Problem Statement:

Write a program to decide the variety of grain-based on the different parameters of the data. (Attached herewith a dataset).

I received the data file containing the data for determining the wheat seed variety , the dataset was fairly small containing only 43 records and 7columns. We need to construct a model that will define which wheat grain is on the basis of the given parameters.

I started with the preprocessing of the given data, first visualizing the data, then determining the correlation between different columns one column of asymmetry coefficient which had very low -ve correlation but since the data was very small I decided not to change anything, next I checked for the null values there were not present any. The grain variety column had object as data, I converted those in numeric values.

Then I started with building the logistic regression model by taking all columns except the grain variety, after building it LR model gave an accuracy of 0.904762.

Then I built the naive Bayes model , it gave an accuracy of 0.976190.

Then I built the decision tree model, it gave an accuracy of 0.950000.

Then, SVM gave an accuracy of 0.931034

Then, KNN classifier gave an accuracy of 0.905836

Comparing these all scores we can come to the conclusion that we can use the NB Gaussian model as it gave the highest accuracy of 0.976190 among all the models.now for the given parameters we can define what type of wheat grain is it from Kama wheat, Rosa wheat, and Canadian wheat.