Wytyczne do projektu zaliczeniowego – Systemy Big Data i NoSQL

1. **Cel Projektu**
   Praktyczne wykorzystanie narzędzi Big Data oraz baz NoSQL do przetwarzania dużych zbiorów danych oraz zastosowanie minimum dwóch metod uczenia maszynowego do analizy i modelowania danych.
   Projekt powinien pokazywać umiejętności:
   - pracę z dużym zbiorem danych (Big Data – minimum **1 GB**),
   - przetwarzanie danych w wybranym środowisku Big Data,
   - wykorzystanie dwóch różnych metod ML,
   - poprawną interpretację wyników.

2. **Wybór zbioru danych**
   Każdy zespół 2,3 osobowy lub indywidualnie wybiera jeden duży zbiór danych (np. z Kaggle, UCI, Google Dataset Search).
   Minimalny rozmiar zbioru: 1 GB (po rozpakowaniu).
   Dodatkowe zasady:
   - Zbiór musi zawierać dane, które umożliwią zastosowanie analiz ML.
   - Każdy projekt musi być **unikalny na poziomie roku**, obowiązuje zasada *kto pierwszy, ten lepszy*.
   - Tematy/dane są zgłaszane prowadzącemu na adres mail ljankowski@agh.edu.pl z linkiem do zbioru i zatwierdzane są przed rozpoczęciem realizacji.

   Przykładowe źródła (przykładowe):
   - Kaggle Datasets,
   - Open Data (np. NYC, Chicago),
   - Enron Email Dataset,
   - Yelp Open Dataset,
   - GUS.

3. **Wymagania technologiczne**
   Zespół powinien wykorzystać **Python** oraz co najmniej jedno narzędzie/system poznanych na zajęciach, np.:
   - Apache Spark (PySpark),
   - Hadoop (HDFS, MapReduce),
   - MongoDB lub inna baza NoSQL (Redis, Cassandra, Neo4j),

   Wymagania minimalne:
   - przynajmniej jeden etap przetwarzania danych musi wykorzystywać technologię Big Data / NoSQL,
   - kod musi być wykonany w sposób reprodukowalny,
   - projekt musi działać na rzeczywistym, dużym zbiorze danych.

4. **Zakres projektu**
   4.1. Wstęp i opis danych
   - danych,
   - wielkość i struktura zbioru,
   - charakter danych i problemu.

   4.2. Przygotowanie i przetwarzanie danych
   - wczytanie danych do wybranego systemu (Spark/Hadoop/NoSQL),
   - oczyszczenie danych, łączenie, filtrowanie,
   - ewentualny podział na train/test.

   4.3. Analiza danych
   - statystyki opisowe,
   - przykładowe wizualizacje (jeśli możliwe dla dużych danych).

   4.4. Dwie różne metody uczenia maszynowego np.:
   - regresja liniowa + random forest,
   - SVM + k-means,

- klasyfikacja + regresja logistyczna,
- PCA + XGBoost.

Wymagania:

- zastosowanie dwóch metod ML,
- przedstawienie metryk jakości (np. accuracy, RMSE, precision/recall itp.).

4.5. Wyniki i interpretacja

- wnioski z analizy,
- porównanie modeli,
- ocena jakości danych.

4.6. Podsumowanie

- napotkane problemy techniczne,
- rekomendacje,
- możliwe rozszerzenia projektu.

5. Harmonogram
   - **Zgłoszenie tematu:** do 06.12.2025
   - **Oddanie/prezentacje projektu:** do 11.01.2026

6. Forma oddania
   - Raportu PDF zawierającego: opis danych, metody, wyniki.
   - Kodu (np. Jupyter Notebook) – z komentarzami.
   - (Opcjonalnie) plików konfiguracyjnych lub docker-compose, jeśli były używane.

7. Kryteria oceny

Kryterium	Punkty	Opis
Praca z dużym zbiorem danych (>1 GB)	20	Poprawne przetwarzanie, efektywność narzędzi
Wykorzystanie technologii Big Data/NoSQL	20	Użycie Spark/Hadoop/MongoDB itp.
Dwie metody ML	20	Zastosowanie + interpretacja wyników
Raport i dokumentacja	20	Przejrzystość, opis metod, wykresy
Kod + powtarzalność	10	Jakość kodu, komentarze
Prezentacja projektu	10	Zrozumienie tematu, pytania

Maksymalnie: 100 punktów, ocena wg skali AGH

8. Wymogi formalne i zasady

- Każdy student tworzy projekt samodzielnie lub w 2–3 osobowym zespole.
- Zbiór danych musi być unikalny
- Wersja finalna projektu musi dać się uruchomić.