# K-Nearest Neighbors

**Instructions:**

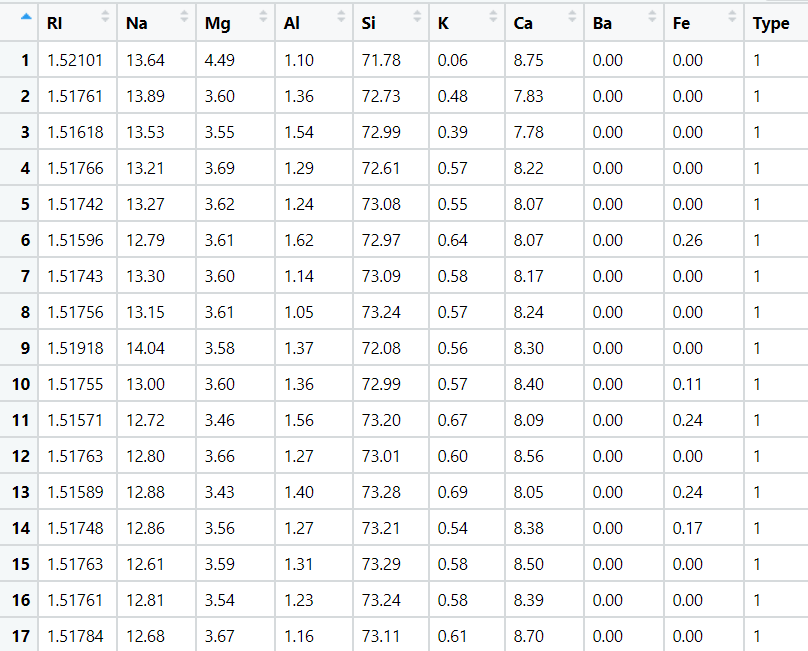
Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: Prajay B. Urkude Batch ID: 16092021**

**Topic: K-Nearest Neighbors**

1. 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.

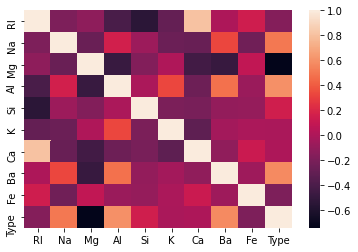


Ans: - Business objective:-

To classify the glass type based on the different features of the glass.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Features** | **Description** | **Type** | **Relevance** |
| RI | Refractive Index | Quantitative, Ratio | Relevant |
| Na | Sodium | Quantitative, Ratio | Relevant |
| Mg | Magnesium | Quantitative, Ratio | Relevant |
| Ai | Aluminum | Quantitative, Ratio | Relevant |
| Si | Silicon | Quantitative, Ratio | Relevant |
| K | Potassium | Quantitative, Ratio | Relevant |
| Ca | Calcium | Quantitative, Ratio | Relevant |
| ba | Barium | Quantitative, Ratio | Relevant |
| Fe | Iron | Quantitative, Ratio | Relevant |
| Type | Type of glass | Quantitative, Ratio | Relevant |

* Importing the libraries : pandas, numpy, matplotlib, seaborn , sklearn .
* From sklearn we import KNeighborsClassifier function and from sklearn.model\_selection package we import train\_test\_split function.
* Loading the datasets and checking for the null values, duplicate values, dropping the unwanted columns for analysis , exploratory data analysis .
* Data Exploration and visualization.

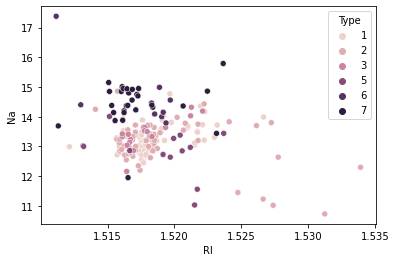


We can notice that Ca and K values don't affect Type that much.

Also Ca and RI are highly correlated, this means using only RI is enough.

So we can go ahead and drop Ca and also K (will perform later)

* Scatter plot of two features, and pairwise plot



Suppose we consider only RI, and Na values for classification for glass type.

