Glass Identification Using Machine Learning

Nadeem Khan

RMIT University

s3764302@student.rmit.edu.au

Qihang Li

RMIT University

s3625051@student.rmit.edu.au

23.05.2019

Table of Contents

EXECUTIVE SUMMARY

The main goal in this report is to investigate the classification of a fragment of glass by implementing K Nearest Neighbour (K-NN) classifier and Decision Tree and get a recommendation for criminological investigators after comparing the performance and accuracy of those two Classification models. The dataset used to analysed is obtained from the UCI Machine Learning Repository Website. And K-NN classifier and Decision Tree are implemented to analyse by using a python script. Overall, the results indicate that K-NN has better performance and higher accuracy rate. It is recommended that using K-NN classifier for glass identification will help criminological investigators solve cases more effectively and efficiently.

Keywords----- Classify glass, K Nearest Neighbour (KNN) Model, Decision Tree

1 INTRODUCTION

It's universally acknowledged that tiny evidence can help investors to make a positive effort on the criminological investigation, such as tracking suspects by identifying the type of small segments of glass the left in the crime scene correctly. Therefore, the study of the classification of types of glass was motivated. [1] Machine learning, as the technology which is expected to bring huge changes to the world, makes it possible to quickly and automatically to analyze bigger and more complex data faster and more accurately, by extracting knowledge from raw data to get reliable and reasonable patterns. The classification of the type of glass fragments is also benefited a lot from machine learning's development.

The main purpose of this report is to classify correctly a small fragment of glass, based upon its main chemical component measured by using machine learning technology. The algorithms implemented to recognize the potential pattern and classify using K Nearest Neighbour (K-NN) classifier and Decision Tree.

This report also investigates what parameters can enhance the classification accuracy rate of K-NN and Decision Tree models. We also compare two models based on the following terms: confusion matrix, classification error rate, precision, recall and F1-Score. Finally, we give recommendations based on the results.

The remainder of the paper is structured as follows. In Section 2, we describe the methodology we used, including data preparation, data exploration and data modelling. Section 3 describes the results of the analysis. We discuss the results in Section 4 and

conclude in Section 5 with a summary of our findings and recommendations.

2 METHODOLOGY

We perform data preparation to get the dataset from the UCI Machine Learning Repository website. Then we explore each column of the dataset to figure out the meaning of each attribute and get the descriptive statistics of each attribute. Besides, we explore the relationship between all pairs of attributes to address plausible hypothesis. In the data modelling process, which is the most significant part of our research, we split the data into a training set and a test set by different percentages. After that, K-NN model and Decision Tree model are implemented. We choose the appropriate values for each parameter in those two models based on the accuracy rates. Finally, Those two models are compared by the following terms: Confusion Matrix, Classification Error Rate, Precision, Recall and F1-Score.

2.1 Data Preparation and Exploration

2.1.1 Dataset Description

The dataset determines the type of glass based its chemical components and its refractive index. There are 214 instances in the dataset. It has 10 attributes and from attribute 3 to 10 is the weight percentage in the corresponding oxide which are oxides of Sodium (Na), Magnesium (Mg), Aluminum (Al), Silicon (Si), Potassium (K), Calcium (Ca), Barium (Ba) and Iron (Fe). Attribute 2 is the refractive index of the glass fragment.

The type attributes are represented by a number from 1 to 7. And they are building windows float processed, building window non-float processed, vehicle windows float processed, vehicle windows non-float processed (none in this dataset), containers, tableware and headlamps.

2.2 Data Modelling

The way we perform data modelling is that split the dataset into a training set and a testing set at the following ratio:

- 50% for training and 50% for testing
- 60% for training and 40% for testing
- 80% for training and 20% for testing

For each of the training/testing split, train the model by selecting appropriate values of each parameter in the model.

1. For the K Nearest Neighbour model, the k is the significant parameter influences the

performances. The k-value means the number of closest neighbours the model will consider. The way we use to choose proper k-value is comparing the error rate within different k values and constructing the graph of the relationship between error rate and k-value.

2. For the Decision Tree model, the parameters to pre-tune to optimize the performance are max_depth, criterion. The max_depth indicates how deep the decision tree can be. By increasing the depth of the tree, the tree can have more splits and capture more information about the data. The criterion parameter provides the ability to use the different-different attribute selection measure. [1]

3 RESULTS

3.1 Results of Data Preparation and Exploration

Based on the observation of descriptive Statistics graph, we find that 0 value appears in the Mg, K and Ca elements' descriptive statistics. That indicates for some types of glass, they don't contain Mg, K and Ca elements. That could be an important criterion for the classification of glass type. The descriptive statistics graph and portion graph are below. Also concentration by weight in glass is very high for silicon as its the main component of glass(about 70 percentage).

