# K-Nearest Neighbors

**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: DHEERAJ MISHRA Batch ID:** DS\_01072021

**Topic: K-Nearest Neighbors**

**Grading Guidelines:**

**1. An assignment submission is considered complete only when correct and executable code(s) are submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered for evaluation.**

**2. Assignments submitted after the deadline will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline.
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of problems are completed.

(OR)

* + All assignments are submitted after the deadline.
* **Grade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of the problems are completed.

(OR)

* + Less than 80% of problems in the assignments are submitted after the deadline.
* **Grade D: (>= 60 and < 70):**
  + Assignments submitted after the deadline and with 50% or less problems.
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of problems in the assignments are submitted after the deadline.

(OR)

* + Less than 30% of problems in the assignments are submitted before the deadline.
* **Grade F: (< 50):** No submission (or) malpractice.

**Hints:**

1. **Business Problem**
   1. **What is the business objective?**
   2. **Are there any constraints?**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



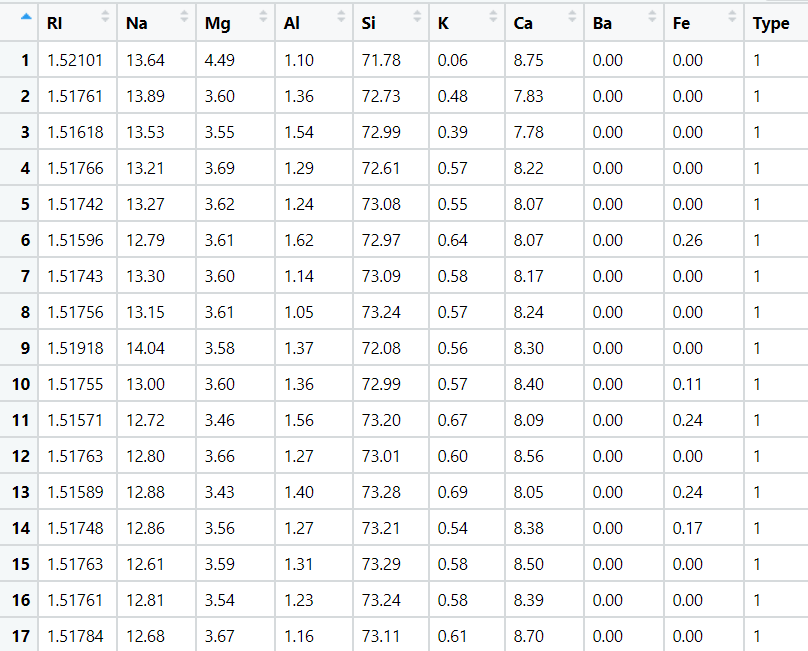
**2.1 Make a table as shown above and provide information about the features such as its data type and its relevance to the model building. And if not relevant, provide reasons and a description of the feature.**

1. **Data Pre-processing**

**3.1 Data Cleaning, Feature Engineering, etc.**

1. **Exploratory Data Analysis (EDA):**
   1. **Summary.**
   2. **Univariate analysis.**
   3. **Bivariate analysis.**
2. **Model Building**
   1. **Build the model on the scaled data (try multiple options).**
   2. **Perform KNN and use cross validation techniques to get optimum K value.**
   3. **Train and test the model and perform cross validation techniques. Compare accuracies, precision and recall and explain them in the documentation.**
   4. **Briefly explain the model output in the documentation.**

1. **Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**
2. A glass manufacturing plant uses different earth elements to design new glass materials based on customer requirements. For that, they would like to automate the process of classification as it’s a tedious job to manually classify them. Help the company achieve its objective by correctly classifying the glass type based on the other features using KNN algorithm.



