



Politechnika  
Wrocławska

Wydział Informatyki i Telekomunikacji

Semestr letni 2023/2024

Uczenie maszynowe PN 11:15-13:00

# Łączenie technik Under-Sampling i Over-Sampling dla zbioru danych niezbalansowanych

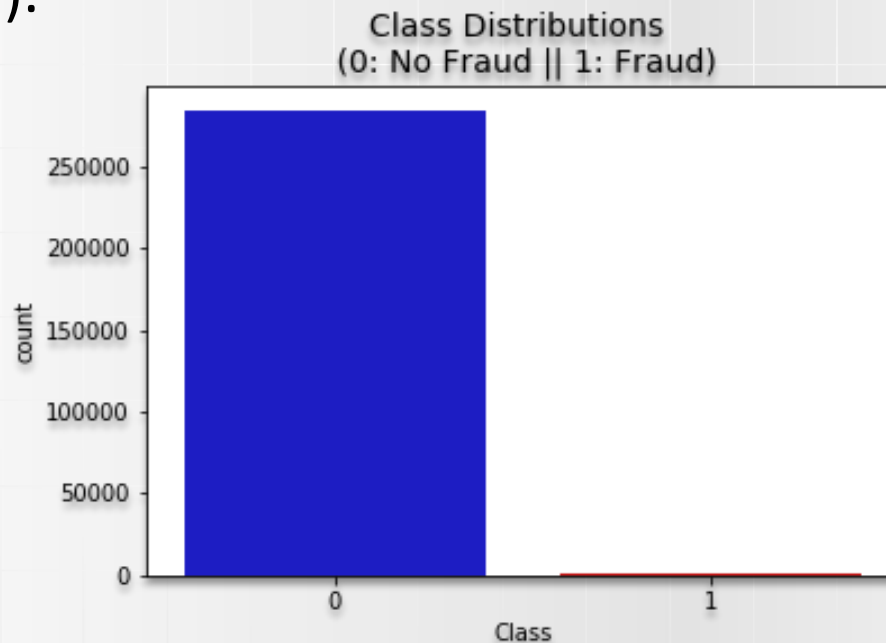


HR EXCELLENCE IN RESEARCH

Weronika Belniak 249048

# Dane niezbalansowane

Dane są niezbalansowane jeśli klasy nie są w przybliżeniu równo liczne (klasa mniejszościowa zawiera wyraźnie mniej przykładów niż inne klasy).

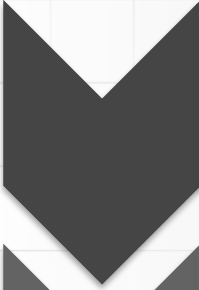


Nieuczciwe transakcje (fraudy)

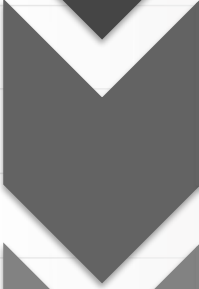
Medycyna (rzadkie choroby)

Spam

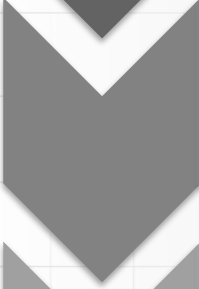
# Dane niezbalansowane



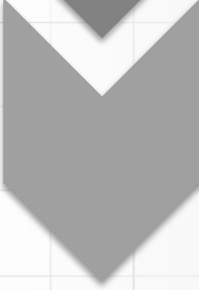
Delikatnie niezbalansowany zbiór  
jedna klasa jest pomiędzy 20% a 40%



