufacturing. Measured using the weight percent in the corresponding oxide.
Fe: Iron	float	Element used in glass manufacturing. Measured using the weight percent in the corresponding oxide.
Type_of_glass (class attribute)	int	Class label from 1-7 representing each glass type

Table 1: Column/Attribute information

Upon inspecting the dataset using '*info()*' function which provides the column name with its corresponding datatype and non-null count, it can be inferred that the dataset doesn't have any missing values. According to the dataset description and dataset inspection, it reveals that the dataset doesn't have records for one glass type which is 4 - vehicle_windows_non_float_processed. Since the dataset itself doesn't have one glass type it is not considered for this study.

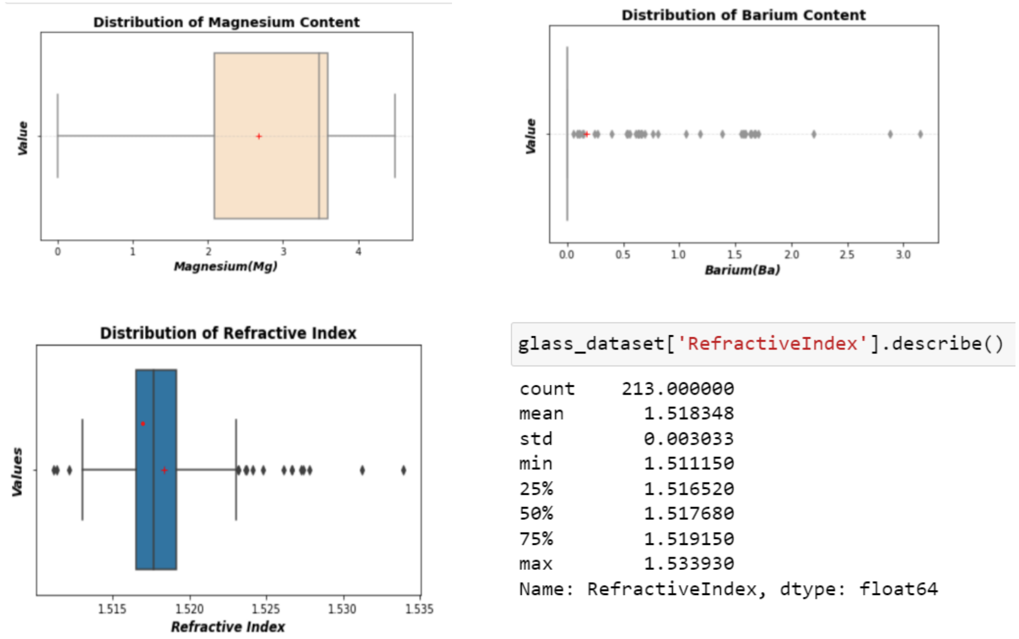
Type_of_glass	Class label
building_windows_float_processed	1
building_windows_non_float_processed	2
vehicle_windows_float_processed	3
vehicle_windows_non_float_processed	4
Containers	5
Tableware	6
headlamps	7

Table 2: Class/Target attribute information

Data Preparation is one of the crucial steps in the Data Science process. It includes cleansing, transforming, and combining raw data. Accurate and complete data is essential for algorithms to learn effectively and produce reliable results. Data Preparation mainly focuses on removing errors in data so that data becomes a true and consistent representation of the processes it originates from. The errors could be mistakes during data entry, missing values, impossible values, redundant white spaces, and outliers.

For the chosen glass dataset, the errors and corresponding actions made are given below.

Error check	Used function/method	Results and action
Missing values	isna()	No missing values.
Duplicate values	isduplicated()	1 duplicate row was found and it was dropped using the drop_duplicates() function.
Data entry error	value_counts()	No data entry error
Redundant white spaces	strip()	No redundant white spaces
Outliers	boxplot()	No outliers



```
glass_dataset['RefractiveIndex'].describe()
```

```
count    213.000000
mean       1.518348
std        0.003033
min        1.511150
25%        1.516520
50%        1.517680
75%        1.519150
max        1.533930
Name: RefractiveIndex, dtype: float64
```

Figure 1 Sample graphs for outlier detection and statistical information.

