Podstawy Przetwarzania Danych

Wykład 1: Wprowadzenie

dr inż. Marcin Luckner
mluckner@mini.pw.edu.pl

Wydział Matematyki i Nauk Informacyjnych

Wersja 1.1
18 października 2021

Fundusze Europejskie Wiedza Edukacja Rozwój
Rzeczpospolita Polska
Politechnika Warszawska
Unia Europejska Europejski Fundusz Społeczny

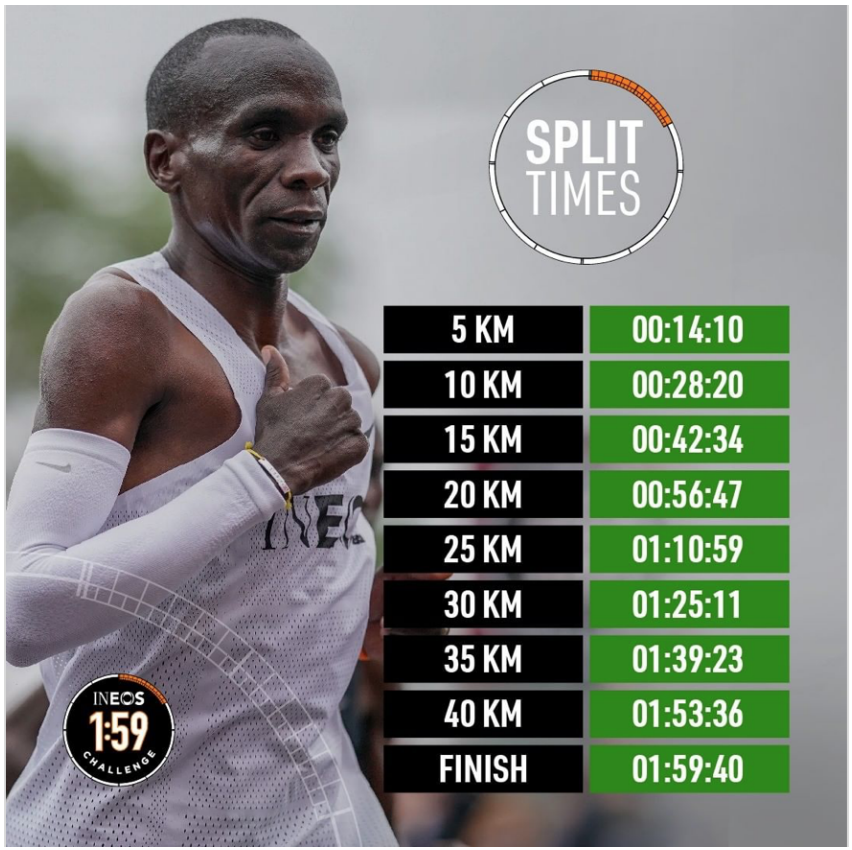
Projekt „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca” współfinansowany jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Zadanie 10 pn. „Modyfikacja programów studiów na kierunkach prowadzonych przez Wydział Matematyki i Nauk Informacyjnych”, realizowane w ramach projektu „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca”, współfinansowanego jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Cele przedmiotu

