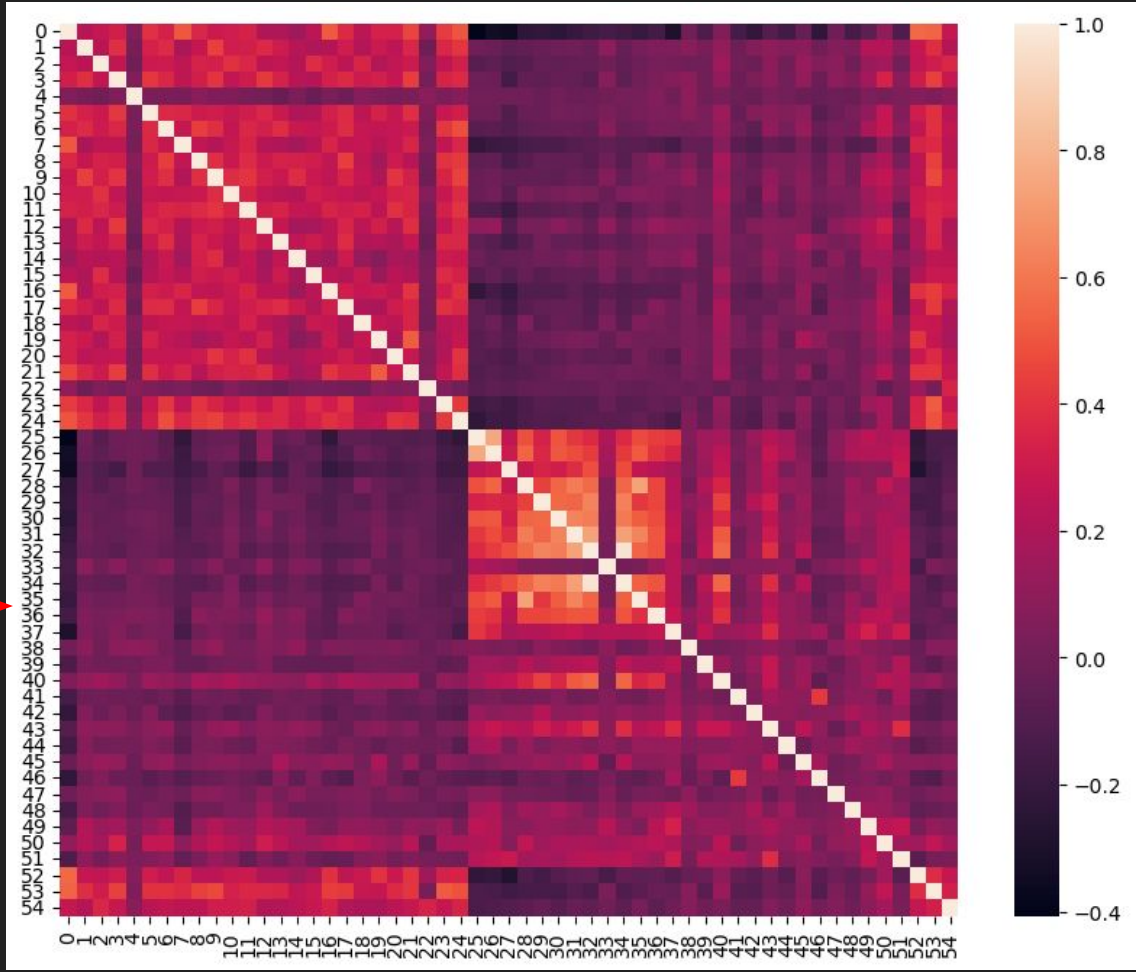
Spam Classification - Naive Bayes

What’s in the data?