Key observations of outlier detection: In Figure 1, '+' Indicates the mean value of each distribution. The horizontal line in the middle of each box represents the median of the particular element content. The box itself contains the middle 50% of the data (interquartile range). The whiskers extend to the lowest and highest data points that are not considered outliers. Most of the columns don't have many outliers. However, the Element Barium has all of its values falling around 0, which shows that this element's weight percent in corresponding oxide is very less. Since this dataset represents the chemical composition of different glass types, with each element contributing to the glass manufacturing process, and it consists of only 9 features for predicting the final value, these minor outliers are not addressed. Added to that, the composition of these 9 features varies according to the type of glass and the manufacturer/guidelines. So collaborating with Subject Matter Experteise(SME) and Business owners is essential to conclude. According to official sources [4], the refractive index of glass ranges from 1.50-1.54. From the graph and statistical information, the minimum and maximum values of the Refractive Index are 1.511150 and 1.533930 Hence this is in the correct range and doesn't have any data entry errors/outliers.

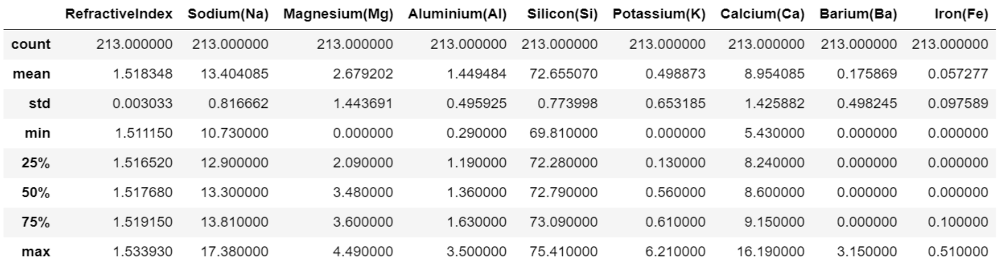
Data preparation also involves transforming data suitable for further analysis. The column Id_number is dropped as it is just an identifier and does not represent any useful information. The target variable "Type_of_glass" is stored in int datatype and each number represents one glass type. While finding correlations and during the data modeling stage this attribute should be treated as a categorical variable instead of a continuous variable. So, to avoid any misleading, this attribute is converted to object datatype using '*astype('str')*' function.

Data Exploration:

Individual Features:

For exploring individual columns, '*describe()*' attribute of Python provides useful insights about each column in the dataset like minimum and maximum values, mean, standard deviation, count of value, and median (Refere Figure 2). Some of the useful insights from it are as follow:

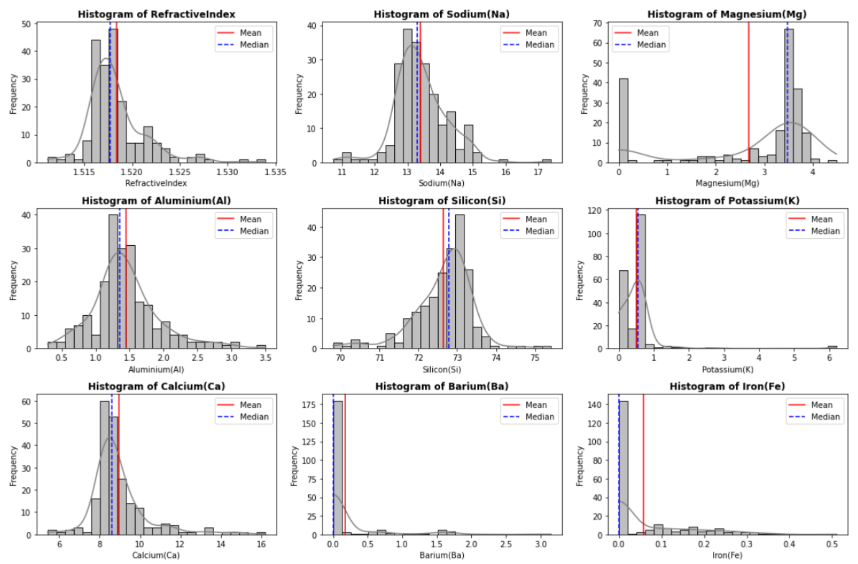
- Refractive Index values lie around the range of 1.5 which is the ideal value for the Refractive Index of glass according to Physics.
- Silicon (Si) concentration is higher compared to other elements.