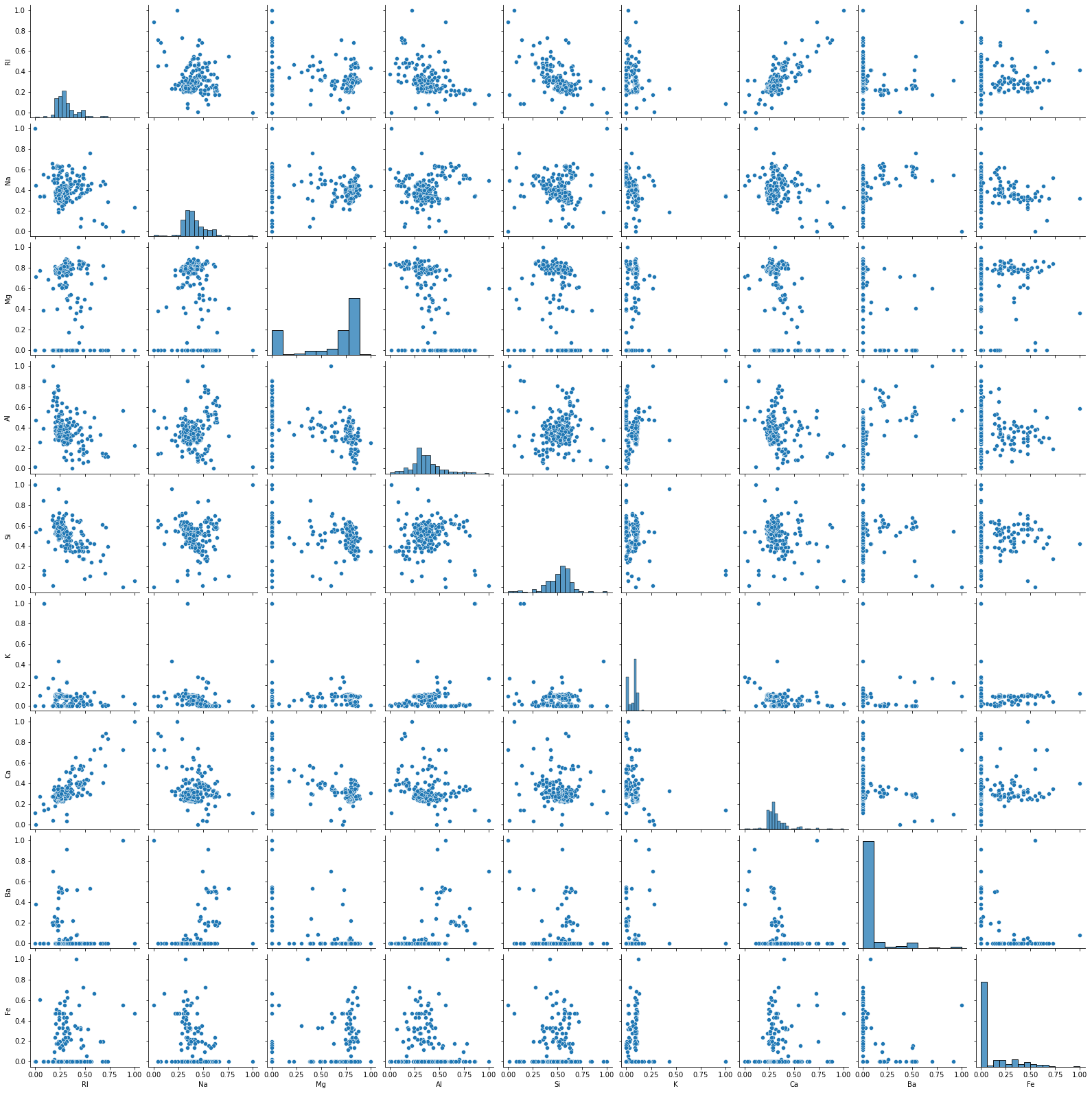
1.  BUSINESS OBJECTIVE:-

Maximize mapping of glass type

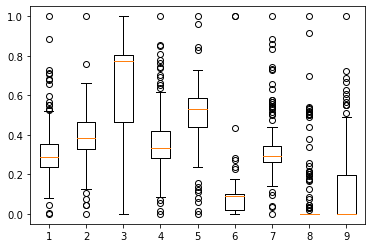
1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| RI | Refractive index of glass | Continuous | Relevant |
| Na | Amount of sodium | Continuous | Relevant |
| Mg | Amount of magnesium | Continuous | Relevant |
| Al | Amount of aluminium | Continuous | Relevant |
| Si | Amount of silicon | Continuous | Relevant |
| K | Amount of potassium | Continuous | Relevant |
| Ca | Amount of calcium | Continuous | Relevant |
| Ba | Amount of barium | Continuous | Relevant |
| Fe | Amount of iron | Continuous | Relevant |
| Type | Glass type | Discrete | Relevant |

