Aplicatie Practica 2

Popescu Mara 3B3

Abordare teoretică

1. **Analiza setului de date**
   Datele disponibile în link includ informații despre turism, cum ar fi:
   - Țara (Country)
   - Categorie de activitate (Category: Nature, Historical, etc.)
   - Vizitatori (Visitors)
   - Venit (Revenue)

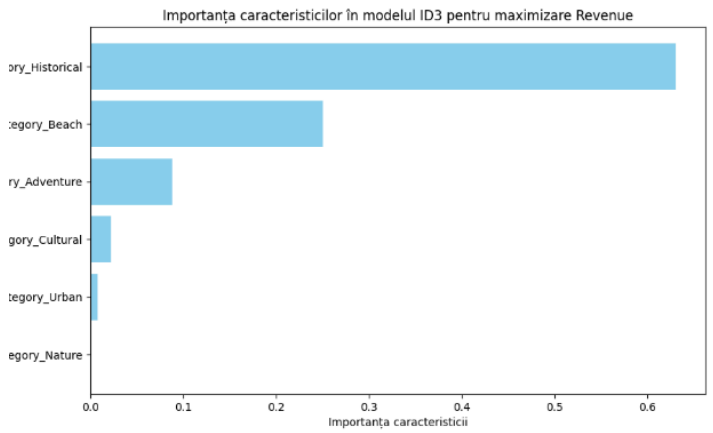
   Obiectivul este să generăm o ierarhie a categoriilor de activități pentru o anumită țară, în funcție de contribuția lor la maximizarea venitului (Revenue) și/sau a venitului per vizitator (Revenue/Visitors).

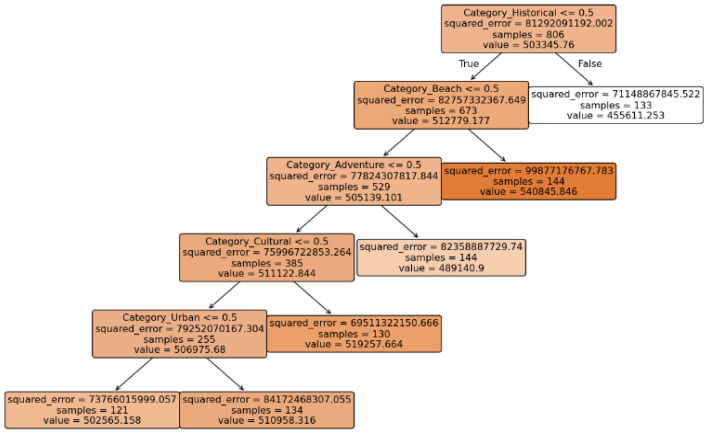
2. **Prelucrarea datelor**
   - Eliminarea valorilor lipsă și curățarea datelor.
   - Generarea de caracteristici suplimentare, cum ar fi:
     - Venit pe cap de vizitator: `Revenue_per_visitor = Revenue / Visitors`.
   - Crearea unui subset de date pentru țara de interes.
3. **Selecția algoritmilor**
4. **Metodologia de evaluare**
   - Metrici de performanță:
     - R^2, MAE, MSE pentru predicția venitului.
     - Corelația dintre predicțiile modelului și veniturile reale pentru ierarhizare.
   - Cross-validare pentru a asigura generalizarea.
5. **Crearea ierarhiei**
   După antrenarea modelelor, se vor calcula scoruri pentru fiecare categorie în funcție de țara fixată. Aceste scoruri pot fi folosite pentru a stabili un ranking.

Am testat fiecare algoritm studiat, folosind implementarile din libraria sklearn. Toate rezultatele sunt pentru Revenue/Visitor pentru India.