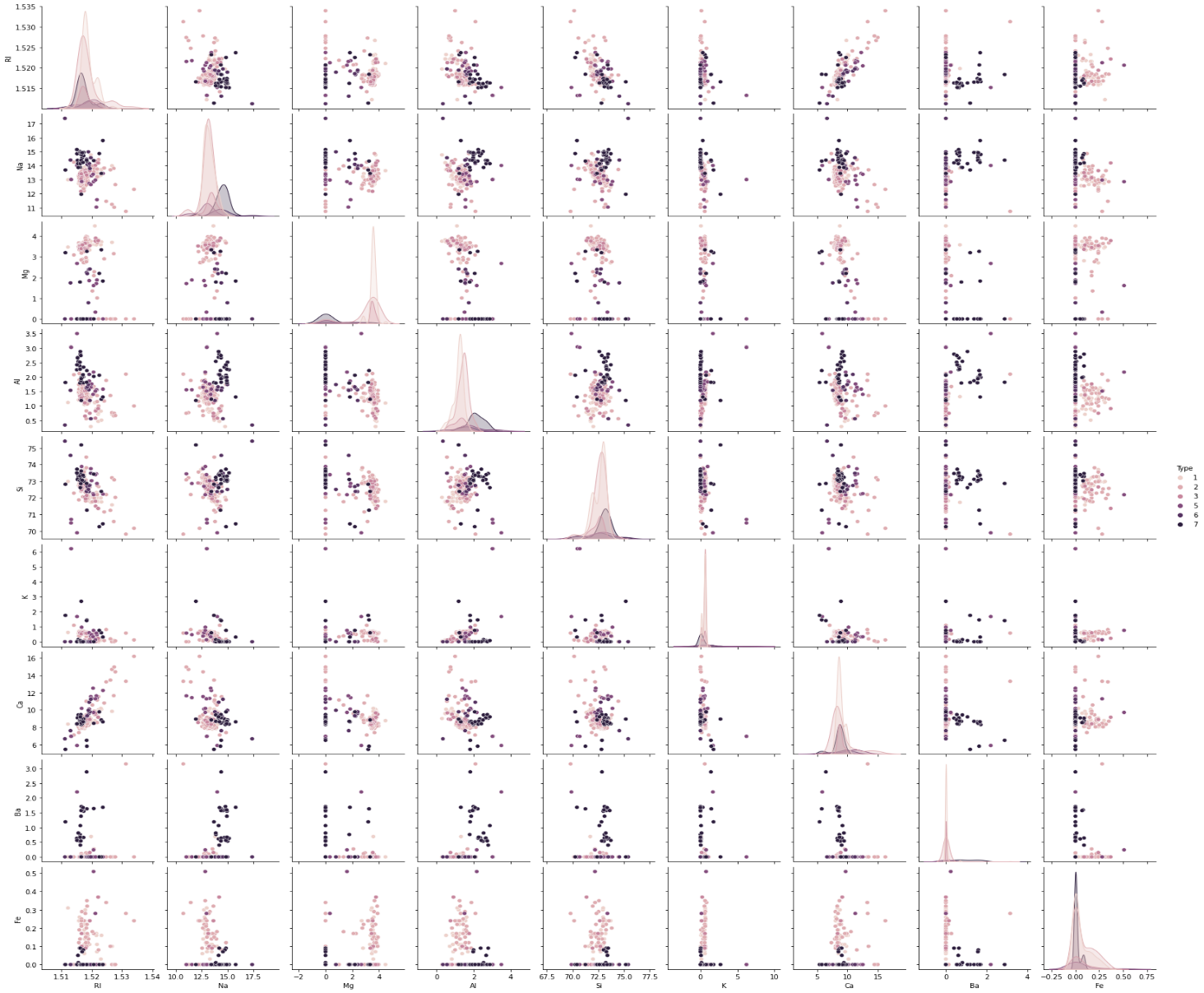
From the above plot, We first calculate the nearest neighbors from the new data point to be calculated.

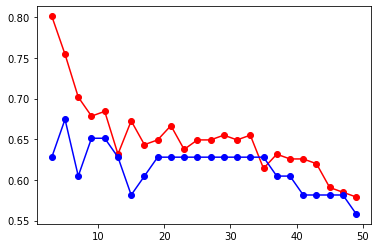
If the majority of nearest neighbors belong to a particular class, say type 4, then we classify the data point as type 4.

But there are a lot more than two features based on which we can classify. So let us take a look at pairwise plot to capture all the features.

