Analiza dataset-ului *Tourism Dataset* pentru maximizarea profitului

Bocănet Raluca-Andreea

17 Ianuarie 2025

1 Introducere

Scopul acestui raport este de a analiza dataset-ul *Tourism Dataset* pentru a maximiza profitul unui lant hotelier intr-o tara selectata. Obiectivul este de a identifica categoriile tematice care contribuie cel mai mult la venituri si de a propune o strategie de prioritizare a acestora. In aceasta analiza, am ales **USA** ca tara de studiu datorita diversitatii categoriilor tematice si a volumului mare de date disponibile.

2 Metodologie

2.1 Preprocesarea Datelor

Setul de date original contine informatii despre tari, categorii tematice, venituri si numarul de vizitatori. Au fost realizate urmatoarele etape de preprocesare:

- Selectarea datelor pentru tara **USA**.
- Crearea unei noi caracteristici: **Venituri per Vizitator** (*Revenue per Visitor*).
- Transformarea categoriilor tematice in variabile dummy pentru a permite utilizarea acestora.
- Impartirea datelor (80%-20%): Am folosit functia train_test_split din sklearn pentru a imparti setul de date in doua parti: 80% pentru antrenament si 20% pentru testare.

2.2 Modele folosite

Au fost selectate doua modele pentru aceasta analiza:

- **Random Forest Regressor**: este o metoda care foloseste mai multi arbori de decizie pentru a face predictii. Fiecare arbore ia o decizie pe baza unui subset de date, iar rezultatele tuturor arborilor sunt combinate pentru a oferi o predictie finala. Este buna pentru captarea relatiilor complexe si nelineare intre date, oferind de obicei rezultate precise..
- **K-Nearest Neighbors (KNN)**:este o metoda simpla de predictie care se bazeaza pe gasirea celor mai apropiate exemple din datele deja cunoscute. Cand vrei sa prezici o valoare pentru un nou punct de date, KNN cauta "vecinii" cei mai apropiati si face media valorilor acestora pentru a face predictia. Este usor de inteles, dar poate fi mai lent si mai putin eficient pe date mari.

2.3 Metrice de Evaluare

Performanta modelelor a fost evaluata utilizand urmatoarele metrice:

- **Mean Squared Error (MSE)**: o metrica de eroare utilizata pentru a evalua diferentele dintre valori observate si valori prezise.
- **R-squared (R^2)**: masoara proportia variatiei explicate de model.

3 Rezultate Experimentale

3.1 Performanta Modelelor

Tabelul de mai jos prezinta rezultatele obtinute de cele doua modele:

Model	Mean Squared Error (MSE)	R-squared (R^2)
Random Forest	10256.67	0.89
K-Nearest Neighbors	10769.78	0.75

3.2 Importanta Caracteristicilor

Modelul Random Forest a permis identificarea importantei relative a categoriilor tematice. Rezultatele sunt prezentate imaginea de mai jos.

