Formatowanie spamu
Data Mining

Michał Maj i Anna Mieszkalska
album 256556 i 255699

24 kwietnia 2023

Spis treści

1 Analiza opisowa i wizualizacja

1.1 Wstęp

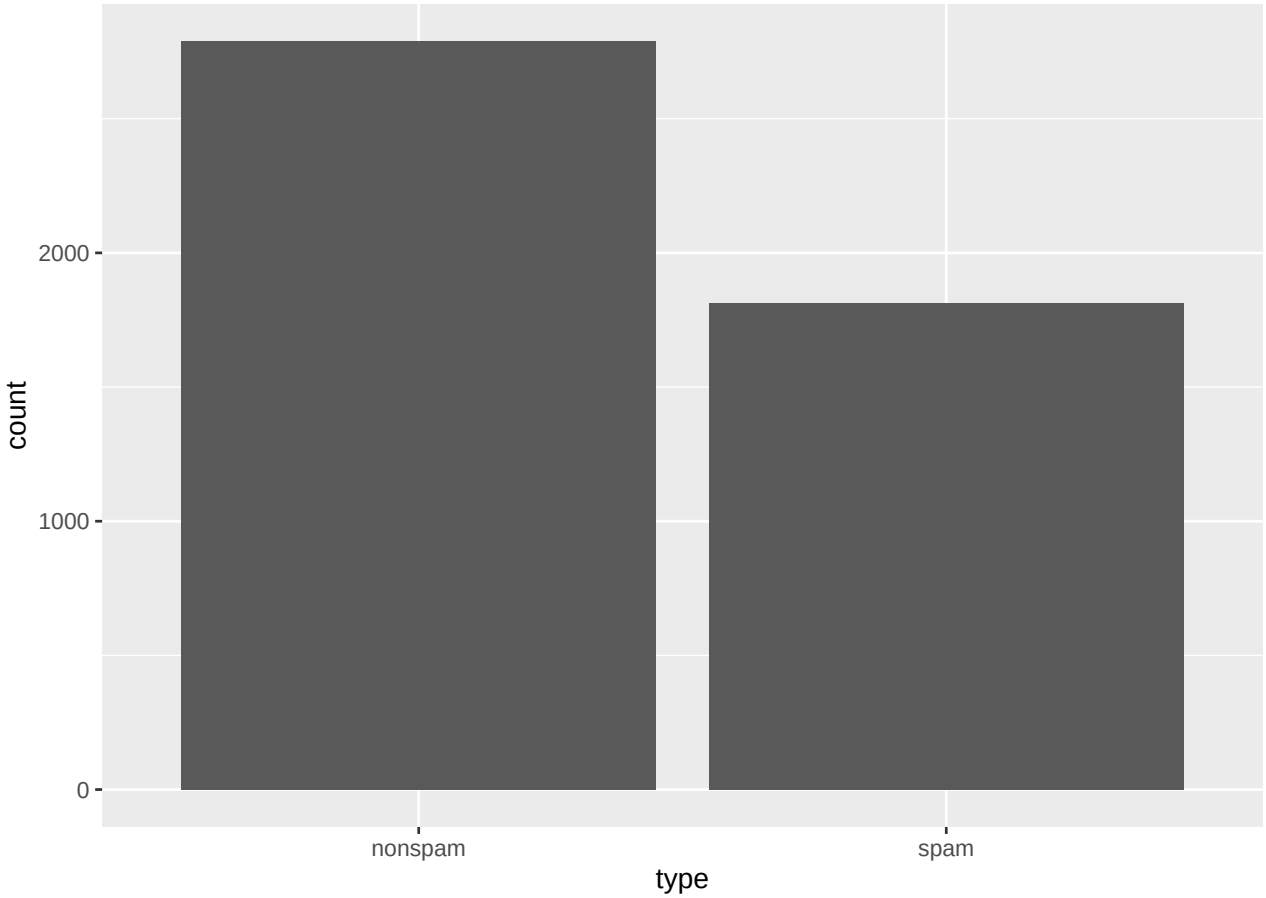
W naszym projekcie będziemy analizować dane o nazwie *Spambase* z biblioteki *kernlab*. Zestaw danych *spambase* jest zbiorem wiadomości e-mail, które zostały przeanalizowane i sklasyfikwoane jako spam lub non-spam. Celem tego zbioru danych jest dostarczenie użytecznych materiałów potrzebnych do analiz i eksploracji w tym zakresie. W tym projekcie użyjemy różnych metod i technik pozyskiwania wiedzy, aby przeanalizować dane *spambase* w celu opracowania modelu klasyfikującego wiadomości e-mail jako spam lub non-spam. Modele opracowane w tym projekcie mogą być przydatne w rzeczywistych serwisach poczt e-mailowych, gdzie problem dostarczania niechcianych wiadomości jest nam powrzechnie znany.

