**Spam recognition data**

**Purpose**

To predict whether an email is spam (advertisements for products/web sites, make money easily, chain letters, pornography, ...) or not.

**Target variable**

* spam: yes (=1), no (=0).

**Predictors**

Variables related to WORD frequency: percentage of times in which the WORD after pct\_ appears in the email = 100 \* (number of times the WORD appears in the e-mail) / total number of words in e-mail. Domain of the variables: [0, 100].

* pct\_make:
* pct\_address:
* pct\_all:
* pct\_3d:
* pct\_our:
* pct\_over:
* pct\_remove:
* pct\_internet:
* pct\_order:
* pct\_mail:
* pct\_receive:
* pct\_will:
* pct\_people:
* pct\_report:
* pct\_addresses:
* pct\_free:
* pct\_business:
* pct\_email:
* pct\_you:
* pct\_credit:
* pct\_your:
* pct\_font:
* pct\_000:
* pct\_money:
* pct\_hp:
* pct\_hpl:
* pct\_george:
* pct\_650:
* pct\_lab:
* pct\_labs:
* pct\_telnet:
* pct\_857:
* pct\_data:
* pct\_415:
* pct\_85:
* pct\_technology:
* pct\_1999:
* pct\_parts:
* pct\_pm:
* pct\_direct:
* pct\_cs:
* pct\_meeting:
* pct\_original:
* pct\_project:
* pct\_re:
* pct\_edu:
* pct\_table:
* pct\_conference:

Variables related to SIGN frequency: percentage of times in which the SIGN after pct\_ appears in the email = 100 \* (number of times the SIGN appears in the e-mail) / total number of words in e-mail. Domain of the variables: [0, 100].

* pct\_semicolon:
* pct\_parenthesis:
* pct\_bracket:
* pct\_exclamatory:
* pct\_dollar:
* pct\_hash:

Variables related to uninterrupted sequences of capital letters. Domain of the variables: [1, +Inf).

* crl\_average = average length of uninterrupted sequences of capital letters
* crl\_longest = longest length of uninterrupted sequences of capital letters
* crl\_total = total length of uninterrupted sequences of capital letters capital letters