Question 4

Introduction

In this question we were given a kaggle challenge named whats cooking. Here the data was about the different cuisine in the world and ingredients used in each cuisine.

So now the data was as a json file which was converted to np array. For data preprocessing, the rows were 0 or 1, 0-> ingredient not present and 1-> ingredient present and a label vector. Here models are taken from sklearn.

Best performing models:

1) **Multinomial Naive Bayes: Acc = 73.94%**

2) **Decision tree: Acc = 61.75%**

Other models tried:

1) Bernoulli Naive Bayes : 71.1%
2) Gaussian Naive Bayes : 40.58%
3) Knn-200 neighbours : 43.93%

Here from the data we can easily interpret that the number of attributes are very high for knn to be effective as knn has curse of dimensionality. Now as the data was yes no type for each cuisine the naive bayes was a good choice and it performs the best.