Zadanie rekrutacyjne, DSC PJATK 2024

Witaj na rekrutacji DSC PJATK mamy tu zadanie, które sprawdzi Wasze podejście do danych, na których operuje uczenie maszynowe. Zadania mogą być robione wedle uznania, na tyle na ile się zdecydujecie. Termin oddania zadania: 11.06.2024 r.

Dataset: NYPD Complaint Data Historic (W razie problemów z pobraniem datasetu można spróbować skorzystać z tego linku: NYPD Dataset)

- Należy pobrać dataset ze strony - można go pobrać ręcznie. Dodatkowe punkty są udzielane za korzystanie z API lub Scrapingu w celu automatycznego pobrania danych. **UWAGA: Dataset waży ponad 2.5 GB!**
- Praca z danymi:
    - Wczytanie dataset'u do odpowiedniej struktury danych, np. DataFrame.
    - Przejrzenie co zawierają dane, do czego mogą być wykorzystane.
    - Wstępna analiza jakości danych.
    - Odrzucenie zbędnych kolumn wraz z uzasadnieniem.
    - Przygotowanie danych do analizy szczegółowej – wedle uznania.
    - Przygotowanie 2 wizualizacji, które opowiadają jakąś historię.
- Przygotowanie małego modelu ML, może być to regresja lub klasyfikacja, jeśli znajdzie się uzasadnienie to można użyć sieci neuronowych, jednak nie będzie to dodatkowo punktowane. Pamiętajcie, że w zależności od wybranego typu modelu dane muszą być odpowiednio przygotowane poprzez różne transformacje. W przypadku uczenia z nadzorem targetem może być jedna z kolumn, która posiada w miarę zbalansowane (ilościowo) wartości.
- Model powinien zostać stworzony zgodnie z dobrymi praktykami oraz przetestowany – jeśli nie ma w zespole wiedzy na ten temat to wystarczy krótki research w internecie, jest dużo materiałów.
- Przygotowanie krótkiej dokumentacji kodu z instrukcją jak uruchomić projekt.
- Przygotowanie krótkiej prezentacji (max 4-6 slajdów z treścią) pokazującej proces waszej pracy, problemów jakie napotkaliście i podsumowanie waszej pracy.
- Podsumowanie wykonanych prac – po zakończeniu prosimy o maila na dsc@pjatk.edu.pl z linkiem do repozytorium z rozwiązaniem zadania. Proszę każdego członka drużyny dołączyć na CC.

Dużym plusem będą dobre praktyki, idempotentność oraz czytelność rozwiązania.

Całość postaramy się sprawdzić w max 2 tydzień po wysłaniu ankiety. Pamiętajcie, że przede wszystkim chodzi o podejście i zaangażowanie, oraz przy okazji zrobienie całkiem fajnego ćwiczenia.

Powodzenia!