Odkrywanie związków w danych wielowymiarowych

Wprowadzenie

Marzena Nowakowska
Wydział Zarządzania i Modelowania Komputerowego
Politechnika Świętokrzyska

Uczenie maszynowe, drążenie danych, odkrywanie wiedzy w danych

Modelowanie to tworzenie modeli układów lub zjawisk o różnym charakterze (fizycznych, gospodarczych, socjologicznych) służące celom badawczym lub naukowym - SJP.

Uczenie maszynowe – budowanie modeli analitycznych z wykorzystaniem specjalistycznego oprogramowania, tzn. takich które można stosować w praktyce wykorzystując komputery.

Drążenia danych (*data mining*) łączy modelowanie klasyczne (wzory matematyczne) i nieklasyczne (**uczenie maszynowe**) obejmując zaawansowane metody do eksploracji i modelowania związków w dużych zbiorach danych.

Całość takich działań badawczych określana jest pojęciem **KDD** - **odkrywania wiedzy w danych** (*konwledge discovery in data*) – 1989 (?).

data engineer → zbierania i przygotowanie danych, eksploracja, budowanie modeli, interpretacja wyników i pozyskiwanie wiedzy ← data scientist

Klasyfikacja typów modeli

Predykcyjne (*predictive models*); przewidują wartość liczbową, np. różne modele regresji, regresyjne drzewa decyzyjne, SSN z nauczycielem, modele szeregów czasowych, SVR (Support Vector Regression).

Klasyfikacyjnych (*classification models*); przewidują wartość opisową, np. klasyfikacyjne drzewa decyzyjne, lasy losowe, SSN z nauczycielem, SVM (Support Vector Machine), regresja logistyczna, k-NN, naiwny klasyfikator Bayesowski).

Eksploracyjne / odkrywcze / opisowe (*exploratory / discovery / descriptive models*) nie przewidują wartości zmiennej; do tej grupy należą:

- modele analizy asocjacji
- modele analizy sekwencji
- modele klasteryzacji (analizy skupień)

Obszary zastosowań KDD

Biznes i marketing; segmentacja klientów, tworzenie profili zakupowych, wykrywanie zależności między produktami lub usługami

Finanse i bankowość; wykrywanie oszustw i anomalii w transakcjach, scoring (ocena punktowa) klientów, prognozowanie cen akcji lub kursów walut

Medycyna i opieka zdrowotna; odkrywanie wzorców chorób, personalizacja terapii, odkrywanie zależności genetycznych (bioinformatyka)

Przemysł i produkcja; predykcyjne utrzymanie ruchu, monitorowanie jakości i wykrywanie wad produkcyjnych

Telekomunikacja i sieci; profilowanie użytkowników sieci, prognozowanie obciążenia sieci, wykrywanie nadużyć i nieautoryzowanego dostępu

Bezpieczeństwo i cyberbezpieczeństwo; wykrywanie prób włamań i ataków, identyfikacja wzorców złośliwego oprogramowania

Administracja publiczna i sektor publiczny; wzorce przestępczości, prognozowanie zagrożeń, analiza danych demograficznych i społecznych

Internet, media, serwisy online; analiza sentymentu w opiniach w mediach społecznościowych, analiza treści w portalach, wykrywanie botów i nadużyć

Transport i logistyka; planowanie zasobów, wczesna wykrywanie opóźnień i ryzyk w logistyce, segmentacja flot i klientów transportowych

Aplikacje algorytmów KDD

Techniki KDD	Typowe zastosowania
Klasteryzacja (analiza skupień)	Segmentacja klientów, Analiza genomów, Grupowanie dokumentów
Klasyfikacja	Ocena ryzyka kredytowego, Diagnoza medyczna Rozpoznawanie obrazu lub tekstu
Regresja	Prognozowanie sprzedaży, Predykcja cen, Szacowanie zapotrzebowania na energię
Redukcja wymiarowości	Wizualizacja danych wielowymiarowych, Przyspieszanie uczenia maszynowego, Eliminacja szumu w danych
Analiza sieci i grafów	Analiza sieci społecznościowych, Analiza powiązań w sieciach przestępczych
Uczenie głębokie (Deep Learning)	Rozpoznawanie obrazów i dźwięku, NLP (analiza tekstu)
Analiza tekstu i tekst mining	Analiza opinii (sentiment analysis), Klasyfikacja dokumentów, Wyszukiwanie informacji
Analiza szeregów czasowych	Prognozowanie popytu, Predykcja cen akcji, Wykrywanie sezonowości

