# Julia smerdel PUM, Lab4

### Najważniejsz widżety:

- **Rozkłady (Distributions):** Tworzy wykresy, które pomagają wizualizować rozkład danych dla wybranych zmiennych. Użytkownicy mogą tworzyć histogramy, wykresy gęstości i wykresy pudełkowe, dzięki czemu mogą lepiej zrozumieć charakterystykę danych oraz wykryć anomalie.

- **Obliczanie Odległości (Distances):** Służy do określania odległości pomiędzy poszczególnymi punktami danych, co jest kluczowe dla analizy podobieństwa między obiektami. Pozwala na wyróżnienie naturalnie występujących grup lub klastrów na podstawie ich relacji przestrzennych.

- **Mapa Odległości (Distance Map):** Ta funkcja generuje graficzną reprezentację odległości między punktami, ułatwiając zrozumienie struktury zbioru danych i pomagając w identyfikacji obszarów z podobnymi atrybutami.

- **Macierz Odległości (Distance Matrix):** Tworzy tabelę przedstawiającą odległości między wszystkimi punktami w zbiorze danych, co umożliwia dogłębną analizę ich wzajemnych powiązań i pomaga w identyfikacji klastrów lub wzorców podobieństwa.

- **T-SNE**: Jest to technika redukcji wymiarowości, która pozwala na efektywną wizualizację skomplikowanych zbiorów danych w przestrzeni o niższej wymiarowości, przy jednoczesnym zachowaniu ich kluczowych właściwości strukturalnych.

- **kNN (k najbliższych sąsiadów):** Wprowadza algorytm k najbliższych sąsiadów do klasyfikacji lub regresji. kNN wykorzystuje wzorce podobieństwa między danymi do prognozowania lub klasyfikacji, co sprawia, że jest to cenione narzędzie w analizie danych.

- **Analiza Składowych Głównych (PCA):** Metoda ta stosuje technikę zmniejszania wymiarów danych, znana jako analiza składowych głównych. Pozwala ona na transformację danych do przestrzeni o zredukowanej liczbie wymiarów, co jest przydatne zarówno dla łatwiejszej wizualizacji, jak i usunięcia mniej istotnych cech, co z kolei może przyczynić się do zwiększenia efektywności algorytmów uczenia maszynowego.

- **Ranking (Rank):** Ten widget przydziela danej pozycji w zbiorze danych określoną pozycję rangową opartą o wybrane atrybuty. Jest to użyteczne dla uporządkowania danych zgodnie z wybranymi kryteriami i wyróżnienia tych, które są najistotniejsze lub mają największe znaczenie w analizie.

- **SVM (support vector machine paradigm):** Widget SVM umożliwia tworzenie modelu klasyfikacyjnego, który oddziela różne kategorie danych na podstawie ich cech. Używa granic decyzyjnych, by maksymalnie różnicować klasy, co pozwala na efektywną analizę wzorców i zachowań w danych. Jest to szczególnie przydatne w identyfikowaniu grup danych o podobnych cechach lub w klasyfikacji danych do określonych kategorii.

-**Discretize** - narzędziem używanym do dyskretyzacji ciągłych atrybutów danych na kategorie. Dyskretyzacja to proces przekształcania danych ciągłych (takich jak liczby rzeczywiste) w dane kategoryczne

## Bone marrow

### Opis

Zbiór danych przedstawia profile ekspresji genów w komórkach jednojądrzastych szpiku kostnego od pacjenta z ostrą białaczką szpikową oraz od dwóch zdrowych dawców, którzy pełnią rolę grupy kontrolnej. Dane zawierają próbki 1000 komórek i 1000 genów charakteryzujących się największą zmiennością.

### Cel

Czy występowanie genu informuje o chorobie.

### Schemat

A diagram of a network

Description automatically generated with medium confidence

### Wyniki

##### knn

A screenshot of a graph

Description automatically generated

##### Pca

A graph with a line and numbers

Description automatically generated

A screenshot of a graph

Description automatically generated

##### Rank

A screenshot of a graph

Description automatically generated

##### Stochastic gradient descent

A screenshot of a computer program

Description automatically generated

A screenshot of a graph

Description automatically generated

##### SVM

A screenshot of a graph

Description automatically generated

## Housing

### opis

Zbiór danych składa się z 506 próbek, z których każda opisuje dane na temat różnych aspektów domostw w różnych lokalizacjach wokół Bostonu

### cel

Celem jest przewidzenie mediany wartości domów zamieszkałych przez właścicieli (MEDV) na podstawie innych cech mieszkań.

### schemat

A diagram of a company

Description automatically generated with medium confidence

### wyniki

Ustawienia sieci neuronowej: jedna warstwa z 10 neuronami.

A screenshot of a computer

Description automatically generated

##### discretize

A screenshot of a computer

Description automatically generated

##### t-sne

A screenshot of a computer

Description automatically generated

##### rank

A screenshot of a graph

Description automatically generated

##### pca

A screenshot of a graph

Description automatically generated

##### kNN

A screenshot of a graph

Description automatically generated

##### stochastic gradient descent

A screenshot of a graph

Description automatically generated

##### SVM

A screenshot of a graph

Description automatically generated

## IRIS

### OPIS

Baza danych składa się z 150 próbek kwiatów irysa, podzielonych na trzy różne gatunki: Iris setosa, Iris virginica i Iris versicolor. Każdy gatunek jest reprezentowany przez 50 próbek.

Dla każdego kwiatu zarejestrowano cztery cechy:

1. Długość działki kielicha (sepal length)

2. Szerokość działki kielicha (sepal width)

3. Długość płatka (petal length)

4. Szerokość płatka (petal width)

### CEL

Zbudowanie modelu, który będzie w stanie na podstawie wymienionych czterech cech kwiatu przewidzieć, do którego z trzech gatunków Iris należy dana próbka.

### SCHEMAT

A diagram of a network

Description automatically generated

### WYNIKI

Ustawienia sieci neuronowej: jedna warstwa z 10 neuronami.

A screenshot of a computer

Description automatically generated

##### RANK

A screenshot of a graph

Description automatically generated

##### PCA

A screenshot of a graph

Description automatically generated

##### SVM

A screenshot of a graph

Description automatically generated

##### KNN

A screenshot of a graph

Description automatically generated

##### STOCHASTIC

A screenshot of a graph

Description automatically generated

## wyniki

Dla BM najlepsze wyniki dało Rank.  
Dla Iris najlepsze wyniki dało kNN.  
Dla Housing trudno określić, która metoda dała najlepsze wyniki. PCA I kNN dają podobne I satysfakcjonujące wyniki.