1. ID3

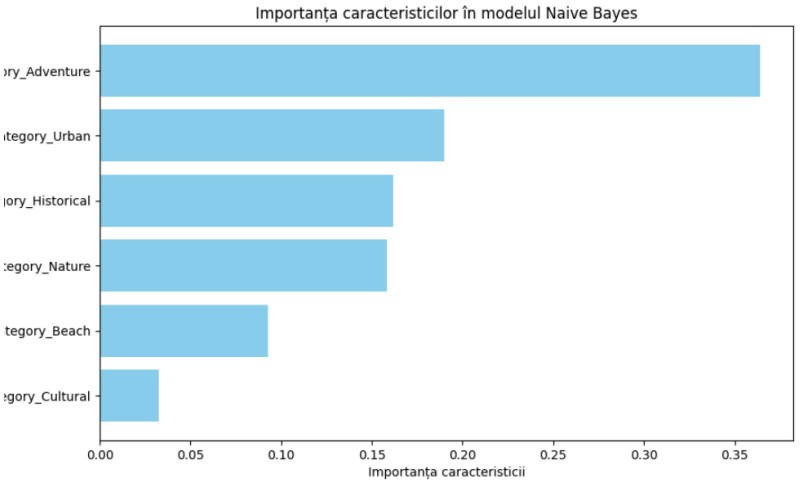
ID3 poate fi utilizat pentru a construi un arbore de decizie bazat pe categorii și pentru a determina ierarhia activităților în funcție de venitul sau venitul per vizitator. Scorul pe setul de testare: 0.0151





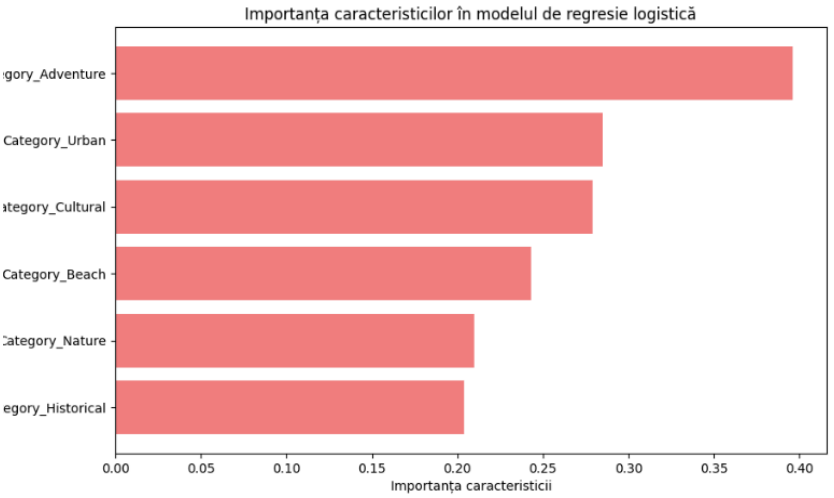
2. Bayes Naiv

Am calculat importanța caracteristicilor pe baza valorilor `theta_` ale modelului Naive Bayes. Pentru fiecare caracteristică, am calculat media valorilor pentru toate clasele (pentru că Naive Bayes calculează un set de parametri pentru fiecare clasă).Scorul pe setul de testare: 0.2333



3. Regresie Logistica

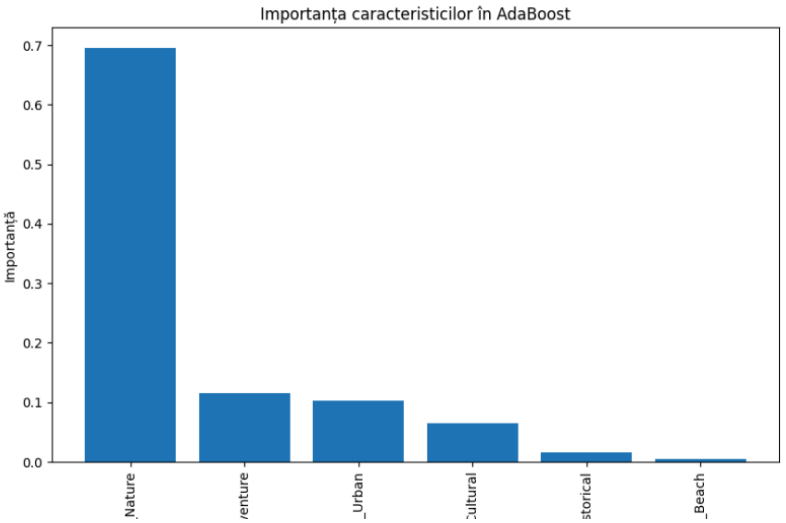
Importanța caracteristicilor în regresia logistică este dată de coeficientul fiecărei caracteristici. Am calculat media valorilor coeficientului pentru fiecare caracteristică, având în vedere că regresia logistică poate să aibă



