**Raport końcowy z projektu prowadzonego w ramach przedmiotu Analiza Big Data**

Celem naszego projektu była konstrukcja narzędzi służących do wykrywania terenów pustynnych w zadanych obszarach Ameryki Północnej. Do realizacji tego zadania wykorzystaliśmy zbiór danych z zasobów NASA, w którego skład wchodzi 56 kolumn: trzy z nich zawierają informacje o długości i szerokości geograficznej (z dokładnością do 1/8 stopnia) oraz dacie, pozostałe natomiast stanowią zbiór cech charakteryzujących dany obszar w podanym okresie czasu. Lokalizacje, którymi dysponowaliśmy to: całe terytorium Stanów Zjednoczonych z wyłączeniem stanu Alaska, południowa część Kanady oraz północna część Meksyku, reprezentowane przez 76 360 par współrzędnych. Okres czasu, z którego pochodzą dane to styczeń 1979 – wrzesień 2023. Łączny czas trwania naszych prac nad projektem to 12 tygodni podzielonych na trzy sprinty. Omówię pokrótce przebieg każdego z nich oraz przedstawię osiągnięcia dokonane przez nasz zespół na ich przestrzeni.

**Sprint 1**

W tym sprincie w głównej mierze zależało nam na zaznajomieniu się ze specyfiką danych oraz zorientowaniu się, które parametry najlepiej charakteryzują tereny pustynne. Dzięki zebraniu i przeanalizowaniu odpowiednich badań naukowych udało nam się wyodrębnić zestaw dziewięciu najważniejszych cech, które stanowiły fundament dalszej eksploracji danych. Przeprowadziliśmy szereg analiz statystycznych takich jak badanie zmian wskaźnika zielonej roślinności w czasie oraz jego korelację z pozostałymi zmiennymi, ale również detekcję sezonowości dla reszty istotnych współczynników. Oprócz samej analizy danych, powstały również dwa użyteczne narzędzia: algorytm umożliwiający przedstawianie wizualizacji na mapie oraz funkcja przypisująca lokalizacjom etykiety miast, o ile do jakiegokolwiek przynależą. Dodatkowo, został utworzony klasyfikator heurystyczny, który na podstawie odpowiedniego układu wartości wyselekcjonowanych cech rozdziela zbiór dostępnych obszarów na rzekomo pustynne i rzekomo niepustynne.

**Sprint 2**

W kolejnym etapie naszych prac położyliśmy nacisk na próbie podziału danych na tereny pustynne i niepustynne z wykorzystaniem bardziej wyrafinowanych metod. Zastosowano techniki klasteryzacji takie jak *k* najbliższych sąsiadów, algorytm Expectation-Maximization czy też *k*-średnich. Grupowań dokonano w kilku wariantach: z użyciem wszystkich wybranych cech, ze wsparciem analizy głównych składowych oraz wykorzystując pojedyncze parametry i ich pary. Wizualizacje sporządzone do ostatniego z wariantów przyniosły najwięcej korzyści, ponieważ wskazały, które czynniki najlepiej kształtują tereny pustynne. W celu weryfikacji jakości predykcji powstał anotowany zbiór danych zawierający 1200 lokalizacji zaetykietowanych w kategoriach: „pustynia”, „step”, „inne”. Dla przeciwwagi, do klasyfikacji terenów na pustynne i niepustynne, stworzyliśmy również algorytm uczenia nadzorowanego bazujący na drzewach decyzyjnych. Ponadto sprawdzono skuteczność klasyfikatora heurystycznego w wersji wytrenowanej na miesiącach od maja do października (oryginalnej) oraz na pozostałych miesiącach i przedstawiono rezultaty na mapach.

**Sprint 3**

Ostatni etap projektu poświęciliśmy podrasowaniu najlepszych modeli uzyskanych w sprincie poprzednim. Zostały podjęte próby ulepszenia grupowania *k*-średnich z wykorzystaniem algorytmu t-SNE, jednak nie przyniosły one oczekiwanych rezultatów, podobnie jak próba posłużenia się algorytmem Birch. Zbudowano i porównano modele regresji logistycznej wytrenowane na danych z miesięcy letnich i zimowych, jednocześnie stosując różne kombinacje cech branych pod uwagę. Podobnie postąpiono stosując SVM z różnymi modyfikacjami funkcji jądra. Do uzyskania lepszych wyników klasyfikacji metodą drzew decyzyjnych zastosowano algorytm lasów losowych oraz Light GBM. Ze względu na zauważalną poprawę, tę samą predykcję przeprowadzono również dla zagregowanych do 3,4 i 5 lat danych. Na koniec przetestowano jak odfiltrowanie terenów miejskich z danych wpływa na wyniki powyżej omówionych klasyfikacji.

**Podsumowanie – przegląd kilku najskuteczniejszych otrzymanych modeli**

* Model Light GBM, wytrenowany na uśrednionych danych z lipca z lat 2020 – 2023 (95% accuracy)
* Model SVM z wielomianową funkcją jądra, wytrenowany na danych z lipca 2019 (93.7% accuracy)
* Model regresji logistycznej wspartej stochastycznym zejściem gradientowym, wytrenowany na danych z lipca 2019 (93.6% accuracy)
* Model *k*-średnich z wykorzystaniem wszystkich składowych, wytrenowany na danych z lipca 2023 (89% accuracy)