* Pairwise plot of all the features

The pair plot shows that the data is not linear and KNN can be applied to get nearest neighbors and classify the glass types

*  Normalization of the data to bring the magnitudes of the all the features on the same scale between 0 to 1.
* Splitting the data in 80:20 ratio, Train: Test
* Apply the KNN model by taking K=21 and fit the model in the test dataset ans check the accuracy. The accuracy of the model is found out = 65%
* Fit the model in the training datasets to check the error and the accuracy found out = 75%.
* The accuracy error is more and the model is underfit so we check the model for different K values.
* For That we iterate the model by taking the the K values between 3 to 50 in the span of 2 and created the accuracy values for each K value and plot the graph for test accuracy and train data accuracy.
* From the graph take the best value of K which gives the good accuracy and less error.



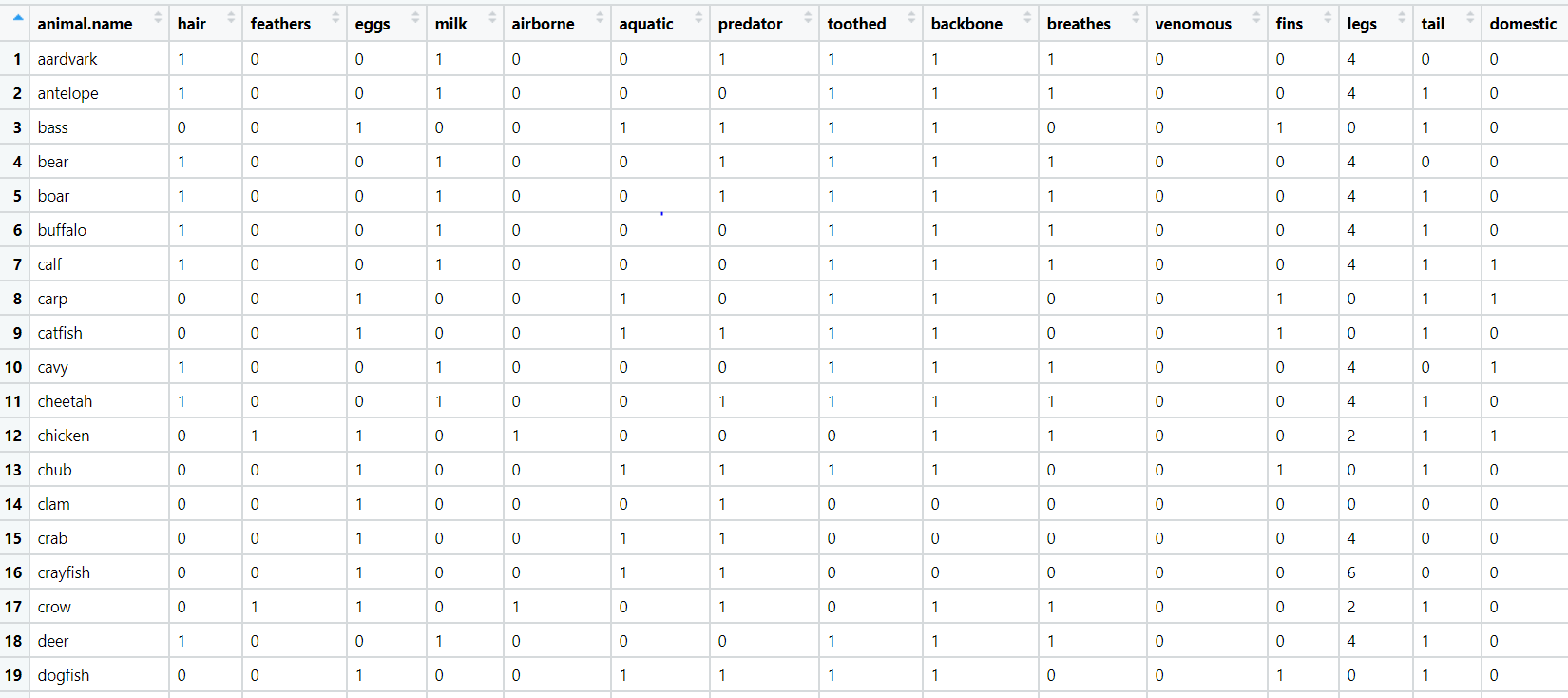
From the above plot, at K= 6 we get the accuracy up to 73% and error up to 2 %.