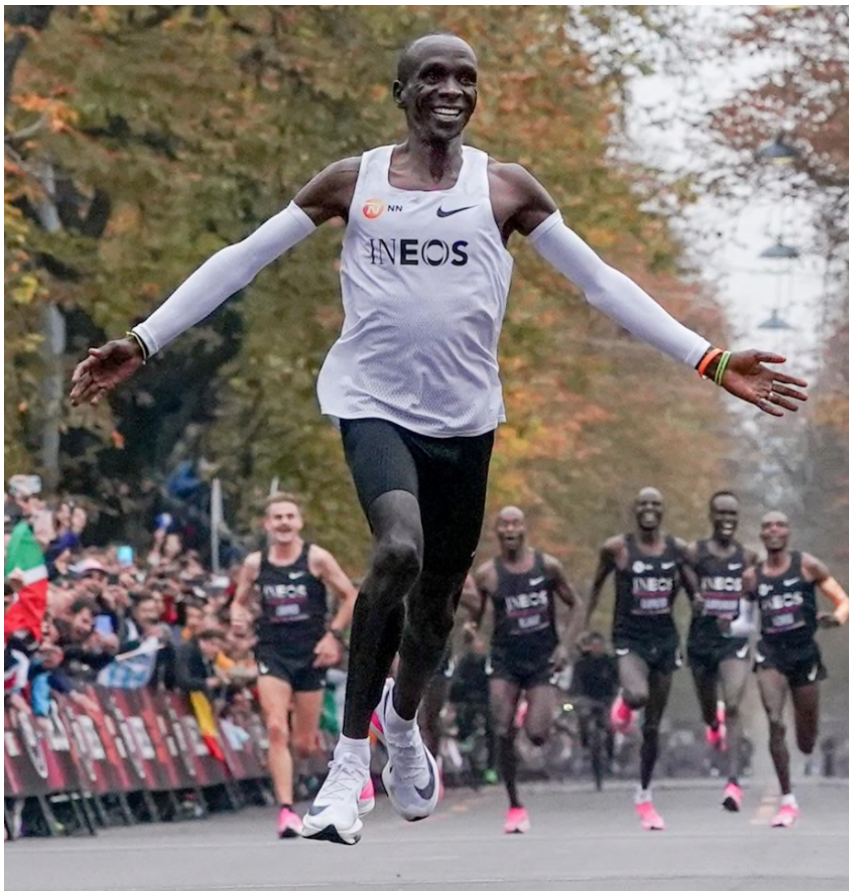
- Celem przedmiotu jest przedstawienie procesu przetwarzania danych w zadaniach uczenia maszynowego.
- Słuchacze mają poznać powody i metody przetwarzania danych wejściowych, sposoby przeprowadzania testów stworzonego rozwiązania i interpretacji wyników.
- Przedmiot ma zapewnić podstawową teoretyczną wiedzę z tego zakresu i umiejętność jej praktycznego zastosowania.

Przełamanie bariery 2:00



- Eliud Kipchoge przebiegł 42195 metrów w 1:59:40.
- Bieg starannie przygotowano, aby ułatwić mu pokonanie bariery 2h.
- Choć jest to niesamowite osiągnięcie Kenijczyka, nie zostało uznane za rekord świata.

