Zadanie 4 (WSI)

Piotr Obst 304090 (grupa 103)

1. **Treść zadania**

Temat 4.1
Zaimplementować naiwny klasyfikator bayesa i zastosować go do klasyfikacji dwóch wybranych zbiorów danych ze strony https://archive.ics.uci.edu/ml/datasets.php. Do oceny klasyfikatora należy użyć walidacji krzyżowej i narysować krzywą ROC.

2. **Sposób uruchomienia**
   - Należy mieć zainstalowanego Pythona 3 oraz matplotlib,
   - Aby uruchomić program, należy wpisać w konsoli „python main.py”. Na niektórych maszynach należy użyć komendy python3 zamiast python.
3. **Przykładowe wyjście programu**

```
loading data
removing duplicates
dropping and categorizing data
calculating ROC curve
10%
20%
29%
40%
50%
60%
70%
80%
90%
100%
```

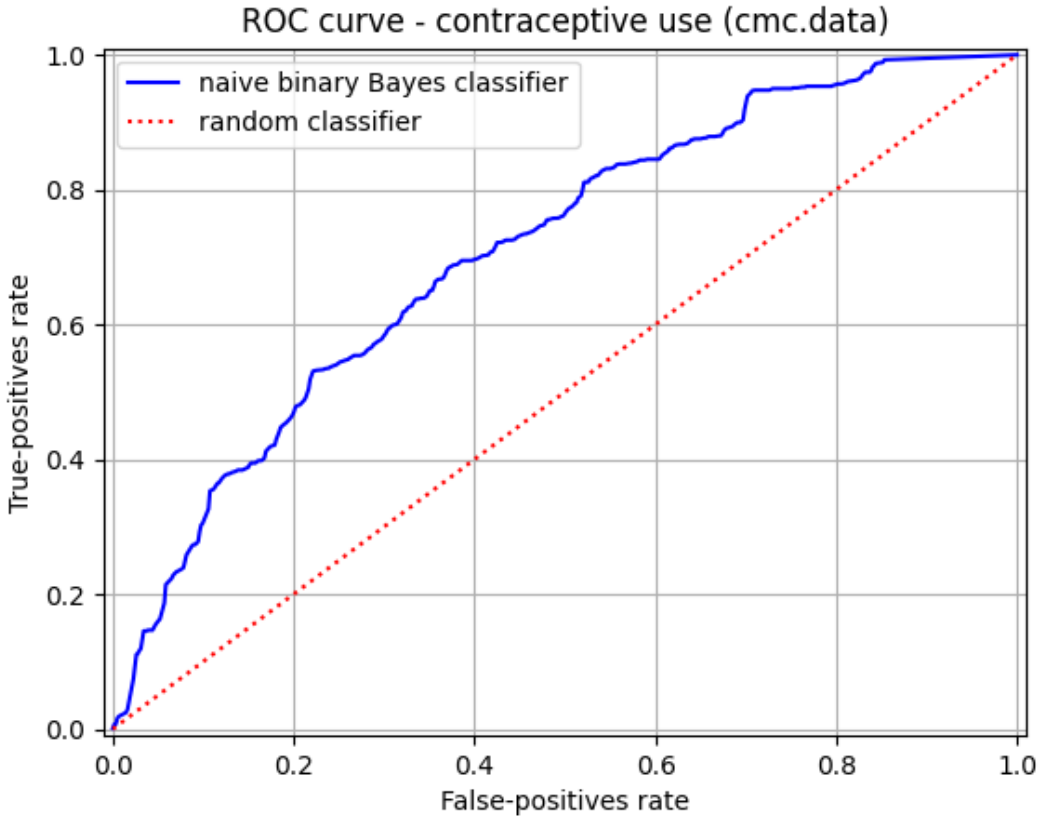
```
Feature:  workclass
                        <=50K >50K
Private.............10454 2996
Self-emp-not-inc....1106  436
Self-emp-inc........294   363
Federal-gov.........363   225
Local-gov...........941   372
State-gov...........559   220
Without-pay.........7     0
Never-worked........0     0
Total...............13724 4612

Feature:  education
                        <=50K >50K
Bachelors...........1724  1314
Some-college........3372  830
11th................611   33
HS-grad.............5098  1000
Prof-school.........91    249
Assoc-acdm..........475   159
Assoc-voc...........590   218
9th.................230   14
7th-8th.............280   21
12th................209   15
Masters.............426   555
1st-4th.............24    1
10th................478   40
Doctorate...........57    161
5th-6th.............49    2
Preschool...........10    0
Total...............13724 4612
```

