Urząd Statystyczny
w Opolu

# NOWOCZESNE TECHNOLOGIE W BADANIU CEN TOWARÓW I USŁUG KONSUMPCYJNYCH

Tomasz Pietras, Związek Marcin

**Urząd Statystyczny w Opolu**

23-24.03.2023 Warszawa

# Plan prezentacji

Pozyskiwanie danych metodą webscrapingu – ceny wyrobów farmaceutycznych

Wykorzystanie uczenia maszynowego – klasyfikacja COICOP

Wykorzystanie uczenia maszynowego – dane z bazy CEP

# Wyroby farmaceutyczne – COICOP 061101

4 apteki

12 kategorii

12 grup COICOP (8-cyfrowe)

Ponad 30 000 produktów

# Wyroby farmaceutyczne – COICOP 061101

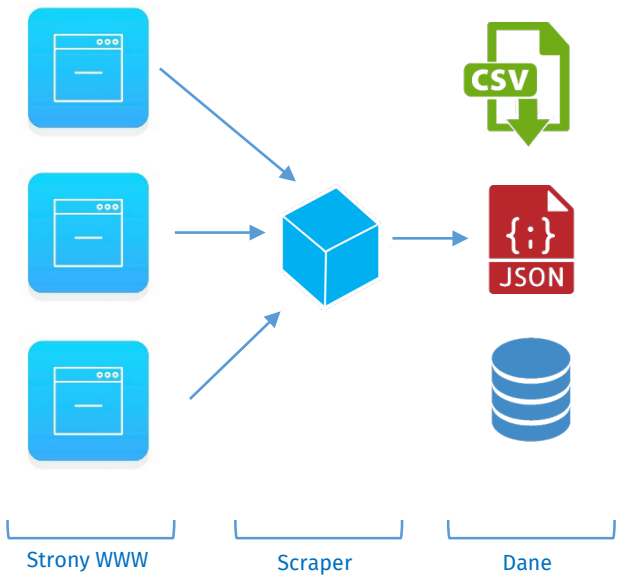
| SYMBOL | NAZWA | OBEJMUJE | NIE OBEJMUJE |
|---|---|---|---|
| 061101 | Wyroby farmaceutyczne | • leki, preparaty medyczne, specyfiki, surowice i szczepionki, witaminy i minerały, olej z wątroby dorsza i halibuta,<br>• odżywki dla sportowców (zalecone przez lekarza), produkty żywnościowe dla dzieci np. Nutramigen, Portagen, przepisane przez lekarza,<br>• zielarskie produkty farmaceutyczne, zioła lecznicze, parafarmaceutyki, olejki eteryczne (do aromaterapii),<br>• soki lecznicze: aloesowy, noni, sok brzozowy,<br>• doustne środki antykoncepcyjne,<br>• wodę utlenioną, płyn do soczewek, gumę antynikotynową, plastry nikotynowe, plastry rozgrzewające,<br>• zakup tlenu do butli,<br>• pozostałe. | • wyrobów weterynaryjnych (093421),<br>• artykułów do higieny osobistej, takich jak np. mydła medyczne (121321).<br><br>UWAGA:<br>Nie należy rejestrować leków nabytych na bezpłatne recepty. |

# Wyroby farmaceutyczne – COICOP 061101

Wyroby farmaceutyczne 061101

| | |
|---|---|
| Artykuły higieniczne | 06110101 |
| Dla diabetyków | 06110102 |
| Niestrawność | 06110103 |
| Odchudzanie | 06110104 |
| Probiotyki | 06110105 |
| Przeciwbólowe | 06110106 |
| Przeziębienie | 06110107 |
| Stawy, mięśnie, kości | 06110108 |
| Uspokajające | 06110109 |
| Wątroba | 06110110 |
| Witaminy i minerały | 06110111 |
| Serce | 06110112 |

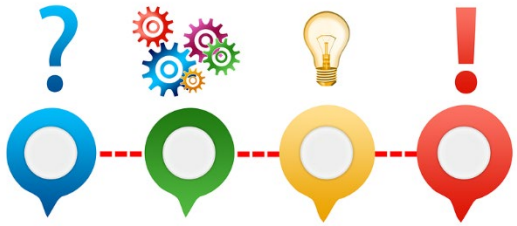
# Roboty do pobierania danych ze stron internetowych




