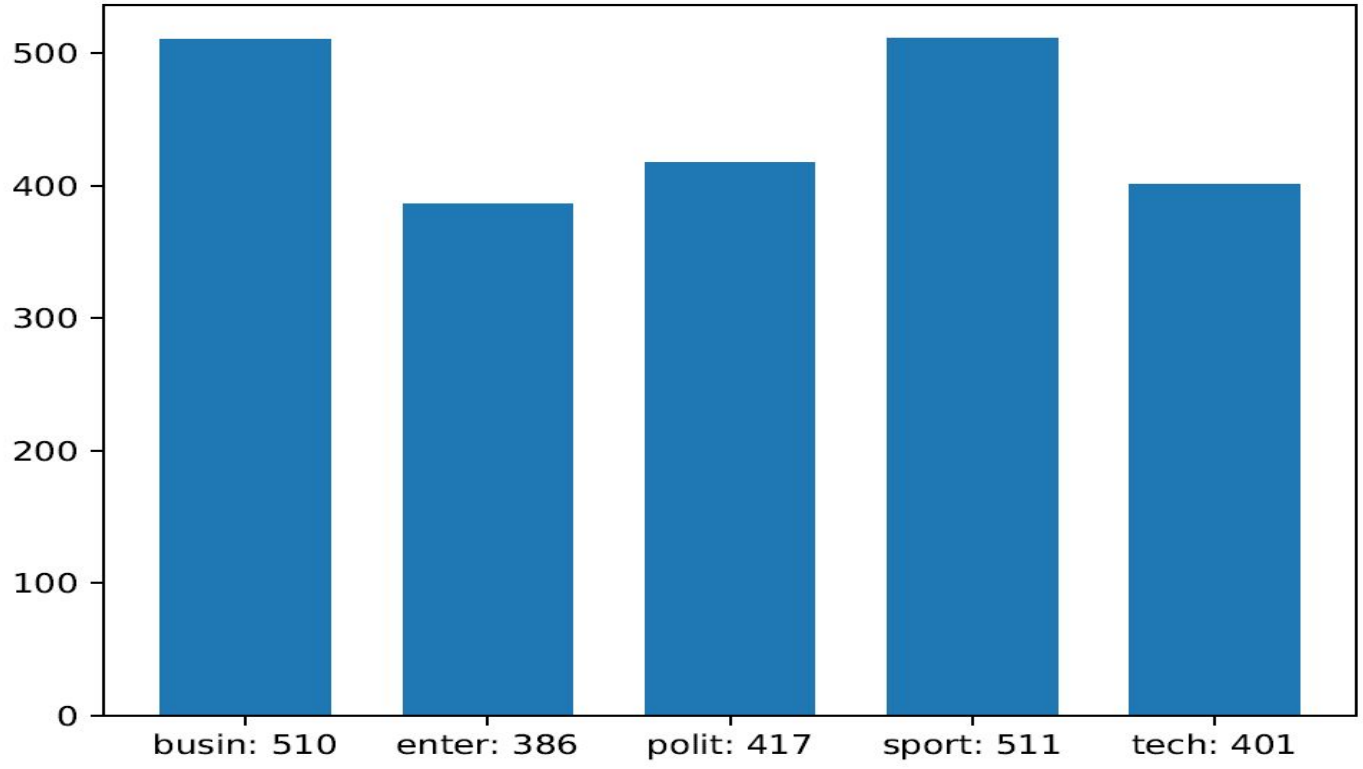
# COMP 472 Project 1

Kimchheng heng 26809413

# Link to GitHub

https://github.com/kimchhengheng/Comp472MiniProj1

# BBC class distribution



# Multinomial Naive Bayes Classifier

- The size of vocabulary is not fixed, it changes every time we run the code
- By running the each model consecutively, the number of feature would be the same
- However the result from try 1 and try 2 is exactly the same
- The model with smoothing, 0.001 and 0.9, has the same result

# Try 1

The confusion matrix

| | | | | |
|---|---|---|---|---|
| 94 | 1 | 0 | 0 | 4 |
| 1 | 72 | 3 | 0 | 2 |
| 3 | 0 | 70 | 0 | 1 |
| 0 | 0 | 0 | 109 | 0 |
| 0 | 0 | 0 | 0 | 85 |

