Polecenia do ćwiczeń nr 3

1) Wczytaj do Pythona plik *churn_pl_klasyfikacja.csv*. Jest to plik z danymi wygenerowanymi w efekcie budowy modelu klasyfikacji zmiennej *Rezygnacja*. Zawiera on trzy nowe zmienne:

   - *Rezygnacja_przewidywane* - przewidywana wartość zmiennej *Rezygnacja*, będąca wynikiem budowy modelu klasyfikującego,
   - *Prawdopodobieństwo_0* – przewidywane prawdopodobieństwo przypisania osoby do grupy osób, które nie zrezygnują z usług firmy (*Rezygnacja* = 0),
   - *Prawdopodobieństwo_1* – przewidywane prawdopodobieństwo przypisania osoby do grupy osób, które zrezygnuje z usług firmy (*Rezygnacja* = 1),
   - *Próba* – przypisanie do próby (0 – testowa, 1 - ucząca).

   Celem jest ocena jakości modelu. Pamiętaj, aby oceny dokonywać w podziale na zbiór uczący i testowy.

   a) Zbuduj macierz pomyłek.

   b) Wyznacz następujące wskaźniki klasyfikacji binarnej:

      - trafność,
      - czułość,
      - swoistość,
      - całkowity współczynnik błędu,
      - wskaźnik FN,
      - wskaźnik FP,

   i podaj ich interpretację.

   c) Jakie jest prawdopodobieństwo, że osoba, którą klasyfikator przypisał do grupy osób, które nie odejdą z firmy, faktycznie z niej nie zrezygnuje?

   d) Jakie jest prawdopodobieństwo, że osoba, którą klasyfikator przypisał do grupy osób, które zrezygnują z usług firmy, faktycznie to zrobi?

   e) Zbuduj krzywą ROC i na jej podstawie oceń jakość modelu.

2) Plik *iris_klasyfikacja.csv* zawiera poza oryginalnymi zmiennymi opisującymi wymiary i gatunek kwiatów irysa także zmienne powstałe w wyniku działania pewnego algorytmu klasyfikacji. Są to zmienne:

   - *Predicted* – przewidywany gatunek irysa (przez pewien algorytm klasyfikacji),
   - *Zbiór* – przydział do zbioru (1- zbiór uczący, 0 – zbiór testowy),
   - *Prawdopodobieństwo1, Prawdopodobieństwo2, Prawdopodobieństwo3* – prawdopodobieństwa wyznaczone przez pewien algorytm klasyfikacyjny, że dany kwiat należy do gatunku odpowiednio 1, 2 i 3.

Dokonaj pełnej oceny jakości modelu:

   a) oblicz odpowiednie współczynniki i podaj ich interpretację,

b) narysuj krzywe ROC, oblicz AUC oraz zinterpretuj otrzymane wyniki.

3) Plik *kraje_szacowanie.csv* zawiera dane dotyczące krajów. Na podstawie zawartych w nim zmiennych zbudowano dwa modele predykcyjne dla zmiennej *internet_użytkownicy* na podstawie wszystkich zmiennych numerycznych, w wyniku czego otrzymano następujące zmienne:

- *internet_użytkownicy_przewidywana1* – szacowana wartość zmiennej *internet_użytkownicy* w modelu 1,
- *internet_użytkownicy_przewidywana2* – szacowana wartość zmiennej *internet_użytkownicy* w modelu 2,
- *Próba* - przypisanie do próby (0 – testowa, 1 - ucząca).

Oceń jakość szacowania zmiennej *internet_użytkownicy* w obu modelach, wykorzystując do tego MAE i MSE.