# Julia smerdel PUM, Lab5

## RF

### 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

### Najważniejsz widżety:

- **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.

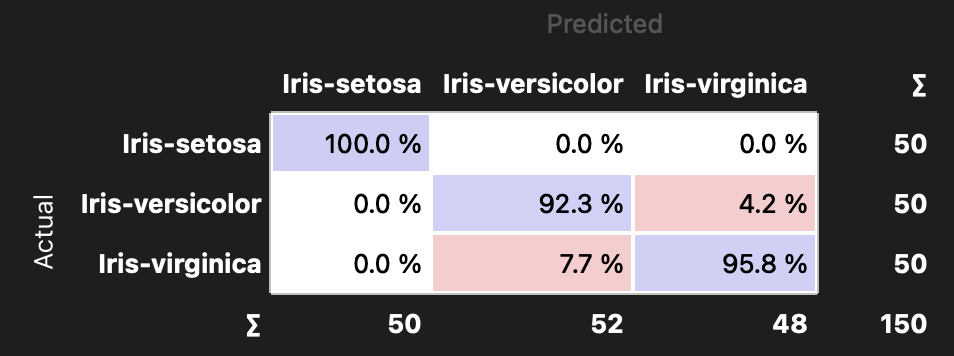
- **Ranking (Rank):** Ten widget przydziela danej pozycji w zbiorze danych określoną pozycję rangową opartą o wybrane atrybuty. Jest to użyteczne dla uporządkowania danych zgodnie z wybranymi kryteriami i wyróżnienia tych, które są najistotniejsze lub mają największe znaczenie w analizie.

- **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 5 sąsiadów i metryki euklidesowej confusion matrix wygląda następująco:



##### Pca

Sieć neuronowa 10, 10.

A screenshot of a graph

Description automatically generated

##### Rank

Sieć neuronowa 10, 10.

A screenshot of a graph

Description automatically generated

##### SVM

A screenshot of a graph

Description automatically generated

##### Neural networks

A screenshot of a graph

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 wyniki

## clustering

### schemat

A diagram of data processing

Description automatically generated

### najważniejsze widżety

**Louvain Clustering –** algorytm wykorzystywany głównie do identyfikacji społeczności w dużych sieciach, oparty na optymalizacji modularności. Modularność to miara siły struktury społeczności w sieci, gdzie wysokie wartości wskazują na dobrze zdefiniowane klastry

**k-Means –** popularny algorytm grupowania, który dzieli zbiór danych na k klastrów, gdzie k jest z góry określoną liczbą grup. Algorytm zaczyna od losowego przypisania punktów jako centroidów klastrów, a następnie iteracyjnie przypisuje każdy punkt danych do najbliższego centroidu, co tworzy klastry. Po przypisaniu wszystkich punktów, centroidy są aktualizowane do średniej wartości wszystkich punktów w klastrze. Proces jet powtarzany do momentu, gdy klastry przestaną się zmieniać

**DBSCAN -** Density-Based Spatial Clustering of Applications with Noise) to algorytm grupowania oparty na gęstości, który może znajdować klastry o dowolnych kształtach oraz obsługiwać punkty odstające. DBSCAN nie wymaga wcześniejszego określenia liczby klastrów, co stanowi dużą zaletę nad innymi metodami, jak k-Means. Algorytm definiuje klastry jako obszary o wysokiej gęstości punktów, oddzielone od innych takich obszarów regionami o niskiej gęstości punktów. Kluczowymi parametrami są 'eps' (maksymalny promień sąsiedztwa) i 'minPts' (minimalna liczba punktów wymagana do utworzenia gęstości). Podejście to jest szczególnie użyteczne, gdy modelujemy złożone struktury przestrzenne.

### wyniki

##### louvain

A screen shot of a computer

Description automatically generated

##### k-means

A screen shot of a computer screen

Description automatically generated

##### DBSCAN

A screen shot of a computer

Description automatically generated

### wyniki

Wyniki klasteryzacji Louvain były zgodne z oczekiwaniami.

Klasyfikacja społeczności w sieci była dobrze zdefiniowana.

Algorytm k-mean skutecznie podzielił dane na wybrane liczby klastrów.

Centroidy klastrów były dobrze zdefiniowane.

Algorytm DBSCAN efektywnie zidentyfikował klastry o różnych kształtach.

Obsługa punktów odstających była znaczącą zaletą.

## Regresja

### schemat

A diagram of a linear regression

Description automatically generated

### najważniejsze widżety

**Formula** - pozwala na tworzenie nowych cech (kolumn) lub modyfikowanie istniejących przy użyciu wyrażeń matematycznych.

**Linear Regression** – pozwala na tworzenie modelu regresji liniowej. Model ten przewiduje wartość zmiennej zależnej na podstawie liniowej kombinacji zmiennych niezależnych

### wyniki

##### linear regression

A screenshot of a computer

Description automatically generated

##### polynomial regression

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

##### non-linear regression

A screenshot of a computer

Description automatically generated

### wyniki

W naszej analizie, model regresji liniowej osiągnął umiarkowane wyniki. Jest to spodziewane, ponieważ regresja liniowa może nie uchwycić złożonych nieliniowych zależności w danych.

Tworzenie wielomianowych cech pozwala modelowi lepiej uchwycić nieliniowe zależności w danych, co prowadzi do poprawy dokładności predykcji.

Wyniki uzyskane za pomocą SVM i sieci neuronowych były lepsze od wyników regresji liniowej, jednak nie były tak dobre jak te uzyskane przy użyciu regresji wielomianowej.