Oprogramowanie KDD

Komercyjne narzędzia KDD

- SAS Enterprise Miner / SAS Viya
- IBM SPSS Modeler
- RapidMiner Studio (wersja komercyjna)
- Alteryx Designer
- Microsoft Azure Machine
- Learning Studio
- Google Vertex AI
- AWS SageMaker
- Databricks
- Oracle Data Mining
- TIBCO Data Science

Free / open-source narzędzia KDD

- KNIME Analytics Platform
- WEKA
- Orange Data Mining
- RapidMiner (wersja Community)
- R (mlr3, caret, tidyverse)
- Python (scikit-learn, pandas, TensorFlow, PyTorch, etc.)
- H2O.ai
- Apache Spark MLlib
- SPMF
- ELKI

Ciekawostka nazewnicza

Specyficzny folklor terminologiczny:

drążenie danych (*Data Mining*)
wydobywanie wiedzy z danych (*Knowledge Discovery in Data*)
kopanie w danych (*Digging into Data*)
eksploracja danych (*Data Exploration*)

Inspiracja → angielski termin *mining* (wydobywanie, kopanie, eksploatowanie)

Mining → nie eksploatowanie złóż węgla (*coal mining*) lecz raczej wydobywanie złotego kruszcu (*gold mining*)

Eksploatowane złoża → **dane** Wydobywany szlachetny kruszec → **wiedza**

Jeden z poszukiwaczy z czasów kalifornijskiej gorączki złota z roku 1849 (ang. *Miner "forty niner"*), o którym m.in. opowiada piosenka *Oh My Darling, Clementine*. (Wikipedia)

In a cavern, in a canyon,	***W grocie, w kanionie,***
Excavating for a mine,	***Poszukując kruszcu (skarbu,***
Dwelt a miner forty-niner	***cennych minerałów),***
And his daughter,	***Żył górnik, 49-latek***
Clementine	***I jego córka Klementyna***