Inferences: -

The best k value was found to be 6.

Dropping ‘Ca’ produced better results by a bit, ‘K’ feature did not affect results in any way.

Also, we noticed that RI and Ca are highly correlated, this makes sense as it was found that the Refractive index of glass was found to increase with the increase in Cao.

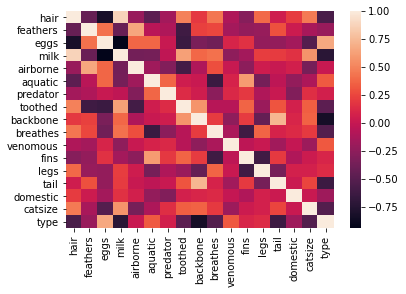
1. A National Zoopark in India is dealing with the problem of segregation of the animals based on the different attributes they have. Build a KNN model to automatically classify the animals. Explain any inferences you draw in the document

Ans: - Business objective:-

To classify the animals based on the different features has the animals.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Features** | **Description** | **Type** | **Relevance** |
| Animal name | Name of animal | Quantitative, Nominal | Relevant |
| Hair | Does animal has hair or not | Quantitative, Nominal | Relevant |
| features | Does animal has feathers or not | Quantitative, Nominal | Relevant |
| Eggs | Does animal gives eggs or not | Quantitative, Nominal | Relevant |
| Milk | Does animal give milk or not | Quantitative, Nominal | Relevant |
| Airborne | Does animal airborne or not | Quantitative, Nominal | Relevant |
| Aquatic | Does animal aquatic or not | Quantitative, Nominal | Relevant |
| Predator | Does animal predator or not | Quantitative, Nominal | Relevant |
| toothed | Does animal has teeth or not | Quantitative, Nominal | Relevant |
| Backbone | Does animal has backbone or not | Quantitative, Nominal | Relevant |
| Breathed | Does animal has hair or not | Quantitative, Nominal | Relevant |
| venomous | Does animal venomous or not | Quantitative, Nominal | Relevant |
| Fins | Does animal has fins or not | Quantitative, Nominal | Relevant |
| Legs | Does animal has legs or not | Quantitative, Nominal | Relevant |
| Tail | Does animal has tail or not | Quantitative, Nominal | Relevant |
| domestic | Does animal domestic or not | Quantitative, Nominal | Relevant |
| Cat size | Cat size of the animal | Quantitative, Nominal | Relevant |
| Type | Size type of animal | Quantitative, Nominal | Relevant |

* Importing the libraries : pandas, numpy, matplotlib, seaborn , sklearn .
* From sklearn we import KNeighborsClassifier function and from sklearn.model\_selection package we import train\_test\_split function.
* Loading the datasets and checking for the null values, duplicate values, dropping the unwanted columns for analysis , exploratory data analysis .
* Data Exploration and visualization.



We can notice that hair and aquatic values don't affect Type that much.

Also hair and milk are highly correlated, this means using only milk is enough.

