**Dataset**:Glass Identification

**Data Set Refference:** <http://archive.ics.uci.edu/ml/machine-learning-databases/glass/>

**Data Description:**

A data frame with 214 observation containing examples of the chemical analysis of 7 different types of glass. The problem is to forecast the type of class on basis of the chemical analysis. The study of classification of types of glass was motivated by criminological investigation. At the scene of the crime, the glass left can be used as evidence (if it is correctly identified!).

**Dataset Information**:

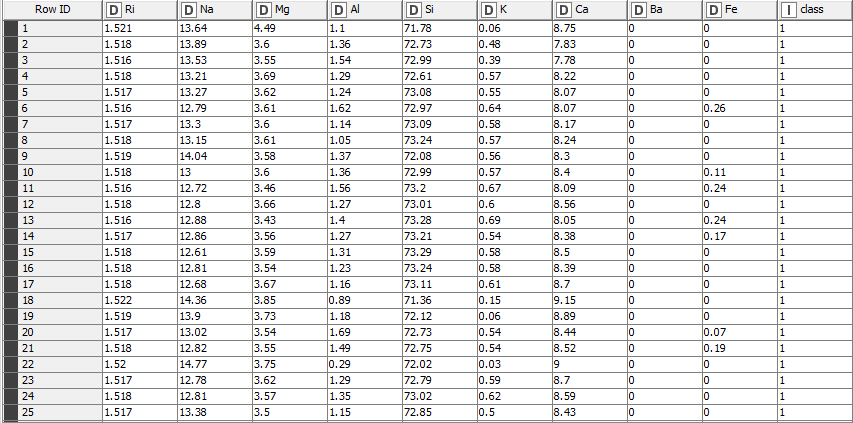
Total Instances: 214

Number of Attributes: 10 (including an Id#) plus the class attribute

Missing Value: None

|  |  |
| --- | --- |
| Name | Type |
| Id number | Integer |
| Refractive index | Double |
| Sodium | Double |
| Magnesium | Double |
| Aluminum | Double |
| Potassium | Double |
| Calcium | Double |
| Barium | Double |
| Iron | Double |
| Type of glass | Integer |

**Data Overview:**



**Steps:**

**Preprocess:**

The type of the attribute to be predicted (class) is integer. To make it prepared for the analysis the attribute was discretized using conditional in Microsoft excel as following:

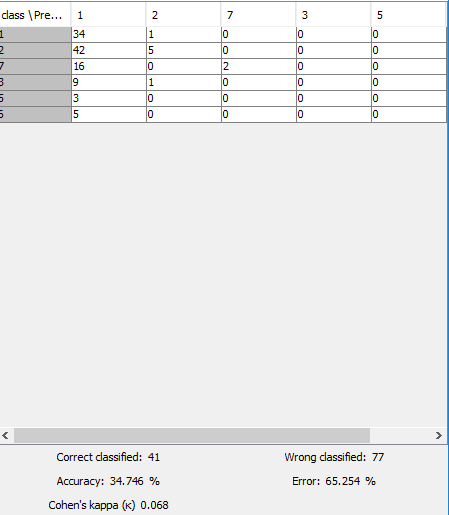
|  |  |
| --- | --- |
| Class | type |
| 1 | Youngbuilding\_windows\_float\_processed |
| 2 | building\_windows\_non\_float\_processe |
| 3 | vehicle\_windows\_float\_processed |
| 4 | vehicle\_windows\_non\_float\_processed (none in this database) |
| 5 | containers |
| 6 | tableware |
| 7 | headlamps |

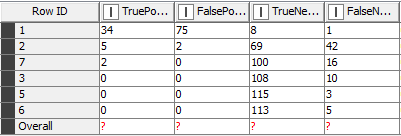
There was no missing value in this dataset hence dealing/discarding missing value was not required.

The dataset was obtained as txt file format. The dataset was transformed into the ARFF format, which is a standard way of representing dataset. After the conversion it was ready to use in knime environment.

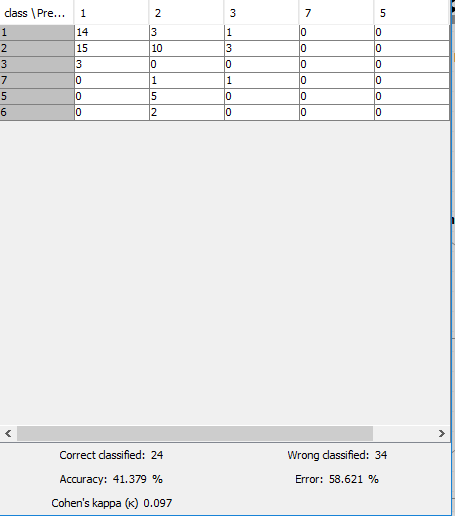
Classification: The Dataset was loaded in knime and following Classifier was tested with Naïve bayes, Decision tree and K-nearest classifier

