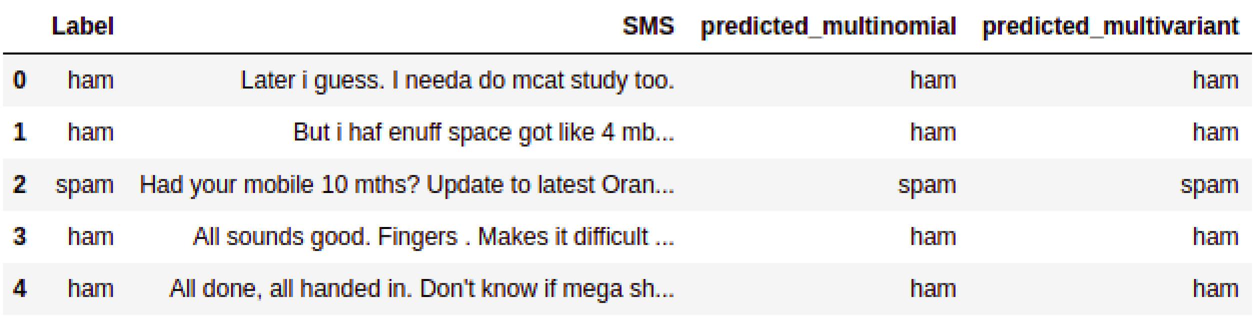
DOCUMENTATION FOR ARTIFICIAL INTELLIGENCE ASSIGNMENT-3

Naive Bayes classifier for Spam filtering-

OUTPUT

Probability of spam :	13.46%
Probability of ham :	86.54%
Number of spam messages(word) is :	15190
Number of ham message(word) is :	57237
Number of unique word are :	0.13458950201884254

Sample output



	Label	SMS	predicted_multinomial	predicted_multivariant
0	ham	Later i guess. I needa do mcat study too.	ham	ham
1	ham	But i haf enuff space got like 4 mb...	ham	ham
2	spam	Had your mobile 10 mths? Update to latest Oran...	spam	spam
3	ham	All sounds good. Fingers . Makes it difficult ...	ham	ham
4	ham	All done, all handed in. Don't know if mega sh...	ham	ham

Prediction using Multinomial

Correct	1100
Incorrect	14
Accuracy	98.74326750448833%

Prediction using Multivariant

Correct	1094
Incorrect	20
Accuracy	98.20466786355476%

Naive Bayes Classifier-

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. Ex- Consider a fictional dataset that describes the weather conditions for playing a game of golf. Given the weather conditions, each tuple classifies the conditions as fit("Yes") or unfit("No") for plaing golf.

-> Assumption behind Naive Bayes Classifier-
The fundamental Naive Bayes assumption is that each feature makes an:

1.independent

2.equal

Bayes Theorem-

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

P(A) - Prior Probability

P(B/A) – Posterior Probability

P(B) – Evidence

All variants of Naive Bayes Classifier-

1. **Bernoulli Naive Bayes:** In the multivariate Bernoulli event model, features are independent booleans (binary variables) describing inputs. Like the multinomial model, this model is popular for document classification tasks, where binary term occurrence(i.e. a word occurs in a document or not) features are used rather than term frequencies(i.e. frequency of a word in the document).
2**. Multinomial Naive Bayes**: Feature vectors represent the frequencies with which certain events have been generated by a multinomial distribution. This is the event model typically used for document classification.
3. **Gaussian Naive Bayes classifier**: In Gaussian Naive Bayes, continuous values associated with each feature are assumed to be distributed according to a **Gaussian distribution**. A Gaussian distribution is also called **Normal distribution**. When plotted, it gives a bell shaped curve which is symmetric about the mean of the feature values.

Time Complexity

- **Training Time**: O(|D|L d + |C||V|))

where L d is the average length of a document in D

-> Assumes V and all D i , n i , and n ij pre-computed in O(|D|L d) time during one pass through all of the data.

-> Generally just O(|D|L d) since usually |C||V| < <|D|L d.

-> |C| |V| = Complexity of computing all probability values (loop over terms and classes).

■ **Test Time:** O(|C| L t)

where L t is the average length of a test document

-> Very efficient overall, linearly proportional to the time needed to just read in all the data.

Multinomial vs Multivariate Bernoulli

1. Multinomial model is almost always more effective in text applications!

2. While classifying a test document

-> Bernoulli model uses binary occurrence information, ignoring the number of occurrences.

-> Multinomial model keeps track of multiple occurrences

-> Bernoulli makes many mistakes while classifying long documents (as it ignores counts).