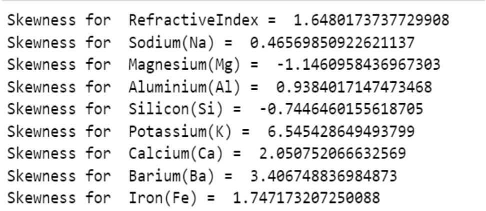


	RefractiveIndex	Sodium(Na)	Magnesium(Mg)	Aluminium(Al)	Silicon(Si)	Potassium(K)	Calcium(Ca)	Barium(Ba)	Iron(Fe)
count	213.000000	213.000000	213.000000	213.000000	213.000000	213.000000	213.000000	213.000000	213.000000
mean	1.518348	13.404085	2.679202	1.449484	72.655070	0.498873	8.954085	0.175869	0.057277
std	0.003033	0.816662	1.443691	0.495925	0.773998	0.653185	1.425882	0.498245	0.097589
min	1.511150	10.730000	0.000000	0.290000	69.810000	0.000000	5.430000	0.000000	0.000000
25%	1.516520	12.900000	2.090000	1.190000	72.280000	0.130000	8.240000	0.000000	0.000000
50%	1.517680	13.300000	3.480000	1.360000	72.790000	0.560000	8.600000	0.000000	0.000000
75%	1.519150	13.810000	3.600000	1.630000	73.090000	0.610000	9.150000	0.000000	0.100000
max	1.533930	17.380000	4.490000	3.500000	75.410000	6.210000	16.190000	3.150000	0.510000

Figure 2: Statistical info of the dataset.

Skewness is a measure of asymmetry or distortion of symmetric distribution. It measures the deviation of the given distribution of a random variable from a symmetric distribution, such as a normal distribution. This measure results can be useful in many ways like understanding the data distribution, selecting appropriate statistical methods, and improving model performance. (Refer Figure 3)





```
Skewness for  RefractiveIndex =  1.6480173737729908
Skewness for  Sodium(Na) =  0.46569850922621137
Skewness for  Magnesium(Mg) =  -1.1460958436967303
Skewness for  Aluminium(Al) =  0.9384017147473468
Skewness for  Silicon(Si) =  -0.7446460155618705
Skewness for  Potassium(K) =  6.545428649493799
Skewness for  Calcium(Ca) =  2.050752066632569
Skewness for  Barium(Ba) =  3.406748836984873
Skewness for  Iron(Fe) =  1.747173207250088
```

Figure 3 Distribution of each feature and skewness measure.

It can be inferred from the above results that, none of the columns are equally distributed. Values of column(element) Aluminium and Calcium are slightly equally distributed. The values of the element Silicon lie in higher concentration compared to other elements followed by Sodium and Calcium. Skewed data can lead to biased or suboptimal models. Understanding skewness allows for appropriate transformations (e.g., log transformation) to normalize the data, thereby improving model performance.

To deeply analyse each feature, visualisation would be a great help. Various plots like kdeplots, violinplots, displot have been used to explore each feature. Appropriate graphs have been plotted along with the statistical information to understand each feature in the dataset. For the below, Refer Figure 2 and Figure 4-13.

(a)Refractive Index(RI): The graph shows multiple peaks indicating refractive index values are not uniformly distributed. Noticeably peaks around 1.515 and 1.520 indicate these values are more common in the dataset. The distribution is likely symmetrical, centered around a refractive index of 1.518. This is because the KDEplot graph is centered at around 1.518 and the mean (1.5183) is close to the median (1.517). The spread of the refractive index is small. This is because the KDEplot graph shows a narrow peak and the standard deviation (0.003) is small

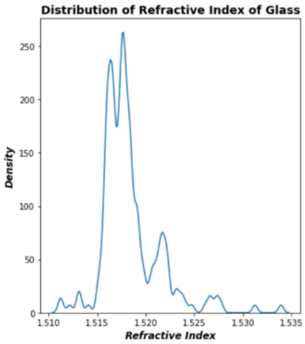
(b) Sodium (Na): The peak (mode) of the distribution is around 13.3 to 13.4, which can also be verified from statistical info results with the mean (13.404) and median (13.300). The mean and median of Sodium (Na) are very close, indicating a fairly symmetric distribution centered around these values. The distribution shows a slight positive skewness (right tail) as the maximum value (17.38) is very far from the mean compared to the minimum value (10.73).

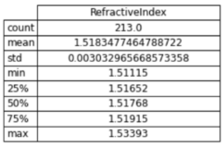
(c)Magnesium (Mg): The magnesium content predominantly clusters at the lower end of the scale, with values around 0 to 2. There is, however, an outlier indicating a significantly higher magnesium content, which may suggest a different glass type where magnesium plays a key role. One group has very low magnesium, while another has significantly higher levels, potentially indicating different types of glass for various applications

(d)Aluminium (Al): The data for Aluminium are mostly concentrated around 1.0 to 1.6 with a median of around 1.36.

(e)Silicon (Si): Silicon, has most of the values around 72 to 74 suggesting it as a common major component of glass. This also indicates its consistent presence in the glass composition of different type of glass.

