# 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: Batch Id:**  **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.**

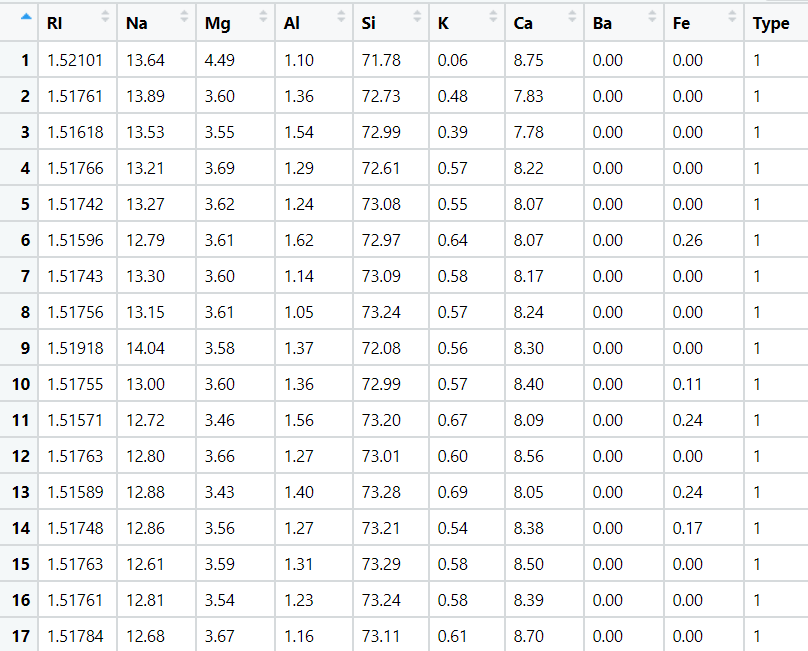
# Note:

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



**Business Problem**

**Objective :-** predict type of glass

**Python code details :**

Data Frame name is df. It has 214 rows and 10 columns.

**Work on each feature of the dataset to create a data dictionary as displayed in the below image:**

Then we create a data frame that’s contain details of each columns ,like- description ,data types ,and save the details named as data\_details .all of them are important .

**Data Pre-processing**

**Data Cleaning and Data Mining.**

Now we check info and describe for df .Check for data types ,unique value and variance .

Then we check for unique value in each columns

:-

RI 178

Na 142

Mg 94

Al 118

Si 133

K 65

Ca 143

Ba 34

Fe 32

Type 6

Dataframe has no missing values in columns .

We have done EDA for each columns and save the details as EDA. covariance for data set save as covariance . historgam and scatter plot for each column all data are normally distributed as well as we check for boxplot .

We have created dataframe with only with (discrete,continuous ,output) and Normalized data frame (considering the numerical part of data) , dummies for discrete data, creating a new df with both df named as X.

**"""**

**5. Model Building**

**5.1 Build the model on the scaled data (try multiple options)**

**5.2 Perform KNN, and use cross validation techniques to get N-neighbors**

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

**5.4 Briefly explain the model output in the documentation. """**

Now we split our data in X\_train, X\_test, Y\_train, Y\_test 75% data on train and 25% test . Preparing a KNN model on training data set with 7 neighbour,then test on test data , Evaluation on Test Data as result 61% are right prediction .

Crosstab

Predictions 1 2 3 5 7

Actual

1 17 3 0 0 0

2 6 10 0 0 1

3 1 0 0 0 0

5 0 2 0 0 1

6 2 1 1 1 0

7 0 1 0 1 6,

Evaluation on Train Dataalso as result 68% are right prediction .