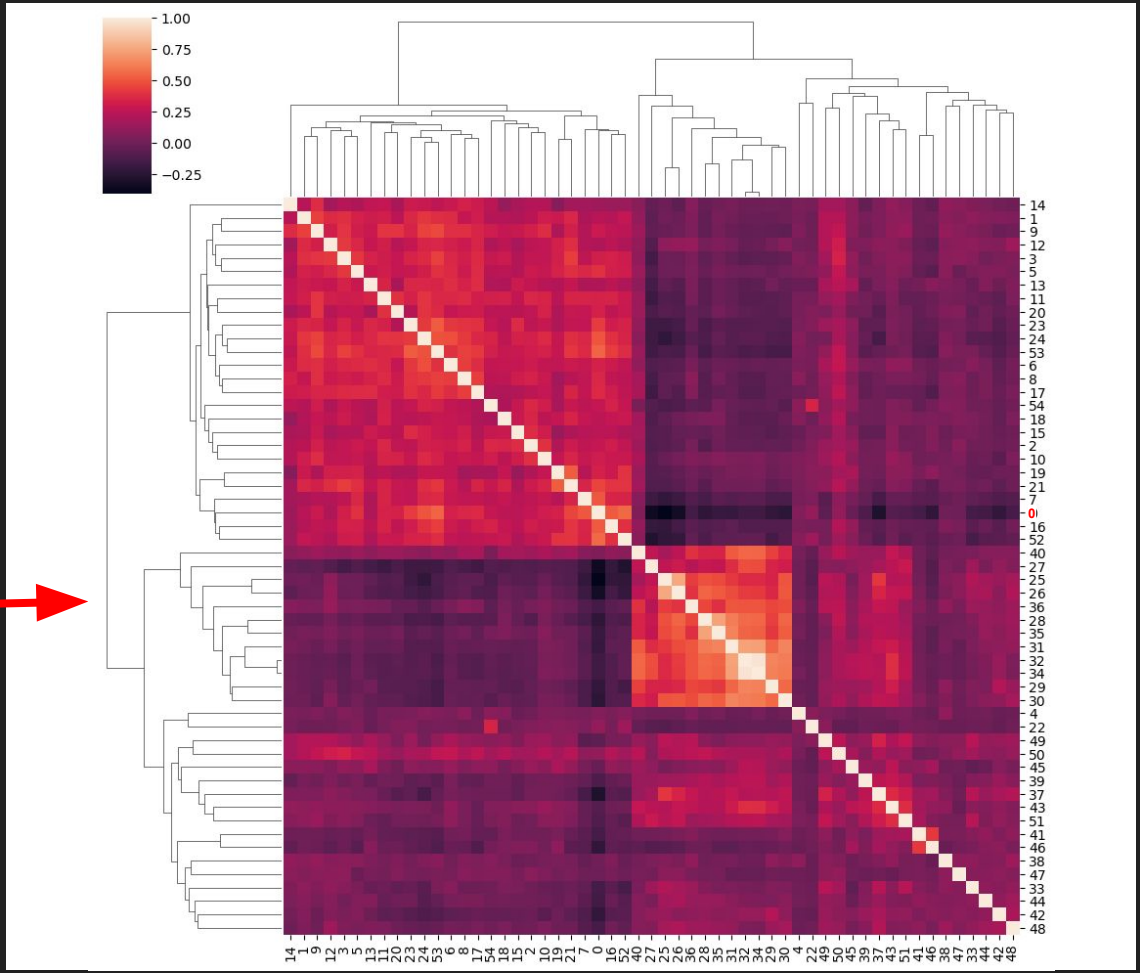
Analysis

Heat Map of all features



Analysis

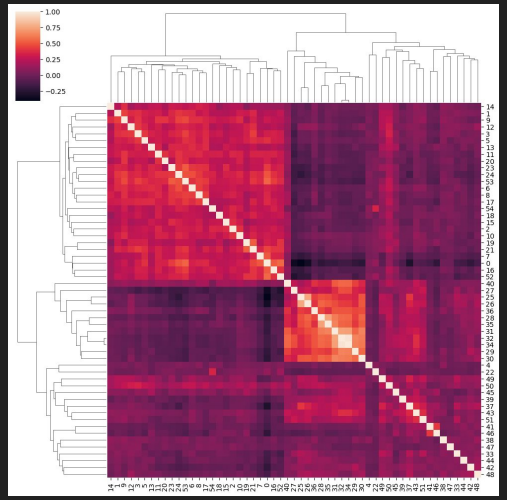
Cluster Map of all features

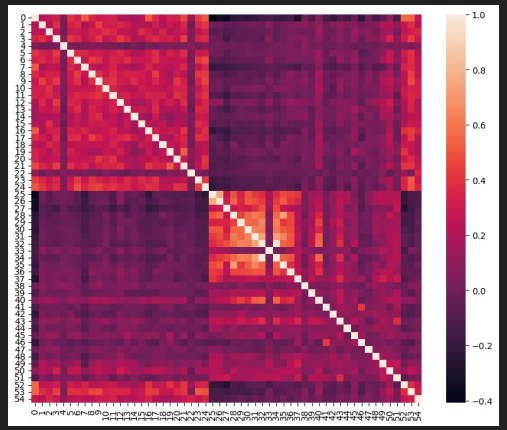


Analysis of data set - why Naive Bayes?

- K-Nearest Neighbours 💀
- Naive Bayes ❓
- Neural Net 💕

Whilst clustermap shows high dependency… I was curious to see NB's efficiency regardless





Naive bayes - How does it work?

Core Assumptions

- All features are independent
- Features are all equally important