(f)Potassium(K): The potassium content is generally low in this dataset, with most values near the lower end of the scale. This suggests that potassium is not a major component in most glasses in this dataset, but a few samples might have higher amounts for specific properties.





	RefractiveIndex
count	213.0
mean	1.5183477464788722
std	0.003032965668573358
min	1.51115
25%	1.51652
50%	1.51768
75%	1.51915
max	1.53393

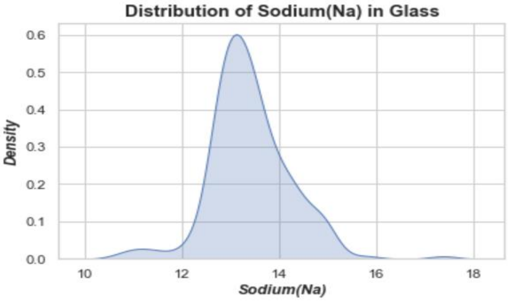
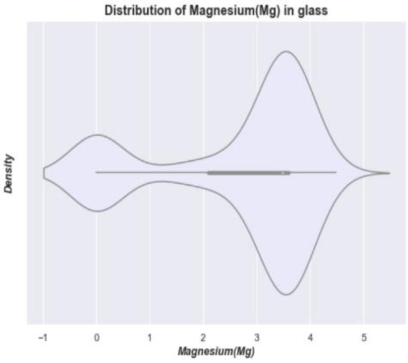
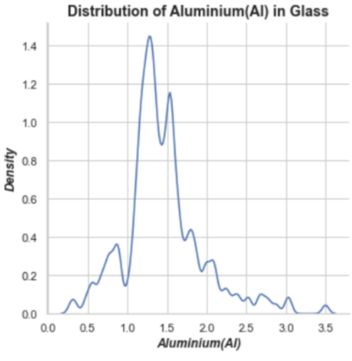


Figure 4: Distribution and statistical information of RI.

Figure 5: Distribution of Sodium





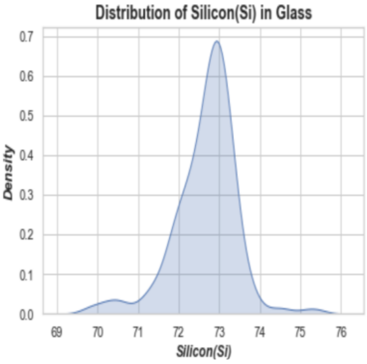


Figure 6: Distribution of Magnesium

Figure 7: Distribution of Aluminium

Figure 8: Distribution of Silicon

(g)Calcium (Ca): Calcium shows a slightly symmetrical broad distribution mainly between 8 and 10.

(h)Barium (Ba): The barium content is generally very low, with a few exceptions, indicating it's not commonly used in the glass types sampled here. The outliers might be specialty glasses with barium to enhance certain properties.

(i)Iron (Fe): Iron content is mostly negligible, but there are a few samples with slightly higher levels. This could be related to color or physical properties of the glass.

(j)Glass_Type: The dataset has imbalance class and it is not uniformly distributed.

