Eksploracja danych

1. **Członkowie zespołu:**

   Piotr Kica
   Maciej Zieliński
   Michał Żelasko

2. **Temat projektu:**

   Analiza zbioru danych Uber Pickups.

3. **Opis projektu:**

   Celem projektu będzie przeanalizowanie danych odnośnie usług wykonywanych przez firmę Uber na obszarze Nowego Jorku w latach 2014-15. W ramach projektu planowane jest znalezienie najbardziej popularnych miejsc i terminów przejazdów (wpływ pory dnia, dnia tygodnia, ewentualnie określonych dni miesiąca, świąt, pór roku etc. na częstotliwość wykonywania usług), porównanie ich z analogicznymi danymi dla innych przewoźników. Dodatkowym celem będzie ustalenie i potencjalne określenie zależności pomiędzy poprzednimi przejazdami klientów i spodziewanym ruchem (oczekujemy, że skoro ktoś jechał z A do B, to istnieje realne prawdopodobieństwo, że będzie wracał z B do A, lub udawał się w jakieś inne miejsce C). Kolejnym elementem może być określenie wpływu faktu, że na dany kierowca operuje tylko na ograniczonym obszarze na efektywność funkcjonowania firmy i liczby wykonanych kursów (na podstawie danych od innych przewoźników).

4. **Wykorzystywany zbiór danych:**

   https://www.kaggle.com/datasets/fivethirtyeight/uber-pickups-in-new-york-city.

   Zawiera następujące dane:

   - dane odnośnie przejazdów z wykorzystaniem firmy Uber od kwietnia do września 2014 roku (z podziałem na miesiące i dokładnymi danymi odnośnie lokalizacji),
   - dane odnośnie przejazdów z wykorzystaniem firmy Uber od stycznia do czerwca 2015 roku (z podziałem na miesiące i dokładność informacji o lokalizacji jest niższa),
   - dane dla innych firm o podobnym profilu działalności, zawierające dane odnośnie firmy, terminu wykonania usługi, lokalizacji, ew. kierowcy itd.
   - zagregowane dane odnośnie ruchu pojazdów z przedsiębiorstw/firm zajmujących się przewozem na terenie Nowego Jorku (z częściowym uwzględnieniem ruchu klasycznych taksówek).

5. **Proponowane modele, algorytmy i metody:**

W pierwszym kroku wykonano wstępną analizę i wizualizację posiadanych danych. Określono popularność usług w zależności od miesiąca, dnia, dnia tygodnia (analiza była prowadzona dla różnych miesięcy) i godziny/pory dnia (analiza była prowadzona dla różnych miesięcy i dni tygodnia). Wyniki analizy zostały przedstawione w postaci dataframe'ów, wykresów słupkowych i histogramów. Wykorzystano biblioteki **pandas, numpy i matplotlib.**
Naniesiono również na mapę Nowego Jorku punkty, w których odbierani byli pasażerowie w zależności od firmy, która wykonywała usługę, miesiąca, dnia tygodnia oraz pory dnia. Do wykonania wizualizacji wykorzystano biblioteki **pandas**, **plotly** (moduł express).
Wyniki przeprowadzonych, wstępnych analiz wraz z ich wizualizacją zostały przedstawione w załączonych plikach (*date_analyzer.ipyn*b oraz *geo_vizualization.ipynb*).

Planowane jest wykonania następujących kroków w celu dalszej analizy posiadanego zbioru danych oraz zastosowania wymienionych poniżej metod (z wyszczególnieniem bibliotek, z których poszczególne metody będą importowane, projekt realizowany będzie przy użyciu języka **python**).

a) Preprocessing danych wejściowych w celu wykrycia kluczowych parametrów dla problemu oraz zbadania korelacji pomiędzy poszczególnymi cechami.
   Metody:
   - PCA,
   - kernel-PCA,
   - t-SNE (choć głównie dla lepszego zrozumienia zbioru danych).

   Planowane wykorzystanie bibliotek:
   - scikit-learn,
   - matplotlib

b) Zastosowanie klasteryzacji do wyszczególnienie rejonów miasta, z których pochodzi najwięcej zamówień na przejazd oraz do znalezienia najintensywniejszych okresów w ruchu miejskim (w szczególności do znalezienia powiązań pomiędzy tymi cechami).
   Metody:
   - kmeans,
   - dbscan (może okazać się skuteczniejszy ze względu na zdolność do eliminacji szumu).

   Planowane wykorzystanie bibliotek:
   - scikit-learn,
   - matplotlib

c) Zastosowanie analizy szeregów czasowych do sprawdzenia następujących cech i prawidłowości w danych:
   - tendencja rozwoju przedsiębiorstw,
   - wahania sezonowe,
   - wahania cykliczne (oczekiwane: cykl dobowy i cykl tygodniowy).

Planowane wykorzystanie biblioteki:

- darts.

d) Z uwzględnieniem dodatkowego zbioru danych Uber&Lyft cab prices stworzenie dodatkowego modelu do predykcji optymalnych/spodziewanych cen usług. Predykcja będzie prowadzona za pomocą następujących metod:
   - regresja(liniowa),
   - prosta sieć neuronowa z docelowym porównaniem jakości predykcji dla danych walidacyjnych i testowych.

   Planowane wykorzystanie biblioteki:
   - scikit-learn do regresji,
   - tensorflow lub pytorch do predykcji z wykorzystaniem sieci neuronowej.

e) *Zbudowanie modelu do przewidywania najlepszych miejsc do *zabierania* klientów w zależności od pory dnia, dnia tygodnia oraz do przewidywania najlepszej pory w zależności od miejsca.
   Planowane wykorzystanie biblioteki:
   - scikit-learn,
   - tensorflow/pytorch

   *Bierzemy pod uwagę możliwość pominięcia/modyfikacji tego podpunktu w zależności od tego czy uda się uzyskać zadowalające i mające sens rezultaty (ze względu na charakter danych ciężko ocenić prawdopodobieństwo stworzenia racjonalnie działającego i nietrywialnego modelu).

6. Wstępna obróbka i wizualizacja danych.
   a. W załączonych plikach przedstawiono wykonane analizy statystyczne i wizualizacje danych.