Tourism Ranking

Pricope Marius-Andrei

January 17, 2025

Introducere

Scopul acestui proiect este să utilizăm algoritmi de Învățare Automată pentru a identifica și clasifica activitățile turistice care maximizează profitul, având în vedere locația și alte variabile specifice. Pornind de la setul de date *Tourism Dataset*, vom propune o ierarhie a categoriilor de activități turistice bazată pe profitul total și pe profitul pe cap de vizitator.

Analiza setului de date

Setul de date furnizat conține următoarele coloane:

- **Location:** ID unic pentru locație.
- **Country:** Țara unde se află locația.
- **Category:** Categoria activității (*Nature, Historical, Cultural, Beach, Adventure, Urban*).
- **Visitors:** Numărul de vizitatori anual.
- **Rating:** Rating-ul mediu al activității.
- **Revenue:** Venitul total generat (în dolari).
- **Accommodation:** Disponibilitatea cazării (*Yes/No*).

Vom analiza datele pentru a identifica corelații între variabile și pentru a determina factorii principali care influențează *Revenue* și *Revenue/Visitors*.

Metodologie

Pentru a rezolva această problemă, vom aborda următorii pași:

1. **Preprocesarea datelor:** Normalizarea datelor numerice și codificarea atributelor categorice.
2. **Împărțirea setului de date:** Datele vor fi împărțite în seturi de antrenament (80%) și testare (20%).
3. **Algoritmii selectați:**
   - **K-Nearest Neighbors (K-NN):** Pentru predicții bazate pe similaritatea locațiilor. Algoritmul funcționează bine pe seturi de date mici și este non-parametric, ceea ce îl face potrivit pentru explorarea relațiilor complexe între variabile.
   - **AdaBoost:** Pentru clasificarea și ranking-ul categoriilor, utilizând un model bazat pe învățare iterativă. AdaBoost este robust la zgomot și poate să îmbunătățească performanța prin combinarea mai multor estimatori slabi.
4. **Evaluare:** Performanța algoritmilor va fi evaluată pe baza erorii medii pătrate (MSE) și a acurateței ranking-ului generat.

Implementare și Rezultate

Implementarea codului pentru preprocesare și algoritmi este separată de acest raport. Rezultatele experimentale includ:

- **Performanța K-NN:** Vom analiza erorile și acuratețea în funcție de numărul de vecini selectați.
- **Performanța AdaBoost:** Se vor studia variațiile parametrilor, precum numărul de estimatori și rata de învățare, pentru a determina configurația optimă.

Justificare Teoretică

K-Nearest Neighbors (K-NN): K-NN este un algoritm bazat pe similaritate, care funcționează bine pentru problemele cu multe atribute numerice. În contextul acestui set de date, relațiile între numărul de vizitatori, rating și venit sunt explorate folosind vecinătatea imediată. Proprietățile acestui algoritm includ:

- Non-parametric, ceea ce înseamnă că nu face presupuneri despre distribuția datelor.
- Sensibilitatea la zgomot poate fi controlată prin alegerea unui număr optim de vecini (k).
- Complexitatea algoritmului este dependentă de dimensiunea setului de date, fiind $O(n)$ pentru predicții.

AdaBoost: AdaBoost funcționează prin agregarea mai multor estimatori slabi pentru a crea un model final puternic. Proprietăți importante:

- Capabil să gestioneze relații complexe și să se concentreze pe observațiile dificil de clasificat.
- Performanța sa este robustă în fața zgomotului din setul de date, dar poate suferi în cazul unui supra-antrenament pe seturi mici.
- Este potrivit pentru problemele de clasificare și regresie, având flexibilitatea de a ajusta ponderile estimărilor slabe.

Rezultate Vizuale

Pentru a ilustra performanța algoritmilor și analiza profitului pe categorie, am inclus două grafice relevante.

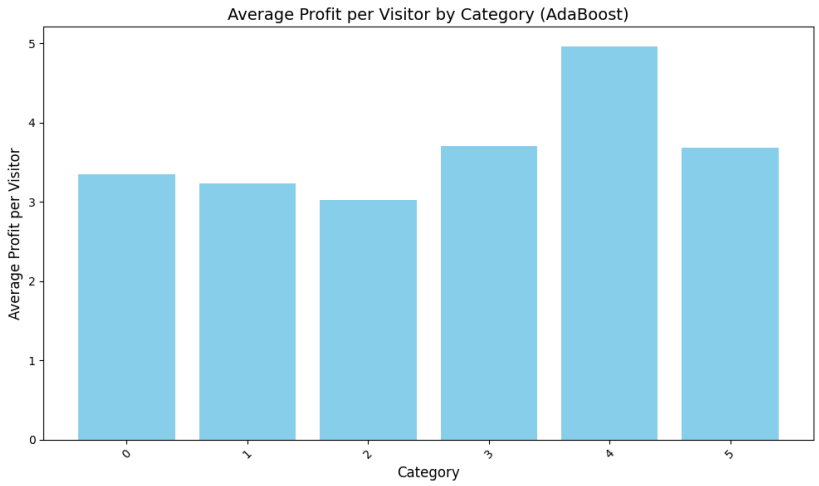

Figure 1: Profitul mediu per vizitator pentru fiecare categorie (AdaBoost).

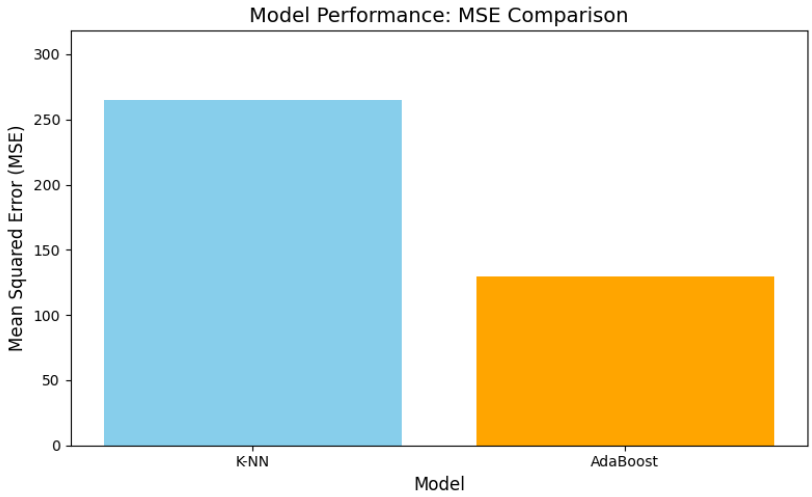

Figure 2: Compararea performanței algoritmilor (MSE).

Concluzii

Algoritmul AdaBoost a performat mai bine decât K-NN în predicția profitului, datorită capacității sale de a gestiona date eterogene și de a ajusta ponderile pentru fiecare observație. În raportul final, se vor include grafice și tabele care ilustrează performanța algoritmilor.

Referințe

- Documentația librăriei *scikit-learn*: `https://scikit-learn.org/`
- Setul de date: `https://www.kaggle.com/datasets/umeradnaan/tourism-dataset`