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
count	214.000000	214.000000	214.000000	214.000000	214.000000	214.000000	214.000000	214.000000	214.000000
mean	1.518365	13.407850	2.684533	1.444907	72.650935	0.497056	8.956963	0.175047	0.057009
std	0.003037	0.816604	1.442408	0.499270	0.774546	0.652192	1.423153	0.497219	0.097439
min	1.511150	10.730000	0.000000	0.290000	69.810000	0.000000	5.430000	0.000000	0.000000
25%	1.516523	12.907500	2.115000	1.190000	72.280000	0.122500	8.240000	0.000000	0.000000
50%	1.517680	13.300000	3.480000	1.360000	72.790000	0.555000	8.600000	0.000000	0.000000
75%	1.519157	13.825000	3.600000	1.630000	73.087500	0.610000	9.172500	0.000000	0.100000
max	1.533930	17.380000	4.490000	3.500000	75.410000	6.210000	16.190000	3.150000	0.510000

Figure 1. Descriptive Statistics

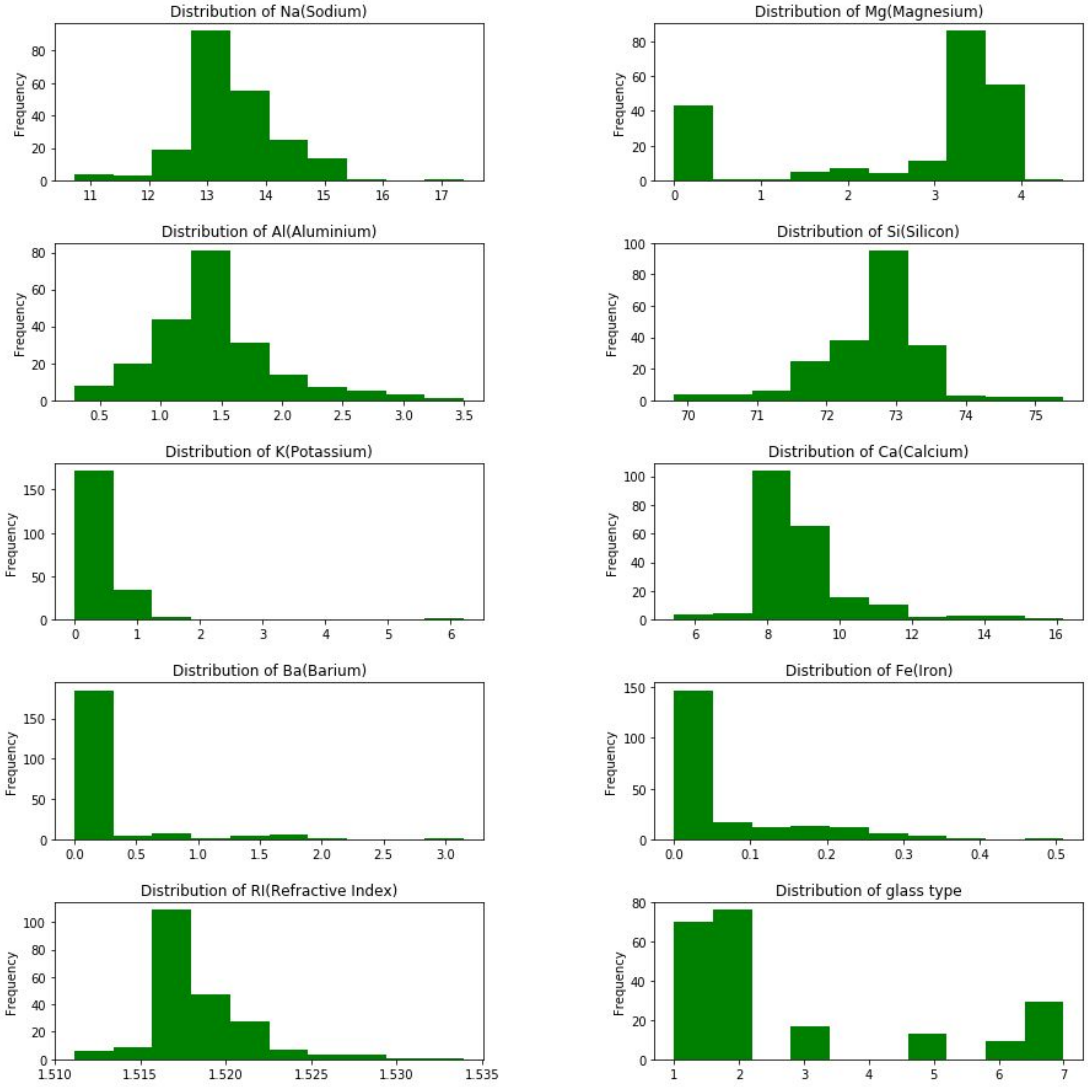


Figure 3. The proportion of Each Type

Above histograms of each column shows their distribution and in case of glass type, it shows that observation of glass type 3,5 and 6 are very less compared to other observations of glass type.

In order to observe the relationship between the pairs of attributes, we generate the Scatter Matrix graph, which is shown below.

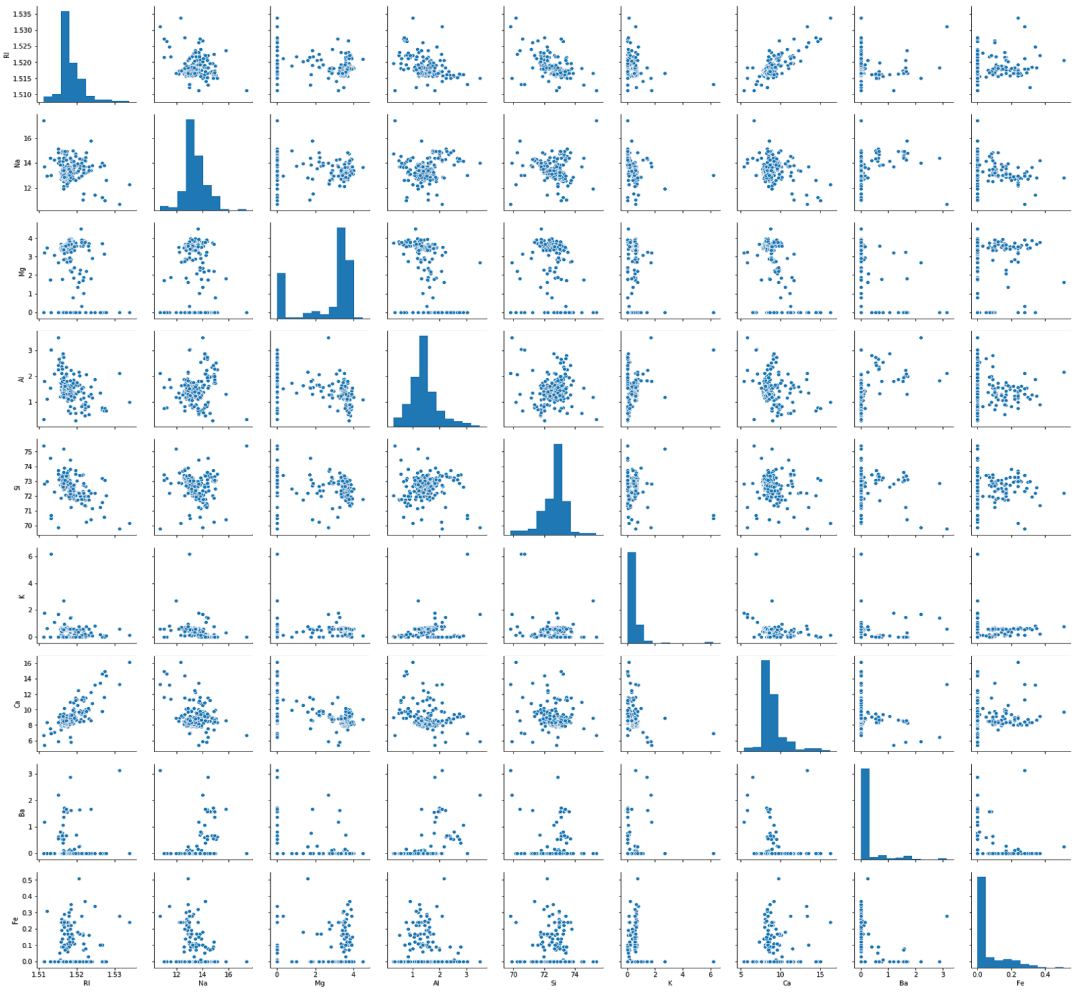
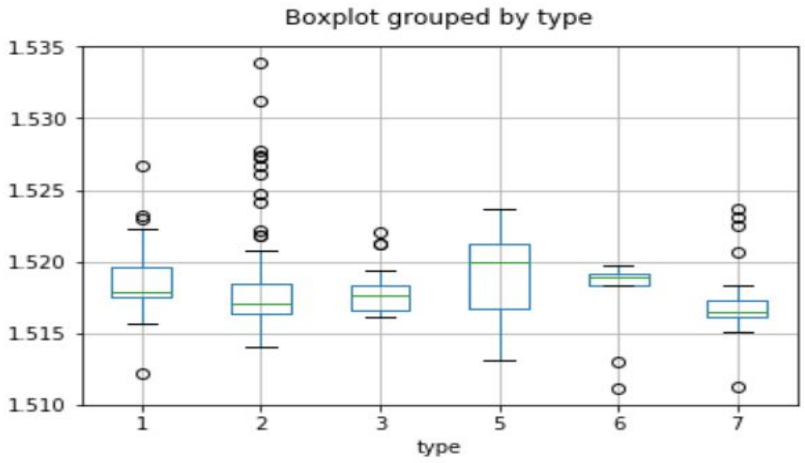


