

Tema Practică

Ștefan Doagă

January 16, 2025

1 Introducere

În acest studiu, abordăm problema decizională a unui proprietar de lanț hotelier, care dorește să deschidă un nou hotel într-o țară specificată, pentru a maximiza veniturile (*Revenue*) și/sau venitul pe cap de vizitator (*Revenue/Visitors*). Folosind datele disponibile din setul de date [Tourism Dataset](#), scopul este să generăm o ierarhie a celor șase categorii tematice de activități (*Category*) care maximizează aceste metrice. Pentru această sarcină, implementăm algoritmul k-Nearest Neighbors (kNN), utilizând metodologia de învățare supervizată.

1.1 Structura și attributele datasetului

Setul de date conține următoarele attribute principale:

- **Location:** Un șir aleator de zece caractere care reprezintă numele locației.
- **Country:** Țara în care se află locația turistică (7 valori posibile).
- **Category:** Categoria activităților tematice (Nature, Historical, Cultural, Beach, Adventure, Urban).
- **Visitors:** Numărul de vizitatori anuali (valoare numerică).
- **Revenue:** Venitul generat anual (valoare numerică).
- **Rating:** Un număr de tip float aleatoriu în intervalul 1,0 și 5,0.

2 Algoritmul k-Nearest Neighbors

2.1 Motivație

Algoritmul kNN a fost selectat datorită următoarelor avantaje:

- Este un algoritm intuitiv și interpretabil, adecvat pentru probleme de regresie.
- Permite modelarea relației complexe dintre *Visitors*, *Revenue* și *Category*.
- Este eficient pentru seturi de date de dimensiuni moderate, precum cel utilizat.

2.2 Implementarea modelului

Implementarea modelului kNN a urmat pașii de mai jos:

1. **Pregătirea datelor:** Datele au fost împărțite în două subseturi:
 - Setul de antrenare: Datele din toate țările, exceptând țara țintă.
 - Setul de testare: Datele corespunzătoare țării țintă.
2. **Antrenarea modelului:** Un model kNN pentru regresie a fost configurat cu $k = 5$ vecini.
3. **Predicții:** Modelul a fost utilizat pentru a prezice *Revenue/Visitors* pentru fiecare categorie tematică.

3 Rapoarte

În cadrul experimentelor s-au folosit rapoarte convenabile pentru împărțirea setului complet de date în set de antrenament și set de testare/validare (80/20 și 90/10)

Ex: Australia

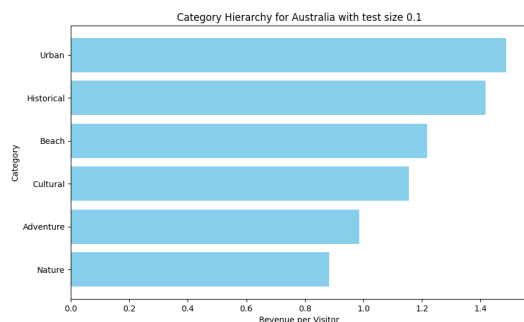


Figure 1: 90/10

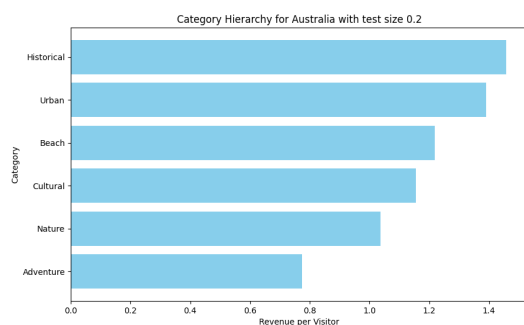


Figure 2: 80/20

4 Concluzii și selecția finală

Pe baza experimentelor:

- kNN oferă un echilibru excelent între performanță, interpretabilitate și eficiență computațională.
- Complexitatea redusă a kNN îl face potrivit pentru implementarea manuală, conform cerințelor.

Astfel, **kNN** a fost selectat pentru implementarea finală. Proprietățile teoretice și performanța susțin alegerea sa drept cea mai potrivită metodă pentru problema studiată.