METODY EKSPLORACJI DANYCH
Laboratorium. Klasyfikacja na podstawie klasyfikatora bayesowskiego i najbliższego sąsiedztwa

Zadanie 1

Mieszkańcy pewnej małej kamienicy, mieszczącej się zaraz tuż przy plaży morskiej, zaczęli od paru dni z ciekawością obserwować treningi grupy zawodników siatkówki plażowej. Mieszkańcy zauważyli, że ta grupa graczy niekiedy dzieli się na dwa zespoły w celu rozegrania meczu. Niestety obserwatorzy nie znają zamiarów trenujących, ale odnotowali, że gracze grają w różnych warunkach pogodowych. Swoje obserwacje odnotowali w tabeli.

Tabela 1. Obserwacje mieszkańców

Nr obserwacji	Siła wiatru	Zachmurzenie	Odczuwalna temperatura	Zagrano mecz
1	silny	pochmurnie	zimno	nie
2	silny	pochmurnie	ciepło	nie
3	brak	słonecznie	ciepło	tak
4	brak	słonecznie	gorąco	nie
5	słaby	pochmurnie	gorąco	tak
6	słaby	słonecznie	ciepło	tak
7	brak	pochmurnie	zimno	nie
8	silny	słonecznie	zimno	tak
9	brak	pochmurnie	gorąco	tak
10	silny	pochmurnie	ciepło	tak

Dzisiaj jest ciepło i słonecznie, ale wieje silny wiatr. Wobec tego mieszkańcy kibicujący grze zastanawiają się czy gracze pojawią się na plaży, aby rozegrać swój mecz treningowy, czy tylko wykonać ćwiczenia.

Proszę

1) najpierw rozwiązać zadanie z wykorzystaniem naiwnego klasyfikatora Bayesa,
2) potem metody najbliższego sąsiedztwa. W metodzie najbliższego sąsiedztwa proszę przyjąć najpierw odległość euklidesową, potem miejską.
3) otrzymane wyniki klasyfikacji proszę porównać ze sobą. Na podstawie przeprowadzonych porównań **sformułować własne wnioski**.
4) wyniki pracy zawrzeć w postaci **sprawozdania**. Do sprawozdania proszę dodać jako załączniki wszystkie pliki z obliczeniami.
5) zadanie zrealizować w narzędziu wybranym przez siebie.

Zadanie 2

Załóżmy, że chcemy zbudować algorytm odfiltrowania spamu z otrzymanych wiadomości pocztą lub za pomocą czatu. W spamie występują treści wg. określonych słów kluczowych, których statystyka występowania została zawarta w tabeli wraz z dokonaną klasyfikacją wg. pewnego[1] algorytmu.

Tabela 2. Statystyka występowania słów kluczowych w wiadomościach

Nr wiad.	Słowa kluczowe					Klasyfikacja:
	pieniądz	darmowy	bogaty	nieprzyzwoicie	tajny	spam
1	nie	nie	tak	nie	tak	tak
2	tak	tak	tak	nie	nie	tak
3	nie	nie	nie	nie	nie	nie
4	nie	tak	nie	nie	nie	tak
5	tak	nie	nie	nie	nie	nie
6	nie	tak	nie	tak	tak	tak
7	nie	tak	nie	tak	nie	tak
8	nie	nie	nie	tak	nie	tak
9	nie	tak	nie	nie	nie	nie
10	nie	nie	nie	nie	tak	nie
11	tak	tak	tak	nie	tak	tak
12	tak	nie	nie	nie	tak	tak
13	nie	tak	tak	nie	nie	nie
14	tak	nie	tak	nie	tak	???

W zadaniu

a) opierając się na klasyfikacji bayesowskiej proszę określić status wiadomości/posta nr 14 na podstawie występowania lub niewystępowania słów kluczowych.
b) Przedyskutować wynik otrzymanego rezultatu. Przedyskutować, czy naiwna klasyfikacja bayesowska jest dobrą metodą wykrywania spamu. Proszę uzasadnić swoje wnioski.
c) wyniki swojej pracy proszę zawrzeć w **sprawozdaniu**. Do sprawozdania proszę dodać jako załączniki wszystkie pliki z obliczeniami.

[1] nieznanego