# Julia smerdel PUM, Lab3

## Bone marrow mononuclear cells with AML

### 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

### 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.

### Wyniki

##### knn

Dla 10 sąsiadów i metryki euklidesowej confusion matrix wygląda następująco:

A screenshot of a graph

Description automatically generated

##### Pca

A graph with a line and numbers

Description automatically generated

Trzeba również dodać sieć neuronową, np 100 neuronów w warstwie 1 i 100 w drugiej.

A screenshot of a cell phone

Description automatically generated

##### Rank

Trzeba również dodać sieć neuronową, np 100 neuronów w warstwie 1 i 100 w drugiej.

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 computer

Description automatically generated

### wnioski

* Najlepsze wyniki dało rank
* Najgorsze wyniki dało kNN (pewnie dlatego, że jest najprostsze)
* Rozmiar Sieci Neuronowych nie ma większego wpływu na jakość klasyfikacji
* Podział na cechy powinien być robiony tylko w sytuacjach gdzie wśród danych numerycznych można jasno wyróżnić kilka odrębnych cech