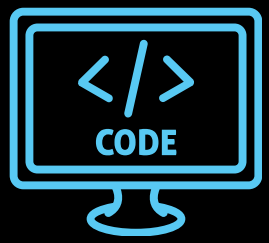

DeepDive™

Projekty matematyczne



Projekty programistyczne

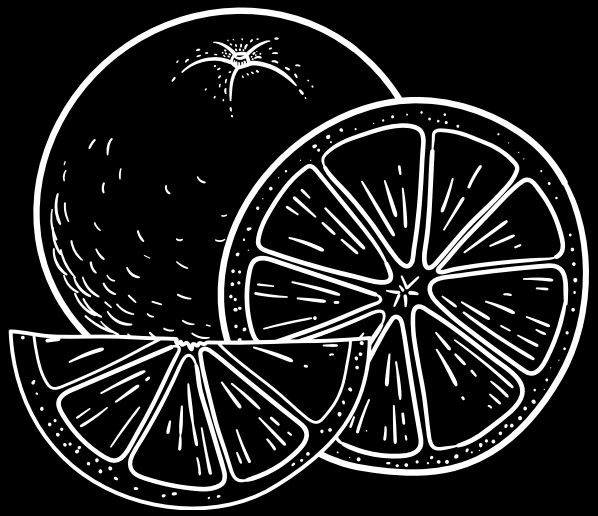
Aplikacje webowe

Artykuły popularnonaukowe

Wydarzenia

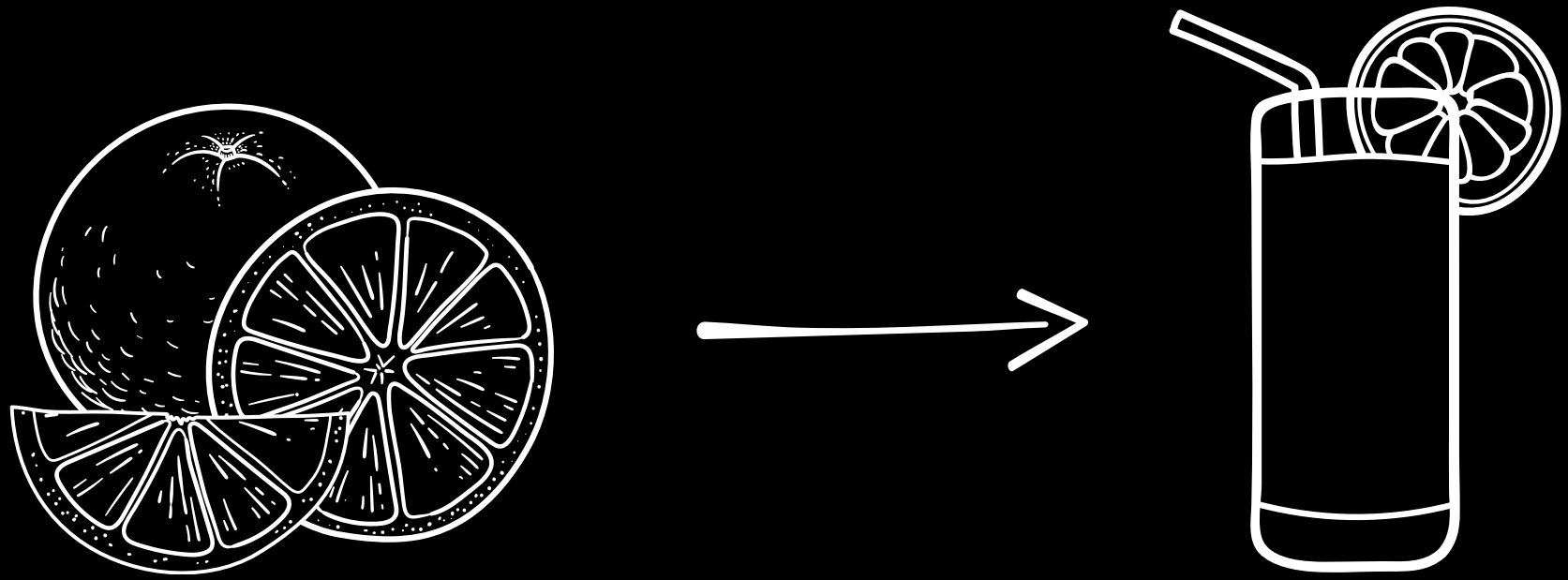
Integracje

DeepDive™



DeepDive™

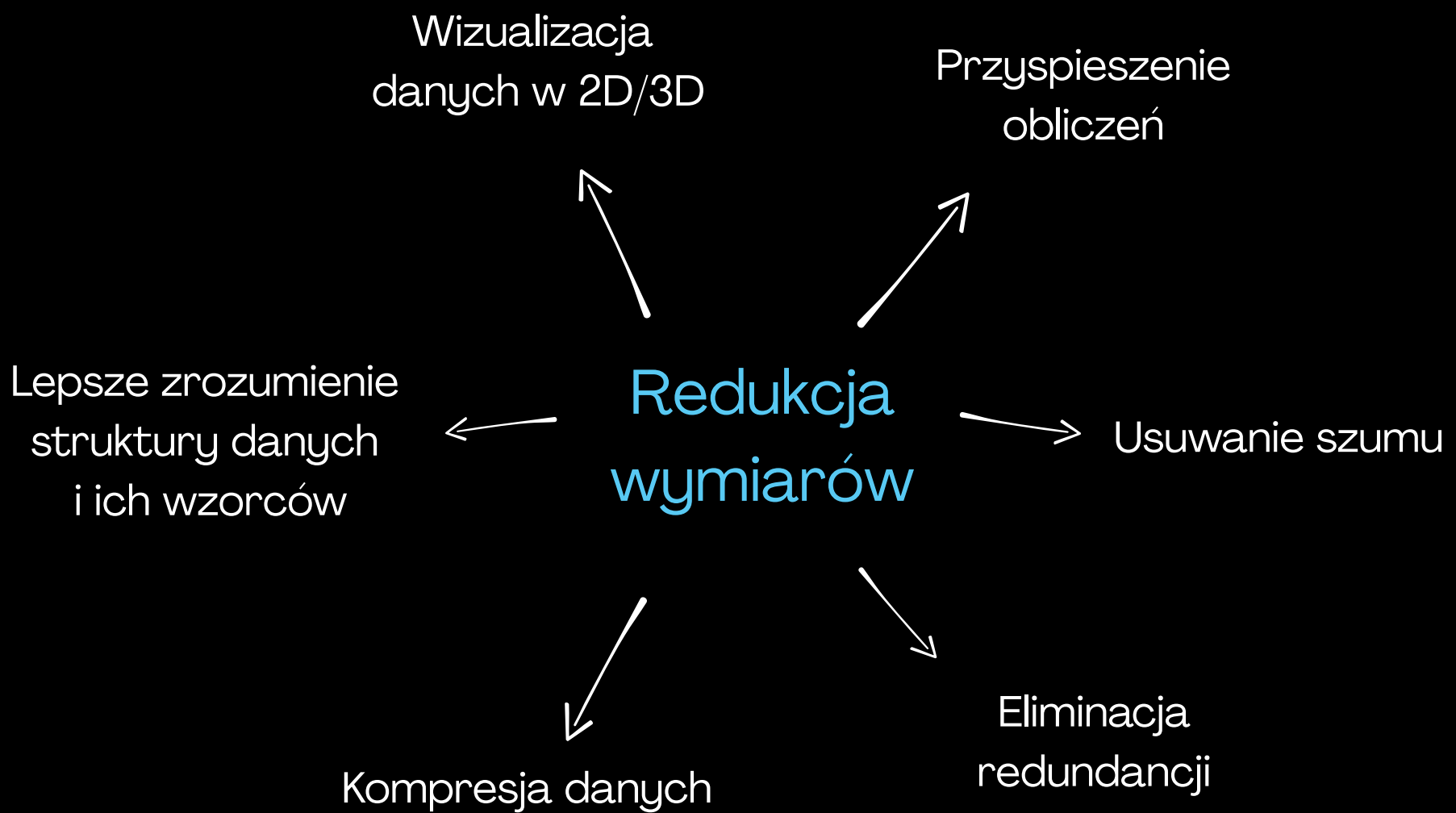
Redukcja wymiarów = wyciskarka



DeepDive™

Po co nam redukcja wymiarów?

DeepDive™


Wizualizacja danych w 2D/3D
Przyspieszenie obliczeń
Lepsze zrozumienie struktury danych i ich wzorców
Redukcja wymiarów
Usuwanie szumu
Kompresja danych
Eliminacja redundancji

DeepDive™

Klątwa wielowymiarowości

DeepDive™

Klątwa wielowymiarowości

Wraz ze wzrostem liczby wymiarów:

dane stają się coraz rzadsze,

odległości między punktami tracą sens,

modele uczą się wolniej i łatwiej przeuczyć model,
rośnie koszt obliczeniowy.