coeficienti diferiți pentru fiecare clasă.
Scorul pe setul de testare: 0.9333

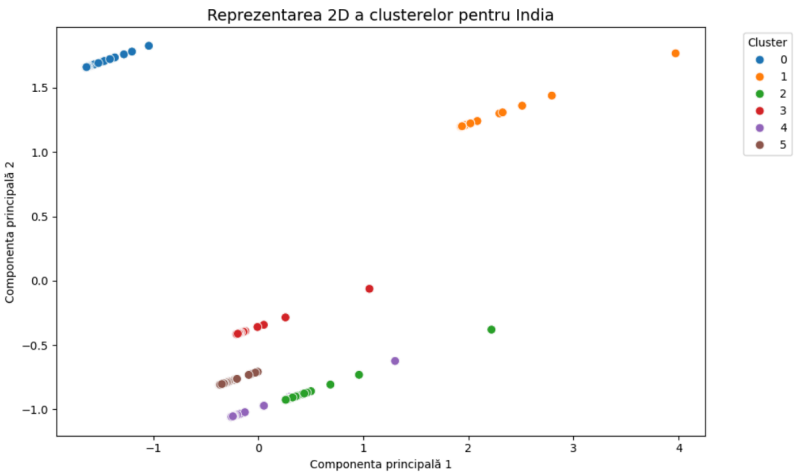
4. AdaBoost

R² de -0.0347 și **MSE de 67.33** sugerează că modelul nu se comportă bine. Un **R² negativ** indică o potrivire slabă cu datele, iar **MSE-ul mare** sugerează erori semnificative în predicțiile modelului.



5.Clusterizare K-Means

Ajusted Rand Index pe setul de testare: 1.0
Ierarhizare:
Adventure
Urban
Historical
Beach
Nature
Cultural



Algoritmul ales

Comparatie

Metoda	Scor	Avantaje	Dezavantaje
ID3	0.0151	-usor de interpretat -bun pentru date discrete	-performanta slaba -sensibil la supraincarcare -necesita preprocesare pentru variabile continue
Bayes Naiv	0.2333	-simplu si rapid de implementat -potrivit pentru date independente	-presupune independenta caracteristicilor(nere alist in acest context) -performanta redusa
Regresie Logistica	0.9333	-performanta ridicata -robusta -coeficienti interpretabili pentru analiza importantei	-limitata la relatii liniare intre caracteristici si probabilitati
AdaBoost	R2=−0.0347 MSE=67.33	-combina mai multe metode slabe pentru a imbunatati performanta	-nu gestioneaza bine predictia in acest task
K-Means	Ajusted Rand Index: 1.0	-grupare nesupravegheata clara -rezultat perfect pe test	-posibila supraincarcare -nu este conceput pentru clasificare predictiva

Regresia Logistica

In urma testelor, regresia logistica pare sa fie metoda care functioneaza cel mai bine pentru task-ul dat, scorul de 93% fiind semnificativ mai mare decat scorul celorlalte metode. Coeficientii modelului ofera un mod direct de a evalua importanta fiecarei categorii, acest lucru fiind esential pentru ierarhizarea in functie de contributia lor la maximizarea revenue/visitor.

