

STROKE PREDICTION



O udarach

Według Światowej Organizacji Zdrowia (WHO) udary są **drugą** w kolejności **najczęstszą przyczyną śmierci** i **trzecią prowadzącą do niepełnosprawności** oraz odpowiedzialną za **11% wszystkich zgonów**.

Nasza analiza ma na celu określenie czy pacjentowi grozi udar i pozwala stwierdzić, czy potrzebna jest dodatkowa diagnostyka, aby temu zapobiec.



Wielkość zbioru i zmienna celu, wybór metryk

Zbiór posiada **5510** rekordów i **11** kolumn.

Zmienną celu jest numeryczna kolumna **Stroke**, która przyjmuje wartości 0-1.

Wartość zmiennej celu (Stroke)	Liczebność wartości
0	4861 (95%)
1	249 (5%)

Metryki:

- f1-score + macierz pomyłek
- Recall
- Precision

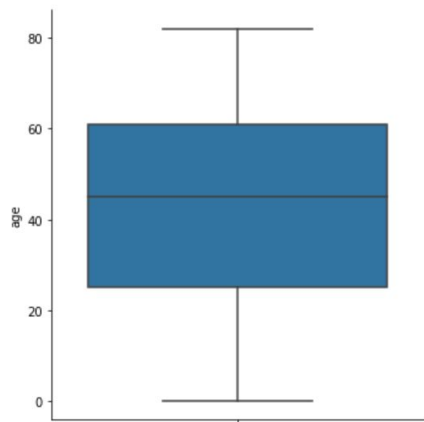


Zmienne niezależne i ich typy

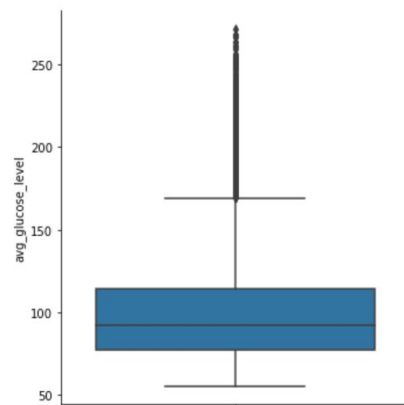
Nazwa zmiennej	typ	braki danych	min	max
gender	kategoryczna	-	-	-
age	ilościowa	-	<u>0.08</u>	<u>82</u>
hypertension	kategoryczna 0-1	-	-	-
heart_disease	kategoryczna 0-1	-	-	-
ever_married	kategoryczna	-	-	-
work_type	kategoryczna	-	-	-
Residence_type	kategoryczna	-	-	-
avg_glucose_level	ilościowa	-	55.12	271.74
bmi	ilościowa	201 / 4%	10.3	97.6
smoking_status	kategoryczna	-	-	-