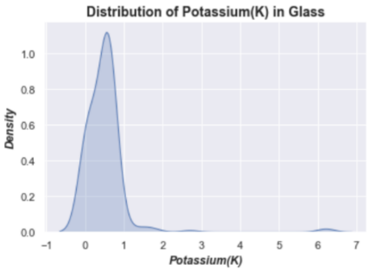


Figure 9: Distribution of Potassium

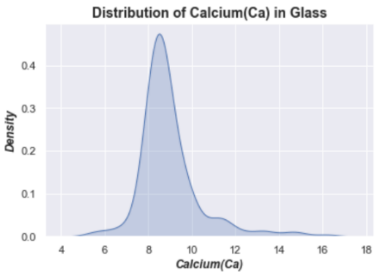


Figure 10: Distribution of Calcium

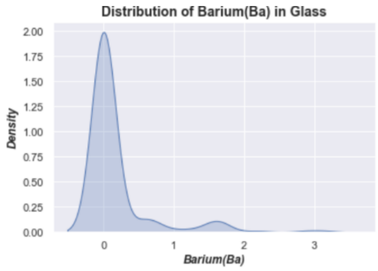


Figure 11: Distribution of Barium

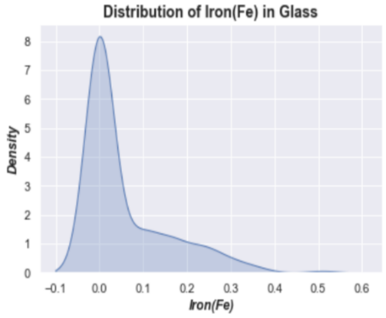


Figure 12: Distribution of Iron

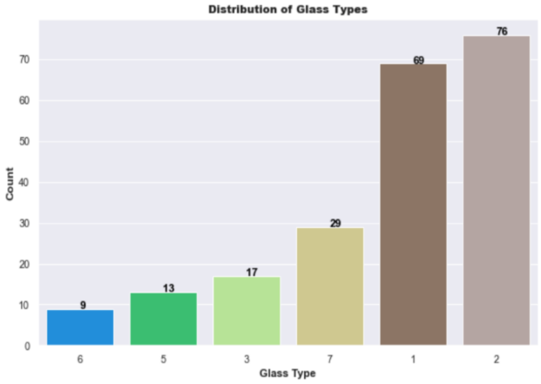
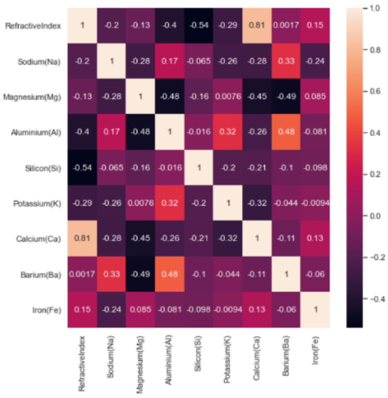


Figure 13: Distribution of Glass Type values

Pair of attributes:

To explore the relationship between features, a pairplot is plotted using heatmap which provides the correlation between features present in the dataset. (Refer Figure 14). There is a high correlation between Calcium and Refractive Index. There is moderate correlation between (Silicon and Refractive Index),(Barium and Magnesium), (Aluminium and Magnesium) and (Aluminium and Barium). When each pairs are explored, various hypothesis was derived. Valuable hypotheses are mentioned below.



	RefractiveIndex	Sodium(Na)	Magnesium(Mg)	Aluminium(Al)	Silicon(Si)	Potassium(K)	Calcium(Ca)	Barium(Ba)	Iron(Fe)
RefractiveIndex	1	-0.2	-0.13	-0.4	-0.54	-0.29	0.81	0.0017	0.15
Sodium(Na)	-0.2	1	-0.28	0.17	-0.065	-0.26	-0.28	0.33	-0.24
Magnesium(Mg)	-0.13	-0.28	1	-0.48	-0.16	0.0076	-0.45	-0.49	0.085
Aluminium(Al)	-0.4	0.17	-0.48	1	-0.016	0.32	-0.26	0.48	-0.081
Silicon(Si)	-0.54	-0.065	-0.16	-0.016	1	-0.2	-0.21	-0.1	-0.098
Potassium(K)	-0.29	-0.26	0.0076	0.32	-0.2	1	-0.32	-0.044	-0.0094
Calcium(Ca)	0.81	-0.28	-0.45	-0.26	-0.21	-0.32	1	-0.11	0.13
Barium(Ba)	0.0017	0.33	-0.49	0.48	-0.1	-0.044	-0.11	1	-0.06
Iron(Fe)	0.15	-0.24	0.085	-0.081	-0.098	-0.0094	0.13	-0.06	1

Figure 14: Pairplot for all continuous value columns

Hypothesis_1: When Calcium content increases, the refractive index also tends to increase. This is arrived from the weak positive correlation between the two variables. (Refer Figure 15)

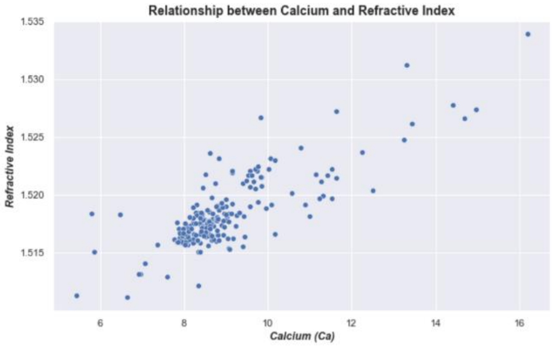


Figure 15: relationship between Calcium and Refractive Index

Hypothesis_2: As the Silicon content increases, the Refractive Index tends to decrease slightly. This is arrived from the slight negative correlation between these two values. (Refer Figure 16)