Eliud Kipchoge

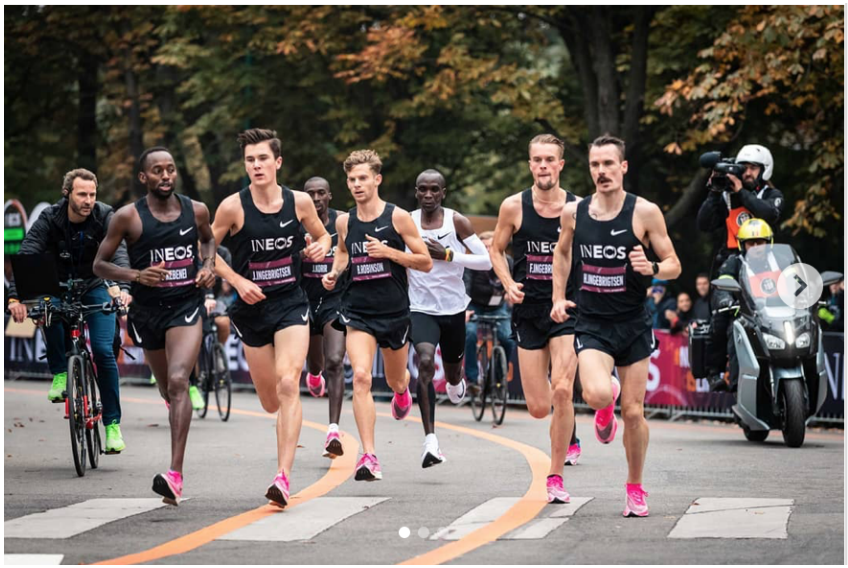


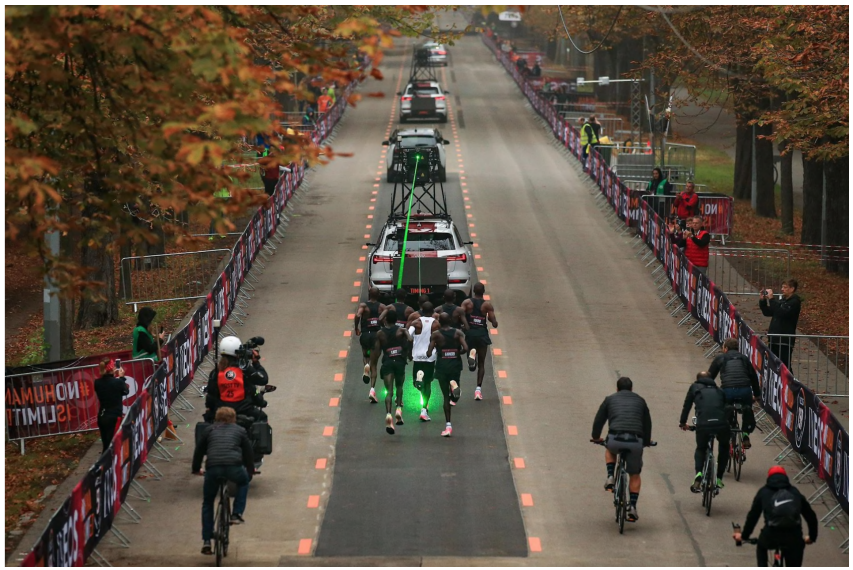
- Najlepszy maratończyk wszech czasów.
- Zdobył podium w 11 ze swoich 12 biegów długodystansowych.
- Ustanowił maratoński rekord świata 2:01:39.
- Dane
  - wiek 34 lata,
  - wzrost 167 cm,
  - waga 52 kilogramy.

Trasa

- Trasę wyznaczono po płaskim terenie.
- Czekano na optymalne warunki pogodowe.
- Wiedeń położony jest na odpowiedniej wysokości nad poziomem morza.

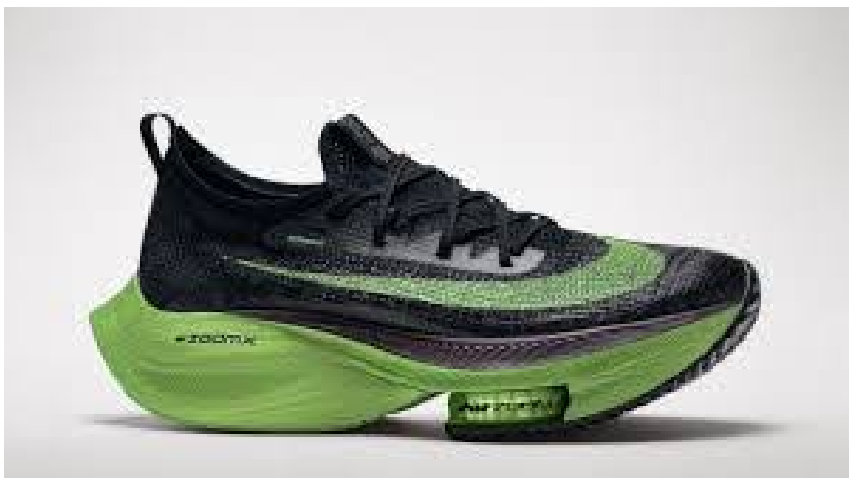
Bieg





- Biegacze wokół Nigeryjczyka nadawali mu tempo i ograniczali opór powietrza.
- Samochód elektryczny wyznaczał laserem ścieżkę biegu.
- Rowerzyści na bieżąco podawali płyny biegaczowi.
- Na całej trasie biegu mógł liczyć na doping.

Buty



- Eliud biegł w butach Nike Vaporfly 4%.
- Zbudowane z kombinacji super miękkiej pianki i płytek węglowych.
- Mają one zapewniać zwiększenie wydajności ruchu o cztery procent.

Treści wykładów I

- Wstępne przetwarzanie danych
  1. Dobór typów i normalizacja danych
  2. Redukcja wymiarowości
  3. Redukcja zaszumienia danych
  4. Selekcja cech
  5. Braki w danych
- Przygotowanie eksperymentu naukowego
  6. Próbkowanie danych
  7. Tworzenie środowiska testowego
- Ocena wyników eksperymentu
  8. Miary oceny wyników modelu
  9. Metodologia oceny wyników modelu
  10. Porównywanie modeli
  11. Wizualizacja wyników
- Analiza wpływu przebiegu eksperymentu na wyniki

Treści wykładów II

12. Analiza procesu uczenia modelu
13. Analiza wpływu danych na wyniki modelu

- Zaawansowane przetwarzanie danych
  14. Manifold learning
  15. Przetwarzanie danych jakościowych

Laboratoria i projekt

- Podczas laboratoriów studenci uczą się jak dokonywać eksploatacji danych, aby móc przeprowadzić analizę wpływu danych na wyniki działania modelu.
- Realizując projekt uczą się praktycznego przetwarzania danych i analizy wpływu przetwarzania na działanie modelu.
- Rozwiązywanie zadań pod nadzorem opiekuna. Samodzielne szukanie sposobu przetworzenia danych, aby maksymalizować parametry jakościowe osiągalne przez zadany model predykcyjny.