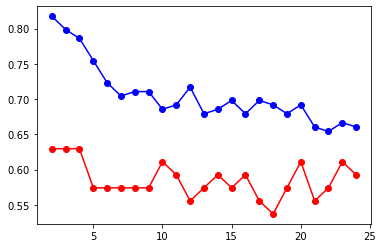
1. DATA CLEANSING :-
2. Dataset consists of 10 colums and 214 rows
3. Obtained duplicates rows and removed
4. All data types are of form int64 and float64
5. No null values found in each column
6. From describe function mean , median and standard deviation obtained
7. Outliers detected but retained
8. Variance for each column obtained
9. Scaling is done through normalization techniques
10. EDA:-
11. From pair plot analysis



1. Box plot for outliers



1. MODEL BUILDING:-
2. Splitting data to train part for 75% and test part for 25%
3. Model builded by KNeighbours classifiers
4. Test accuracy = 0.6296296296296297
5. Train accuracy = 0.7987421383647799
6. Tunnig the model to get optimum no for k
7. It is right fit model



1. BENEFITS :-

From above information we can assign the type of glass on basis of new dataset .

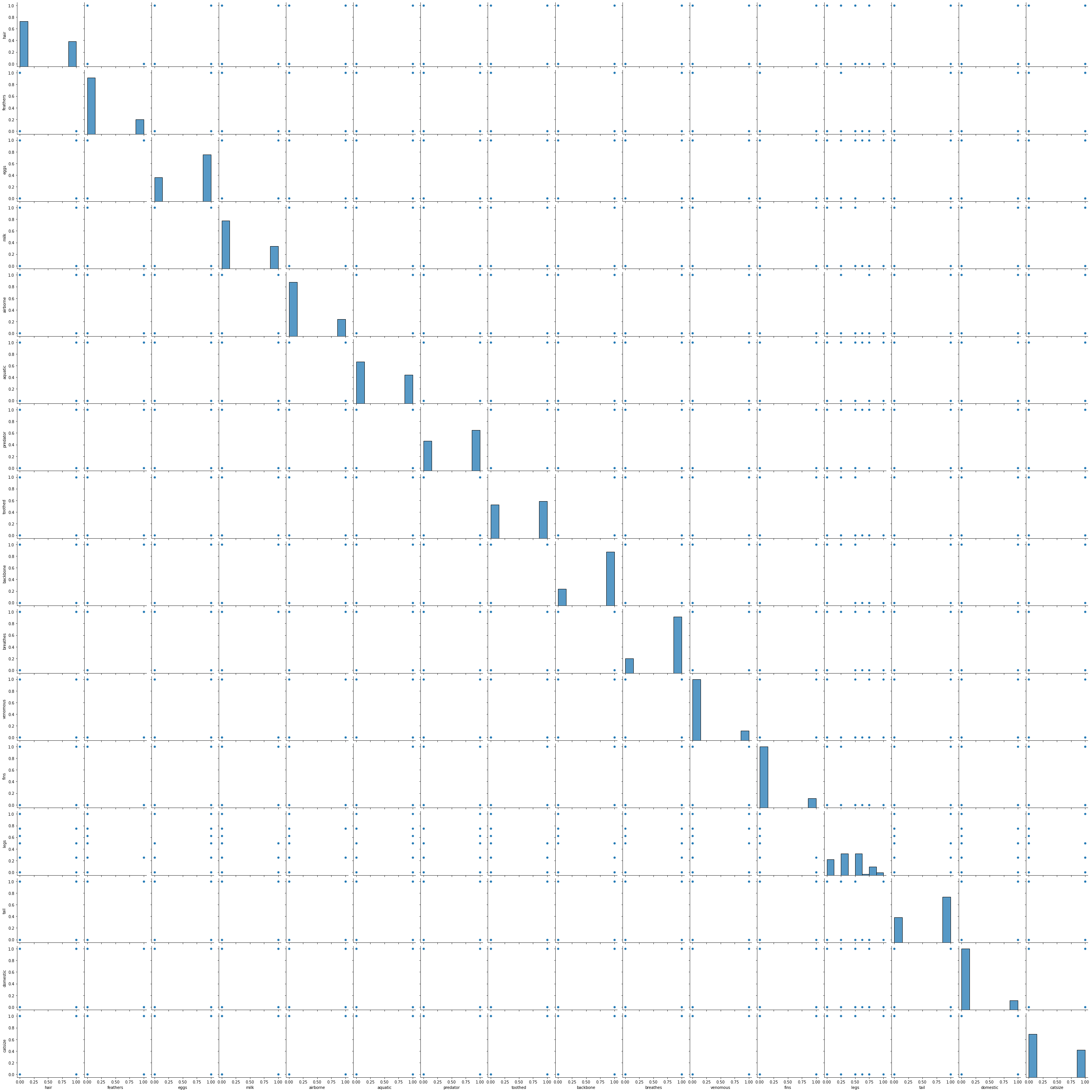
1. A National Park in India is dealing with the problem of segregation of its species based on the different attributes they have. Build a KNN model to automatically classify new species. Explain any inferences you draw in the documentation.
2. BUSINESS OBJECTIVE:-