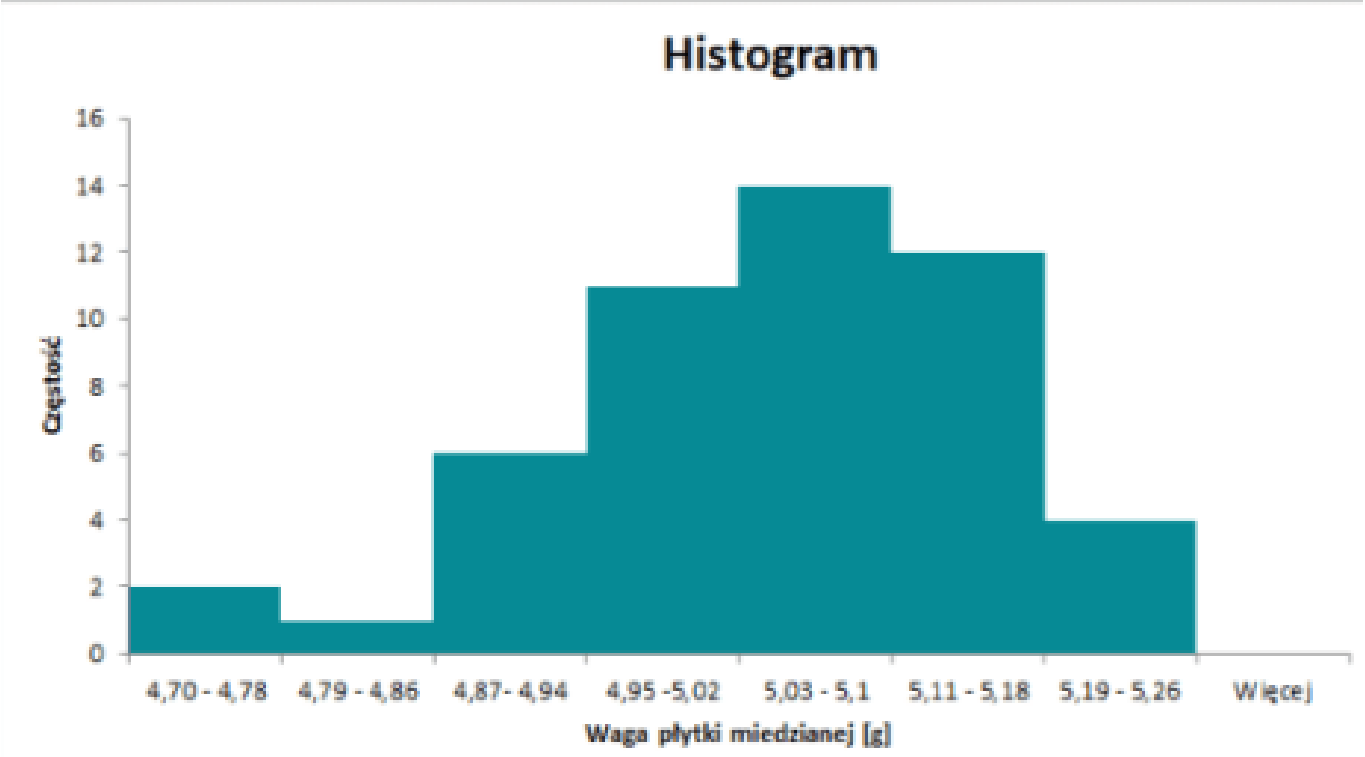
Wymiarowość danych

Wymiar danych to pojęcie z zakresu informatyki, statystyki, analizy danych i uczenia maszynowego; oznacza liczbę cech (atrybutów, kolumn, zmiennych) opisujących pojedynczy punkt danych.
Dane wielowymiarowe to dane opisane przez kilka atrybutów (co najmniej dwa).

Dane jednowymiarowe (x_i)

5,11
5,03
4,90
4,98
4,87
4,93
4,99
5,01
5,13
5,15
5,09
5,08
5,24
5,26
5,17
5,09
5,11
4,97
4,84
4,75
5,02

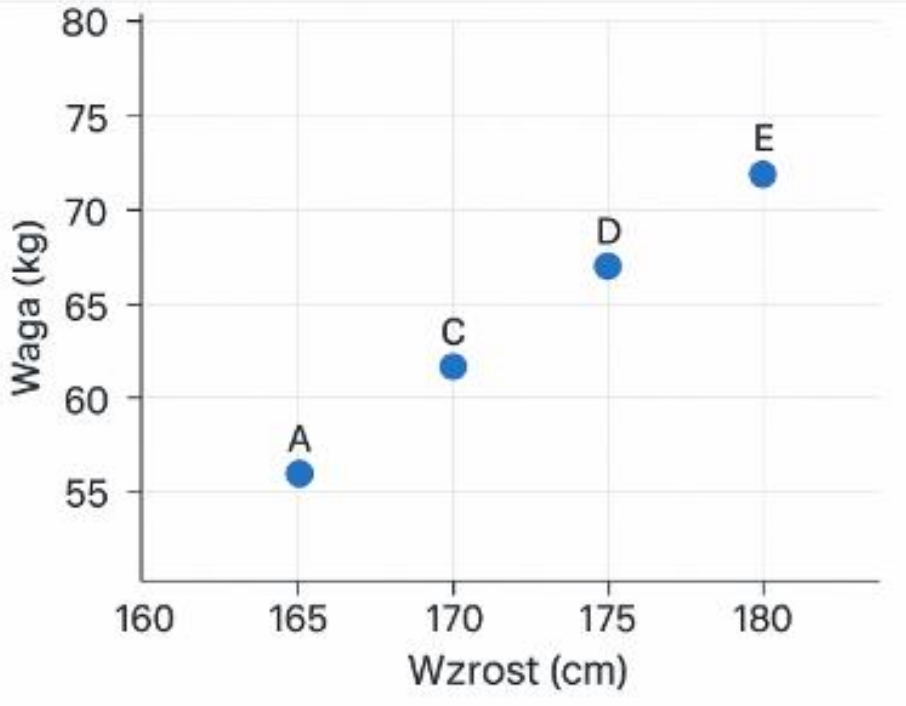
Wartości numeryczne mogą być odłożone na osi liczbowej z zachowaniem odległości, tekstowe – nie można określić ich położenia na osi



Dane dwuwymiarowe (x_i , y_i)

