DAT565/DIT407 Assignment 3

Ola Bratt
ola.bratt@gmail.com

Patrick Attimont
patrickattimont@gmail.com

2024-02-xx

This paper is addressing the assignment 3 study queries within the *Introduction to Data Science & AI* course, DIT407 at the University of Gothenburg and DAT565 at Chalmers. The main source of information for this project is derived from the lectures and Skiena [1].

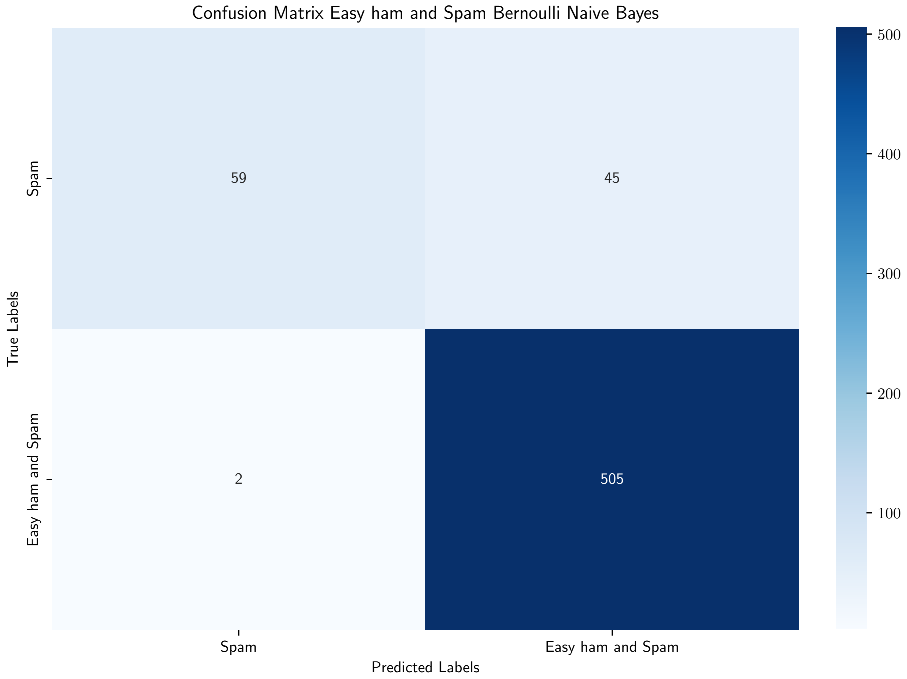
Problem 1: Spam and Ham

Problem 2: Preprocessing

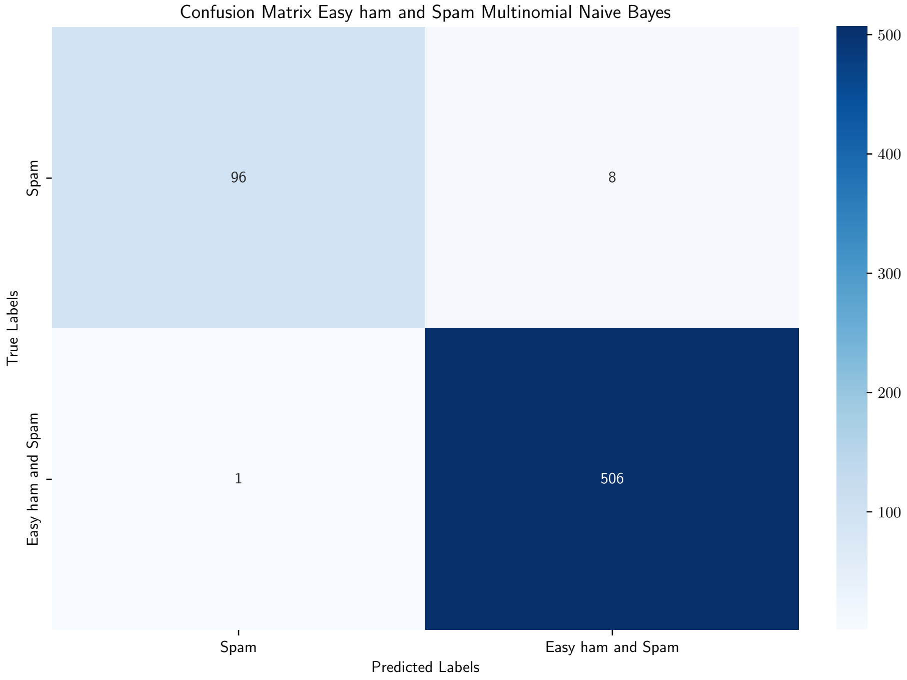
Problem 3: Easy Ham

Model	accuracy	precision	recall	F1 score
Multinomial Naive Bayes	0.9852700490998363	0.9844357976653697	0.9980276134122288	0.99118511
Bernoulli Naive Bayes	0.9230769230769231	0.9181818181818182	0.9960552268244576	0.95553453