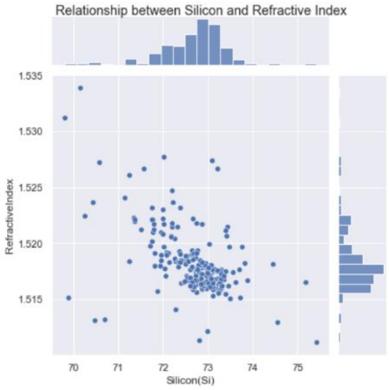


Figure 16: Relationship between Silicon and Refractive Index

Hypothesis_3: Higher levels of Aluminium correspond to lower levels of Barium and vice versa. This could suggest that in glass manufacturing, barium might be used as an alternative or supplement to aluminium in certain glass types, possibly due to their different impacts on properties like density and refractive index. (refer Figure 17)

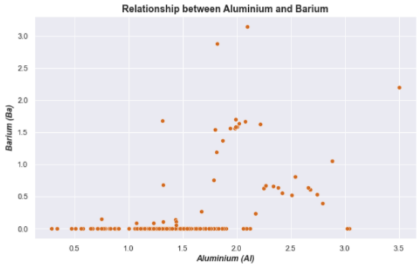


Figure 17: Relationship between Aluminium and Barium

Hypothesis_4: All the glass type have similar refractive index which is around 1.51, suggesting this as ideal value. (Refer Figure 18)

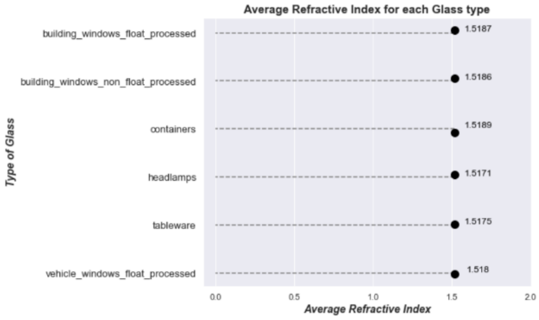
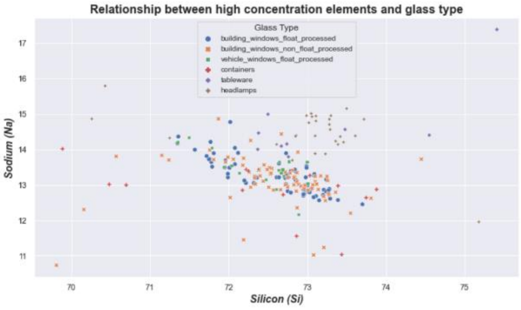


Figure 18: Relationship between each Glass Type and Refractive Index

Hypothesis_5: Increase in Silicon content could decrease the Sodium content. This is arrived from the weak negative correlation between these two values.



Hypothesis_6: There is no correlation or relationship between Barium & Magnesium pair and Aluminium & Magnesium pair. (Refer Figure 19)

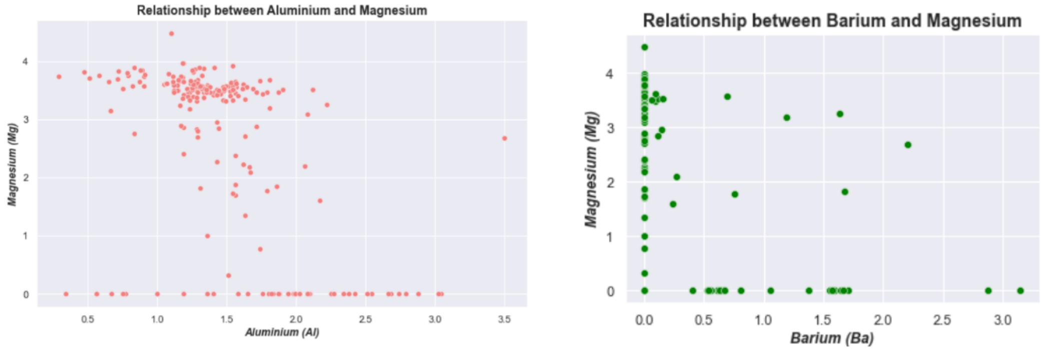


Figure 19: No correlation pairs

Data Modelling:

In this section, data is trained by building 2 classification models based on KNN classification and Decision Tree. This involves various steps like standardisation, parameter tuning, feature selection, model training, validating the model and evaluation.

(a)Standardisation:

Standardization is performed to bring down all the features on the same scale. By bringing down all the features to same scale the model treat each feature as same. Here Min-Max scalar is used which scales data to a fixed range, typically 0 to 1, by subtracting the minimum value and dividing by the range (max-min). This is performed on the input sample X (without the target variable) inorder to prevent data-leakage.