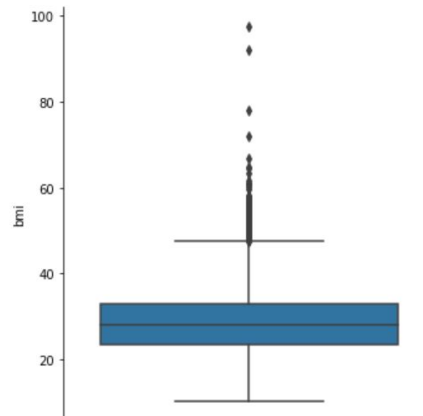
WIEK



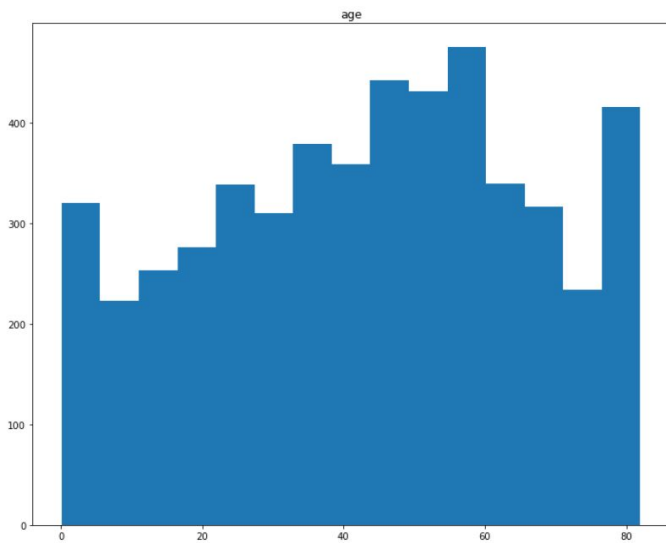
ŚREDNI POZIOM GLUKOZY



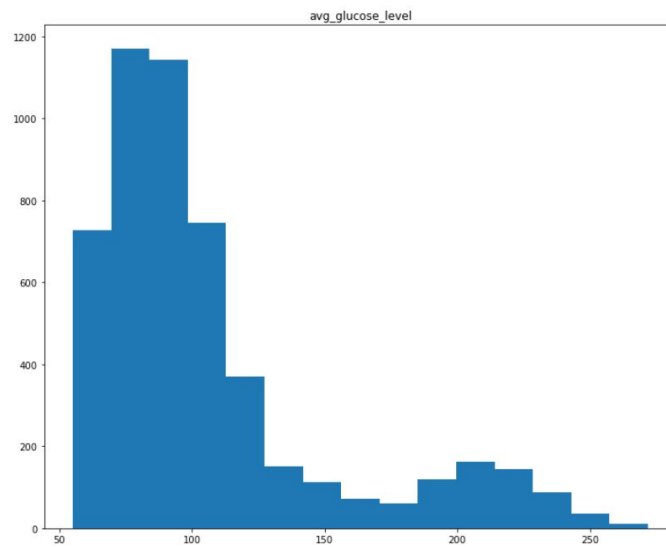
BMI



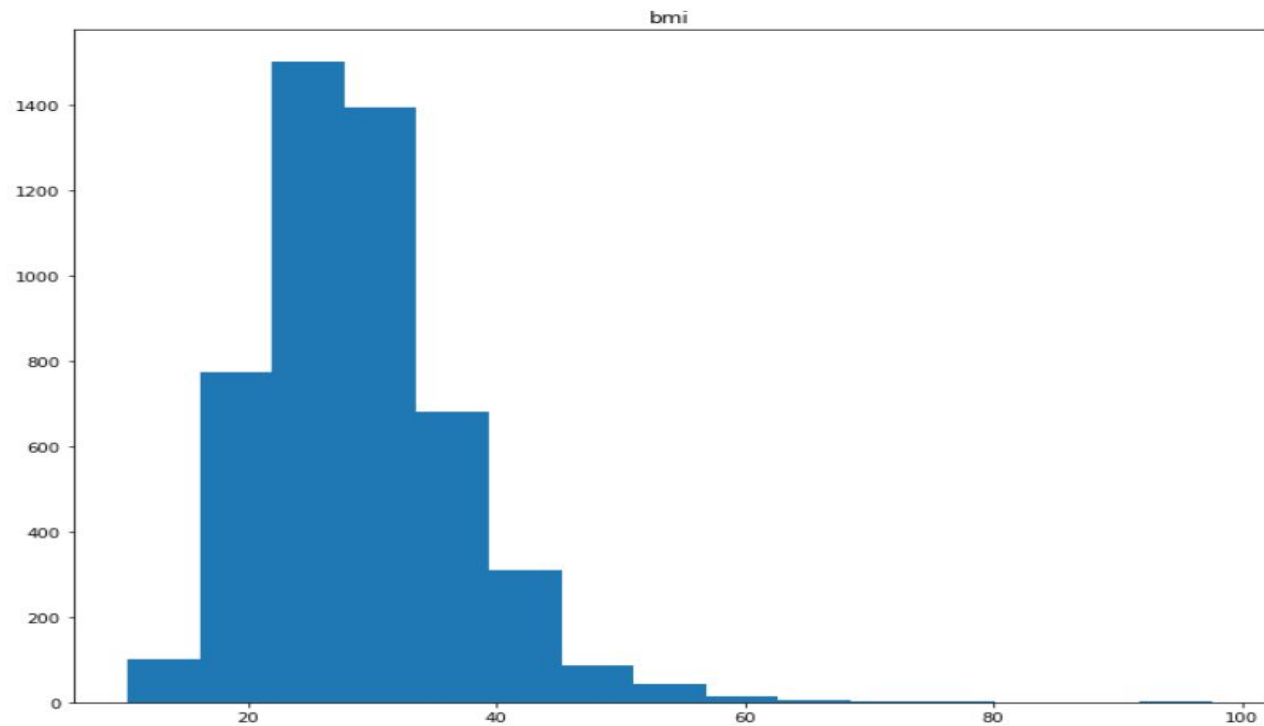
WIEK



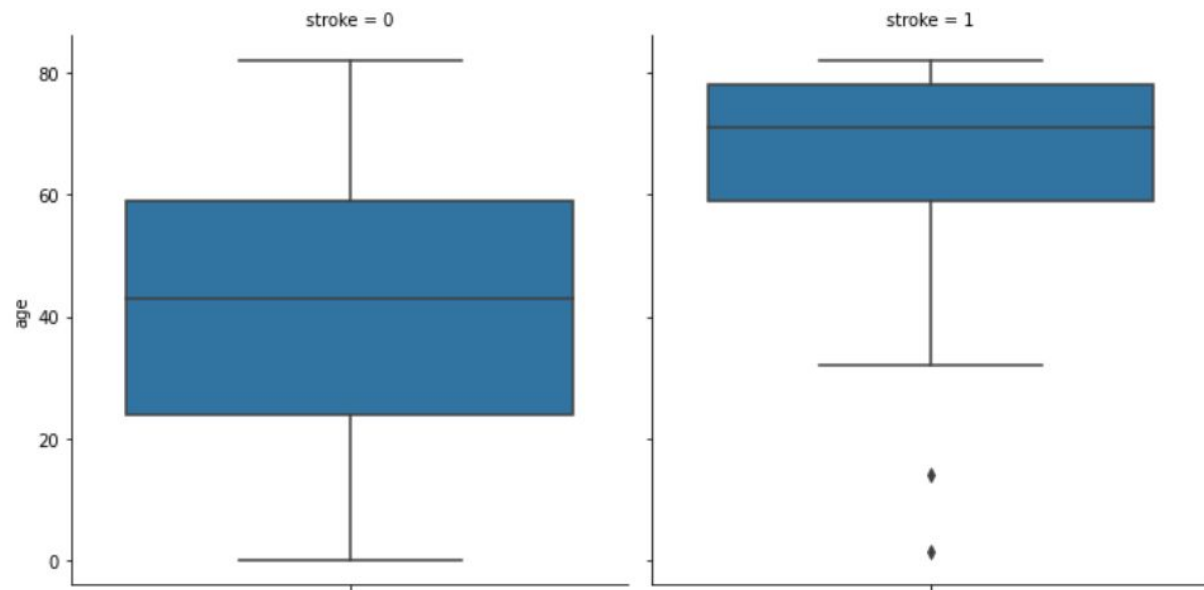
ŚREDNI POZIOM GLUKOZY



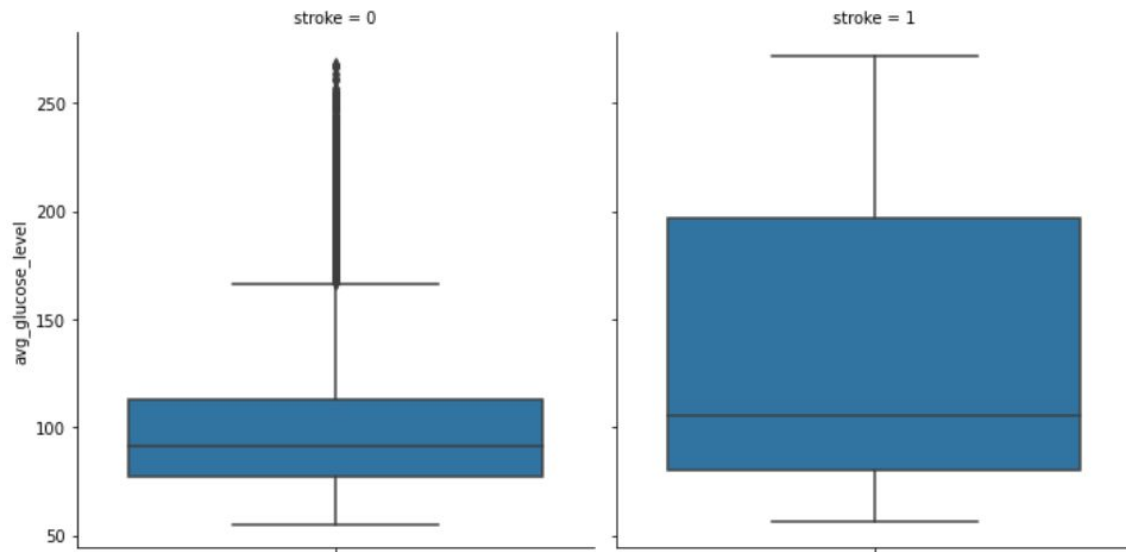
BMI



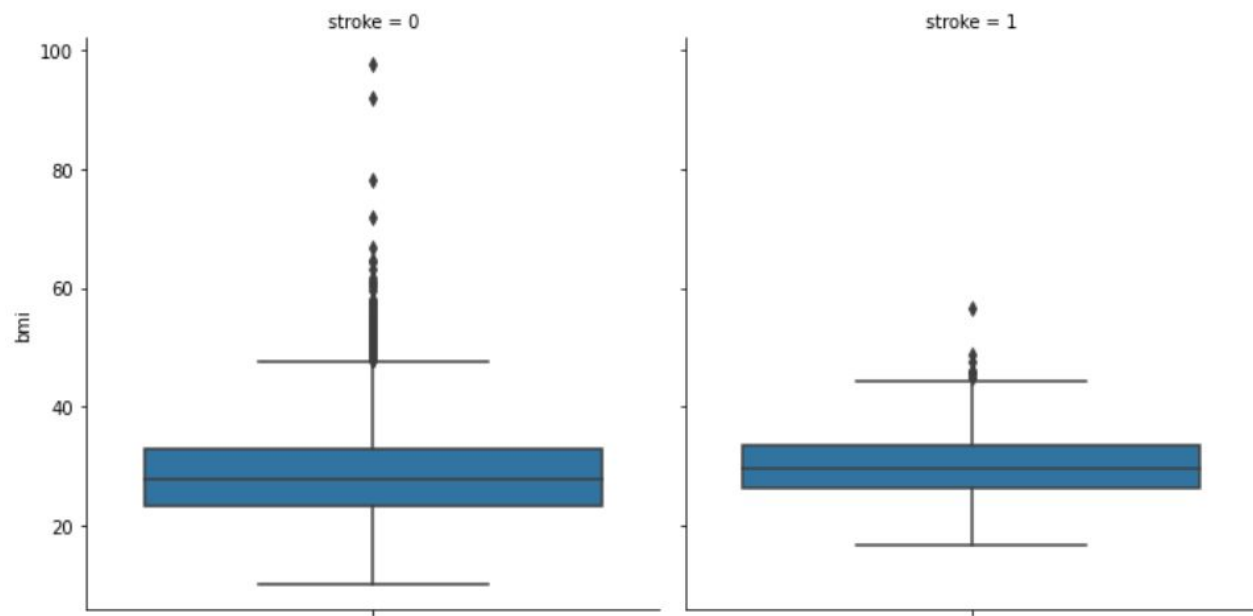
WIEK



ŚREDNI POZIOM GLUKOZY



BMI



Zmienne niezależne - inne spostrzeżenia

- **gender** - 1 wartość **Other**.
- **age** - do pierwszych 2 lat życia dane przedstawione są w miesiącach, przy czym brane jest, że 1 miesiąc = 0.08.
- **avg_glucose_level** - poziom jest zależny od tego, czy był mierzony po posiłku czy na czczo, jaką metodą był mierzony, a także normy różnią się u osób starszych, dzieci oraz kobiet w ciąży. Wartość ponad 200 mg/dL wskazuje na cukrzycę. Nasza największa wartość w zbiorze jest poniżej liczby, którą znaleźliśmy w źródłach, więc nasze dane uznaliśmy za poprawne.
- **bmi** - wartość powyżej 40 oznacza otyłość III stopnia. Dane uznaliśmy za prawidłowe, gdyż przykładowo osoba o wzroście 175 cm i wadze 300 kg posiada bmi 98 (zbliżone do maksymalnej wartości w naszym zbiorze), więc są to przypadki możliwe (najgrubszy człowiek świata, który trafił do KRG, ważył około 600 kg).



Zmienne niezależne - podsumowanie

Po analizie i w wyniku One-Hot-Encodingu zbiór ostatecznie zawierał **23** zmienne.