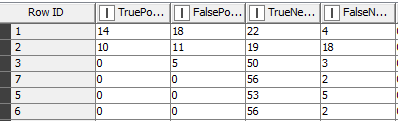
**Naïve Bayes:**



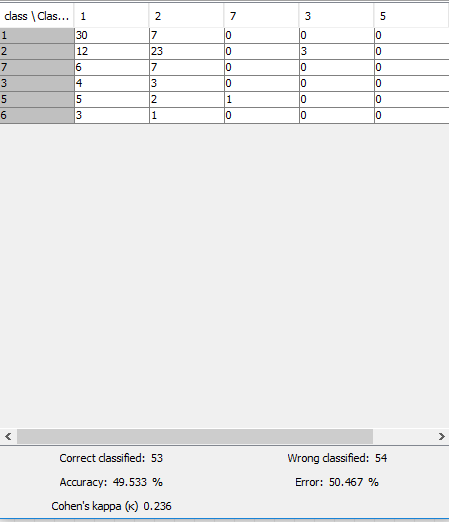


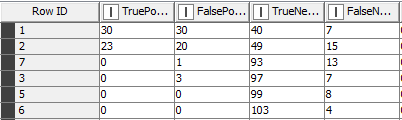
**Decission Tree:**





**K nearest:**





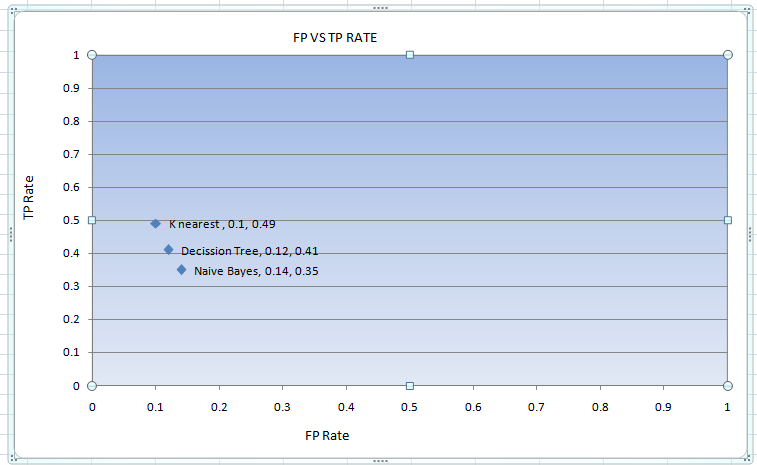
**Classifier Rate:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| naïve bayes | tp | fn |  | Fp | tn |
|  | 34 | 1 |  | 75 | 8 |
|  | 5 | 42 |  | 2 | 69 |
|  | 2 | 16 |  | 0 | 100 |
|  | 0 | 10 |  | 0 | 108 |
|  | 0 | 3 |  | 0 | 115 |
|  | 0 | 5 |  | 0 | 113 |
| total | 41 | 77 |  | 77 | 513 |
| tprate | 0.35 |  | fp rate | 0.14 |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| decision tree | tp | fn |  | Fp | tn |
|  | 14 | 4 |  | 18 | 22 |
|  | 10 | 18 |  | 11 | 19 |
|  | 0 | 3 |  | 5 | 50 |
|  | 0 | 2 |  | 0 | 56 |
|  | 0 | 5 |  | 0 | 53 |
|  | 0 | 2 |  | 0 | 56 |
| total | 24 | 34 |  | 34 | 256 |
| tp rate | 0.41 |  | fp rate | 0.12 |  |
|  |  |  |  |  |  |
| knn | tp | fn |  | Fp | tn |
|  | 30 | 7 |  | 30 | 40 |
|  | 23 | 15 |  | 20 | 49 |
|  | 0 | 13 |  | 1 | 93 |
|  | 0 | 7 |  | 3 | 97 |
|  | 0 | 8 |  | 0 | 99 |
|  | 0 | 4 |  | 0 | 103 |
| total | 53 | 54 |  | 54 | 481 |
| tp rate | 0.49 |  | Fp rate | 0.10 |  |
|  |  |  |  |  |  |
|  | Accuracy | fprate | tp rate |
| naïve | 34.746 | 0.14 | 0.35 |
| decision | 41.379 | 0.12 | 0.41 |
| knn | 49.533 | 0.10 | 0.49 |

**Distance :**

|  |  |
| --- | --- |
| Distance | Value |
| Naïve | 0.66 |
| Decission Tree | 0.60 |
| KNN classifier | 0.52 |

**ROC Graph:**



From ROC graph here it is seen that by using k-nearest neighborclassifier there TP rate is higher and FP is lower than other classifier. Here also accuracy is higher than other classifier and distance is smaller than other classifier. In this data set there only concern the class value classification. So for considering all of these It is said that k-nearest neighborclassifier is suitable for this data set.

**Summary:**

The accuracy may change if the other numeric attributes can be discretized or taking another sampling. Further improvement can be done by observing the correlation among attributes and merging them wisely.