- Częstotliwość pobierania danych
- Liczba robotów
- Rozwiązania techniczne związane z pobieraniem danych
- Oczyszczenie zbioru danych
- Problemy z danymi scrapowanymi

# Webscraping - pozyskiwanie danych

- Wybór metody
- Zastosowanie
- Sposób działania
- Zalety i Wady



# Wykorzystanie ML w danych scrapowanych

Uczenie maszynowe przypisuje symbole Coicop na podstawie wyuczonego wzorca

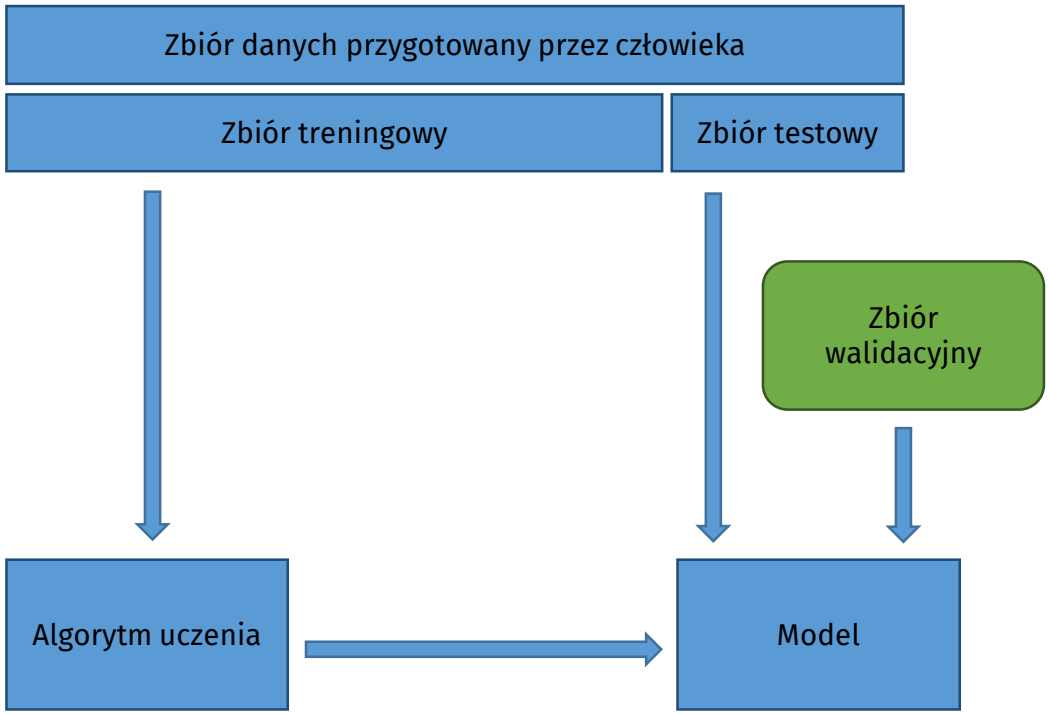
# Wektoryzacja

- Do pracy z danymi tekstowymi trzeba zastosować przekształcenie danych w macierz.
- Wektoryzacja to stworzenie macierzy w której poszczególne słowa stanowią nazwy kolumn, a wartościami są cyfry 1,0 informujące o występowaniu słowa w symbolu coicop. Każdy wiersz to indywidualna nazwa produktu
- Operacja ta pomaga komputerowi nauczeniu się wzorca produktu oraz przypisania mu prawidłowego coicopu

| R | S | T | U | V | W | X | Y | Z | AA |
|---|---|---|---|---|---|---|---|---|---|
| ml | mobilat | multiwitamina | multiwitaminar | nabłyszczający | natura | olejek | olejek odziei | optima | panty |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | | | | | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | | | | | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| | | | | | | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | | | | | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | | | | | | | | |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | | | | | | | | |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| | | | | | | | | | |
| 0 | 0 | 1 | | 0 | 0 | 0 | 0 | 0 | 0 |

# Przygotowanie danych do ML

- Po pobraniu danych zbiór był przygotowywany ręcznie – nadano mu symbole coicop ręcznie była to jednorazowa operacja. Zbiór posiadał około 10 tysięcy rekordów. Na dzień dzisiejszy zbiór ten zawiera 38 tysięcy produktów.
- Zbiór ten został podzielony na 2 części:
- Zbiór uczący liczący 30 tysięcy produktów
- Zbiór testowy liczący 8 tysięcy produktów
- Zbiór walidacyjny jest to zbiór danych które zostały pobrane podczas innego scrapowania.