W wyniku usunięcia braków danych zbiór ostatecznie posiadał **4909** rekordów.

Dodatkowo zmienną **bmi** podzieliliśmy na kategorie i przeprowadziliśmy drugą turę uczenia najlepszych modeli, podmieniając zmienną ilościową bmi na porządkową.



Transformacja zmiennej bmi

BMI	Ryzyko chorób towarzyszących otyłości
< 18.5	minimalne, ale zwiększony poziom wystąpienia innych problemów zdrowotnych
< 25	minimalne
< 30	średnie
< 35	wysokie
< 40	bardzo wysokie
>= 40	ekstremalny poziom ryzyka

https://pl.wikipedia.org/wiki/Wska%C5%BAnik_masy_cia%C5%82a



Oversampling, Undersampling, SMOTE

Jako że nasz zbiór danych jest niezbalansowany zdecydowaliśmy się na użycie technik balansujących i przetestowanie ich skuteczności na naszym zbiorze:

- **Oversampling** - polega na stworzeniu kopii wierszy klasy rzadziej występującej
- **Undersampling** - polega na usunięciu wierszy klasy częściej występującej
- **SMOTE** - generuje syntetyczne dane dla klasy mniejszościowej poprzez minimalne zmiany wartości już istniejących punktów

Ze względu na małą liczbę jedynek w zbiorze zrezygnowaliśmy z Undersamplingu.



Standaryzacja zmiennych, redukcja wymiarów

Po zamianie zmiennych kategoriycznych na liczbowe dokonaliśmy **standaryzacji** zmiennych. Do wszystkich modeli użyliśmy tak przeskalowanych danych, również do drzew decyzyjnych czy lasów losowych, które nie potrzebują spełnienia tego założenia.

Ze względu na małą liczbę zmiennych pominęliśmy krok redukcji wymiarów i nie stosowaliśmy PCA.



Budowa modeli

Wykorzystaliśmy 6 algorytmów:

- Drzewo decyzyjne
- Las losowy
- XGBoost
- SVM
- KNN
- Regresja logistyczna

Dla każdego algorytmu wytypowaliśmy zestaw parametrów i **GridSearchem** wyznaczyliśmy najlepsze modele wg **f1-score** i **macierzy pomyłek** dla każdego z 2 sposobów zbalansowania zbioru (Oversampling, SMOTE).

Dodatkowo w regresji logistycznej w przypadku zmiennych dopełniających się (np. female, male) eliminowaliśmy z modelu jedną zmienną ze względu na założenie braku współliniowości zmiennych.



TOP 3 najlepszych modeli



XGBOOST

Dla danych z oversamplingu

Parametry modelu: max_depth: 3, booster: gbtree, learning_rate: 0.05 criterion: entropy, max_features: 3, n_estimators: 50, min_impurity_decrease: 0.3, min_samples_split: 100, min_samples_leaf: 50, min_child_weight: 20, reg_lambda: 0.1, reg_alpha: 0

	precision	recall	f1-score	support
0	0.99	0.74	0.84	929
1	0.15	<u>0.83</u>	<u>0.26</u>	53
accuracy			0.74	982
macro avg	0.57	0.78	0.55	982
weighted avg	0.94	0.74	0.81	982

[[684 245]
[9 44]]
ROC score: 0.783232122184536

Dla danych ze SMOTE

	precision	recall	f1-score	support
0	0.98	0.72	0.83	929
1	0.14	<u>0.79</u>	<u>0.23</u>	53
accuracy			0.72	982
macro avg	0.56	0.75	0.53	982
weighted avg	0.94	0.72	0.80	982

[[665 264]
[11 42]]
ROC score: 0.754138148140626

SVM

Dla danych z oversamplingu

```
Parametry modelu: C: 1, kernel: poly, gamma: auto, degree: 1
precision    recall  f1-score   support

   0         0.99    0.73    0.84     929
   1         0.15    0.85    0.26      53

 accuracy          0.74     982
 macro avg         0.57    0.79    0.55     982
weighted avg         0.94    0.74    0.81     982

[[678 251]
 [  8  45]]
ROC score: 0.7894368056542844
```

