Akademia Ekonomiczno-Humanistyczna w Warszawie

SPRAWOZDANIE

INTELIGENTNA ANALIZA DANYCH

**LAB2**

**WALIDACJA KRZYŻOWA, STRATYFIKACJA,**

**DRZEWA LOSOWE I NAIWNY KLASYFIKATOR BAYESA**

07.11.2021

JOANNA PRAJZENDANC

36358

MIŁOSZ SAKOWSKI

36381

**Spis treści**

[1. Cel i przebieg ćwiczenia 3](#_Toc29772)

[2. Definicje i założenia 3](#_Toc13946)

[2.1. Wyjaśnienie pojęć 3](#_Toc23313)

[2.2. Przygotowanie bazy danych 3](#_Toc28633)

[i. Importowanie bazy danych i oryginalna zawartość 3](#_Toc7684)

[ii. Docelowa baza danych 3](#_Toc12652)

[3. Walidacja krzyżowa i stratyfikacja 3](#_Toc13095)

[3.1. Zadanie #1 3](#_Toc30990)

[i. Treść polecenia 3](#_Toc2373)

[ii. Rozwiązanie 3](#_Toc10574)

[4. Klasyfikator drzewa losowego 4](#_Toc13270)

[4.1. Zadanie #2 4](#_Toc32327)

[i. Omówienie kodu 4](#_Toc8963)

[ii. Pytania 4](#_Toc8720)

[4.2. Zadanie #3 4](#_Toc12503)

[i. Treść polecenia 4](#_Toc9646)

[ii. Rozwiązanie 4](#_Toc21720)

[4.3. Wnioski 4](#_Toc2746)

[5. Naiwny klasyfikator bayesa 4](#_Toc6567)

[5.1. Zadanie #4 4](#_Toc11031)

[i. Treść polecenia 4](#_Toc3155)

[ii. Rozwiązanie 5](#_Toc30568)

[5.2. Wnioski 5](#_Toc23423)

[6. Porównanie klasyfikatorów 5](#_Toc2600)

[7. Wnioski końcowe 5](#_Toc19602)

# Cel i przebieg ćwiczenia

Celem ćwiczenia było utrwalenie wiedzy w zakresie klasyfikatorów. W pierwszej kolejności należało zapoznać się z problemem niezbalansowanego zbioru danych i sposobów rozwiązania go za pomocą walidacji krzyżowej i stratyfikacji w przypadku klasyfikatora binarnego. Następnym krokiem było zapoznanie się z klasyfikatorem drzewa losowego oraz naiwnym klasyfikatorem bayesa.

# Definicje i założenia

## Wyjaśnienie pojęć

W sprawozdaniu pojawiają się następujące pojęcia:

* niezbalansowana baza danych
* walidacja krzyżowa
* stratyfikacja
* klasyfikator binarny
* klasyfikator drzewa losowego
* naiwny klasyfikator bayesa

## Przygotowanie bazy danych

Podczas wykonywania zadań omówionych w sprawozdaniu, do rozwiązania użyto tej samej bazy danych z informacjami o pożarach w parku XX.

### Importowanie bazy danych i oryginalna zawartość

### Docelowa baza danych

# Walidacja krzyżowa i stratyfikacja

## Zadanie #1

### Treść polecenia

Proszę pobrać inny zbiór danych, zmienić klasy w taki sposób żeby zbiór był niezbalansowany oraz przetestować drzewo decyzyjne stosując 5-krotną cross walidację ze stratyfikacją

### Rozwiązanie

# Klasyfikator drzewa losowego

## Zadanie #2

### Omówienie kodu

### Pytania

#### Czy uzyskane wyniki są zadowalające?

#### Czy model został przetestowany w prawidłowy sposób?

Nie, ponieważ istnieje ryzyko, że w zbiorze trenującym są same 0 i żadnych 1 albo w zbiorze testującym nie ma żadnych 1 więc nie zbadamy skuteczności modelu w dopasowywaniu 1

## Zadanie #3

### Treść polecenia

Proszę pobrać dowolny zbiór danych ze strony <https://archive.ics.uci.edu/ml/index.php.> Następnie proszę podzielić zbiór na dane trenujące i testujące, wytrenować 5 modeli lasów losowych z różną maksymalną głębokością i liczbą drzew w lesie, porównać wyniki. Proszę o sporządzenie sprawozdania z wnioskami.

### Rozwiązanie

#### 2 drzewa o głębokości równej 2

#### 10 drzew o głębokości równej 2

#### 30 drzew o głębokości równej 2

#### 50 drzew o głębokości równiej 2

#### 70 drzew o głębokości równej 2

#### 10 drzew o głębokości równiej 1

#### 10 drzew o głębokości równiej 3

#### Porównanie wyników

## Wnioski

# Naiwny klasyfikator bayesa

## Zadanie #4

### Treść polecenia

Proszę pobrać dowolny zbiór danych ze strony <https://archive.ics.uci.edu/ml/index.php> Następnie proszę podzielić zbiór na dane trenujące i testujące, wytrenować i przetestować klasyfikator GaussianNB. Proszę o sporządzenie sprawozdania z wnioskami.

### Rozwiązanie

## Wnioski

# Porównanie klasyfikatorów

# Wnioski końcowe