Table 1: Precision and accuracy for Easy Ham and Spam



(a) Easy ham vs spam, Bernoulli Naive Bayes



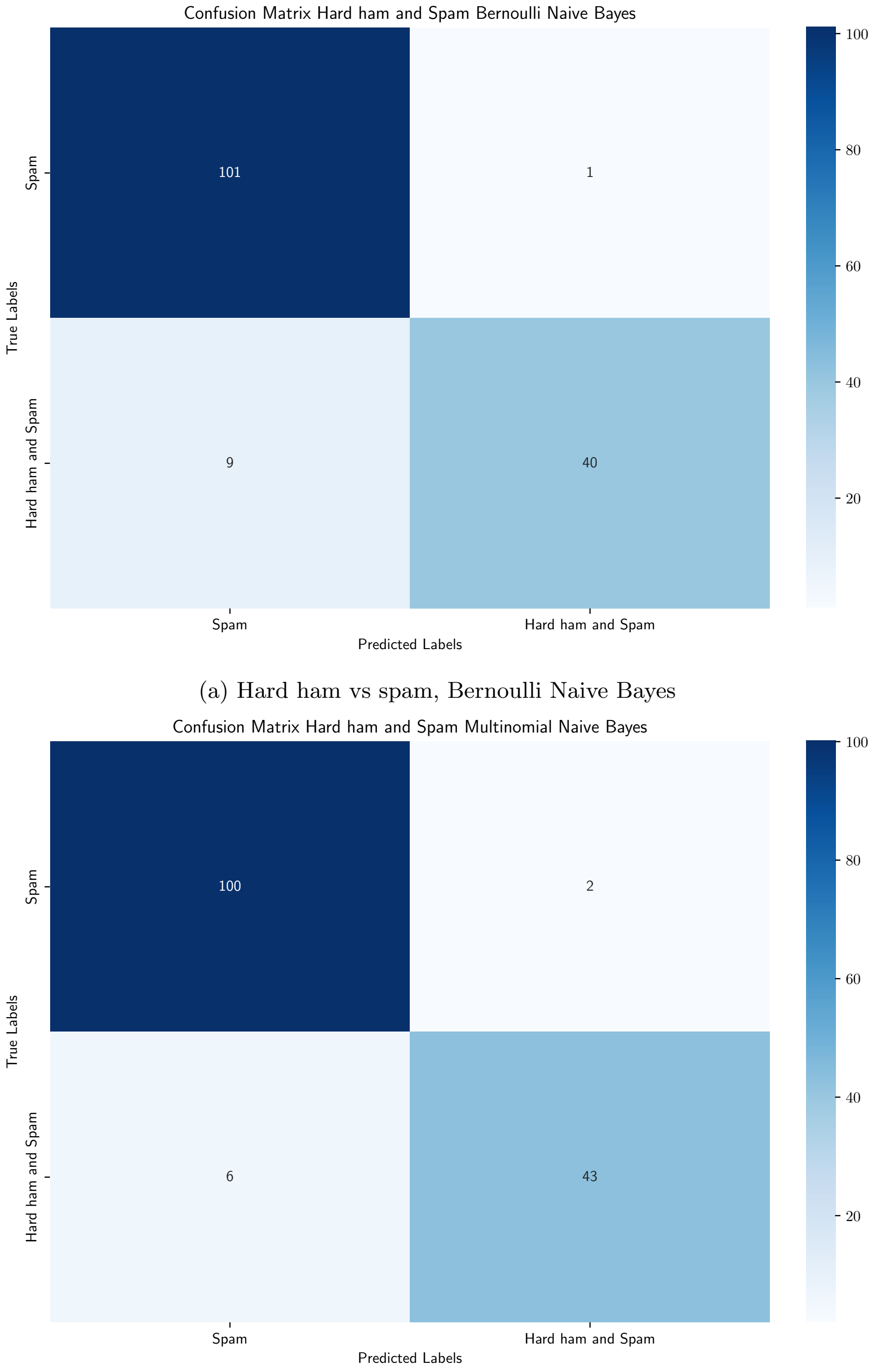
(b) Easy ham vs spam, Multinomial Naive Bayes

Figure 1: Confusion matrixes of easy ham and spam

Model	accuracy	precision	recall	F1 score
Multinomial Naive Bayes	0.9470198675496688	0.9555555555555556	0.8775510204081632	0.9148936
Bernoulli Naive Bayes	0.9337748344370861	0.975609756097561	0.8163265306122449	0.8888888

Table 2: Precision and accuracy for Hard Ham and Spam

Problem 3: Hard Ham



(a) Hard ham vs spam, Bernoulli Naive Bayes

(b) Hard ham vs spam, Multinomial Naive Bayes

Figure 2: Confusion matrixes of hard ham and spam

References

[1] Steven S Skiena. *The Data Science Design Manual.* Retrieved 2024-01-20. 2024. URL: `https://ebookcentral.proquest.com/lib/gu/detail.action?docID=6312797`.

Appendix: Source Code