**Artificial Intelligence**

Group 3B

Rick Fontein s1483870

Niels Geuze s1478494

Mark van Kruistum s1460536

Kasper de Kruiff s1479733

Rens Kruining s1351524

**A:**

The results of the male/female blogs is with smoothing k=1 66% and with k=0.00001 74% which was the best we found, with the ham/spam emails with k=1 the accuracy is 98.97% which was also the highest we got if you go either lower or higher it won’t change or will lower, with k=0.001 the accuracy is 98.63%. The script that gives these results is MachineLearning.java in package A, you can run it without doing anything and on the top you can see 2 variables that you can change to adjust the smoothing (BLOG\_SMOOTHING and EMAIL\_SMOOTHING).

**B:**

NBC:

Male/Female train/test: Spam/Ham 10- folds cross-validation

a b <-- classified as a b <-- classified as

18 7 | a = F 2366 46 | a = ham

10 15 | b = M 21 460 | b = spam

Accuracy = 0.66 Accuracy = 0.977

Precision = 0.662 Precision = 0.978

Recall = 0.66 Recall = 0.977

F-measure = 0.659 F-measure = 0.977

**C**

J48:

Male/Female train/test: Spam/Ham 10- folds cross-validation

a b <-- classified as a b <-- classified as

15 10 | a = F 2358 54 | a = ham

8 17 | b = M 64 417 | b = spam

Accuracy = 0.64 Accuracy = 0.959

Precision = 0.641 Precision = 0.959

recall = 0.64 recall = 0.959

F-Measure = 0.639 F-Measure = 0.959

BayesianLogisticRegression:

Male/Female train/test: Spam/Ham 10- folds cross-validation

a b <-- classified as a b <-- classified as

18 7 | a = F 2407 5 | a = ham

11 14 | b = M 11 470 | b = spam

Accuracy = 0.64 Accuracy = 0.994

Precision = 0.644 Precision = 0.994

recall = 0.64 recall = 0.994

F-Measure = 0.638 F-Measure = 0.994

As you can see for the Male/Female data the NBC is the best classifier although it is close, this is because it is very hard to say if an email is male or female this gives a lot of variation and logistic regression is not very good when that happens. Decision trees are in general not ideal for text which you can see in the results as well since it is the worst in all cases. This is why for determining if a certain text is written by a Male or Female NBC is the best solution out of these three. Spam and Ham is a lot easier to separate, the line between them is a lot clearer which makes it so that Logistic regression becomes really good and you can see that in the result as it is nearly perfect, while the others still have errors >2% of the time.