Umiarkowanie niezbalansowany zbiór  
jedna klasa jest pomiędzy 5% a 20%

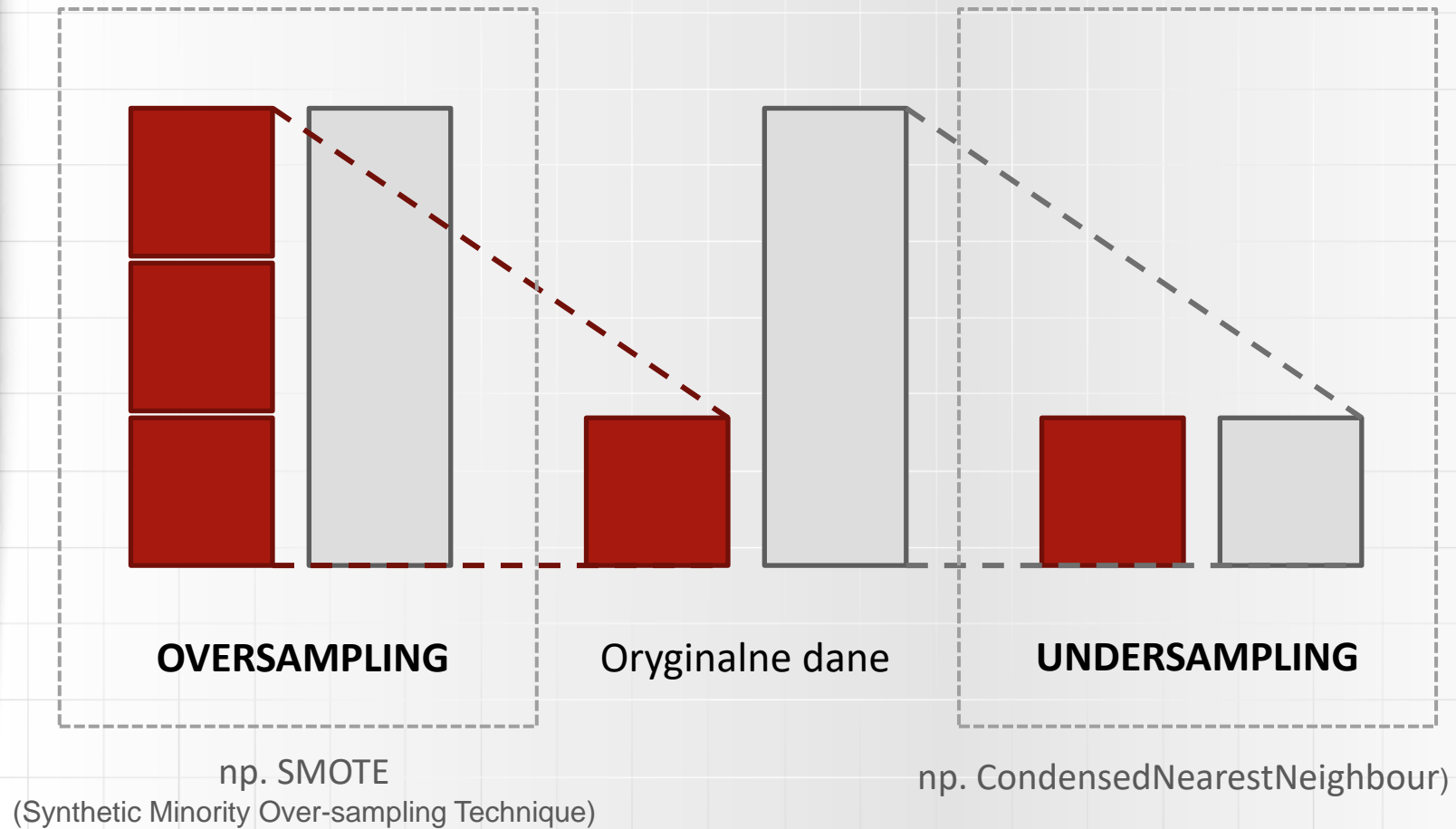


Silnie niezbalansowany zbiór  
jedna klasa między 0.1% a 5%



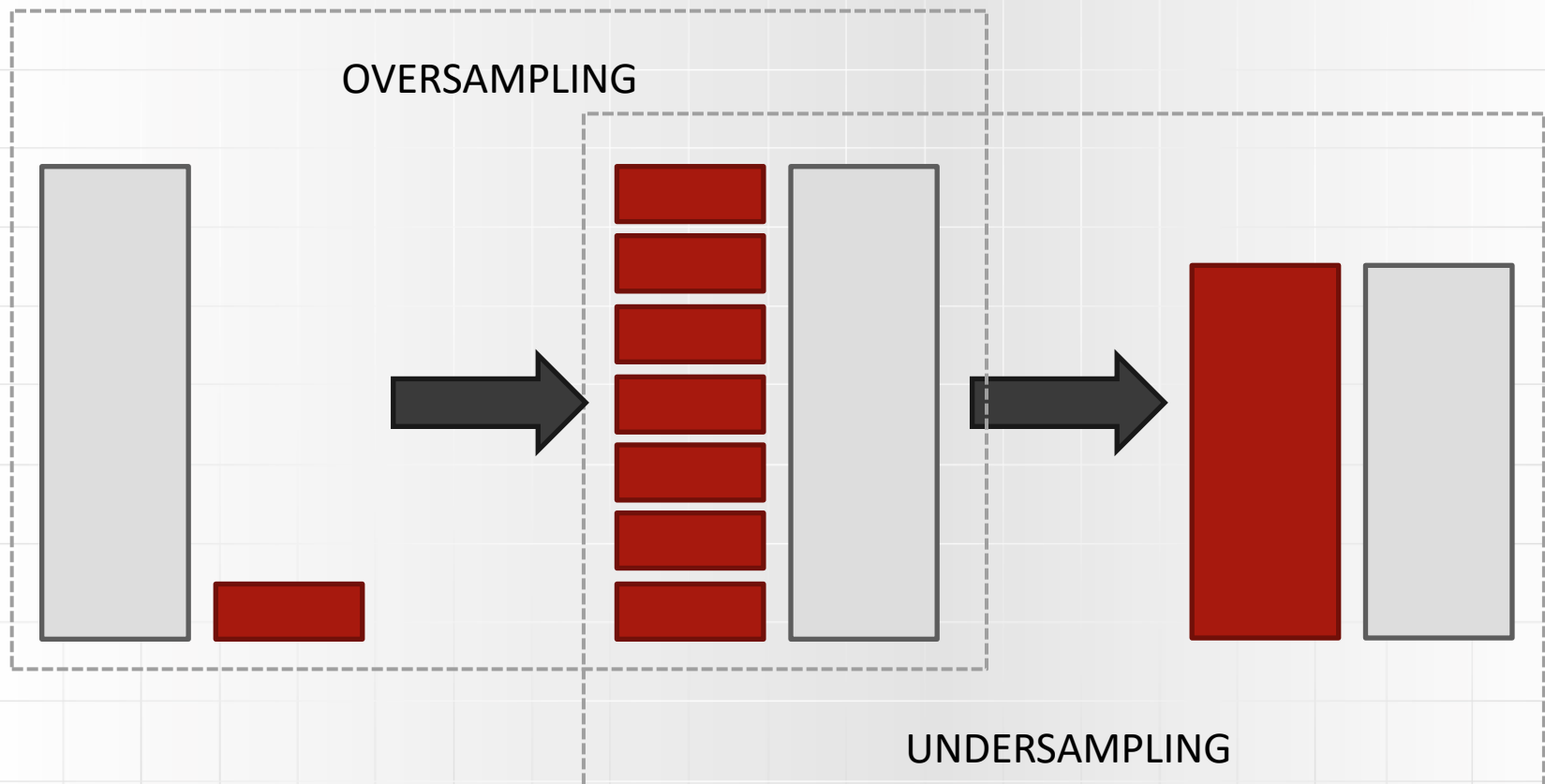
Ekstremalnie niezbalansowany zbiór  
jedna klasa poniżej 0.1% (czyli mniej niż 1 na 1000)

# Undersampling i Oversampling



# Undersampling i Oversampling

- łączenie obu technik



# Źródła

1. Le, Tuong & Vo, Minh & Vo, Bay & Lee, Mi & Baik, Sung (2019) „A Hybrid Approach Using Oversampling Technique and Cost-Sensitive Learning for Bankruptcy Prediction”
2. Mohammed, Roweida & Rawashdeh, Jumanah & Abdullah, Malak (2020) „Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results”
3. <https://imbalanced-learn.org/stable/references/combine.html>\*
4. <https://eqibuana.medium.com/how-to-deal-with-imbalanced-data-in-classification-tasks-1046e5be0e0>\*
5. <https://mirosławmamczur.pl/niezbilansowane-dane/>\*
6. <https://hersanyagci.medium.com/random-resampling-methods-for-imbalanced-data-with-imblearn-1fbba4a0e6d3>\*
7. <https://www.mastersindatascience.org/learning/statistics-data-science/undersampling/>\*