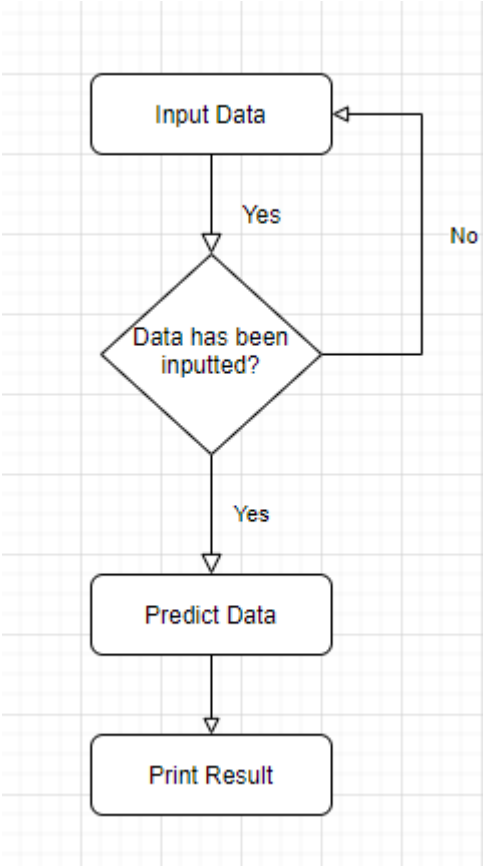
Muhammad Ariq Naufal

Dictionary / Corpus that been used:

1. Multinomial Naive Bayes
2. Logistic Regression

Flowchart

First you need input the data like the write comment or news title and the data that has been inputted will be printed the result of predicted input data.



I tested the data using 3 supervised learning method:

1. Multinomial Naive Bayes
2. Logistic Regression

SPAM FILTER CASE

In the spam filter case, I am using Multinomial Naive Bayes method to predict the spam filter, because it has better prediction like in the example below.

Naive bayes predict spam better:

```
In [112]: 1 test_comment('sdadasfdasffasda')
Out[112]: array([1], dtype=int64)
```

While Random Forest and Logistic Regression predict it is not a spam:

Logistic Regression

```
In [37]: 1 test_comment('dasdasfsadfasfaf')
Out[37]: array([0], dtype=int64)
```

The accuracy score shows the logistic regression has better score than naive bayes. But since the naive bayes can predict spam better I choose the naive bayes to predict the mail message.

NEWS CASE

For the news case i choose the logistic regression to predict the category of the news. Because the logistic regression have better accuracy score than naive bayes

Logisitic Regression:

	precision	recall	f1-score	support
Business	0.90	0.90	0.90	5343
Entertainment	0.95	0.97	0.96	7241
Medical	0.94	0.86	0.90	2138
Technology	0.91	0.90	0.91	4939
avg / total	0.92	0.92	0.92	19661

Naive Bayes:

	precision	recall	f1-score	support
Business	0.89	0.88	0.88	5343
Entertainment	0.95	0.96	0.95	7241
Medical	0.91	0.86	0.88	2138
Technology	0.87	0.89	0.88	4939
avg / total	0.91	0.91	0.91	19661