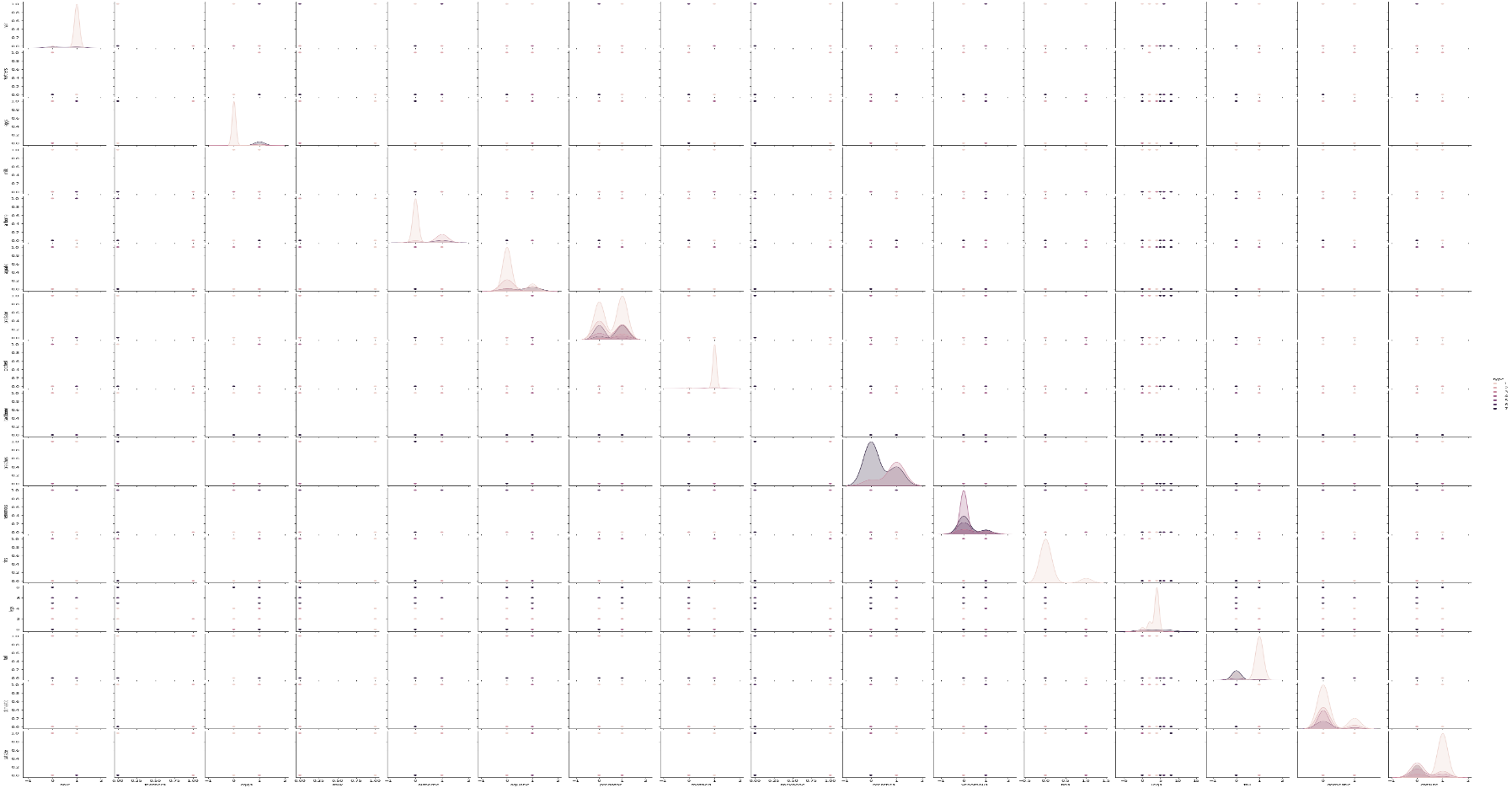
So we can go ahead and drop hair and also (will perform later)

* Scatter plot of two features, and pairwise plot

But there are a lot more than two features based on which we can classify. So let us take a look at pairwise plot to capture all the features.

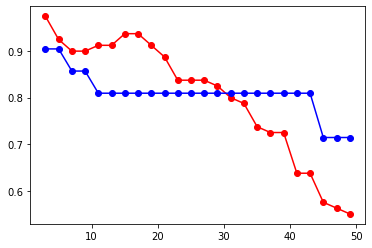
* Pairwise plot of all the features

The pair plot shows that the data is not linear and KNN can be applied to get nearest neighbors and classify the glass types



The pair plot shows that the data is not linear and KNN can be applied to get nearest neighbors and classify the glass types

* Normalization of the data to bring the magnitudes of the all the features on the same scale between 0 to 1.
* Splitting the data in 80:20 ratio, Train: Test
* Apply the KNN model by taking K=21 and fit the model in the test dataset ans check the accuracy. The accuracy of the model is found out = 80%
* Fit the model in the training datasets to check the error and the accuracy found out = 88%.
* The accuracy error is more and the model is underfit so we check the model for different K values.
* For That we iterate the model by taking the the K values between 3 to 50 in the span of 2 and created the accuracy values for each K value and plot the graph for test accuracy and train data accuracy.
* From the graph take the best value of K which gives the good accuracy and less error.



From the above plot, at K= 5 we get the accuracy up to 90% and error up to 2 %.

The best k value was found to be 5.

Also, we noticed that hair and milk are highly correlated, this makes sense as it was found that as the animal has hair that gives milk.