1.2 Opis danych

Zbiór danych *spambase* wyodrębnia 58 cech, które oznaczają częstość występowania danego znaku bądź słowa w jednym e-mailu. Pierwsze 48 zmiennych dotyczy występowania konkretnych słów, następne 6 występowania znaków, a kolumny 55-57 dotyczą średniej, najdłuższej i całkowitej długości wielkich liter. Ostatnia zmienna *type* odpowiada za określenie typu e-maila jako spam lub non-spam, zatem będziemy rozważać dwie klasy Zbiór ten składa się z 4601 obserwacji (wiadomości e-mail).

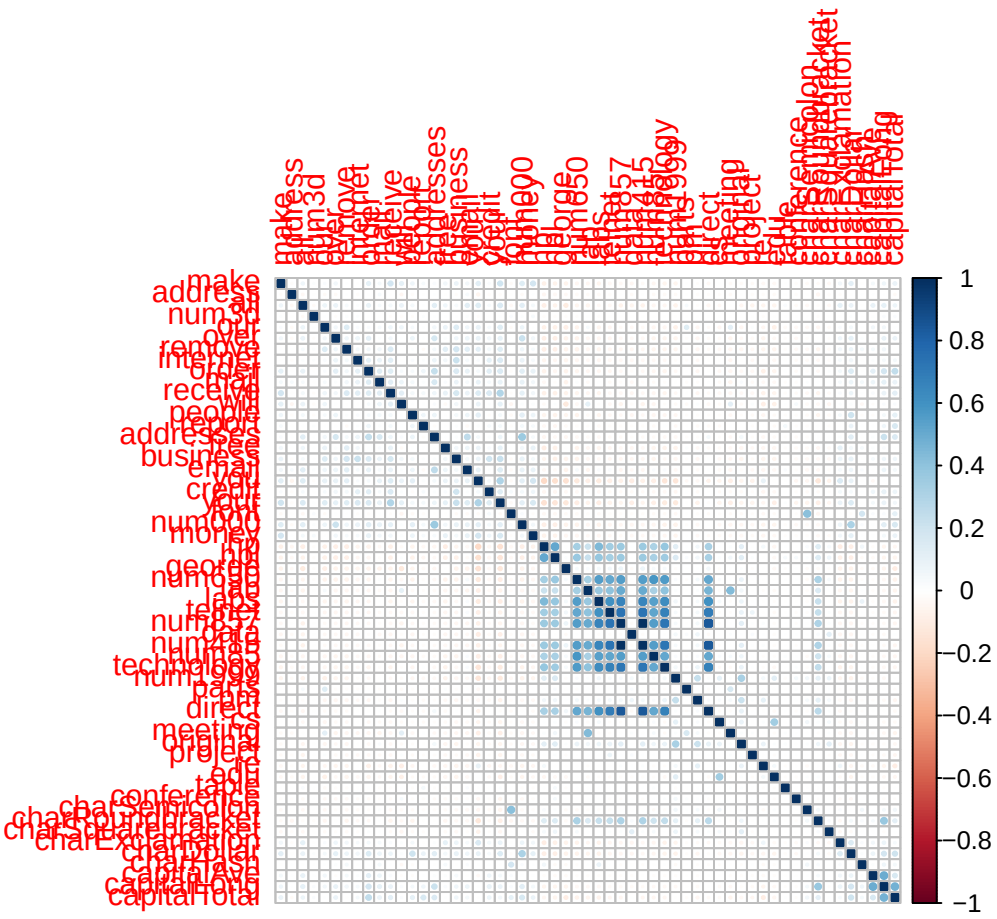
Tabela 1: Rozkład klas.

Var1	Freq
nonspam	2788
spam	1813



Ilość e-maili, które zostały sklasyfikowane jako spam wynosi 1813 (tabela 1), a ilość tych, które nie są spamem wynosi 2733, zatem klasa spam stanowi prawie 40% całości, więc dane są dość zbalansowane

Za pomocą funkcji `str` mamy, że wszystkie zmienne są typu *numeric*, oczywiście oprócz zmiennej *type*, która jest typu *factor*. Patrząc do tabeli **??** widzimy, że wszystkie typy zmiennych zostały określone prawidłowo. Funkcja `is.na()` mówi nam, że nasze dane nie posiadają żadnych wartości NA, należy jednak sprawdzić, czy w tym przypadku nie są one kodowane inaczej.



Podsumowując, mamy:

- $n = 4601$ (liczba przypadków),
- $p = 58$ (liczba cech),
- $K = 2$ (licza klas),
- 0 wartości brakujących