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: Kavali Rakesh

Batch Id: DSWDMCON 18012022

Topic: K-Nearest Neighbor

1. Business Problem

1.1. Objective

1.2. Constraints (if any)

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

Name of Feature	Description	Type	Relevance
ID	Customer ID	Quantitative, Nominal	Irrelevant, ID does not provide useful information

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:

3. Data Pre-processing

3.1 Data Cleaning, Feature Engineering, etc.

4. Exploratory Data Analysis (EDA):

4.1. Summary

4.2. Univariate analysis

4.3. Bivariate analysis

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.

6. 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

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Type
1	1.52101	13.64	4.49	1.10	71.78	0.06	8.75	0.00	0.00	1
2	1.51761	13.89	3.60	1.36	72.73	0.48	7.83	0.00	0.00	1
3	1.51618	13.53	3.55	1.54	72.99	0.39	7.78	0.00	0.00	1
4	1.51766	13.21	3.69	1.29	72.61	0.57	8.22	0.00	0.00	1
5	1.51742	13.27	3.62	1.24	73.08	0.55	8.07	0.00	0.00	1
6	1.51596	12.79	3.61	1.62	72.97	0.64	8.07	0.00	0.26	1
7	1.51743	13.30	3.60	1.14	73.09	0.58	8.17	0.00	0.00	1
8	1.51756	13.15	3.61	1.05	73.24	0.57	8.24	0.00	0.00	1
9	1.51918	14.04	3.58	1.37	72.08	0.56	8.30	0.00	0.00	1
10	1.51755	13.00	3.60	1.36	72.99	0.57	8.40	0.00	0.11	1
11	1.51571	12.72	3.46	1.56	73.20	0.67	8.09	0.00	0.24	1
12	1.51763	12.80	3.66	1.27	73.01	0.60	8.56	0.00	0.00	1
13	1.51589	12.88	3.43	1.40	73.28	0.69	8.05	0.00	0.24	1
14	1.51748	12.86	3.56	1.27	73.21	0.54	8.38	0.00	0.17	1
15	1.51763	12.61	3.59	1.31	73.29	0.58	8.50	0.00	0.00	1
16	1.51761	12.81	3.54	1.23	73.24	0.58	8.39	0.00	0.00	1
17	1.51784	12.68	3.67	1.16	73.11	0.61	8.70	0.00	0.00	1

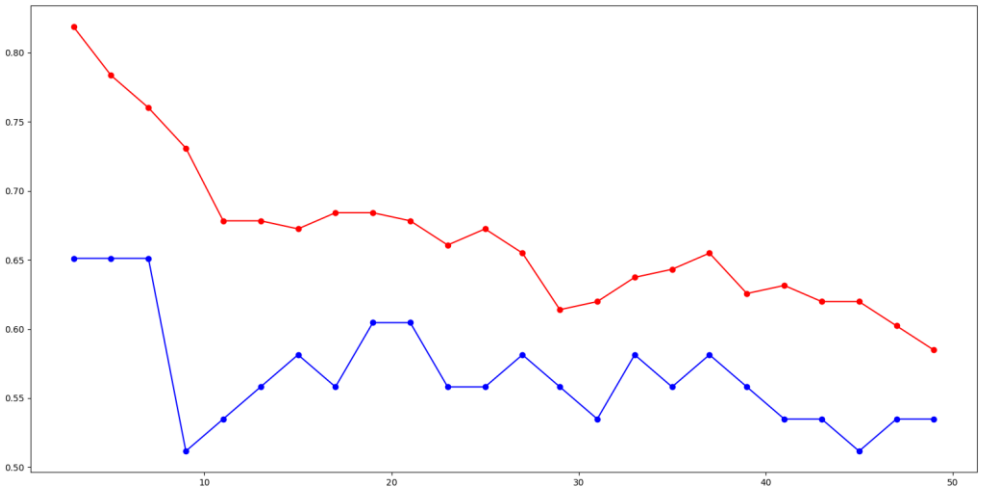
Sol:
Business Objective: To identify the new type of glass to which type by using the KNN model.

Constraints: Lack of analysis of the glass data.

Data Type: the given data is the chemical composition percentage of certain chemical and the type of the glass.

Data Pre-Processing: After summarizing the complete data I got to know that the normalization of the data has to be done because different variable have the different values so the data can be used after doing the normalization only.

KNN model for the data set: I have done the KNN model for the given data set by splitting the complete data into training and testing data as 70% and 30% of the original data in Python.
I have made the model with a 70% Accuracy approximately to locate a new glass to the given data. The below graph shows the witness of K value which is the output of Python



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.

	animal.name	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins	legs	tail	domestic
1	aardvark	1	0	0	1	0	0	1	1	1	1	0	0	4	0	0
2	antelope	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0
3	bass	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0
4	bear	1	0	0	1	0	0	1	1	1	1	0	0	4	0	0
5	boar	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0
6	buffalo	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0
7	calf	1	0	0	1	0	0	0	1	1	1	0	0	4	1	1
8	carp	0	0	1	0	0	1	0	1	1	0	0	1	0	1	1
9	catfish	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0
10	cavy	1	0	0	1	0	0	0	1	1	1	0	0	4	0	1
11	cheetah	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0
12	chicken	0	1	1	0	1	0	0	0	1	1	0	0	2	1	1
13	chub	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0
14	clam	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
15	crab	0	0	1	0	0	1	1	0	0	0	0	0	4	0	0
16	crayfish	0	0	1	0	0	1	1	0	0	0	0	0	6	0	0
17	crow	0	1	1	0	1	0	1	0	1	1	0	0	2	1	0
18	deer	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0
19	dogfish	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0

Sol:

Business Objective: To identify the new type of animal to which category by using the KNN model.

Constraints: Lack of analysis of the animals data.

Data Type: the given data is the type of animal and the characteristics of the animal with ratings to their respective characteristic properties. All the data is used for the analysis except animal name.

Data Pre-Processing: After summarizing the complete data I got to know that the normalization of the data has to be done because different variable have the different values so the data can be used after doing the normalization only.

KNN model for the data set: I have done the KNN model for the given data set by splitting the complete data into training and testing data as 70% and 30% of the original data in Python.
I have made the model with a 90% Accuracy approximately to locate a new animal to the given data. The below graph shows the witness of K value which is the output of Python