DeepDive™

Metody liniowe

PCA
(Principal Component Analysis)

najpopularniejsza, przekształca
dane w nowe, ortogonalne osie
maksymalnej wariancji

LDA
(Linear Discriminant Analysis)

uwzględnia przynależność do klas
optymalna dla klasyfikacji

ICA
(Independent Component Analysis)

szuka komponentów statystycznie
niezależnych

DeepDive™

Metody nieliniowe

t-SNE
(t-distributed Stochastic Neighbor Embedding)

świetna do wizualizacji, zachowuje lokalną strukturę danych

UMAP
(Uniform Manifold Approximation and Projection)

szybsze od t-SNE, zachowuje więcej globalnej struktury

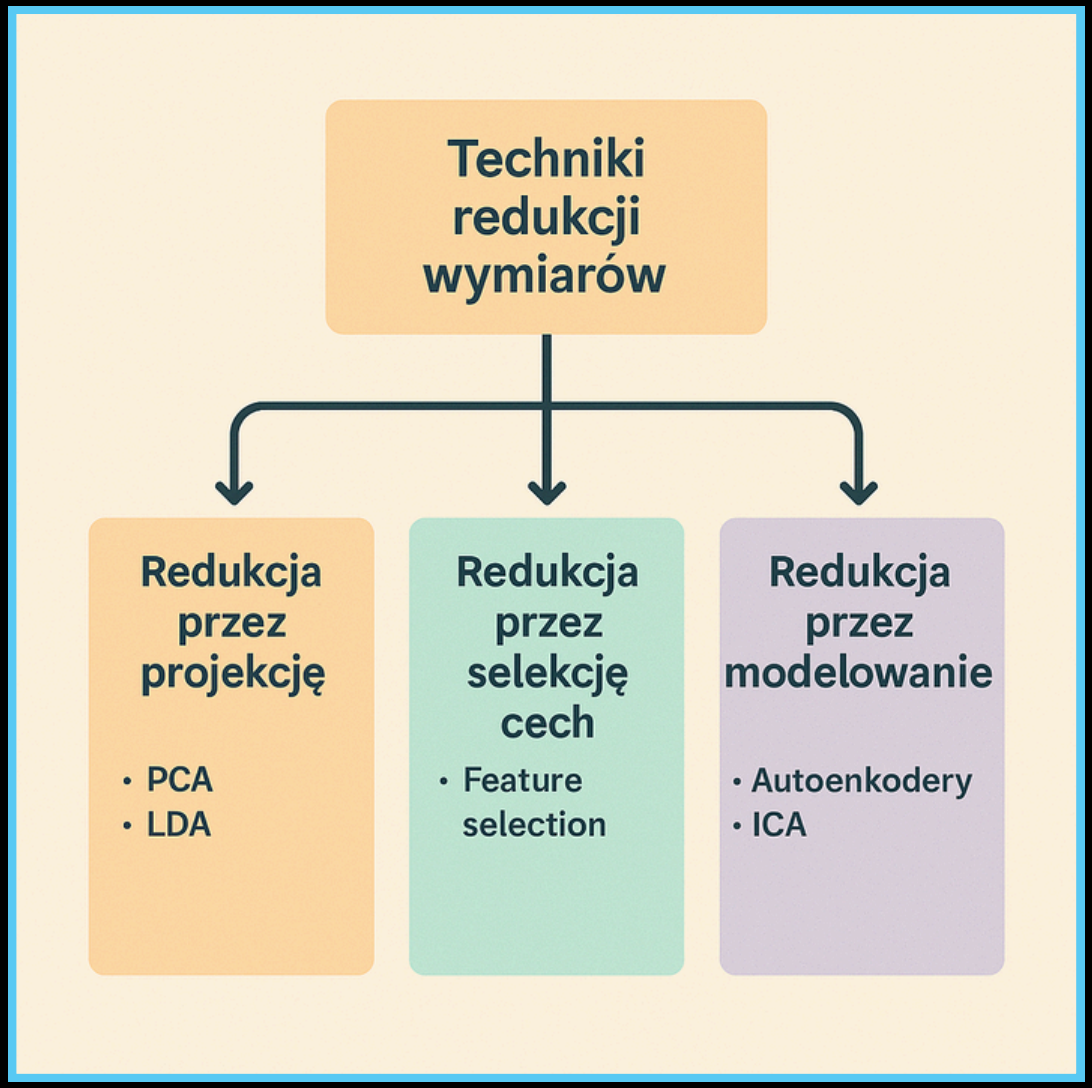
Isomap
oparta na geodezyjnych odległościach na kolektorze danych

Metody oparte o uczenie maszynowe

Autoenkodery (**Autoencoders**)
sieci neuronowe uczące się kompresji danych

Feature selection (np. LASSO, SelectKBest)
wybierają najważniejsze cechy bez tworzenia nowych

DeepDive™


Techniki redukcji wymiarów
Redukcja przez projekcję
• PCA
• LDA
Redukcja przez selekcję cech
• Feature selection
Redukcja przez modelowanie
• Autoenkodery
• ICA

Jak działa PCA?