(b) Parameter turning for classifiers:

The important parameters in KNN classification algorithm which decides the output are

- n_neighbors – number of nearest neighbours to consider for classification

- p (Distance metric) – determines how similarity between data points is calculated. Possible values are 1 (Manhattan distance) or 2 (Euclidean distance)
- weights – to assign weights votes to the nearest neighbors. Possible values are distance or uniform.

The important parameters in KNN classification algorithm which decides the output are,

- criterion – specifies the quality of split at each node – possible values are gini, entropy
- max_depth – depth of the decision tree – integer
- min_samples_split – minimum samples required to split a node - integer • max_features – number of features to consider when looking for best split - int, float or {“auto”, “sqrt”, “log2”}, default=None
- min_samples_leaf- minimum number of samples required to be at a leaf node – integer

The parameters for both the classifiers are tuned by passing different combinations of these values and arrived at the best parameters. Basic for loop approach is used.

(c)K-fold cross validation

K-fold cross vcalidation is performed by randomly dividing data into a training set with X% of the observations and keeping the rest as a holdout data set.

The key parameter is k, which determines on how many folds the data is splitted. This approach is performed using k-fold and cross_val_score. This is found by passing different values and calculating the accuracy score using cross_val score.

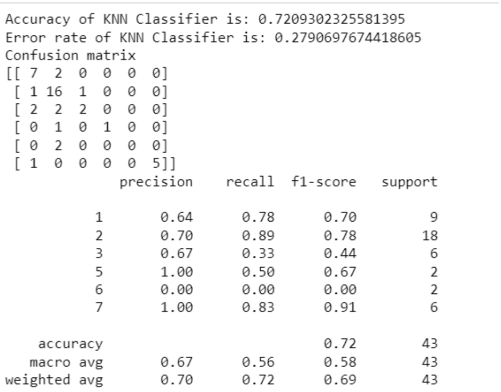
The best value is then used to predict the output.

(d)Feature selection

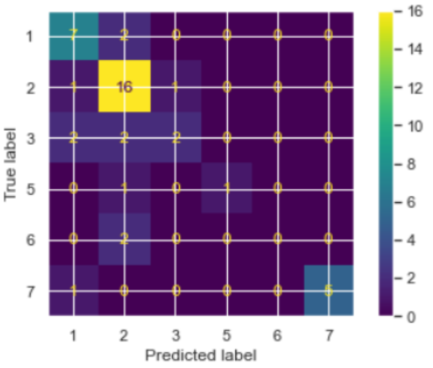
Hill climbing algorithm helps to determine the best features that is useful for output prediction. This is used to scale down the features and then classification is performed to get optmised results.

Results:

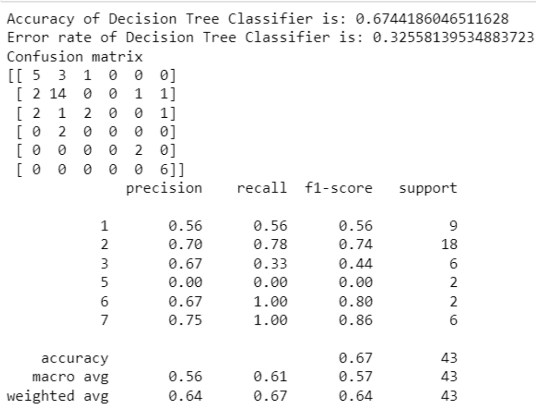
Results of KNN classifier using optimised dataset using feature selection and parameter tuning:



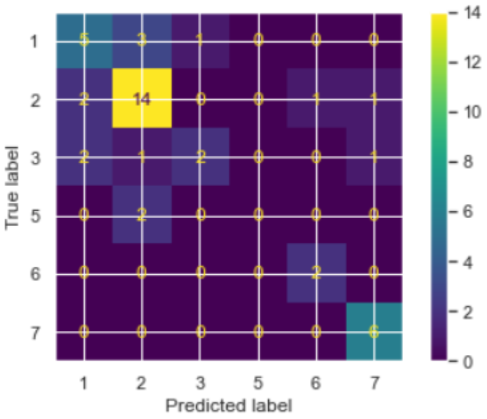
```
Accuracy of KNN Classifier is: 0.7209302325581395
Error rate of KNN Classifier is: 0.2790697674418605
Confusion matrix
[[ 7  2  0  0  0  0]
 [ 1 16  1  0  0  0]
 [ 2  2  2  0  0  0]
 [ 0  1  0  1  0  0]
 [ 0  2  0  0  0  0]
 [ 1  0  0  0  0  5]]
              precision    recall  f1-score   support

           1       0.64      0.78      0.70         9
           2       0.70      0.89      0.78        18
           3       0.67      0.33      0.44         6
           5       1.00      0.50      0.67         2
           6       0.00      0.00      0.00         2
           7       1.00      0.83      0.91         6

    accuracy                           0.72        43
   macro avg       0.67      0.56      0.58        43
weighted avg       0.70      0.72      0.69        43
```