Implementare

Pentru a rezolva task-ul am implementat regresia logistica astfel:

- Am implementat functia sigmoida, folosita in regresia logistica pentru a mapa valorile continue intr-un interval intre 0 si 1.
- Am implementat functia de calculare a costului, care utilizeaza urmatoarea formula:

$$J(\theta) = -\frac{1}{m} * \sum[ylog(h) + (1 - y)log(1 - h)]$$

- Am implementat functia de gradient descent pentru a optimiza parametrii theta
- Functia predict transforma probabilitatile generate de functia sigmoid in predictii binare
- Functia feature_importance evalueaza importanta fiecarei caracteristici pe baza valorii coeficientului asociat
- Pe partea de prelucrare a setului de date am filtrat randurile corespunzatoare unei anumite tari (in cazul meu, India), am transformat valorile categoriale in variabile binare, am impartit datele in set de antrenare si de testare(80-20) si am adaugat o coloana de 1-uri pentru termenul liber.

Rezultatele pe care le-am obtinut cu implementarea mea au fost un pic mai slabe decat cele obtinute cu sklearn:

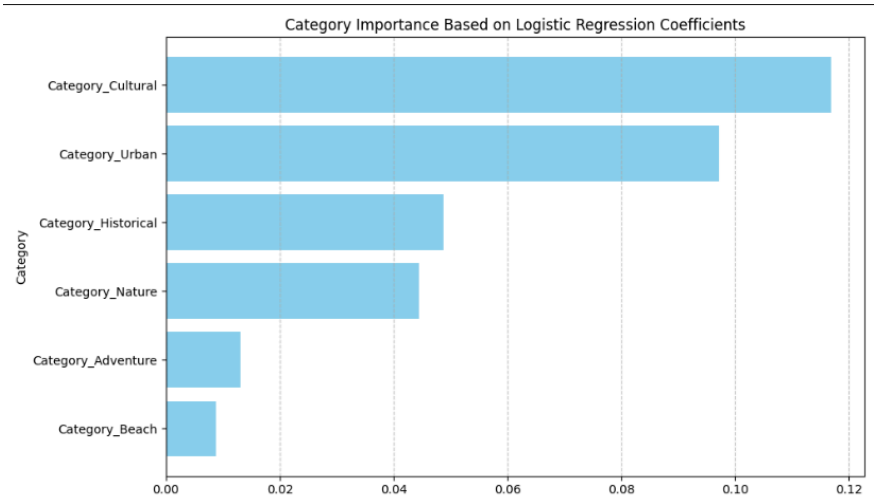
Rezultatele maxime au fost obtinute pentru learning rate 0.1, 1000 iteratii:

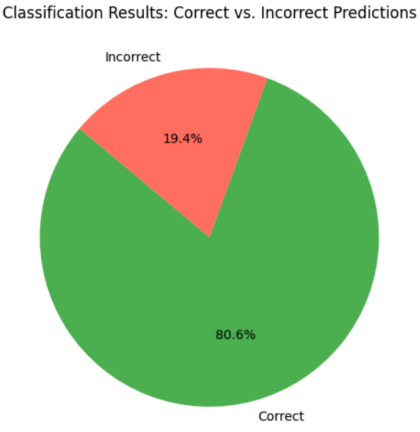
Acuratete: 80.56%

Ranking:

1. Category_Cultural: 0.1169
2. Category_Urban: 0.0971
3. Category_Historical: 0.0487
4. Category_Nature: 0.0444
5. Category_Adventure: 0.0131
6. Category_Beach: 0.0087

Rezultate cross-validare:
[80.55, 73.74, 69.83, 78.77, 69.83]
Mean Accuracy: 74.55%





Am incercat sa imi apropii modelul de cel de la sklearn prin aplicarea regularizarii L2, ajustand functia de cost si gradientul astfel:

$$J(\theta) = -\frac{1}{m}\sum[ylog(h) + (1 - y)log(1 - h)] + \frac{\lambda}{2m}\sum\theta^2$$

$$gradient = \frac{1}{m}X^T(h - y) + \frac{m}{\lambda}\theta$$

Insa nu am reusit sa obtin acuratete mai mare de 80%.