Cross table

Predictions 1 2 3 5 6 7

Actual

1 43 6 1 0 0 0

2 15 43 0 1 0 0

3 10 4 2 0 0 0

5 0 4 0 3 0 3

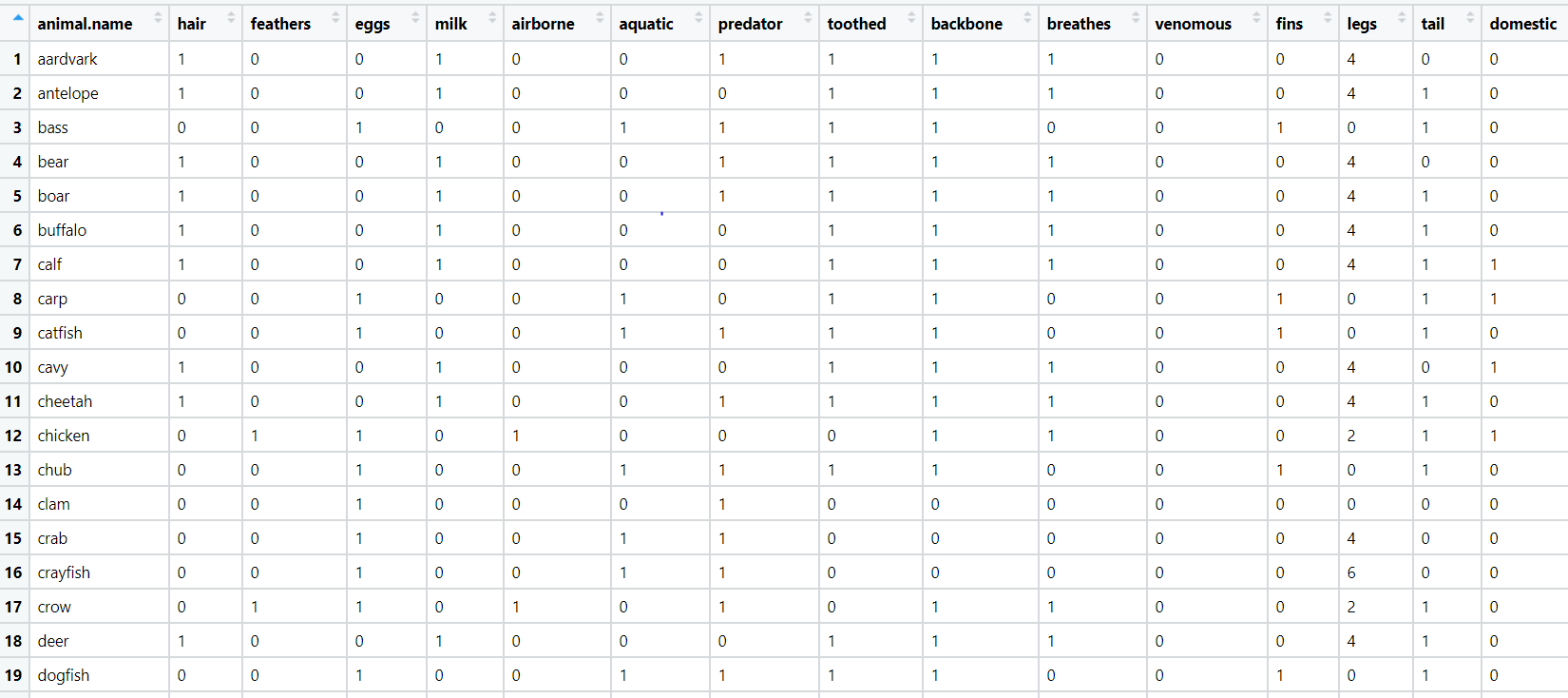
6 1 1 0 0 1 1

7 1 2 0 0 1 17

Then we run a model within a for loop with different k value from 3 to 17 only for odd numbers. And plot them for right fit model .

* used library :- pandas ,numpy , networkx, sk learn ,matplotlib

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.

**Business Problem**

**Objective :-** predict type of animal

**Python code details :**

Data Frame name is df. It has 214 rows and 10 columns.

**Work on each feature of the dataset to create a data dictionary as displayed in the below image:**

Then we create a data frame that’s contain details of each columns ,like- description ,data types ,and save the details named as data\_details .all of them are important .

**Data Pre-processing**

**Data Cleaning and Data Mining.**

Now we check info and describe for df .we found that animal name in each row is different so we removed that column. Check for data types ,unique value and variance .

Then we check for unique value in each columns

:-

animal name 100

hair 2

feathers 2

eggs 2

milk 2

airborne 2

aquatic 2

predator 2

toothed 2

backbone 2

breathes 2

venomous 2

fins 2

legs 6

tail 2

domestic 2

catsize 2

type 7

Dataframe has no missing values in columns .

We have done EDA for each columns and save the details as EDA. covariance for data set save as covariance . historgam and scatter plot for each column all data are normally distributed as well as we check for boxplot .

We have created dataframe with only with (discrete,continuous ,output) and Normalized data frame (considering the numerical part of data) , dummies for discrete data, creating a new df with both df named as X.

**"""**

**5. Model Building**

**5.1 Build the model on the scaled data (try multiple options)**

**5.2 Perform KNN, and use cross validation techniques to get N-neighbors**

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

**5.4 Briefly explain the model output in the documentation. """**

Now we split our data in X\_train, X\_test, Y\_train, Y\_test 75% data on train and 25% test . Preparing a KNN model on training data set with 7 neighbour,then test on test data , Evaluation on Test Data as result 80% are right prediction .

Crosstab

Predictions 1 2 4 6 7

Actual

1 9 0 2 0 0

2 0 5 0 0 0

4 0 0 4 0 0

6 0 0 0 2 0

7 0 0 0 3 1 Evaluation on Train Dataalso as result 88% are right prediction .

Cross table

Predictions 1 2 3 4 5 6 7

Actual

1 29 0 0 1 0 0 0

2 0 15 0 0 0 0 0

3 1 0 1 2 1 0 0

4 0 0 0 9 0 0 0

5 0 0 0 0 4 0 0

6 0 0 0 0 0 6 0

7 0 0 0 2 1 1 2

Then we run a model within a for loop with different k value from 3 to 17 only for odd numbers. And plot them for right fit model .