Łukasz Stępień

09.03.2023r.

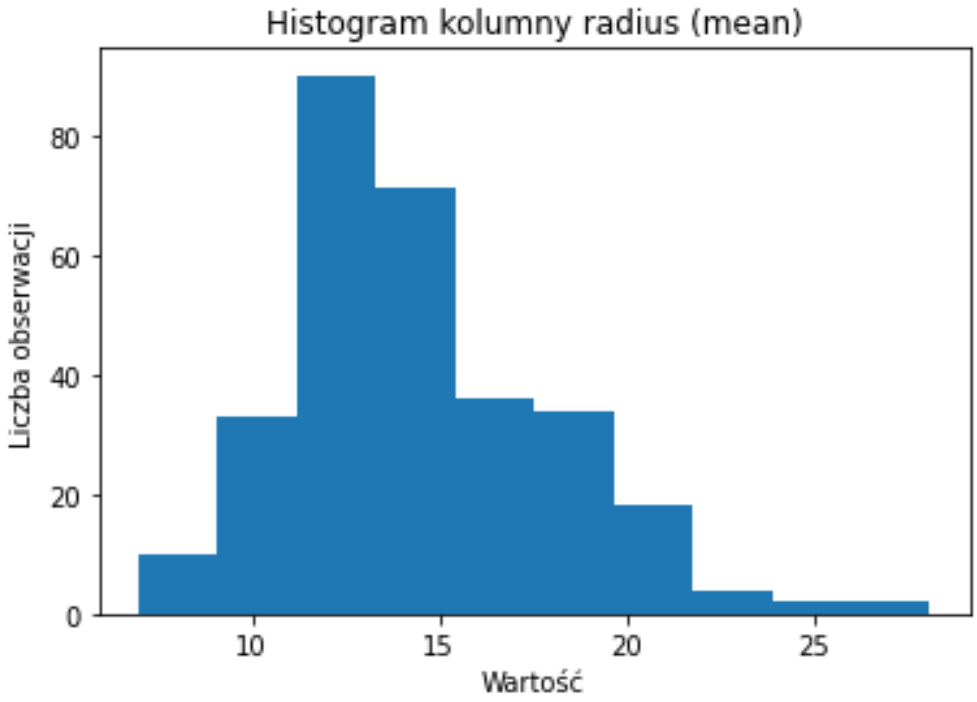
Laboratorium 2

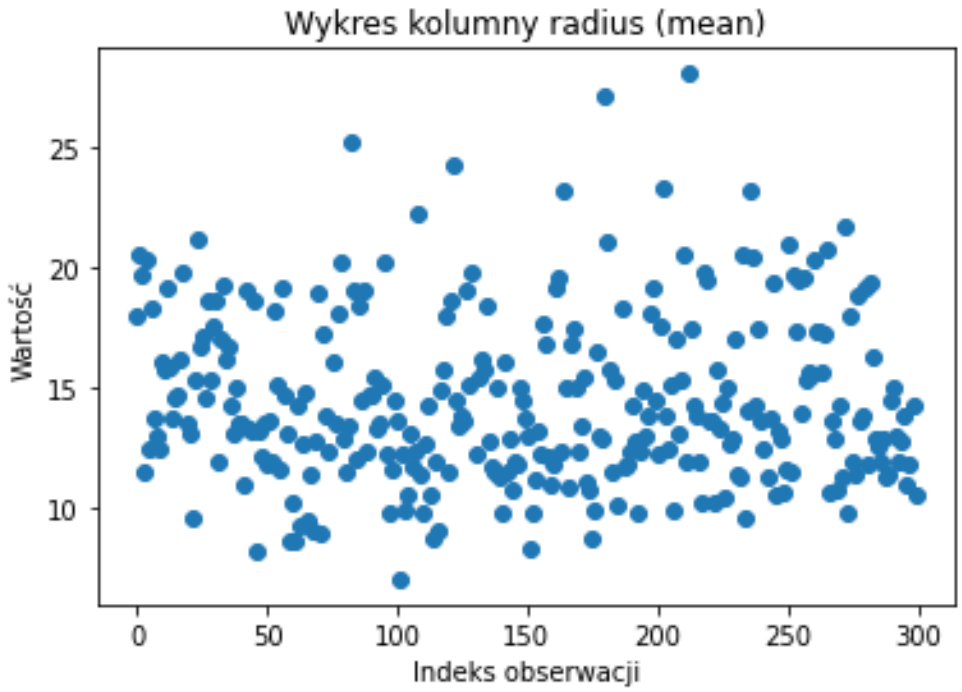
Metoda najmniejszych kwadratów

1. Temat zadania:

Zastosuj metodę najmniejszych kwadratów do predykcji, czy nowotwór jest złośliwy (ang. *malignant*) czy łagodny (ang. *benign*). Nowotwory złośliwe i łagodne maja różne charakterystyki wzrostu. Istotne cechy to m. in. promień i tekstura. Charakterystyki te wyznaczane są poprzez diagnostykę obrazowa i biopsje. Do rozwiązania problemu wykorzystaj bibliotekę pandas, typ DataFrame oraz dwa zbiory danych: breast-cancer-train.dat oraz breast-cancer-validate.dat.

2. Wykres i histogram kolumny „radius (mean)”:





3. Opis programu

Implementacja programu została podzielona na podpunkty, podobnie jak w oryginalnej treści zadania:

- a) Otwórz zbiory breast-cancer-train.dat i breast-cancer-validate.dat używając funkcji pd.io.parsers.read csv z biblioteki pandas,
- b) Stwórz histogram i wykres wybranej kolumny danych przy pomocy funkcji hist oraz plot,
- c) Stwórz reprezentacje danych zawartych w obu zbiorach dla liniowej i kwadratowej metody najmniejszych kwadratów,
- d) Stwórz wektor b dla obu zbiorów,
- e) Znajdź wagi dla liniowej oraz kwadratowej reprezentacji najmniejszych kwadratów,
- f) Oblicz współczynnik uwarunkowania,
- g) Sprawdź jak dobrze otrzymane wagi przewidują typ nowotworu.

4. Wyniki:

Współczynniki uwarunkowania:

Reprezentacja liniowa	1345082.9797889264
Reprezentacja kwadratowa	2671169.1557143712

Algorytm dla reprezentacji liniowej (po prawej stosunek ilości błędnych predykcji do liczby pacjentów):

Liczba przypadków fałszywie dodatnich	0	0,000%
Liczba przypadków fałszywie ujemnych	4	0,013%

Algorytm dla reprezentacji kwadratowej (po prawej stosunek ilości błędnych predykcji do liczby pacjentów):

Liczba przypadków fałszywie dodatnich	1	0,003%
Liczba przypadków fałszywie ujemnych	10	0,033%

5. Wnioski:

Program umożliwia poprawne wczytywanie plików .dat oraz tworzenie na ich podstawie histogramów oraz wykresów. Współczynnik uwarunkowania jest lepszy dla metody liniowej niż kwadratowej, lecz są tego samego rzędu wielkości (10^7). Jest on dosyć wysoki, więc zadanie jest słabo uwarunkowane. Jednak ostateczne wyniki można uznać za zadawalające. Błąd predykcji we wszystkich przypadkach nie przekracza 1%.