Results of Decision Tree classifier using optimised dataset using feature selection and parameter tuning:

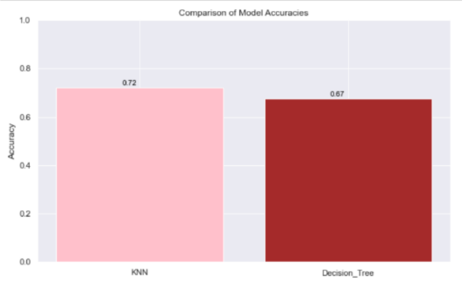


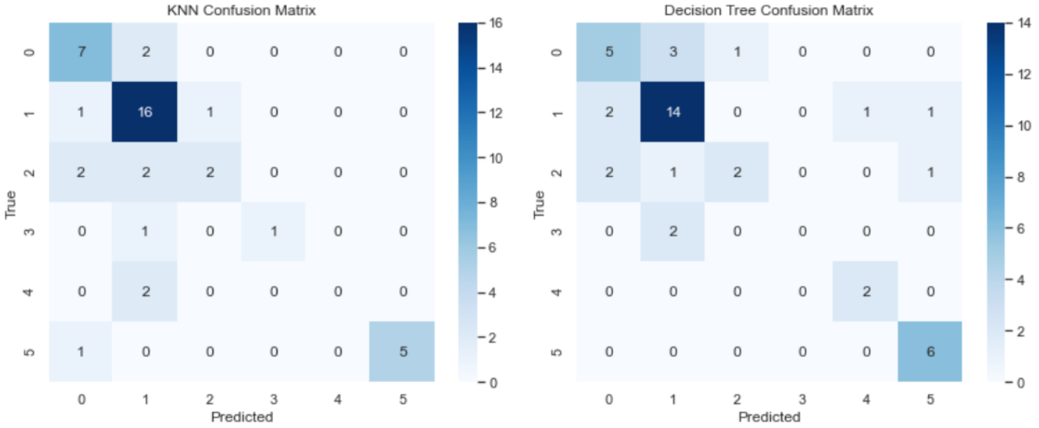
```
Accuracy of Decision Tree Classifier is: 0.6744186046511628
Error rate of Decision Tree Classifier is: 0.32558139534883723
Confusion matrix
[[ 5  3  1  0  0  0]
 [ 2 14  0  0  1  1]
 [ 2  1  2  0  0  1]
 [ 0  2  0  0  0  0]
 [ 0  0  0  0  2  0]
 [ 0  0  0  0  0  6]]
              precision    recall  f1-score   support

           1       0.56      0.56      0.56         9
           2       0.70      0.78      0.74        18
           3       0.67      0.33      0.44         6
           5       0.00      0.00      0.00         2
           6       0.67      1.00      0.80         2
           7       0.75      1.00      0.86         6

    accuracy                           0.67        43
   macro avg       0.56      0.61      0.57        43
weighted avg       0.64      0.67      0.64        43
```



Discussion:

By comparing the results of the KNN classifier and Decision Tree classifier, the KNN performs better than Decision Tree. So this can be selected as a good algorithm for the Glass Identification dataset.





Conclusion:

The data modelling and presentation for the classification problem using Glass Identification dataset shows good results for the chosen algorithms.

References:

Absolute indices of refraction list, index of refraction of various materials. (n.d.). Retrieved May 29, 2024, from https://www.physlink.com/reference/indicesofrefraction.cfm

B. German. (1987). Glass identification [dataset]. [object Object]. https://doi.org/10.24432/C5WW2P

Polyanskiy, M. N. (2024). Refractiveindex.info database of optical constants. Scientific Data, 11(1), 94. https://doi.org/10.1038/s41597-023-02898-2

Refractive index. (2024). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Refractive_index&oldid=1220231635

Seaborn. Kdeplot—Seaborn 0. 13. 2 documentation. (n.d.). Retrieved May 29, 2024, from https://seaborn.pydata.org/generated/seaborn.kdeplot.html

Skewness. (n.d.). Corporate Finance Institute. Retrieved May 29, 2024, from https://corporatefinanceinstitute.com/resources/data-science/skewness/

Referred from Practical Data Science with Python course materials