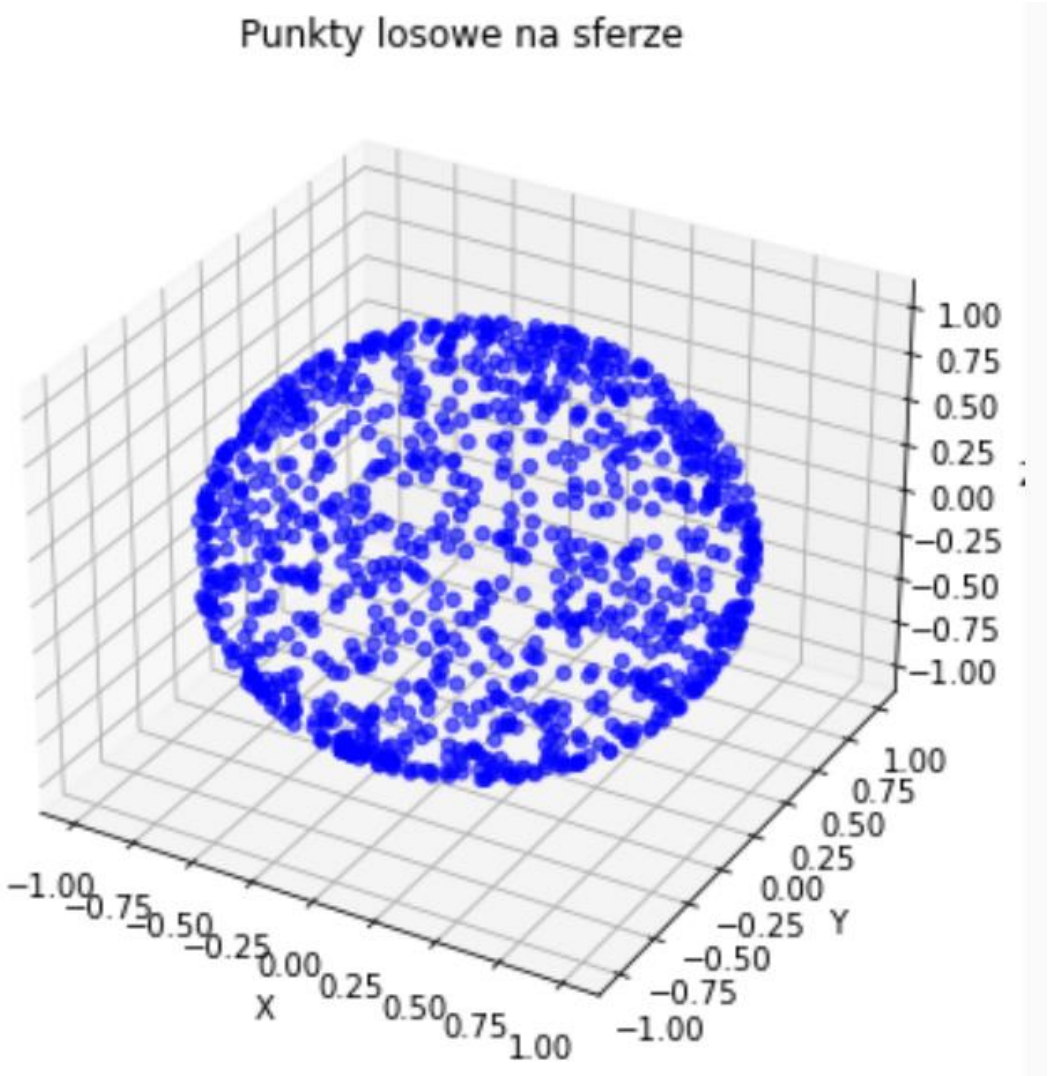
Numeryczne – odłożone na płaszczyźnie w układzie współrzędnych o dwóch osiach

Osoba	Wzrost (cm)	Waga (kg)
A	160	55
B	170	65
C	165	60
D	175	70
E	180	75



Dane trójwymiarowe (x_i , y_i , z_i)

Dane: (x, y, z) takie, że $x^2 + y^2 + z^2 = r^2$



Dane o wymiarze większym niż 3

W strukturze docelowej do analiz mają postać tabelaryczną.

Wiek	Lokalizacja	Zakupy	Czas_online [h]
23	Warszawa	elektronika, książki	3.5
35	Kraków	odzież, kosmetyki	2.0
42	Gdańsk	artykuły spożywcze, AGD	4.2
29	Wrocław	gry komputerowe, elektronika	5.1
51	Poznań	meble, sprzęt ogrodniczy	1.8
38	Lublin	książki, artykuły dziecięce	2.7
47	Katowice	kosmetyki, biżuteria	3.0
31	Szczecin	odzież sportowa, elektronika	4.8

Inne przykłady

- Wyniki badań medycznych (ciśnienie, cholesterol, BMI)
- Sensory IoT (czas, temperatura, ciśnienie, wilgotność, zapylenia, hałas)

Problemy

- Wysoka wymiarowość
- Redundancja informacji
- Zanieczyszczenia, nawet w danych eksperymentalnych
- Brak intuicji wizualnej (nie można danych zilustrować dla wymiaru > 3D)

System SAS

Statistical Analysis System – jedno z najważniejszych narzędzi do analizy danych w projektach biznesowych, badawczych lub przemysłowych, wykorzystywany przez profesjonalistów w organizacjach otoczenia społeczno-gospodarczego.