DeepDive™

1. Standaryzacja danych
2. Obliczenie macierzy kowariancji/korelacji
3. Wyznaczenie wektorów i wartości własnych
4. Wybór składowych głównych
5. Transformacja danych

DeepDive™

Case study

Dane z Biura Statystyk Pracy ze Stanów Zjednoczonych.
Przedstawiają tygodniowe zarobki w 2020 roku z podziałem na stany.

DeepDive™

Macierz korelacji

	Transformowana liczba pracowników	Transformowana mediana zarobków	Transformowana liczba kobiet pracujących	Mediana zarobków kobiet	Transformowana ilość mężczyzn pracujących	Mediana zarobków mężczyzn
Transformowana liczba pracowników	1.000000	0.287842	0.999026	0.300256	0.999373	0.262250
Transformowana mediana zarobków	0.287842	1.000000	0.284447	0.969466	0.290151	0.969635
Transformowana liczba kobiet pracujących	0.999026	0.284447	1.000000	0.299753	0.996842	0.260402
Mediana zarobków kobiet	0.300256	0.969466	0.299753	1.000000	0.300320	0.901497
Transformowana ilość mężczyzn pracujących	0.999373	0.290151	0.996842	0.300320	1.000000	0.263151
Mediana zarobków mężczyzn	0.262250	0.969635	0.260402	0.901497	0.263151	1.000000

DeepDive™

Wektory i wartości własne

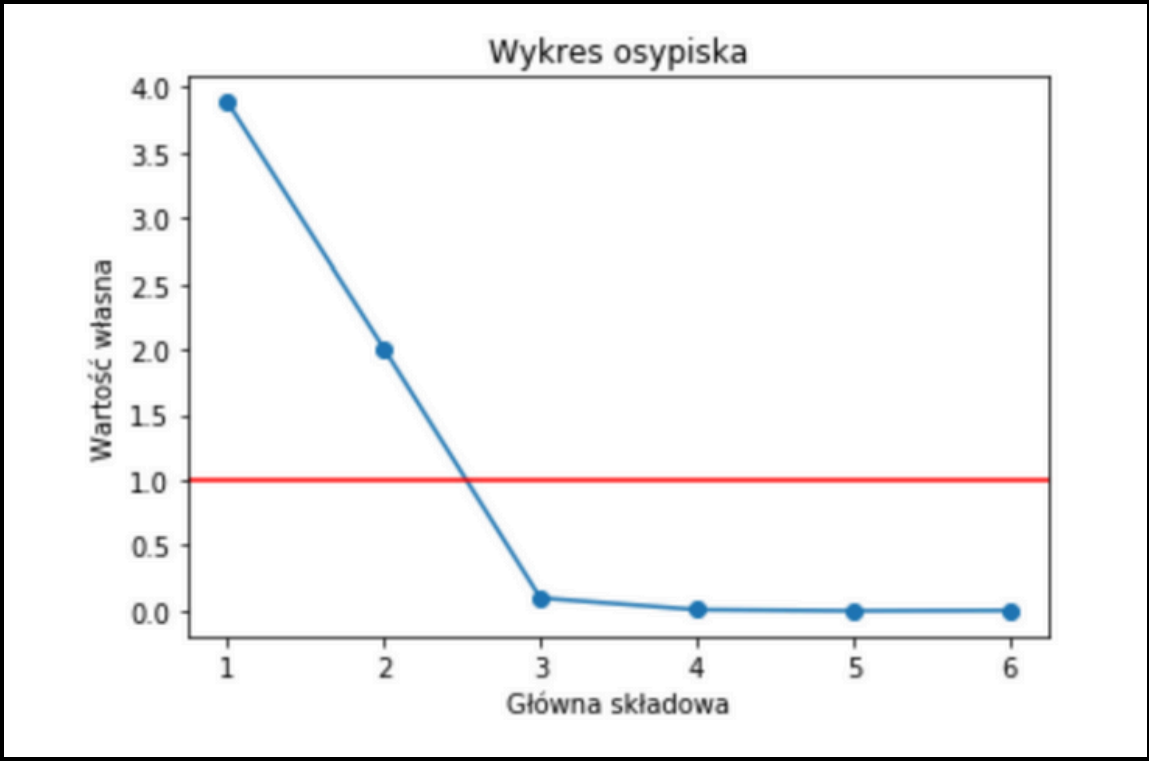
	Wektory własne					
Transformowana liczba pracowników	-0.419575	-0.396923	0.010384	0.019246	-0.814499	0.050215
Transformowana mediana zarobków	-0.403576	0.425013	-0.003109	0.735537	-0.002831	-0.339790
Transformowana liczba kobiet pracujących	-0.418716	-0.397564	0.000062	-0.305947	0.361194	-0.665248
Mediana zarobków kobiet	-0.398912	0.406232	-0.709910	-0.373295	0.000774	0.180341
Transformowana ilość mężczyzn pracujących	-0.419611	-0.395872	0.016430	0.284961	0.454001	0.616135
Mediana zarobków mężczyzn	-0.388015	0.426637	0.704017	-0.380078	0.002155	0.165289
Wartości własne	3.884728	2.006370	0.098145	0.008493	0.000001	0.002262
% Wyjaśnianej wariancji	64,75 %	33,43 %	1,64 %	0,14 %	0,00025 %	0,04 %