Maximize mapping of animal type

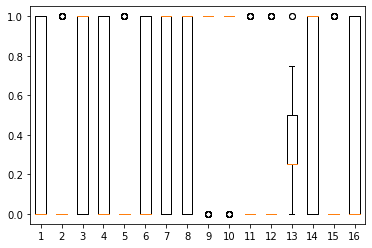
1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| animal name | Name of animal | Text ,strings | Not elevant |
| hair | Hair present or not | Binary | Relevant |
| feathers | Feathers present or not | Binary | Relevant |
| eggs | Gives egg or not | Binary | Relevant |
| milk | Gives milk or not | Binary | Relevant |
| airborne | Fly or not | Binary | Relevant |
| aquatic | Aquatic or not | Binary | Relevant |
| predator | Animals eats others | Binary | Relevant |
| toothed | Having teeth or not | Binary | Relevant |
| backbone | Having backbone or not | Binary | Relevant |
| breathes | Breathes or not | Binary | Relevant |
| venomous | Venomous or not | Binary | Relevant |
| fins | Fins or not | Binary | Relevant |
| legs | Having legs or not | Binary | Relevant |
| tail | Having tail or not | Binary | Relevant |
| domestic | Domestic or not | Binary | Relevant |
| catsize | Cat size or not | Binary | Relevant |
| type | Type of animal | Discrete | Relevant |

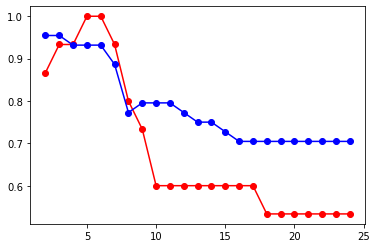
1. DATA CLEANSING :-
2. From dataset nominal column dropped
3. Dataset consists of 18 colums and 101rows
4. Obtained duplicates rows and removed
5. All data types are of form int64 and object
6. No null values found in each column
7. From describe function mean , median and standard deviation obtained
8. Outliers detected but retained
9. Variance for each column obtained
10. Scaling is done through normalization techniques
11. EDA:-
12. From pair plot analysis



1. Box plot for outliers



1. MODEL BUILDING:-
2. Splitting data to train part for 75% and test part for 25%
3. Model builded by KNeighbours classifiers
4. Test accuracy = 0.9333333333333333
5. Train accuracy = 0.9545454545454546
6. Tunnig the model to get optimum no for k
7. It is right fit model



1. BENEFITS :-

From above information we can assign the type of animals on basis of new dataset .