4. **Dodatkowe informacje**
   - Wykresy ROC zapisywane są w folderze ‘graphs’,
   - Program można uruchomić w trybie ‘gadatliwym’: ‘python main.py -v’,
   - Kod pisałem w oparciu o standard Pep8 oraz starałem się, aby był jak najbardziej uniwersalny i stabilny,
   - Zgodnie z ustaleniami, starałem się pisać kod samodokumentujący się oraz to zadanie będzie jeszcze omawiane na konsultacjach, więc raczej nie ma potrzeby opisywania kodu w tym raporcie.

5. **Wybrane zbiory danych i odpowiadające nim krzywe ROC**
   - Wybór metody antykoncepcji
     https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice

     Ilość wpisów: 1473
     atrybuty:
     a) wiek żony – numeryczny, ale zamieniłem na kategorie będące przedziałem wieków
     b) edukacja żony – kategoryczny
     c) edukacja męża – kategoryczny
     d) liczba dzieci – numeryczny,
     ale zamieniłem na kategorie 0=0, 1=1, … 8=8, 9=(9 lub więcej)
     e) religijność żony – kategoryczny
     f) czy żona pracuje – kategoryczny
     g) zawód męża – kategoryczny
     h) standard życia – kategoryczny
     i) ekspozycja na media – kategoryczny
     j) metoda antykoncepcji – kategoryczny – atrybut klasowy.
     1=brak antykoncepcji, 2=długotrwała, 3=krótkotrwała
     połączyłem atrybuty 2 i 3 w jeden atrybut ‘2’, ponieważ zadanie tego wymagało

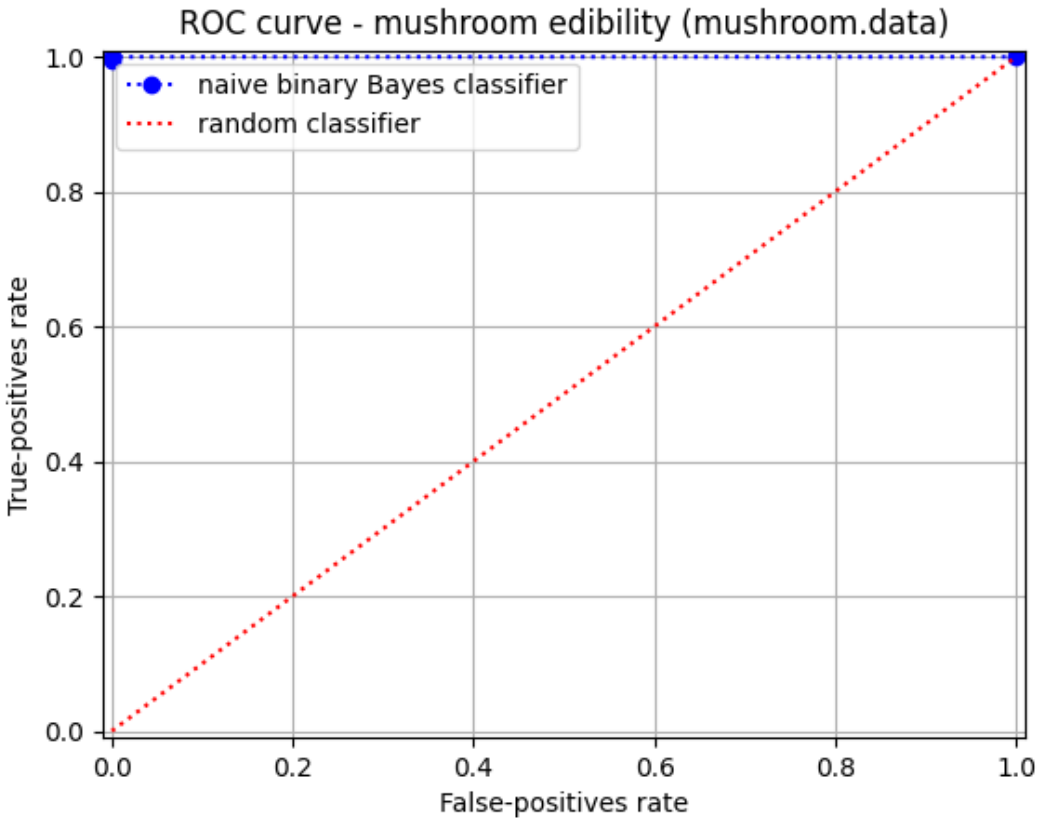
     Krzywa ROC:



- Jadalność grzybów
  https://archive.ics.uci.edu/ml/datasets/Mushroom

  Ilość wpisów: 8124
  atrybuty (wszystkie są kategoryczne)
  a) jadalność – atrybut klasowy,
  e=jadalny, p=trujący/jadalność nieznana/niezalecany
  b) kształt kapelusza
  c) powierzchnia kapelusza
  d) kolor kapelusza
  e) rany
  f) odór
  g) połączenie blaszek
  h) odległość między blaszkami
  i) wielkość blaszek
  j) kolor blaszek
  k) kształt trzonka
  l) korzeń
  m) powierzchnia trzonka nad pierścieniem
  n) powierzchnia trzonka pod pierścieniem
  o) kolor nad pierścieniem
  p) kolor pod pierścieniem
  q) typ osłonki
  r) kolor osłonki
  s) liczba pierścieni
  t) typ pierścienie
  u) kolor zarodników
  v) populacja
  w) środowisko

  Krzywa ROC: (bardzo dobre wytrenowanie, bardzo przewidywalny zbiór)



- Zarobki - https://archive.ics.uci.edu/ml/datasets/Adult
  Przygotowałem ten zbiór, ponieważ zbiór dotyczący jadalności grzybów jest zbyt przewidywalny.

  Ilość wpisów: 48842
  atrybuty:
  a) wiek – numeryczny – zamieniłem na kategorie (przedziały wiekowe)
  b) rodzaj pracy – kategoryczny
  c) waga wpisu – numeryczny – trudne do zamienienia na kategorie i mało znaczące, więc wyrzuciłem ten atrybut
  d) edukacja – kategoryczny
  e) ‘education-num’ – numeryczny – nie znalazłem opisu tego parametru, więc również wyrzuciłem
  f) stan cywilny – kategoryczny
  g) zawód – kategoryczny
  h) związek / rodzina – kategoryczny
  i) rasa – kategoryczny
  j) płeć – kategoryczny
  k) przychody – numeryczny – trudne do zamienienia na kategorie – wyrzuciłem
  l) wydatki – numeryczny – trudne do zamienienia na kategorie – wyrzuciłem
  m) liczba godzin pracy w tygodniu – numeryczny – zamieniłem na kategorie – przedziały
  n) kraj pochodzenia – kategoryczny – zawiera dużo kategorii, więc dla uproszczenia wyrzuciłem wszystkie wpisy, poza tymi z ‘United-States’
  o) roczne zarobki – atrybut klasowy

  Krzywa ROC: (algorytm udało się wytrenować lepiej, niż w przypadku zbioru dotyczącego metod antykoncepcji)