# Topic: K-Nearest Neighbor

**Instructions**

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name: Shahina Athar**

**Batch Id: 10122020**

**Topic: K-Nearest Neighbor**

1. **Business Problem**
   1. **Objective**
   2. **Constraints (if any)**
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, if not relevant provide reasons and provide description of the feature.**

**Using R and Python codes perform:**

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 N-neighbors**
   3. **Train and Test the data and perform cross validation techniques, compare accuracies, precision and recall and explain about them.**
   4. **Briefly explain the model output in the documentation.**

1. **Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.**

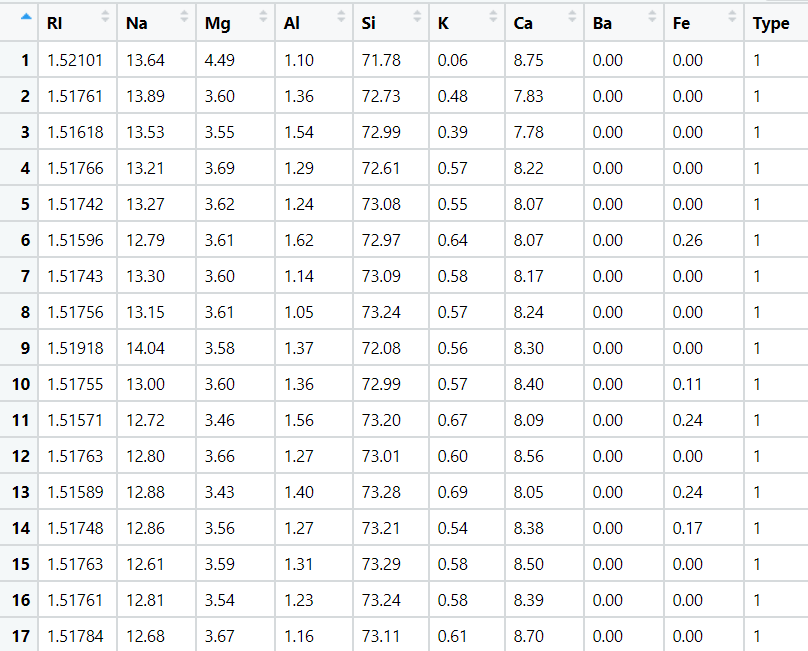
# Note:

The assignment should be submitted in the following format:

* R code
* Python code
* Code Modularization should be maintained
* Documentation of the model building (elaborating on steps mentioned above)

**Problem Statement: -**

A glass manufacturing plant, uses different Earth elements to design a new glass based on customer requirements for that they would like to automate the process of classification as it’s a tedious job to manually classify it, help the company reach its objective by correctly classifying the Earth elements, by using KNN Algorithm



* 1. **Train and Test the data and perform cross validation techniques, compare accuracies, precision and recall and explain about them.**

**Ans: This is MultiClass Classification Problems: Output has (1,2,3,5,6,7) category:**

**precision recall**

**1 0.71 0.86**

**2 0.69 0.60**

**3 0.00 0.00**

**5 0.33 0.33**

**6 0.00 0.00**

**7 0.78 0.88**

Precision : (True Positive/True Positive + False Positive)

Recall: TruePositives / (TruePositives + FalseNegatives)

On the basis of graph, we find that k=15 is the best k value (shortest distance) where, test accuracy = 67%, train accuracy = 66% they both are close enough.

