Analiza setului de date și deducerea metodologiei

1 Analiza setului de date și deducerea metodologiei

Setul de date analizat conține următoarele caracteristici principale: **Accommodation_Available**, **Category**, **Country**, **Rating**, **Revenue** și **Visitors**. Pentru a înțelege relația dintre aceste variabile, s-a generat un tabel de corelații care prezintă relațiile numerice dintre ele. Rezultatele sunt analizate mai jos.

	Accommodation_Available	Category	Country	Rating	Revenue	Visitors
Accommodation_Available	1.000	0.000	0.000	0.027	0.017	0.024
Category	0.000	1.000	0.000	0.000	0.021	0.000
Country	0.000	0.000	1.000	0.028	0.016	0.000
Rating	0.027	0.000	0.028	1.000	0.000	-0.010
Revenue	0.017	0.021	0.016	0.000	1.000	0.009
Visitors	0.024	0.000	0.000	-0.010	0.009	1.000

Figure 1: Tabelul de corelații pentru variabilele din setul de date.

1.1 Interpretarea tabelului de corelații

Rezumatul observațiilor cheie este prezentat în continuare:

- **Accommodation_Available:** Prezintă o corelație foarte slabă cu toate celelalte variabile, cu o valoare maximă de **0.027** față de *Rating*. Acest lucru sugerează că disponibilitatea unităților de cazare nu influențează semnificativ alte variabile analizate.
- **Category:** Nu are corelații notabile (majoritatea valorilor fiind **0.000**). Aceasta indică faptul că tipul de activitate turistică nu are o influență directă semnificativă asupra metricilor financiare sau asupra numărului de vizitatori.
- **Country:** Prezintă corelații foarte slabe, cu valori maxime de **0.028** față de *Rating* și **0.016** față de *Revenue*. Acest lucru evidențiază o influență minimă a locației geografice asupra veniturilor și evaluărilor.

- **Rating:** Corelează slab pozitiv cu *Accommodation_Available* (**0.027**) și *Country* (**0.028**), dar are o corelație ușor negativă cu *Visitors* (**-0.010**). Aceasta poate indica un comportament atipic, în care evaluările mai mari nu atrag neapărat mai mulți vizitatori.
- **Revenue:** Corelează foarte slab cu toate celelalte variabile, cu o valoare maximă de **0.021** față de *Category*. Aceasta sugerează că venitul total nu este influențat puternic de vreo caracteristică specifică analizată.
- **Visitors:** Prezintă o corelație slabă cu *Revenue* (**0.009**), ceea ce sugerează că numărul de vizitatori nu influențează semnificativ venitul per vizitator sau venitul total.

1.2 Utilizarea algoritmilor de învățare automată

Pentru a construi un model predictiv pentru **Revenue**, am folosit algoritmul **Regresie Liniară**. După antrenarea și evaluarea modelului, am obținut următorul rezultat pentru eroarea medie pătratică (MSE):

$$\textbf{MSE} = 79904781029.4081$$

Acest rezultat sugerează că modelul de regresie liniară oferă o performanță acceptabilă, dar în continuare ar putea fi îmbunătățit prin fine-tuning sau prin utilizarea unor algoritmi mai complexi, cum ar fi K-Nearest Neighbors sau alte tehnici de învățare automată.

1.3 Concluzii asupra alegerii algoritmului

Pentru problema analizată, **Regresia Liniară** a fost considerată cel mai adecvat algoritm datorită următoarelor motive:

- Este simplu, interpretabil și bine adaptat problemelor liniare. Corelațiile slabe din date sugerează că nu sunt necesare modele nelineare complicate.
- Este rapid și eficient în antrenare și inferență, ceea ce reprezintă un avantaj, având în vedere dimensiunea redusă a setului de date.
- Coeficienții rezultatului permit extragerea unei ierarhii clare a activităților turistice (*Category*) în funcție de impactul lor asupra *Revenue* și *Revenue/Visitors*.

Pe baza analizelor experimentale, **Regresia Liniară** a fost selectată ca algoritm optim pentru predicția veniturilor și ierarhizarea activităților turistice în contextul țărilor analizate.