Najważniejsze cechy

- Zaawansowana analiza danych
- Zarządzanie i przetwarzanie danych (w tym import/eksport plików innych formatów)
- Wbudowany język programowania (4GL)
- Raportowanie i wizualizacja
- Bezpieczeństwo i zarządzanie użytkownikami
- Zastosowania korporacyjne
- Oferta gotowych rozwiązań („pod klucz”)

Zasoby systemu SAS

Biblioteki - organizowanie dostępu przez referencję

Pliki:

- zbiory danych (DATA, VIEW); *.*sas7bdat*
- zasoby typu ACCESS (dostęp do danych w formatach innych niż SAS)
- zbiory plików dostępne poprzez typ CATALOG (np. makra, definicje formatów); *.sas7bcat,
- programy w języku 4GL; *.sas

Biblioteki systemu SAS

Biblioteka (*Library*) – referencja identyfikująca folder na dysku. Foldery te zawierają zasoby tematyczne, najczęściej pliki z danymi SAS.

Wybrane biblioteki standardowe (wbudowane):

- SASHELP – zasoby do samodzielnego uczenia
- SASUSER – folder plików stałych użytkownika
- WORK – folder plików tymczasowych użytkownika; po zakończeniu sesji z SAS'em WORK jest czyszczony automatycznie

Niezależnie od tego użytkownik może tworzyć swoje własne biblioteki.

Organizacja dostępu – poprzez operację skojarzenie za pomocą menu, do którego sposób dostępu zależy od modułu SAS.

Nazwa biblioteki nie może przekraczać ośmiu znaków. Składa się z ciągu znaków zaczynających się od litery bez spacji i znaków specjalnych.

Dane systemu SAS

Zbiór z danymi w formacie systemu SAS mają rozszerzenie *sas7bdat*.
Są przechowywane w bibliotekach - dostęp referencyjny:

nazwa_biblioteki.nazwa_zbioru

Przykłady:

zbiór danych: SASUSER.**NKWDL** plik: **nkwdl**.sas7bdat
zbiór danych: SAS_Dane.**oceny** plik: **oceny**.sas7bdat

Struktura zbioru danych taka jak struktura tabeli w bazie danych - stąd inna nazwa zasobu w SAS: tabela. Dwa typy zasobów: DATA i VIEW.

Nazwa zbioru danych w SAS

- maksymalna długość: 32 znaki
- musi zaczynać się literą lub podkreśleniem
- może zawierać litery, cyfry i podkreślenia
- nie może zawierać spacji ani znaków specjalnych

Typy danych SAS

Dwa typy danych:

- **tekstowy** do 200 znaków: litery, cyfry i znaki specjalne,
- **liczbowy** w postaci zmiennoprzecinkowej na 8 bajtach (64 bity): cyfry, litera E dla notacji naukowej, kropka jako separator części ułamkowej, znak minus przed liczbą.

Dana typu **daty** jest reprezentowana jako liczba dni od 1960-01-01.

Dana typu **czasu** jest reprezentowane jako liczba sekund jaka upłynęła od północy do podanego czasu.

Co to jest SEMMA?

Sampling - próbkowanie

Exploration - wstępne wizualne lub statystyczne opracowanie danych

Manipulation - modyfikowanie i uaktualnianie danych

Modelling - modelowanie z wykorzystaniem EM, np.: sieci neuronowe, drzewa decyzyjne, techniki statystyczne, analizy szeregów czasowych

Assessment - ocena wyników i porównanie modeli

System SAS Enterprise Miner

SAS EM – sztandarowy produkt SAS (pakiet/moduł uczenia maszynowego)
SAS EG – SAS Enterprise Guide do prostszych analiz z intuicyjnym interfejsem
W obu przypadkach → **graficzne środowisko typu „drag and drop”**, dzięki czemu użytkownicy mogą tworzyć przepływy analityczne (tzw. *process flow diagrams*) bez konieczności pisania kodu.

- SAS EG tworzy jeden plik (*.*egp*)
- SAS EM tworzy cały pakiet plików i katalogów dla jednego projektu (*.*emp*).

SAS EM to potężne rozwiązanie wspierające proces podejmowania decyzji opartych na danych. Obsługuje różne techniki analizy danych, zgodnie z opracowana przez SAS Institute metodologią SEMMA.