Mr Thomas Bayes

At the complete basics :

Naive Bayes classifier works by calculating and comparing the probability of classes based on probabilities of features.

Most likely class for a new message is $\hat{y} = \max_{class} P(class|message)$

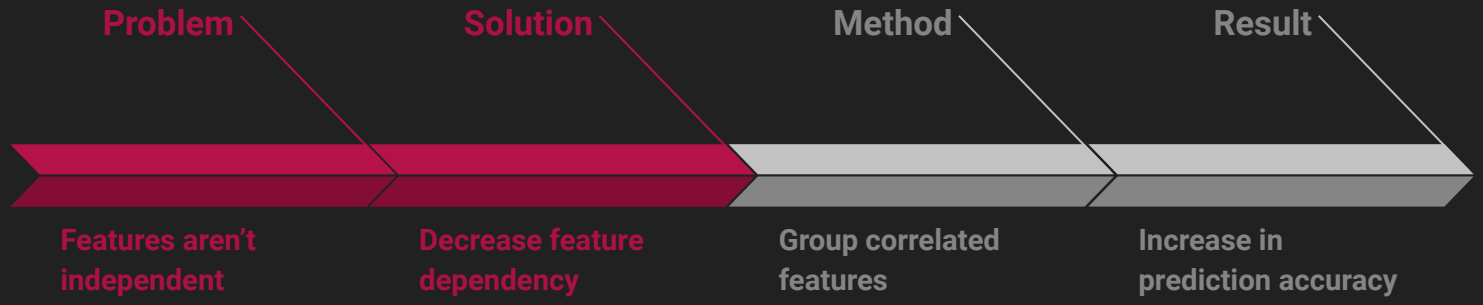
Which simplifes down

$$\hat{y} = \max_{class} P(class) \cdot P(message|class)$$

Optimisations - Logarithmic

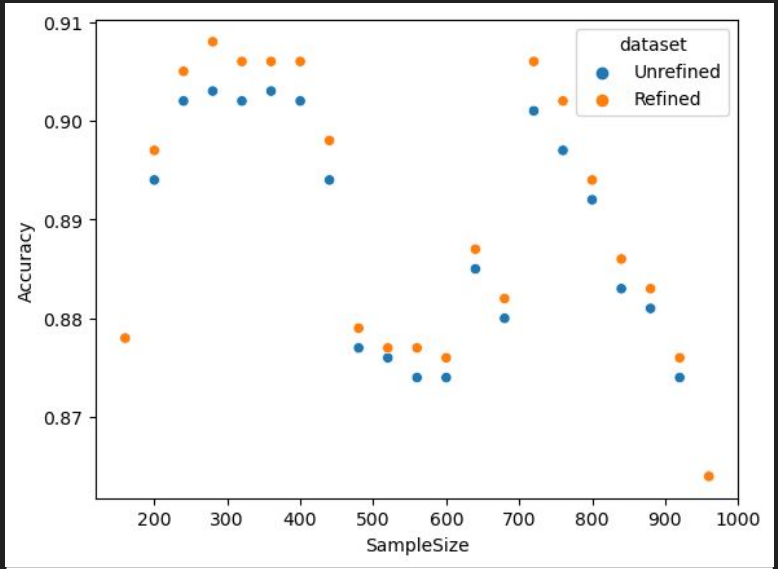
Logarithmic - reduce information loss from underflow of multiplying probabilities