# Try 2

The confusion matrix

| | | | | |
|---|---|---|---|---|
| 94 | 1 | 0 | 0 | 4 |
| 1 | 72 | 3 | 0 | 2 |
| 3 | 0 | 70 | 0 | 1 |
| 0 | 0 | 0 | 109 | 0 |
| 0 | 0 | 0 | 0 | 85 |

# Try 3

The confusion matrix

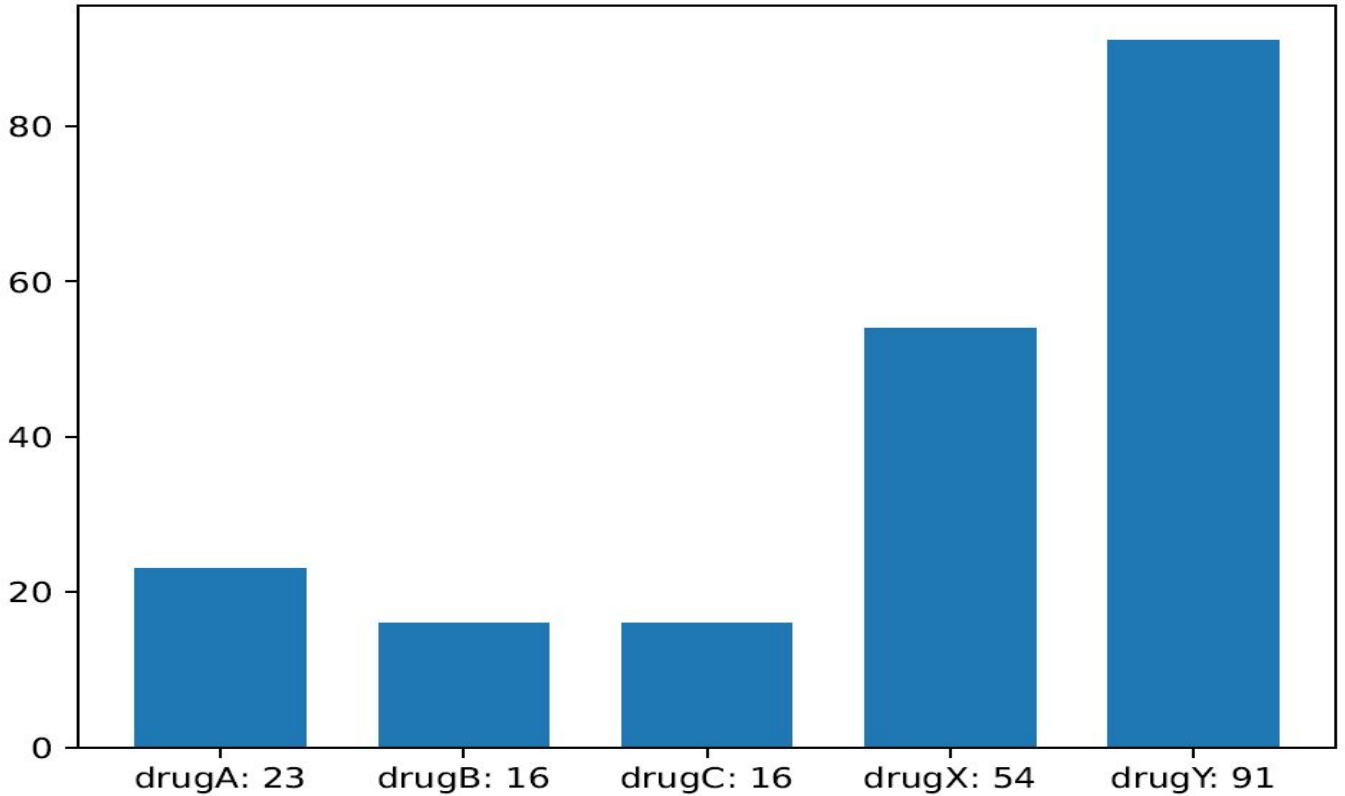
| | | | | |
|---|---|---|---|---|
| 92 | 1 | 2 | 0 | 4 |
| 0 | 76 | 2 | 0 | 0 |
| 2 | 0 | 71 | 0 | 1 |
| 0 | 0 | 0 | 109 | 0 |
| 1 | 0 | 0 | 0 | 84 |

