Raport z Laboratorium

Wersja: Podstawowa (4.0)

Tsimafei Kotski 158740

Vasil Kusmartsev 156202

Grupa: L2

Celem laboratorium było zaimplementowanie algorytmu generacji drzewa decyzyjnego od podstaw, bez użycia gotowych bibliotek.

1. Obliczanie entropii

Entropia mierzy niepewność w zbiorze danych. Dla całego zbioru Titanic entropia wyniosła 0.9710, co wskazuje na wysoką niepewność w danych.

2. Obliczanie entropii warunkowej

Entropia warunkowa to średnia entropia po podziale danych według wybranego atrybutu.

3. Obliczanie Information Gain

Information Gain mierzy redukcję entropii po podziale danych.

4. Obliczanie Intrinsic Information

Intrinsic Information normalizuje Information Gain względem liczby wartości atrybutu. Zapobiega to preferowaniu atrybutów z wieloma wartościami.

5. Obliczanie Gain Ratio

Gain Ratio to stosunek Information Gain do Intrinsic Information.

6. Wyniki dla wszystkich atrybutów

Pclass | IG: 0.0817 | II: 1.3702 | GR: 0.0596

Sex | IG: 0.3915 | II: 0.9710 | GR: 0.4032

Age | IG: 0.0093 | II: 1.4907 | GR: 0.0063

SibSp | IG: 0.0407 | II: 1.6191 | GR: 0.0251

Parch | IG: 0.0166 | II: 1.1326 | GR: 0.0146

Wybrany jako najlepszy został atrybut Sex, który miał najwyższy Gain Ratio równy 0.4032. Oznacza to, że płeć najlepiej separuje klasy survived i not survived.

7. Rekurencyjne budowanie drzewa

Algorytm działa rekurencyjnie wybierając najlepszy atrybut i dzieląc dane na podgrupy. Proces kończy się gdy wszystkie przykłady są jednej klasy lub osiągnięta zostanie maksymalna głębokość.

8. Graficzna wizualizacja

Drzewo zostało zwizualizowane graficznie z kolorowymi węzłami. Niebieskie węzły to pytania, zielone i czerwone to liście z decyzjami.

Wnioski

Program poprawnie implementuje wszystkie wymagane funkcje. Atrybut Sex okazał się najbardziej istotny dla przeżycia na Titanicu, co odzwierciedla historyczny fakt zasady kobiety i dzieci pierwsze.