DeepDive™

Macierz ładunków czynnikowych

	PC1	PC2	PC3	PC4	PC5	PC6
Transformowana liczba pracowników	-0.826970	-0.795437	-0.825277	-7.862446e-01	-0.827041	-0.764767
Transformowana mediana zarobków	-0.562228	0.602016	-0.563135	5.754130e-01	-0.560739	0.604316
Transformowana liczba kobiet pracujących	0.003253	-0.000974	0.000020	-2.224010e-01	0.005147	0.220555
Mediana zarobków kobiet	0.001774	0.067786	-0.028195	-3.440211e-02	0.026261	-0.035027
Transformowana ilość mężczyzn pracujących	-0.000991	-0.000003	0.000439	9.412193e-07	0.000552	0.000003
Mediana zarobków mężczyzn	0.002388	-0.016162	-0.031641	8.577637e-03	0.029305	0.007862

Wybór ilości głównych składowych



kryterium osypiska (1966)

kryterium Kaisera (1960)

kryterium minimalnego zasobu wyjaśnianej zmienności (>70%)

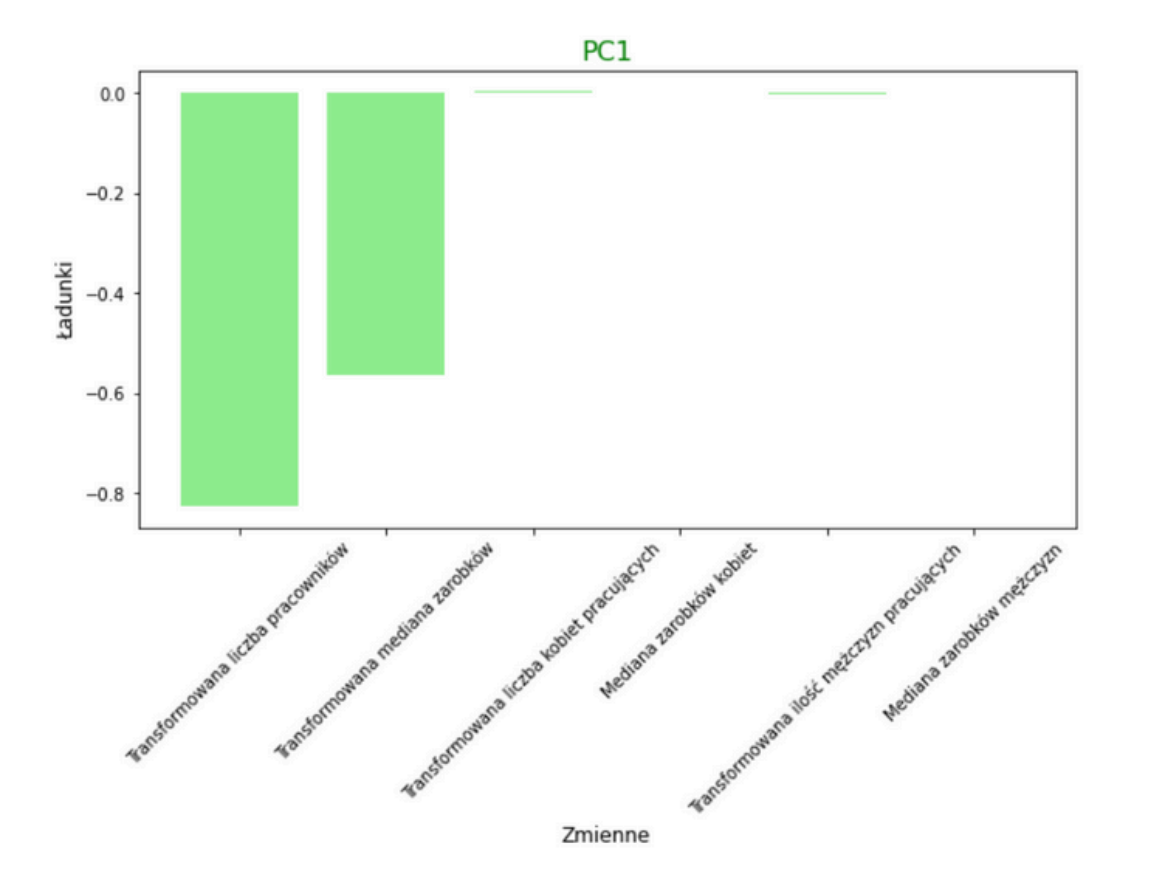
64.75 % + 33.43 % = 98.18 %

DeepDive™

Wybór ilości głównych składowych

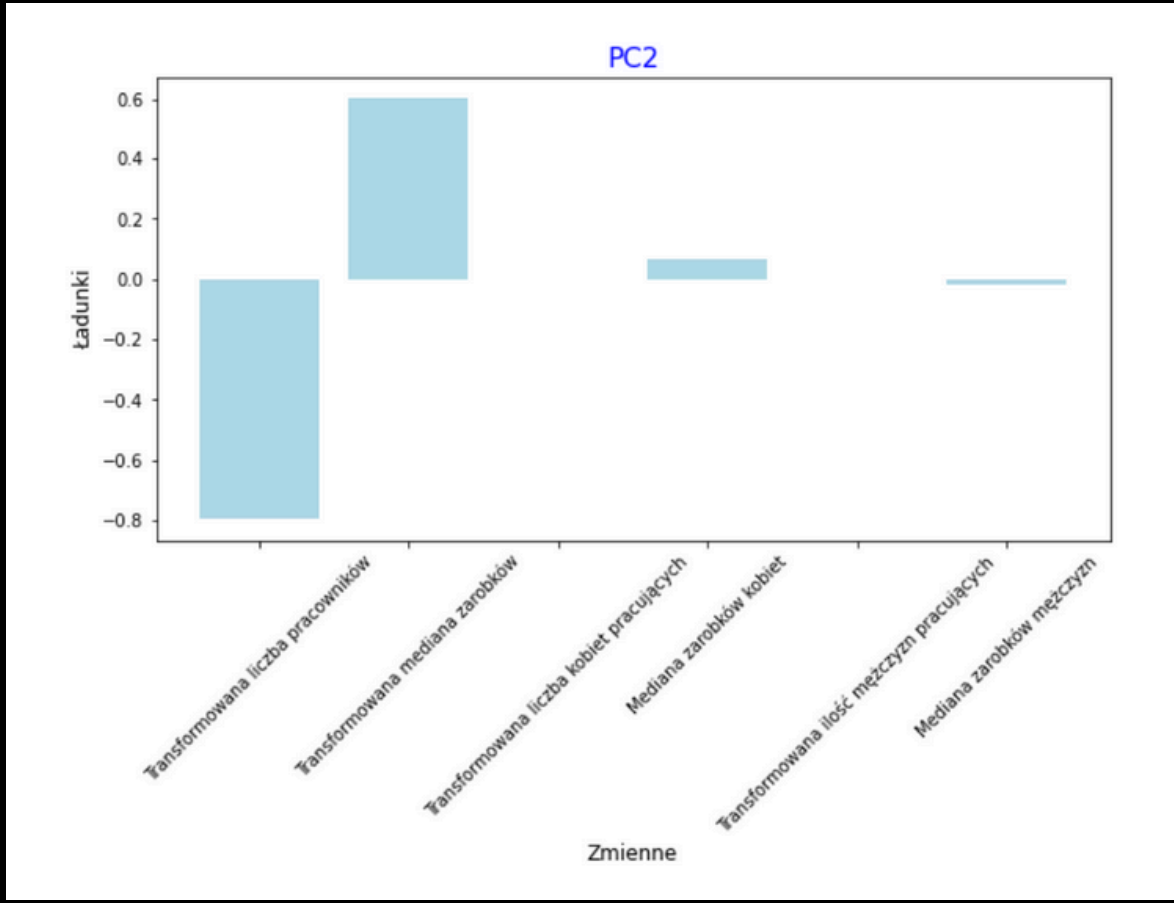
	PC1	PC2
Transformowana liczba pracowników	-0.826970	-0.795437
Transformowana mediana zarobków	-0.562228	0.602016
Transformowana liczba kobiet pracujących	0.003253	-0.000974
Mediana zarobków kobiet	0.001774	0.067786
Transformowana ilość mężczyzn pracujących	-0.000991	-0.000003
Mediana zarobków mężczyzn	0.002388	-0.016162