# Wyniki algorytmów

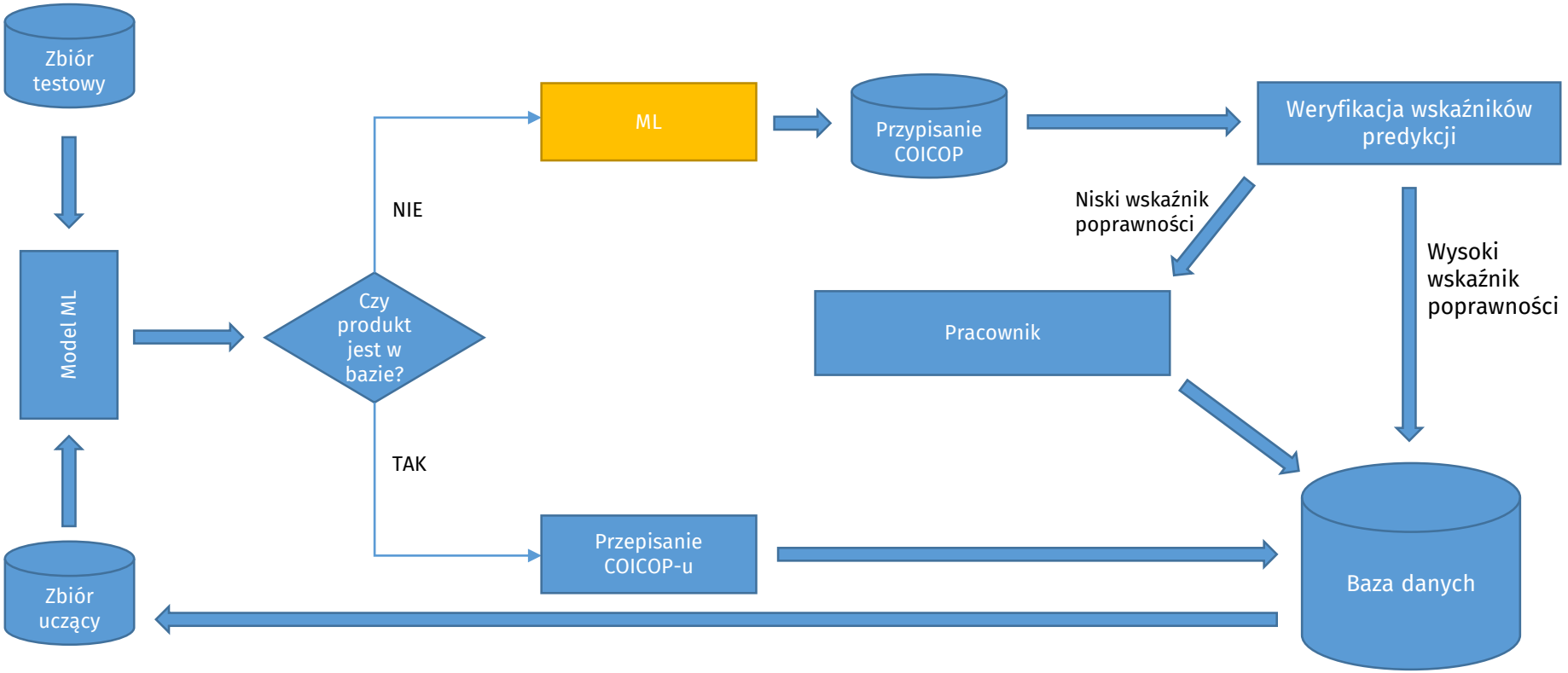
## Logistyczna Regresja

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 000000 | 0.96 | 0.95 | 0.95 | 4637 |
| 06110101 | 0.90 | 0.79 | 0.84 | 243 |
| 06110102 | 0.85 | 0.73 | 0.78 | 321 |
| 06110103 | 0.79 | 0.65 | 0.71 | 271 |
| 06110104 | 0.78 | 0.63 | 0.70 | 632 |
| 06110105 | 0.96 | 0.86 | 0.91 | 608 |
| 06110106 | 0.93 | 0.94 | 0.93 | 482 |
| 06110107 | 0.93 | 0.90 | 0.91 | 1539 |
| 06110108 | 0.86 | 0.82 | 0.84 | 1277 |
| 06110109 | 0.83 | 0.73 | 0.78 | 485 |
| 06110110 | 0.93 | 0.73 | 0.82 | 437 |
| 06110111 | 0.81 | 0.93 | 0.87 | 4611 |
| 06110112 | 0.76 | 0.67 | 0.72 | 537 |
| accuracy | | | 0.88 | 16080 |
| macro avg | 0.87 | 0.79 | 0.83 | 16080 |
| weighted avg | 0.88 | 0.88 | 0.88 | 16080 |