Dla danych ze SMOTE

```
precision    recall  f1-score   support

   0         0.98    0.72    0.83     929
   1         0.14    0.79    0.24      53

 accuracy          0.73     982
 macro avg         0.56    0.76    0.54     982
weighted avg         0.94    0.73    0.80     982

[[673 256]
 [ 11  42]]
ROC score: 0.7584438531998294
```




Regresja logistyczna

Dla danych z oversamplingu

	precision	recall	f1-score	support
0	0.99	0.74	0.84	929
1	0.15	<u>0.83</u>	<u>0.26</u>	53
accuracy			0.74	982
macro avg	0.57	0.78	0.55	982
weighted avg	0.94	0.74	0.81	982

[[685 244]
[9 44]]

ROC score: 0.7837703353169365




Dla danych ze SMOTE

	precision	recall	f1-score	support
0	0.98	0.75	0.85	929
1	0.15	<u>0.79</u>	<u>0.25</u>	53
accuracy			0.75	982
macro avg	0.57	0.77	0.55	982
weighted avg	0.94	0.75	0.82	982

[[694 235]
[11 42]]

ROC score: 0.7697463289802384





Najlepsze modele

Po zmianie 'bmi' na zmienna kategoryczną dwa modele poprawiły nieco swoje wyniki.

XGBoost (oversampling)

Parametry modelu: max_depth: 5, booster: gbtree, learning_rate: 0.05 criterion: entropy, max_features: 3, n_estimators: 50, min_impurity_decrease: 0.3, min_samples_split: 100, min_samples_leaf: 50, min_child_weigh: 20

	precision	recall	f1-score	support
0	0.98	0.79	0.88	929
1	0.17	<u>0.75</u>	<u>0.28</u>	53
accuracy			0.79	982
macro avg	0.58	0.77	0.58	982
weighted avg	0.94	0.79	0.85	982

training_time = 0.177

[[737 192]
[13 40]]
ROC score: 0.7740215691451551



Regresja logistyczna (oversampling)

	precision	recall	f1-score	support
0	0.99	0.74	0.85	929
1	0.16	<u>0.85</u>	<u>0.27</u>	53
accuracy			0.75	982
macro avg	0.57	0.79	0.56	982
weighted avg	0.94	0.75	0.82	982

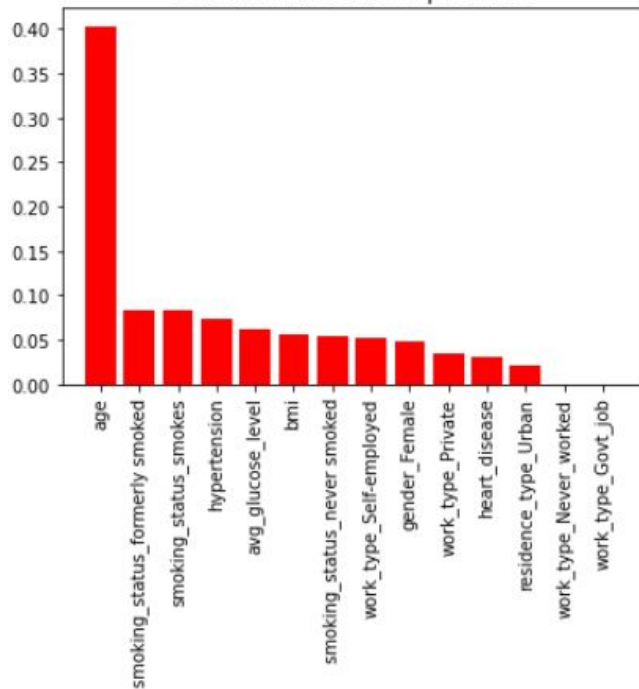
training_time = 0.029

[[688 241]
[8 45]]
ROC score: 0.7948189369782886



Najważniejsze zmienne

XGBClassifier feature importances



Regresja logistyczna - features coefficients:

age 1.808 —
hypertension 0.177 —
heart_disease 0.036
avg_glucose_level 0.156
bmi 0.167
gender_Female -0.018
work_type_Govt_job -0.38 —
work_type_Never_worked -0.298 —
work_type_Private -0.55 —
work_type_Self-employed -0.586 —
residence_type_Urban 0.028
smoking_status_formerly smoked 0.061
smoking_status_never smoked -0.015
smoking_status_smokes 0.139



DZIEKUJEMY ZA UWAGĘ