Figure 4. Scatter Matrix for all columns

Plausible hypothesis: is there relationship between different elements for glass classification

After exploring the graph of chemical elements and reflective index, we can find that there is a positive correlation relationship between Calcium (Ca) and reflective index. Besides, Aluminum (Al) and Silicon (Si) have negative correlation relationship with reflective index. The rest of the elements don't show an obvious relationship with the reflective index. That indicates that the contents of Calcium, Aluminum and Silicon could be a significant index to classify glass as different types of glass have a different level of the reflective index.

Figure 5. Boxplot of RI and Type



Different types of glass have different Reflective Index (RI) range and average which is clearly shown by the graph above. As a result, the reflective index can make an effect on classifying the type of glass. Besides, we can find that the range and average of RI of non-floated type glass are much lower than floated glass. Therefore, the other plausible hypothesis is that the float process can influence the reflective index.

3.2 Results of Data Modelling

3.2.1 K-NN Model

After collecting the error rates within K-value from 1 to 20. The result clearly shows that the appropriate K-value is 1, as the error rate is lowest within K-value 1. (See Figure 6). For other Splits same process is repeated for finding best k values.

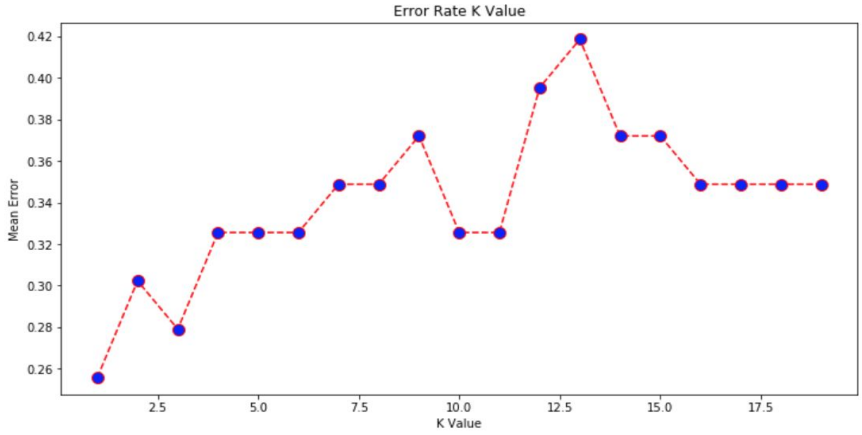


Figure 6. Error Rate With Different K Value

1. 50% data for training and 50% for testing

```
Confusion Matrix:
[[31  2  3  0  0  0]
 [ 1 28  1  2  1  0]
 [ 4  4  2  0  0  0]
 [ 0  1  0  4  0  0]
 [ 0  0  0  0  4  0]
 [ 1  0  0  0  0 18]]
Classification Report:
              precision    recall  f1-score   support

           1       0.84      0.86      0.85        36
           2       0.80      0.85      0.82        33
           3       0.33      0.20      0.25        10
           5       0.67      0.80      0.73         5
           6       0.80      1.00      0.89         4
           7       1.00      0.95      0.97        19

   micro avg       0.81      0.81      0.81       107
   macro avg       0.74      0.78      0.75       107
weighted avg       0.80      0.81      0.80       107

Classification Error Rate:
0.1869158878504673
```

