**

Drzewo decyzyjne

Wydział: Zarządzania

Kierunek studiów: Informatyka i ekonometria

Rok studiów: II

Rodzaj studiów: Studia stacjonarne

Autorzy: Julia Baranowska, Aleksander Folfas, Julia Głowa

# Opis projektu

W projekcie przedstawiono algorytm ID3 –Quinlana, który tworzy drzewo decyzyjne. Program został napisany w języku Python przy wykorzystaniu poniższych bibliotek:

* numpy;
* pandas;
* math;
* networkx;
* matplotlib.pyplot;

Algorytm przedstawiony w projekcie miał za zadanie odpowiedzieć na pytanie czy osoba o danych cechach dostanie się na staż informatyczny, a jeśli tak – to odpowiada na jaki konkretny staż się dostanie.

Przesłankami w projekcie są:

1. Status studenta.
2. Poziom wykształcenia – brak, licencjat, magister.
3. Doświadczenie - akademickie, brak, komercyjne.
4. Poziom zaawansowania języka angielskiego – podstawowy, zaawansowany, biegły.
5. Język programowania – Python, R, Java, HTML.

Owocem przesłanek jest jedna konkluzja, której opcjami są Data Analyst, Software Developer, UX/UI Designer lub brak stażu.

Algorytm ID3-Quinlana polega na tworzeniu drzew decyzyjnych z danych treningowych. Algorytm dokonuje tego poprzez wybieranie atrybutów, które najlepiej podzielą zbiór danych na mniejsze podzbiory o podobnych właściwościach. Kolejnym krokiem algorytmu jest tworzenie węzłów drzewa decyzyjnego, które reprezentują wybrane atrybuty. Każdy węzeł ma swoją etykietę, która odpowiada wartościom tego atrybutu. Następnie algorytm tworzy poddrzewa dla każdej wartości atrybutu, z których każde reprezentuje jedną gałąź w drzewie. Gałęzie są tworzone do momentu otrzymania pojedynczej konkluzji.

# Funkcje zastosowane w projekcie

Funkcja *„entropy”* ma za zadanie obliczyć wartość entropii dla kolejnych możliwości przesłanek.

Następnie funkcja *„information\_gain”* oblicza przyrost informacji, na podstawie której podzielimy tabelę.

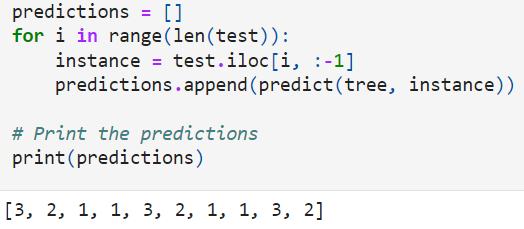
Funkcja *„best\_feature”* wyszukuje kolumnę z największym przyrostem informacji.

Funkcja *„build\_tree*” dzieli tablice na podstawie kolumny otrzymane dzięki powyższej funkcji i zwraca drzewo.

Funkcja *„predict”* ma za zadanie dopasować konkluzję dla danych testowych – sprawdzających działanie projektu.

# Wynik dla danych testowych

Dane testowe posiadają informacje na temat 10 różnych. Dla każdej z nich zaprogramowany algorytm zwraca informację na jaki staż się dostaną.



Program przypisał im kolejno staże:

Software developer (3), UX/UI designer (2), Data Analyst (1), Data Analyst (1), Software developer (3), UX/UI designer (2), Data Analyst (1), Data Analyst (1), Software developer (3), UX/UI designer (2).

Przesłanki dla pierwszej z osób prezentują się następująco:



Brak statusu studenta, tytuł magistra, doświadczenie komercyjne, podstawowa znajomość języka angielskiego oraz znajomość języka Java. Tej osobie przypisano staż Software Developer wydaje się być logicznym wyborem.