

Statystyczne Metody Przetwarzania Danych

Projekt grupowy

Opis projektu

Projekt należy wykonać w grupach 2-3 osobowych. Wszystkie osoby z grupie otrzymują wspólną ocenę za całość projektu (Szczegółowy opis i elementy projektu znajduje się poniżej).

Narzędzia

Preferowanym narzędziem wykonania projektu jest Jupyter Notebook - język Python (chyba, że prowadzący laboratorium wyrazi zgodę na inną technologię).

W zadaniu można korzystać z bibliotek:

- numpy,
- pandas,
- scikit-learn,
- scipy,
- matplotlib,
- inne (po uzgodnieniu z prowadzącym).

Wysyłanie projektu

Projekt powinien składać się z następujących elementów:

1. Baza danych (wybrana baza danych).
2. Notatnik z implementacją rozwiązania - plik `*.ipynb` (+ ewentualnie pliki `*.py`).
3. Prezentacja - plik `*.pdf` oraz plik ze źródłem prezentacji (w zależności od tego w czym prezentacja była wykonana może to być plik: `*.tex`, `*.ppt`, `*.pptx`, `*.md`, `*.odp`). Prezentacja powinna trwać maksymalnie 5 minut (po przekroczeniu czasu prezentacja jest przerywana).

Wszystkie pliki powinny zostać spakowane archiwizatorem **zip** i wysłane przez platformę WIKAMP.

Struktura plików jaką należy oddać do oceny (przykład):

```
+-- 📁 [NAZWA_PROJEKTU]
+-- 📁 projekt
+--   📄 projekt.ipynb
+--   📄 opcjonalny_plik.py
+-- 📁 prezentacja
+--   📄 prezentacja.pdf
+--   📄 prezentacja.md
+-- 📁 baza_danych
+--   📄 plik.csv
+--   📄 plik.info
```

Elementy jakie powinny się znaleźć w projekcie:

1. Zbiór danych (minimum 6 klas).

Sugerowana strona ze zbiorami danych (może być inne źródło, jeżeli prowadzący wyrazi zgodę):
<https://archive.ics.uci.edu>.

UWAGA: Skład osobowy grupy oraz wybrany zbiór danych należy zgłosić prowadzącemu - kolejność zgłoszeń ma znaczenie (zabezpieczenie by dwie grupy nie wybrały takiego samego zbioru danych).

2. Podział zbioru danych na treningowy i testowy.

Parametry podziału (proporcje) powinny być konfigurowalne.

3. Wstępna analiza danych.

Ile jest próbek każdej klasy, ile jest cech, czy dane są znormalizowane, czy wymagają normalizacji, czy są inne czynności jakie należy wykonać, statystyki (tylko istotne - takie które wpływają na dalsze kroki).

4. Wstępne przetworzenie danych (*preprocessing*).

- Normalizacja danych,
- podział zbioru treningowego na treningowy i walidacyjny,
- usunięcie błędnych danych.

5. Selekcja cech.

Opis jakie metody selekcji cech zastosowano, dlaczego wybrano taką metodę a nie inną.

6. Klasyfikacja.

Przeprowadź klasyfikację poznanymi metodami klasyfikacji:

- NN,
- k-NN,
- NM i
- k-NM.

Oprócz powyższego:

- przeprowadź analizę wyników,
- wybierz metodę i uzasadnij wybór.

7. Przetestuj wybrane modele na zbiorze testowym.

Wykonaj analizę i przedstaw wyniki badań.

8. Wnioski.

Zasady oceniania

Oprócz wyżej wymienionych elementów w projekcie oceniane są także takie elementy jak:

- metodyka przeprowadzonego eksperymentu,
- obszerność/wnikliwość badań,
- czytelność i przejrzystość przedstawionych wyników,
- wnioski.

Termin

Wyniki należy zaprezentować najpóźniej na ostatnim laboratorium (jest to termin nieprzekraczalny).

Przekroczenie terminu (tj.brak przesłania pracy przed ostatnim laboratorium skutkuje oceną 2 z projektu).