## Ranger

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 000000 | 0.91 | 0.96 | 0.93 | 4637 |
| 06110101 | 0.91 | 0.74 | 0.81 | 243 |
| 06110102 | 0.88 | 0.65 | 0.75 | 321 |
| 06110103 | 0.80 | 0.55 | 0.65 | 271 |
| 06110104 | 0.76 | 0.53 | 0.63 | 632 |
| 06110105 | 0.93 | 0.81 | 0.86 | 608 |
| 06110106 | 0.93 | 0.88 | 0.90 | 482 |
| 06110107 | 0.91 | 0.88 | 0.89 | 1539 |
| 06110108 | 0.87 | 0.77 | 0.81 | 1277 |
| 06110109 | 0.88 | 0.55 | 0.68 | 485 |
| 06110110 | 0.93 | 0.64 | 0.76 | 437 |
| 06110111 | 0.78 | 0.94 | 0.85 | 4611 |
| 06110112 | 0.85 | 0.56 | 0.67 | 537 |
| accuracy | | | 0.85 | 16080 |
| macro avg | 0.87 | 0.73 | 0.78 | 16080 |
| weighted avg | 0.86 | 0.85 | 0.85 | 16080 |

# Uczenie maszynowe



# Uczenie maszynowe – dane z bazy CEP

Samochody osobowe nowe – COICOP  071111
Samochody osobowe używane – COICOP 071121

Ustalanie próby oraz wyliczanie wag – na podstawie danych dotyczących transakcji samochodów osobowych z bazy Centralnej Ewidencji Pojazdów.

Wady?

| | | | |
|---|---|---|---|
| VOLKSKSWAGEN | VOLKSWAGEN/ | VOLKSWAGEN-VW | VOLKWAGEN |
| VOLKSWAGEN (D) | VOLKSWAGEN, VW | VOLKSEAGEN | VOLKSWAGET |

# Uczenie maszynowe – dane z bazy CEP

iMōto 2.0 Wskaźniki Słowniki

CENY
- Dane jednostkowe
- Upload

TABLICE KONTROLNE
- Ilość ofert
- Min-Max
- Puste oferty

CEP
- Dane jednostkowe
- Kosz

## CEP: Dane

Edytuj zaznaczone | Usuń zaznaczone

| Kwartał | Rok | Marka | Marka poprawiona | Model | Model poprawiony |
|---|---|---|---|---|---|
| | | | JEEP | | |
| III Kwartał | 2022 | CHRYSLER | JEEP 76,15 | GRAND CHEROKEE | GRAND CHEROKEE 94,98 |
| III Kwartał | 2022 | CHRYSLER | JEEP 76,15 | GRAND CHEROKEE | GRAND CHEROKEE 94,98 |
| III Kwartał | 2022 | CHRYSLER | JEEP 76,15 | GRAND CHEROKEE | GRAND CHEROKEE 94,98 |
| III Kwartał | 2022 | CHRYSLER | JEEP 92,10 | JEEP CHEROKEE | CHEROKEE 52,75 |
| III Kwartał | 2022 | CHRYSLER | JEEP 92,10 | JEEP CHEROKEE | CHEROKEE 52,75 |
| III Kwartał | 2022 | CHRYSLER | JEEP 88,86 | JEEP GRAND CHEROKEE | GRAND CHEROKEE 88,17 |
| III Kwartał | 2022 | CHRYSLER | JEEP 67,36 | JEEP WRANGLER | WRANGLER 8,57 |
| III Kwartał | 2022 | DAIMLERCHRYSLER | JEEP 52,02 | GRAND CHEROKEE | GRAND CHEROKEE 94,98 |
| III Kwartał | 2022 | DAIMLERCHRYSLER | JEEP 83,65 | JEEP CHEROKEE | CHEROKEE 52,75 |

Wyświetlono: 100, Odfiltrowano: 17333, Zaznaczono: 0, Wszystkich: 2189659

# Dziękujemy za uwagę