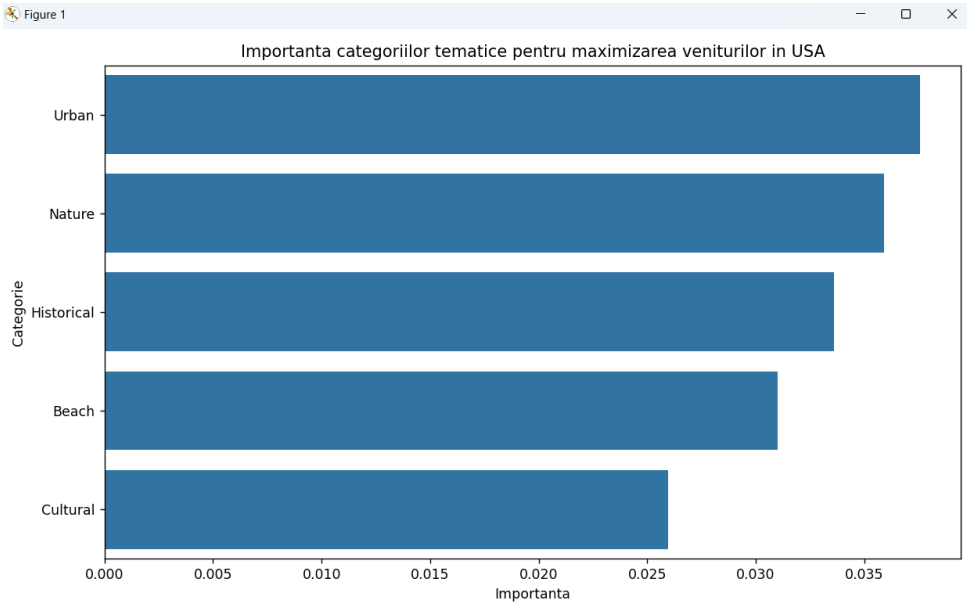
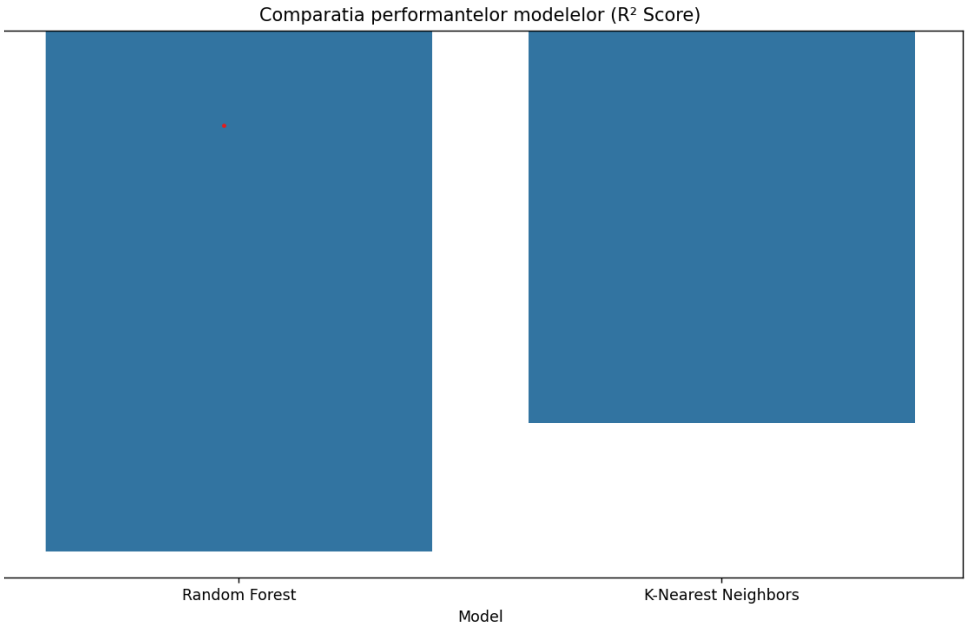


Figure 1: Importanta categoriilor tematice pentru maximizarea veniturilor in USA.

4 Analiza Comparativa

Modelul Random Forest a demonstrat performante mai bune in comparatie cu KNN, atat in ceea ce priveste *MSE*, cat si R^2. Acest lucru se datoreaza capacitatii sale de a captura relatii complexe si de a gestiona variabile irelevante.



5 Justificare teoretica

5.1 Random Forest

Algoritmul Random Forest foloseste mai multi arbori de decizie si tehnica de *bagging* pentru a reduce erorile. Este robust la zgomot si poate gestiona date eterogene.

- Rezistent la zgomot si variabile irelevante.
- Poate lucra cu date mixte.

5.2 K-Nearest Neighbors

KNN este un model non-parametric care face predictii pe baza vecinilor cei mai apropiati. Este sensibil la zgomot si la alegerea hiperparametrului k.

6 Concluzii

Modelul Random Forest a fost selectat datorita performantelor superioare si interpretabilitatii oferite prin importanta caracteristicilor si K-NN deoarece este unul dintre cei mai usori algoritmi de inteles si cu care se clasifica instantele.