2. 60% data for training and 40% for testing

```
Confusion Matrix:
[[27  3  3  0  0  0]
 [ 1 22  1  1  1  0]
 [ 2  3  3  0  0  0]
 [ 0  1  0  3  0  0]
 [ 0  0  0  0  2  0]
 [ 0  1  0  0  0 12]]
Classification Report:
              precision    recall  f1-score   support

           1       0.90      0.82      0.86        33
           2       0.73      0.85      0.79        26
           3       0.43      0.38      0.40         8
           5       0.75      0.75      0.75         4
           6       0.67      1.00      0.80         2
           7       1.00      0.92      0.96        13

   micro avg       0.80      0.80      0.80        86
   macro avg       0.75      0.79      0.76        86
weighted avg       0.81      0.80      0.80        86

Classification Error Rate:
0.19767441860465118
```

3 80% data for training and 20% for testing

```
Confusion Matrix:
[[14  2  3  0  0  0]
 [ 1  9  0  1  1  0]
 [ 1  2  3  0  0  0]
 [ 0  0  0  1  0  0]
 [ 0  0  0  0  1  0]
 [ 0  0  0  0  0  4]]
Classification Report:
              precision    recall  f1-score   support

           1       0.88      0.74      0.80        19
           2       0.69      0.75      0.72        12
           3       0.50      0.50      0.50         6
           5       0.50      1.00      0.67         1
           6       0.50      1.00      0.67         1
           7       1.00      1.00      1.00         4

   micro avg       0.74      0.74      0.74        43
   macro avg       0.68      0.83      0.73        43
weighted avg       0.77      0.74      0.75        43

Classification Error Rate:
0.2558139534883721
```

The results show that the performance of the K-NN model can be optimized by increasing the number of data using for training. That indicates that the K-NN can make effort on classifying the type of glass. However, the performance of classifying glass type 3, 4 and 5 was very less before tuning the parameters as the observations of these types of glass are very less in the dataset.

3.2.2 Decision Tree Model

We perform the same process we did in finding the proper parameters of the K-NN model, which is collecting the error rates within different Max_Depth. After collecting the error rates within Max_Depth from 1 to 15 which is the max tree depth when the testing set is 20% of total data. The result clearly shows that the error rate is less for max depth(4 and 12) but we choose maxdepth = 4 because maxdepth = 12 causing the problem of overfitting the tree (See Figure 7). This same process of tuning the parameter is done with other data spits.

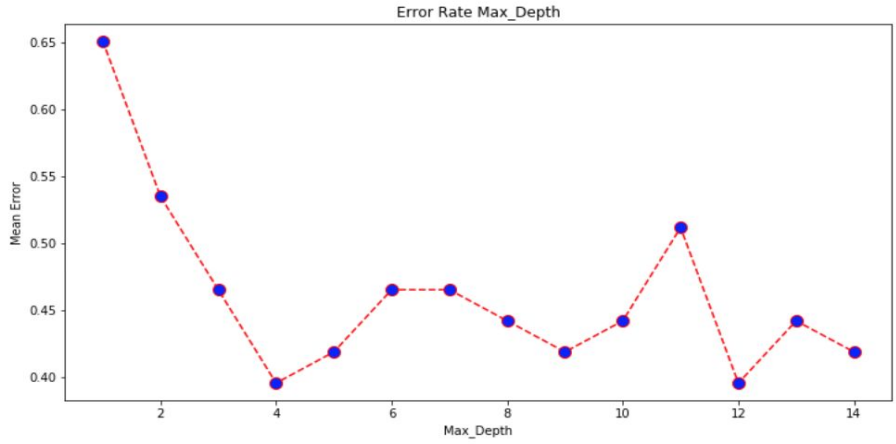


Figure 7. Error Rate With Different Max_Depth

1. 50% for training and 50% for testing:

```
Confusion Matrix:
[[24  9  3  0  0  0]
 [ 5 23  2  2  0  1]
 [ 4  5  1  0  0  0]
 [ 0  0  0  5  0  0]
 [ 0  1  0  0  3  0]
 [ 1  2  0  0  0 16]]
Classification Report:
              precision    recall  f1-score   support

           1       0.71      0.67      0.69        36
           2       0.57      0.70      0.63        33
           3       0.17      0.10      0.12        10
           5       0.71      1.00      0.83         5
           6       1.00      0.75      0.86         4
           7       0.94      0.84      0.89        19

   micro avg       0.67      0.67      0.67       107
   macro avg       0.68      0.68      0.67       107
weighted avg       0.67      0.67      0.67       107

Classification Error Rate:
0.3271028037383178
```