Środowisko SAS EM

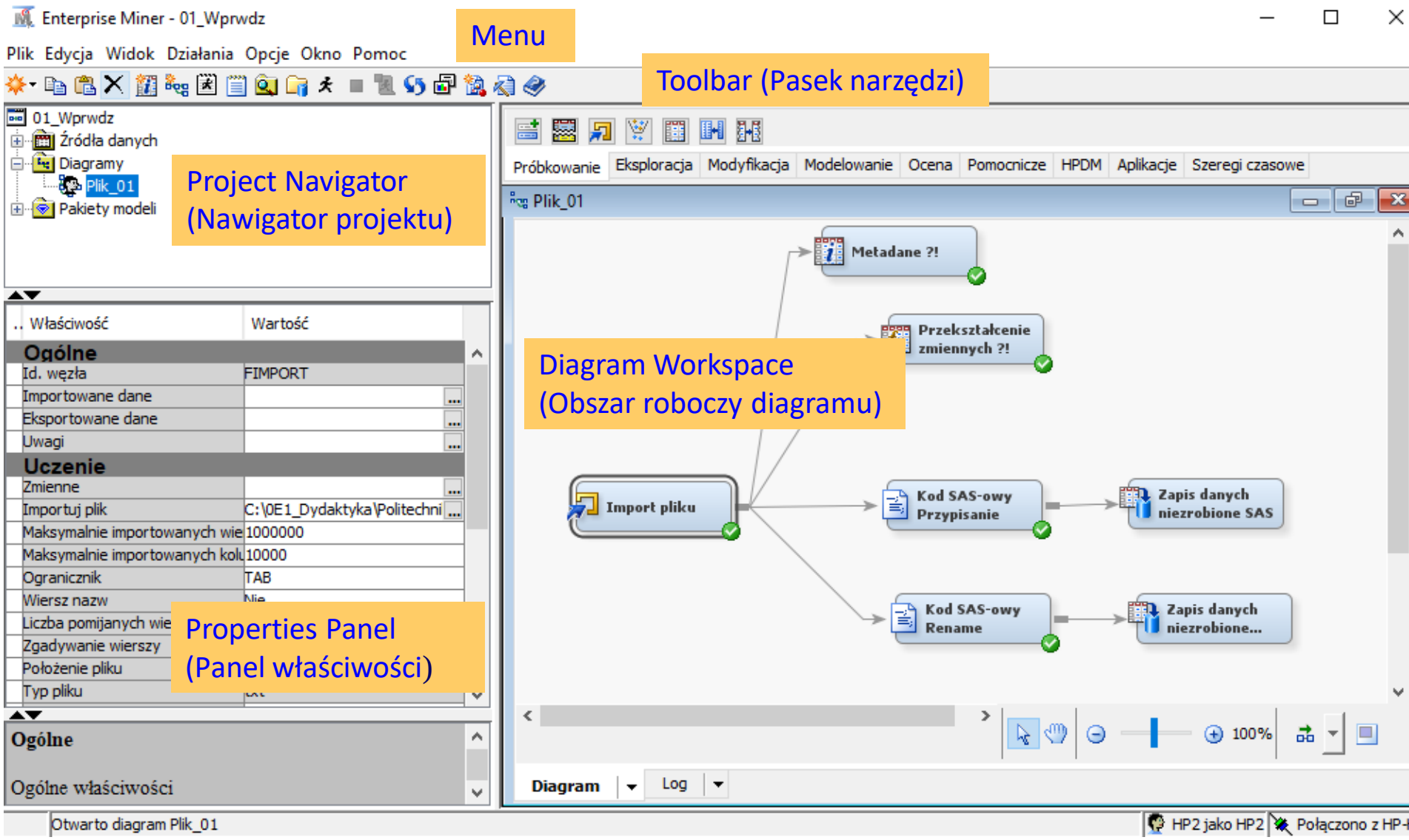
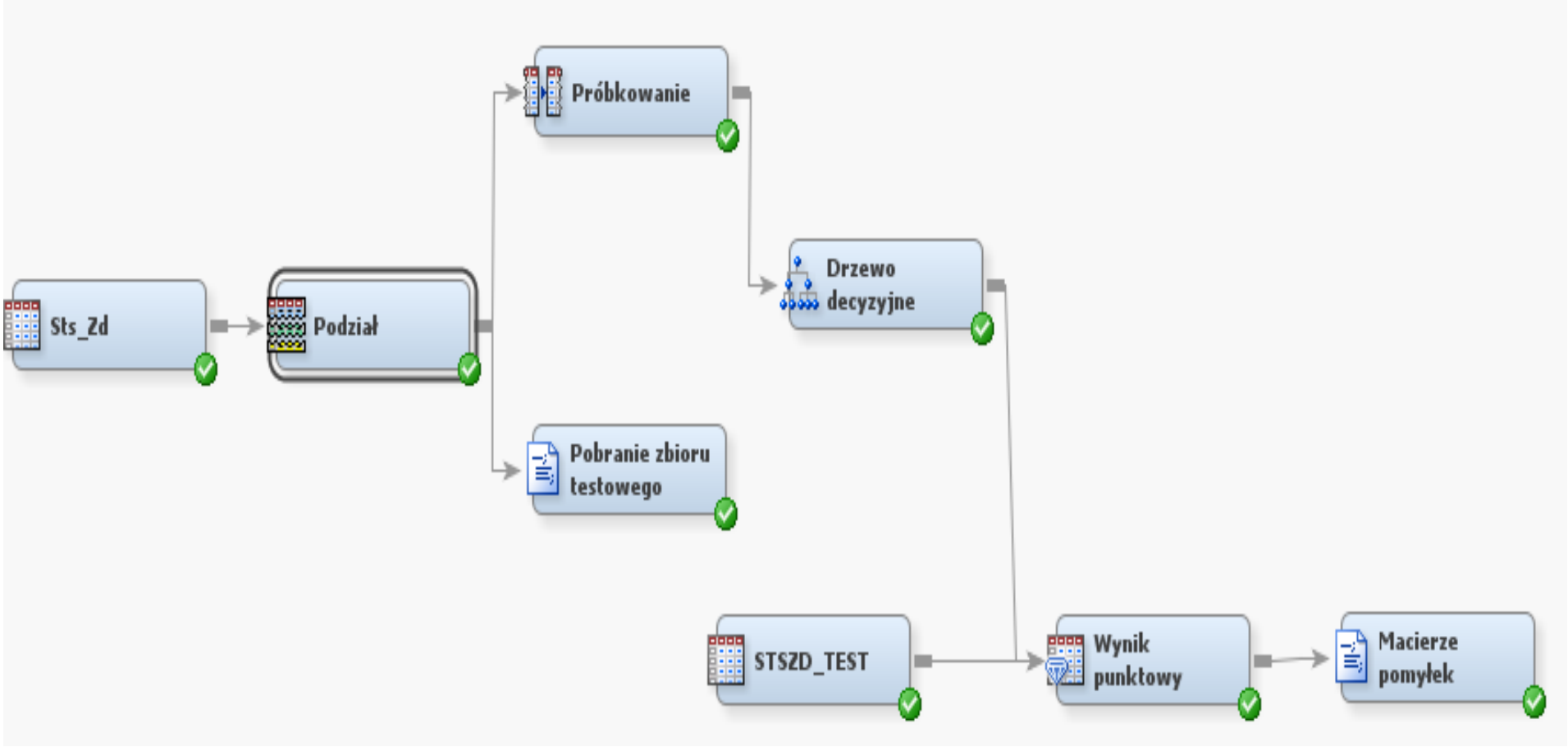