Grouping correlated columns - increase feature independency, good because naive works on assumption of all features independent


Problem
Solution
Method
Result
Features aren't independent
Decrease feature dependency
Group correlated features
Increase in prediction accuracy

Optimisations - Grouping correlated columns

Removing High Feature Correlection	
PMCC threshold **0.7**	PMCC threshold **0.6**
Grouping lead to higher feature independency **LEAD TO** Higher independence and balanced weighting **Result** **Minor Gain in accuracy**	Grouping lead to higher feature independency **BUT** Too much information lost from removal column **Result** **Minor Loss in accuracy**




```
Columns removed to provide Refined data:
[25, 28, 31]
21 Samples Sizes
Refined accuracy total = 18.693 and
Unrefined total = 18.636
```


Is it worth
anything?


IMPROVMENT
IS
IMPROVEMENT


IMPROVMENT
IS
IMPROVEMENT

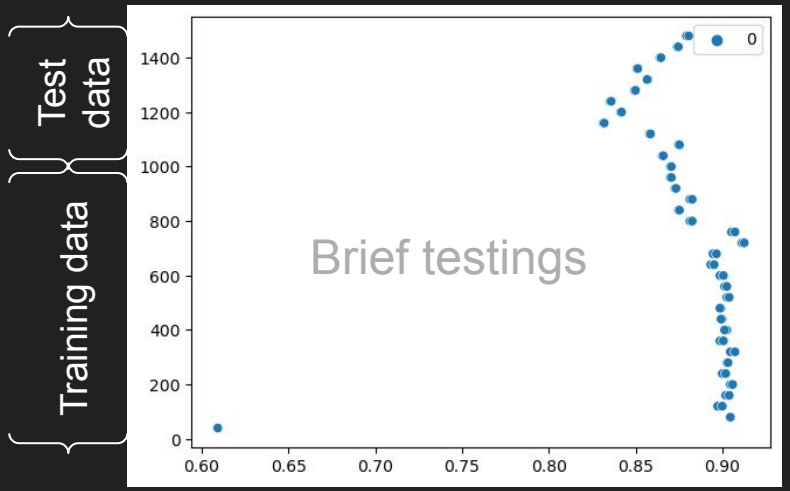

IMPROVMENT
IS
IMPROVEMENT

Accuracy belief

If test data is representative, then accuracy of roughly 0.85

if test data contains anomalies, more akin to 0.8 or 0.75

Hence final bet is **0.8**



If i had more time

Implement a Neural Network

Incorporate Machine Learning into Naive Bayes

- Detect trends alongside single feature probabilities

Difficulty of approach (10%)
How ambitious and complicated? How well is it suited for task? Clearly state the approach you have taken to solve the task and why you have chosen it.
Description of algorithm (10%)
How well understood and explained is algorithm? Describe algorithm you have implemented in terminology supporting mathematical notation
Implementation and Optimisation (10%)
optimisation of your chosen algorithm? Why you made these choices
Explanation and Contextualisation of Results (10%)
How well have you provided a contextualised assessment of your predicted accuracy? Both strengths and limitations. This could include proposing any future work that could address the current limitations of your model.
Presentation and Polish (10%)
How well presented is your video? your video must deliver information clearly.

CM12001: Graded Assignment 2: Spam Filtering | Home (bath.ac.uk)