2. 60% for training and 40% for testing

```
Confusion Matrix:
[[14 17  2  0  0  0]
 [ 1 22  1  2  0  0]
 [ 3  5  0  0  0  0]
 [ 0  1  0  3  0  0]
 [ 0  0  0  0  2  0]
 [ 1  1  0  1  0 10]]
Classification Report:
              precision    recall  f1-score   support

           1       0.74      0.42      0.54        33
           2       0.48      0.85      0.61        26
           3       0.00      0.00      0.00         8
           5       0.50      0.75      0.60         4
           6       1.00      1.00      1.00         2
           7       1.00      0.77      0.87        13

   micro avg       0.59      0.59      0.59        86
   macro avg       0.62      0.63      0.60        86
weighted avg       0.63      0.59      0.57        86

Classification Error Rate:
0.40697674418604646
```

3. 80% for training and 20% for testing

```
Confusion Matrix:
[[ 8 11  0  0  0  0]
 [ 1  9  0  0  1  1]
 [ 1  4  1  0  0  0]
 [ 0  0  0  1  0  0]
 [ 0  0  0  0  1  0]
 [ 0  0  0  0  0  4]]
Classification Report:
              precision    recall  f1-score   support

           1       0.80      0.42      0.55        19
           2       0.38      0.75      0.50        12
           3       1.00      0.17      0.29         6
           5       1.00      1.00      1.00         1
           6       0.50      1.00      0.67         1
           7       0.80      1.00      0.89         4

   micro avg       0.56      0.56      0.56        43
   macro avg       0.75      0.72      0.65        43
weighted avg       0.71      0.56      0.54        43

Classification Error Rate:
0.4418604651162791
```

The results show that the performance of the Decision Tree model is not satisfactory. The error rate is even increasing after adding the training data's quantity. That means it's not actually "learning" and it may not make efforts on glass classification. Besides, selecting an appropriate parameter is not efficient, as the higher value of maximum depth causes overfitting, and a lower value causes underfitting [3].

4 DISCUSSION

The results of data modelling show that the K-NN model performs better on glass classification. As the testing data's quantities' increasing, the error rate is obviously getting lower. Besides, the appropriate parameter which is K-Value is easy to figure out. Not like the K-NN model, the Decision Tree model can't make much effort on classifying the type of glass. And a proper parameter is always inefficient to find because the Max_Depth of the tree is changing as the change of testing data's quantity.

	ACCURACY FOR DIFFERENT DATA SPLITS		
	20% TEST SET	40% TESTSET	50% TEST SET
KNN	74.42	80.24	81.31
DECISION TREE	74.42	59.31	67.83

Figure 8. Comparing accuracy of both the models

Moreover, the classification of type 3 and 5, which are vehicle windows float processed and

containers glass, are always effective. The fact indicates that there need more training data to support the K-NN model to improve performance.

The reason for influencing the performance of those two models on this dataset is that the K-NN model directly learns from the training instances (observation). Not like the K-NN model, Decision Tree first builds a classification model on training dataset before actually doing classification.[4]

5 CONCLUSION

This research aimed to implement two machine learning model, which are K-NN and Decision Tree, to classify the type of glass automatically and compare their performance. The results show that K-NN has a much low error rate on glass classification as the values of each attribute used to train model are numeric values.(figure 8)

Based on the results, we suggest using K-NN model to do glass classification and provide more data to train the model to improve its performance.

Besides, we can choose appropriate training model by figuring the type of values of attributes in the dataset. Decision Tree is applicable for numeric and nominal values. K-NN model is used for numeric values.

6 REFERENCES

[1] Archive.ics.uci.edu. (2019). *UCI Machine Learning Repository: Glass Identification Data Set*. [online] Available at: https://archive.ics.uci.edu/ml/datasets/glass+identification [Accessed 27 May 2019].

[2],[3] DataCamp Community. (2019). *Decision Tree Classification in Python*. [online] Available at: https://www.datacamp.com/community/tutorials/decision-tree-classification-python [Accessed 27 May 2019].

[4] KNN, D. and Richie, R. (2019). *Decision tree vs. KNN*. [online] Data Science Stack Exchange. Available at: https://datascience.stackexchange.com/questions/9228/decision-tree-vs-knn [Accessed 28 May 2019].