Diagram procesu w SAS EM



4GL - wprowadzenie

4GL – Fourth Generation Language wbudowany, deklaratywny język programowania do tworzenia skryptów w środowisku SAS. Dwa rodzaje skryptów:

- Data step – operacje na zbiorach danych: tworzenie, wczytywanie, przekształcenia, przetwarzanie
- Proc step – przywoływanie gotowych procedur do analizy, modelowania, raportowania

Każda instrukcja skryptu kończy się średnikiem

```
DATA pracownicy;
  INPUT imie $ pensja;
  DATALINES;
Jan 4000
Anna 5000
Piotr 4500
;
RUN;
```

```
DATA nowy_zbior;
  SET pracownicy;
  podwyzka = pensja * 1.10;
RUN;

PROC MEANS
DATA=pracownicy;
  VAR pensja;
RUN;
```

Dostęp do danych w SAS EM

Właściwość	Wartość
Ogólne	
Id. węzła	Ids
Importowane dane	
Eksportowane dane	
Uwagi	
Uczenie	
Typ wyniku	Widok
Rola	Uczące
Powtórz przebieg	Nie
Agreguj	Nie
Porzuć zmienne mapujące	Nie
Kolumny	
Zmienne	
Decyzje	
Odśwież metadane	
Doradca	Podstawowy
Opcje zaawansowane	
Dane	
Wybór danych	Źródło danych
Próba	Domyślna
Opcje próby	
Źródło danych	
Źródło danych	
Właściwości źródła danych	
Nowa tabela	
Nazwa tabeli	
Walidacja zmiennych	Ścisła
Rola nowej zmiennej	Odrzuć

Zakładka Próbkowanie → Dane wejściowe

- Określenie zbioru danych i przypisanie mu roli, zdefiniowanie wielkości próby.
- Przypisanie zmiennym (cechom) znaczenie w procesie modelowania.
- Potwierdzenie lub zmiana typu cechy.