Wykres ładunków czynnikowych



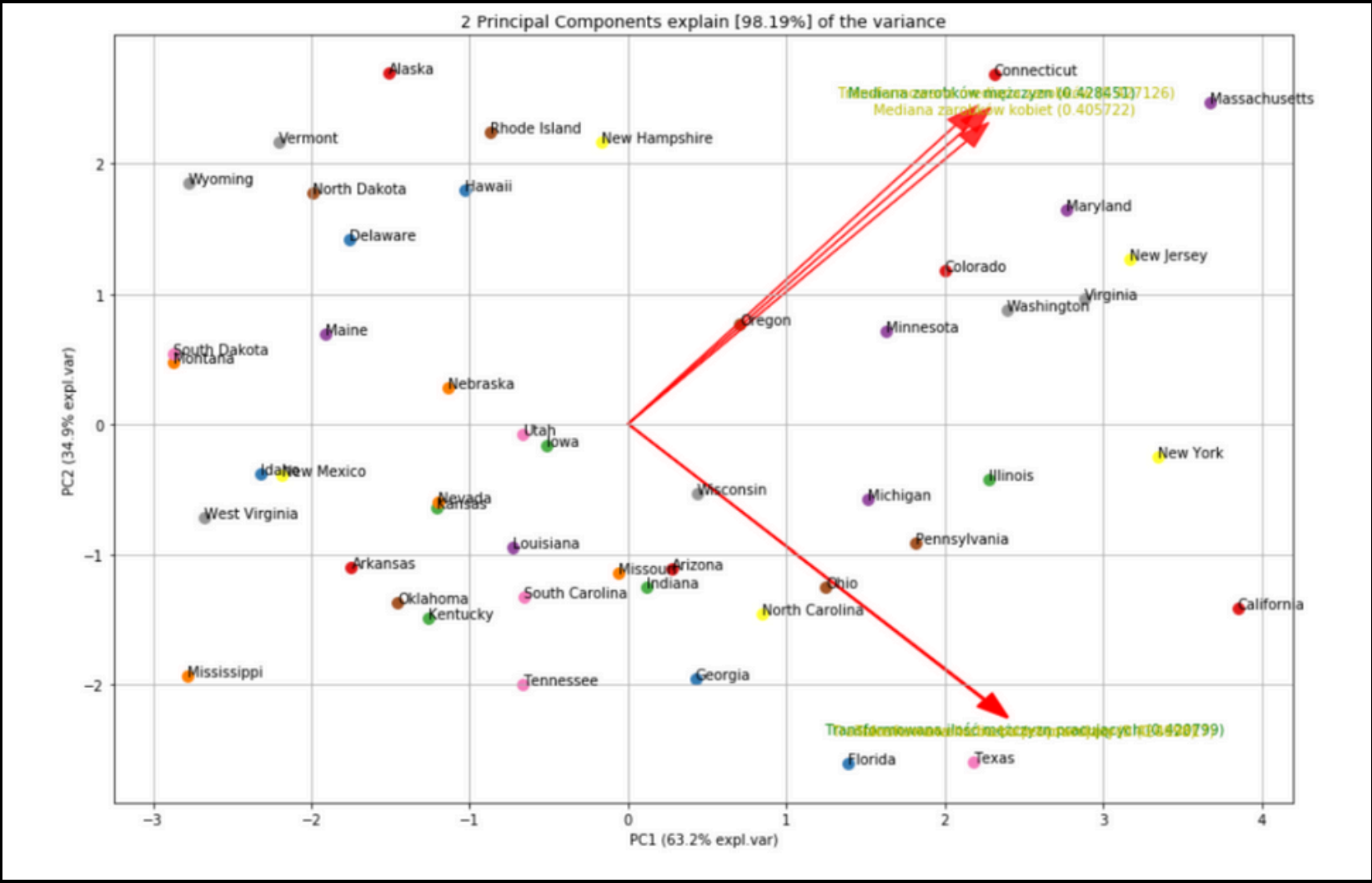
DeepDive™

Wykres ładunków czynnikowych



DeepDive™

Biplot



DeepDive™

Wnioski

PC1 silnie zależy od zarobków (kobiet i mężczyzn) – rozróżnia stany bogatsze i biedniejsze.

PC2 związany z udziałem pracujących mężczyzn.

Connecticut, Massachusetts, New Jersey – wysokie zarobki.

Texas, Florida, Georgia – więcej pracujących mężczyzn, ale niższe zarobki.

🧭 PCA ujawnia naturalne klastry stanów wg cech społeczno-ekonomicznych.

Jak działa t-SNE?

DeepDive™

t-Distributed Stochastic Neighbor Embedding

stochastyczna metoda porządkowania sąsiadów
w oparciu o rozkład t

DeepDive™

zachowuje lokalne struktury danych,
ale niekoniecznie globalne

świetne do wizualizacji, ale nie nadaje się do predykcji

wyniki mogą się różnić przy każdym uruchomieniu
(chyba że ustawi się random_state)

DeepDive™

1. Obliczenie podobieństw między punktami w przestrzeni wysokowymiarowej
2. Losowa inicjalizacja punktów w przestrzeni 2D lub 3D
3. Obliczenie podobieństw między punktami w niskim wymiarze
4. Minimalizacja różnicy między tymi dwoma rozkładami podobieństw
5. Przemieszczanie punktów w 2D, aż rozkłady będą jak najbardziej zbliżone

DeepDive™

Perplexity

Mała wartość (5–30):

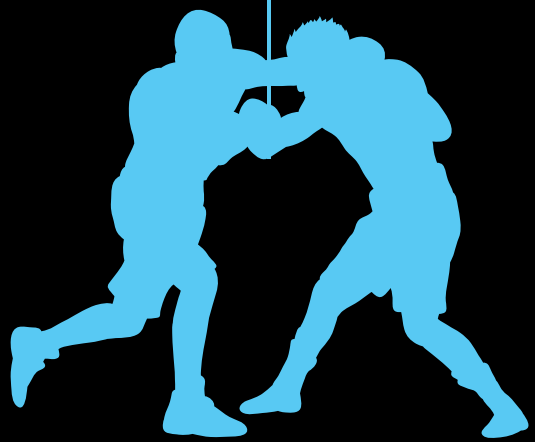