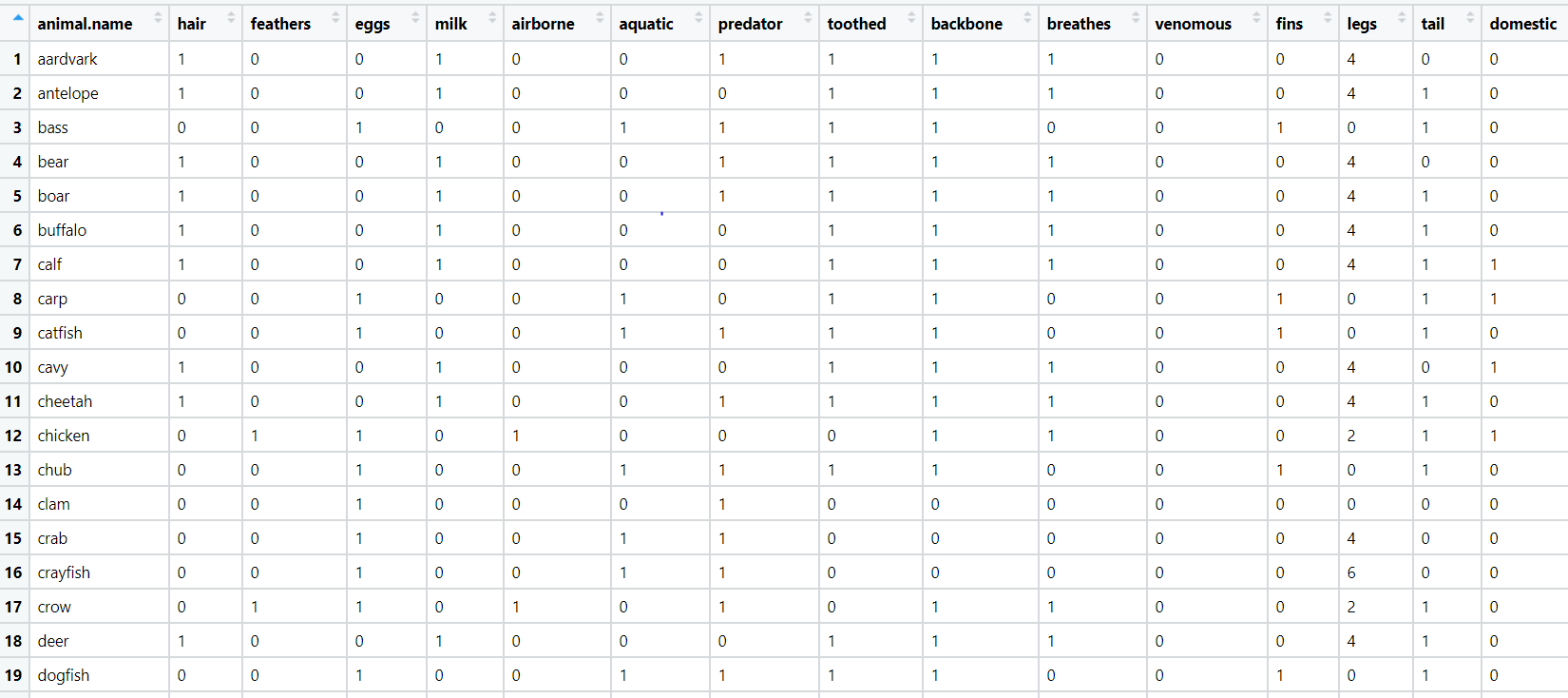
* 1. **Briefly explain the model output in the documentation.**

**Ans:** This problem is Multi Class Classification Problems and we train the algorithm so that we get confusion matrix, from the confusion matrix we can see Precision, recall as well as accuracy, and if we talk about accuracy On the basis of the graph we find that k=15 is the best k value (shortest distance) where, test accuracy = 67%, train accuracy = 66% they both are close enough.

1. **Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.**

**Ans:** After Prepared a Multi class classification model using K-Nearest Neighbour for glass manufacturing plant dataset, We can correctly classifying the Earth elements, By using KNN Algorithm with 67% of accuracy, because of this they can make the different designsof glass.

Problem Statement: -

A National Park, in India is dealing with a problem of segregation of its species based on the different attributes it has so that they can have cluster of species together rather than manually classify them, they have taken painstakingly collected the data and would like you to help them out with a classification model for their business objective to be achieved, by using KNN Algorithm classify the different species and draft your inferences in the documentation.

* 1. **Train and Test the data and perform cross validation techniques, compare accuracies, precision and recall and explain about them.**

**Ans: This is MultiClass Classification Problems: Output has (1,2,4,5,6,7) category:**

**precision recall**

**1 0.91 1.00**

**2 1.00 1.00**

**4 1.00 1.00**

**5 0.00 0.00**

**6 0.00 0.00**

**7 1.00 1.00**

Precision : (True Positive/True Positive + False Positive)

Recall: TruePositives / (TruePositives + FalseNegatives)

On the basis of the graph, we find that k=7 is the best k value (shortest distance) where, test accuracy = 90%, train accuracy = 87% they both are close enough.

* 1. **Briefly explain the model output in the documentation.**

**Ans:** This problem is Multi-Class Classification Problems and we train the algorithm so that we get confusion matrix, from the confusion matrix we can see Precision, recall as well as accuracy, and if we talk about accuracy On the basis of graph we find that k=7 is the best k value (shortest distance) where, test accuracy = 90%, train accuracy = 87% they both are close enough.

1. **Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.**

**Ans:** After Prepared a Multiclass classification model using K-Nearest Neighbour for National Park dataset, We can correctly classify the different species, By using KNN Algorithm with 90% of accuracy, because of this, they can easily segregate its species based on the different attributes.