# Spambase Dataset

**Dataset files**

* <http://archive.ics.uci.edu/ml/machine-learning-databases/spambase/spambase.zip>

**Dataset entry description**

Most of the attributes indicate whether a particular word or character was frequently occuring in the e-mail.

The run-length attributes (Attributes 55-57) measure the length of sequences of consecutive capital letters.

Attribute 1 - 48: type word\_freq\_WORD = percentage of words in the e-mail that match WORD, i.e. 100 \* (number of times the WORD appears in the e-mail) / total number of words in e-mail.

1. word\_freq\_make: continuous.
2. word\_freq\_address: continuous.
3. word\_freq\_all: continuous.
4. word\_freq\_3d: continuous.
5. word\_freq\_our: continuous.
6. word\_freq\_over: continuous.
7. word\_freq\_remove: continuous.
8. word\_freq\_internet: continuous.
9. word\_freq\_order: continuous.
10. word\_freq\_mail: continuous.
11. word\_freq\_receive: continuous.
12. word\_freq\_will: continuous.
13. word\_freq\_people: continuous.
14. word\_freq\_report: continuous.
15. word\_freq\_addresses: continuous.
16. word\_freq\_free: continuous.
17. word\_freq\_business: continuous.
18. word\_freq\_email: continuous.
19. word\_freq\_you: continuous.
20. word\_freq\_credit: continuous.
21. word\_freq\_your: continuous.
22. word\_freq\_font: continuous.
23. word\_freq\_000: continuous.
24. word\_freq\_money: continuous.
25. word\_freq\_hp: continuous.
26. word\_freq\_hpl: continuous.
27. word\_freq\_george: continuous.
28. word\_freq\_650: continuous.
29. word\_freq\_lab: continuous.
30. word\_freq\_labs: continuous.
31. word\_freq\_telnet: continuous.
32. word\_freq\_857: continuous.
33. word\_freq\_data: continuous.
34. word\_freq\_415: continuous.
35. word\_freq\_85: continuous.
36. word\_freq\_technology: continuous.
37. word\_freq\_1999: continuous.
38. word\_freq\_parts: continuous.
39. word\_freq\_pm: continuous.
40. word\_freq\_direct: continuous.
41. word\_freq\_cs: continuous.
42. word\_freq\_meeting: continuous.
43. word\_freq\_original: continuous.
44. word\_freq\_project: continuous.
45. word\_freq\_re: continuous.
46. word\_freq\_edu: continuous.
47. word\_freq\_table: continuous.
48. word\_freq\_conference: continuous.

Attribute 49 - 54: type char\_freq\_CHAR = percentage of characters in the email that match CHAR, i.e. 100 \* (number of CHAR occurrences) / total characters in email

1. char\_freq\_;: continuous.
2. char\_freq\_(: continuous.
3. char\_freq\_[: continuous.
4. char\_freq\_!: continuous.
5. char\_freq\_$: continuous.
6. char\_freq\_#: continuous.

Attribute 55: type capital\_run\_length\_average = average length of uninterrupted sequences of capital letters

1. capital\_run\_length\_average: continuous.

Attribute 56: type capital\_run\_length\_longest = length of longest uninterrupted sequence of capital letters

1. capital\_run\_length\_longest: continuous.

Attribute 57: type capital\_run\_length\_total = sum of length of uninterrupted sequences of capital letters = total number of capital letters in the e-mail

1. capital\_run\_length\_total: continuous.

Attribute 58 (Label): type spam = denotes whether the e-mail was considered spam

1 means Yes (Spam)

0 means No (Not spam)

**Example entry**

0,0.72,0.72,0,0,0,0,1.45,0,0,0.72,0,0,0,0,2.91,0,0.72,1.45,0,0,0,0.72,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.123,0,0.495,0,0,1.525,8,61,1

**Labelling field**

Attribute 58 (Label)

Possible values

* 1 means Spam
* 0 means Not spam

**Questions to answer using ML**

1. Mail is spam or not?
2. How many mails were spam?

**Reference**

* <http://archive.ics.uci.edu/ml/datasets/Spambase>