Literatura

1. Daniel T. Larose, Metody i modele eksploracji danych, PWN, Warszawa, 2017
2. Siegmund Brandt, Analiza Danych, PWN, Warszawa, 2016
3. Odkrywać! Ujawniać! Objaśniać! Zbiór esejów o sztuce prezentowania danych, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa 2014

- Dodatkowa literatura będzie podawana w referencjach do poszczególnych części wykładu

Komunikacja

- Dodatkowe materiały związane z realizacją przedmiotu będą ukazywać się na platformie MS Teams.
- Informacje o ocenach będą przekazywane za pomocą narzędzia Sprawdziany w systemie USOS.
- Pozostałe informacje dotyczące bieżącej realizacji przedmiotu będą przekazywane przy pomocy USOSMAIL.

Oceny cząstkowe

- Laboratoria
  - Ocena czterech zadań punktowanych (40%).
- Projekt
  - Ocena dłuższego projektu (60%) w tym:
    - Ocena uzyskanej jakości wyników predykcji w porównaniu z działaniem modelu operującego na nieprzetworzonych danych (30%).
    - Ocena sposobu przeprowadzenia i dokumentacji eksperymentów porównujących modele (30%).

Zaliczenie przedmiotu

- Do zaliczenia przedmiotu wymagane jest
  - Uzyskanie ponad połowy punktów za każdy element oceny cząstkowej
    - Laboratoria
    - Ocena jakości wyników predykcji
    - Ocena przeprowadzenia i dokumentacji eksperymentów
  - Uzyskanie ponad połowy sumarycznej liczby punktów z przedmiotu

Ocena

- Na podstawie uzyskanych punktów
  - Ponad 90 procent – 5.0
  - Ponad 80 procent – 4.5
  - Ponad 70 procent – 4.0
  - Ponad 60 procent – 3.5
  - Ponad 50 procent – 3.0
- Studenci z 50 procentami i mniej nie zaliczają przedmiotu

Zajęcia laboratoryjne

- Zajęcia laboratoryjne dzielą się na punktowane i niepunktowane.
- Zajęcia punktowane sprawdzają wiedzę z wykładu i niepunktowanych zajęć laboratoryjnych.
- Można poprawić jedne, najniżej ocenione zajęcia punktowane w terminie zajęć poprawkowych.

Zajęcia punktowane

- Zajęcia punktowane wymagają podania odpowiedzi na pytania przekazane studentom w momencie rozpoczęcia zajęć.
- Pojedyncze zajęcia pozwalają na zdobycie maksymalnie czterech punktów.
- Podczas zajęć można korzystać z Internetu i własnego sprzętu komputerowego.
- Na zajęciach punktowanych mogą przebywać tylko osoby przypisane do nich w systemie USOS.

Harmonogram laboratoriów

- Pierwsze dwa laboratoria będą poświęcone zapoznaniu się narzędziami analizy danych w środowisku R.
- Na kolejnych zajęciach przeprowadzone zostaną cztery iteracje modułów tematycznych
  - Dwa zajęcia niepunktowane
  - Zajęcia punktowane podsumowujące moduł.
  - Zajęcia punktowane mogą wymagać wiedzy z poprzednich modułów.
- Zajęcia poprawkowe odbędą się na ostatnich zajęciach semestru

Tematyka projektu

- Projekty dotyczą analizy rzeczywistych danych.
- Zadania sprowadzają się do zadań regresji, klasyfikacji lub klasteryzacji.
- Wszystkie zadania są zadaniami predykcji, więc należy przewidzieć przyszły stan na podstawie danych historycznych.

Sposób realizacji

- Tematy przydzielane są w sposób losowy.
- Realizacja projektów w zespołach trzyosobowych.
- Co najmniej pięć spotkań kontrolnych w terminach ustalonych przez zespół.
- Spotkania kontrolne odbywają się w slotach czasowych, na które zespół zapisuje się zdalnie z wyprzedzeniem.

Projekt „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca” współfinansowany jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Zadanie 10 pn. „Modyfikacja programów studiów na kierunkach prowadzonych przez Wydział Matematyki i Nauk Informacyjnych”, realizowane w ramach projektu „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca”, współfinansowanego jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.