t-SNE skupia się bardziej na lokalnej strukturze danych.
Może uwydatnić małe klastry.
Większe ryzyko szumu i przetrenowania.

Większa wartość (50–100):

t-SNE patrzy szerzej, uwzględnia więcej sąsiadów.
Zachowuje więcej globalnej struktury.
Może spłaszczyć lokalne różnice.

PCA vs t-SNE

PCA	t-SNE
szybki, liniowy, deterministyczny	wolniejszy, nieliniowy, probabilistyczny
zachowuje globalną strukturę (maksymalizuje wariancję)	lepiej oddaje lokalne podobieństwa
szybka redukcja wymiarów	wizualizacja złożonych zbiorów danych



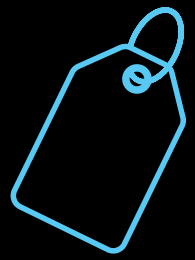
DeepDive™

Jak działa LDA?

DeepDive™

Liniowa analiza dyskryminacyjna

metoda, która znajduje takie kombinacje cech, które najlepiej oddzielają od siebie różne klasy. Jej celem jest zmniejszenie liczby wymiarów, zachowując przy tym maksymalną ilość informacji potrzebną do rozróżnienia kategorii.
Uzyskane nowe cechy mogą być wykorzystane do budowy klasyfikatora lub jako wstępny krok w analizie danych.



super do klasyfikacji i wizualizacji,
jeśli mamy dane z etykietami

zakłada, że dane w klasach
są normalnie rozłożone i mają równą kowariancję

w odróżnieniu od PCA,
LDA maksymalizuje separację klas, a nie wariancję ogólną

DeepDive™

Jak wygląda preprocessing danych?

DeepDive™


DeepDive™