# Try 4

The confusion matrix

| | | | | |
|---|---|---|---|---|
| 92 | 1 | 2 | 0 | 4 |
| 0 | 76 | 2 | 0 | 0 |
| 2 | 0 | 71 | 0 | 1 |
| 0 | 0 | 0 | 109 | 0 |
| 1 | 0 | 0 | 0 | 84 |

# Drug distribution



# NB model

- The class is not distribute well, so the NB model does not provide a good result

NB

The confusion matrix

| | | | | |
|---|---|---|---|---|
| 1 | 0 | 0 | 2 | 1 |
| 0 | 1 | 0 | 3 | 0 |
| 0 | 0 | 0 | 4 | 1 |
| 0 | 0 | 0 | 6 | 5 |
| 0 | 0 | 0 | 4 | 22 |

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 0.25 | 0.40 | 4 |
| 1 | 1.00 | 0.25 | 0.40 | 4 |
| 2 | 0.00 | 0.00 | 0.00 | 5 |
| 3 | 0.32 | 0.55 | 0.40 | 11 |
| 4 | 0.76 | 0.85 | 0.80 | 26 |
| accuracy | | | 0.60 | 50 |
| macro avg | 0.61 | 0.38 | 0.40 | 50 |
| weighted avg | 0.62 | 0.60 | 0.57 | 50 |

# Base-DT

Base DT has the highest performance, I believe because the class is not well distribution. So there is an bias between class. And the test and train just split randomly, so there is a probability that the test set would work very because of the train set.

Base-DT

The confusion matrix

| | | | | |
|---|---|---|---|---|
| 4 | 0 | 0 | 0 | 0 |
| 1 | 3 | 0 | 0 | 0 |
| 0 | 0 | 5 | 0 | 0 |
| 0 | 0 | 0 | 11 | 0 |
| 0 | 0 | 0 | 0 | 26 |

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.80 | 1.00 | 0.89 | 4 |
| 1 | 1.00 | 0.75 | 0.86 | 4 |
| 2 | 1.00 | 1.00 | 1.00 | 5 |
| 3 | 1.00 | 1.00 | 1.00 | 11 |
| 4 | 1.00 | 1.00 | 1.00 | 26 |
| accuracy | | | 0.98 | 50 |
| macro avg | 0.96 | 0.95 | 0.95 | 50 |
| weighted avg | 0.98 | 0.98 | 0.98 | 50 |

# Top-DT

- The second highest score is the Top DT.

Top-DT

The confusion matrix

| | | | | |
|---|---|---|---|---|
| 4 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 5 | 0 |
| 0 | 0 | 0 | 11 | 0 |
| 0 | 0 | 0 | 0 | 26 |

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.50 | 1.00 | 0.67 | 4 |
| 1 | 0.00 | 0.00 | 0.00 | 4 |
| 2 | 0.00 | 0.00 | 0.00 | 5 |
| 3 | 0.69 | 1.00 | 0.81 | 11 |
| 4 | 1.00 | 1.00 | 1.00 | 26 |
| accuracy | | | 0.82 | 50 |
| macro avg | 0.44 | 0.60 | 0.50 | 50 |
| weighted avg | 0.71 | 0.82 | 0.75 | 50 |

# Compare iteration vs Cross validation

NB

average accuracy 0.6

average macro-average F1 0.4

average weighted-average F1 0.568

standard deviation accuracy 0.0

standard deviation macro-average F1 0.0

standard deviation weighted-average F1 0.0

NB

average accuracy 0.6

average macro-average F1 0.3804

average weighted-average F1 0.5435

standard deviation accuracy 0.0671

standard deviation macro-average F1 0.1402

standard deviation weighted-average F1 0.0817

# Compare iteration vs Cross validation

Base-DT

average accuracy 0.98

average macro-average F1 0.9492

average weighted-average F1 0.9797

standard deviation accuracy 0.0

standard deviation macro-average F1 0.0

standard deviation weighted-average F1 0.0

Base-DT

average accuracy 0.99

average macro-average F1 0.9865

average weighted-average F1 0.9896

standard deviation accuracy 0.02

standard deviation macro-average F1 0.0322

standard deviation weighted-average F1 0.0209

# Rerun the step 10 time

- When running all the model inside a loop, the result is of all model is the same, because the program use the same test and train set
- When using cross validate model to run, by split the test and train set into 10 different portion, so each time program using different test and train set the result is different. However, it take very long to run the program since the multilevel network does not converge in 200 iteration.