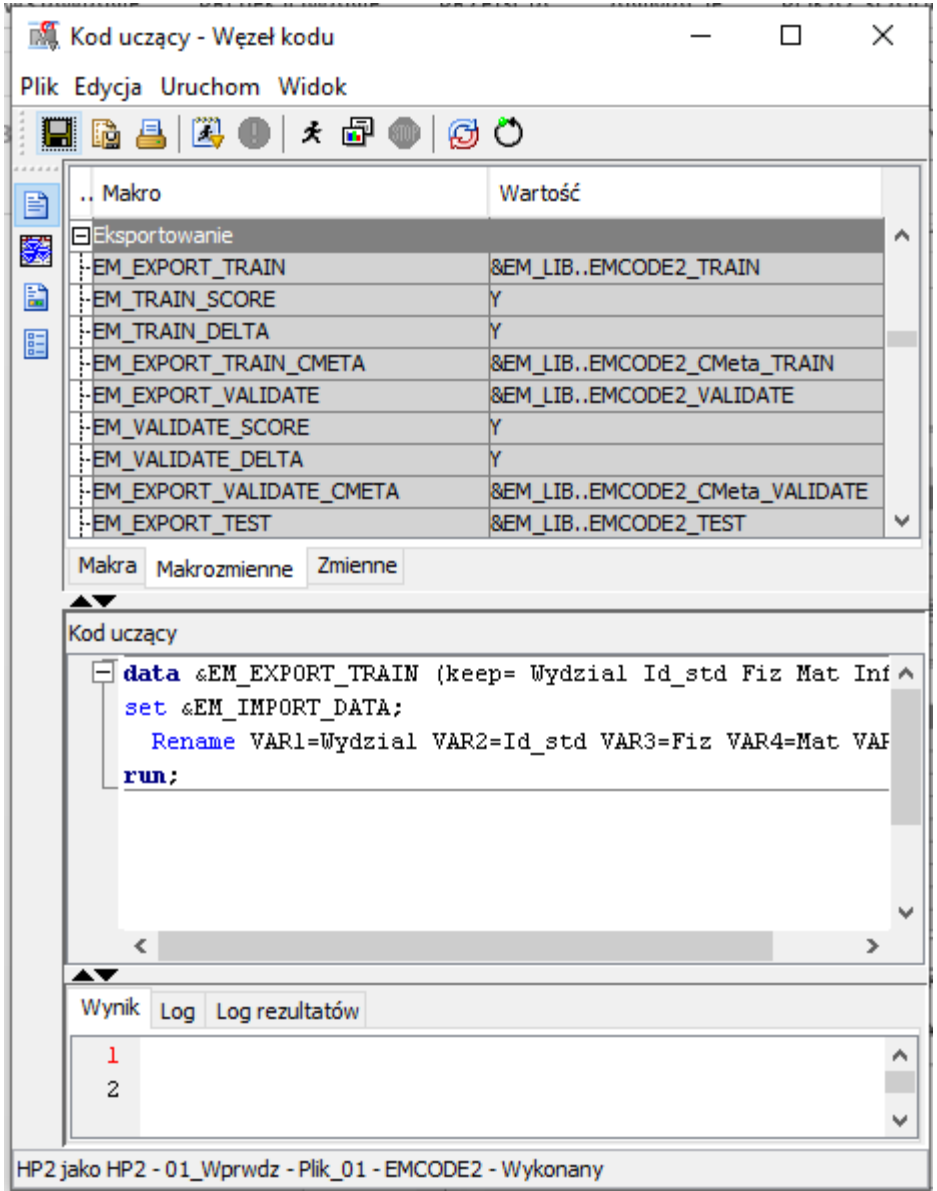
Program w 4GL w SAS EM

.. Właściwość	Wartość
Ogólne	
Id. węzła	EMCODE3
Importowane dane	...
Eksportowane dane	...
Uwagi	...
Uczenie	
Zmienne	...
Edytor kodu	...
Typ narzędzia	Pomocnicze
Potrzebne dane	Nie
Powtórz przebieg	Nie
Użyj prawdopodobieństw a priori	Tak
Wynik punktowy	
Typ doradcy	Podstawowy
Kod publikacji	Publikacja
Format kodu	Datastep
Status	
Czas utworzenia	16.05.24 15:27
Id. przebiegu	
Ostatni błąd	
Ostatni status	
Czas ostatniego uruchomienia	
Czas trwania przebiegu	
Host sieci gridowej	
Węzeł dodany przez użytkownika	Nie

Zakładka pomocnicze → Kod SAS-owy

- Przetwarzania danych poza standardowymi węzłami (np. czyszczenie, transformacja).
- Implementacja niestandardowych algorytmów, obliczeń lub raportów.

Edytor kodu w SAS EM



Makrozmienne

- EM_IMPORT_DATA identyfikuje zbiór danych importowany z węzła poprzedzającego
- EM_EXPORT_TRAIN identyfikuje zbiór danych eksportowany do następnego węzła

Odwołanie do makrozmiennych w kodzie – poprzez operator &, np. &EM_IMPORT_DATA