

# Machine learning pentru recomandarea motocicletelor

Răzvan Iacob

May 2024

## Contents

<b>1</b>	<b>Capitolul 1</b>	<b>3</b>
1.1	INTRODUCERE . . . . .	3
1.2	MOTIVAȚIA ALEGERII BAZEI DE DATE RESPECTIVE . .	3
<b>2</b>	<b>Capitolul 2</b>	<b>3</b>
2.1	CONTEXTUL BAZEI DE DATE ȘI AL PROIECTULUI . . . .	3
2.2	CERINȚE . . . . .	4
2.2.1	Biblioteci utilizate: . . . . .	5
2.3	CE DORIM SĂ OBȚINEM . . . . .	5
2.3.1	Scopul proiectului: . . . . .	5
2.3.2	Obiectivele: . . . . .	5
2.3.3	Rezultatele așteptate: . . . . .	5
2.3.4	Impactul sau beneficiile: . . . . .	5
<b>3</b>	<b>Capitolul 3</b>	<b>6</b>
3.1	Aspecte teoretice relevante, inclusiv state-of-the-art . . . . .	6
3.1.1	Machine Learning: . . . . .	6
3.1.2	Regresia: . . . . .	6
3.1.3	Algoritmii de Regresie: . . . . .	7
3.1.4	Evaluarea Performanței Modelelor: . . . . .	7
3.1.5	Preprocesarea Datelor: . . . . .	8
3.1.6	Starea actuală a domeniului: . . . . .	8
<b>4</b>	<b>Capitolul 3</b>	<b>9</b>
4.1	Implementarea aspectelor teoretice în cadrul proiectului . . . . .	9
4.1.1	Preprocesarea datelor . . . . .	9
4.1.2	Analiza exploratorie a datelor . . . . .	10
<b>5</b>	<b>Testare și validare</b>	<b>13</b>
5.1	Crearea modelelor și alegerea celui potrivit . . . . .	13
<b>6</b>	<b>Rezultate</b>	<b>14</b>

6.1	Vizualizarea Rezultatelor . . . . .	14
<b>7</b>	<b>Utilizarea modelului într-un proiect practic</b>	<b>16</b>
7.1	Salvarea modelului . . . . .	16
7.2	Folosirea modelului . . . . .	16
<b>8</b>	<b>Concluzii</b>	<b>17</b>

# 1 Capitolul 1

## 1.1 INTRODUCERE

În ziua de astăzi, caracterizată de un val imens al datelor și un acces facil la informații, modul nostru de a aborda cercetarea și dezvoltarea tehnologică s-a transformat. Domeniul machine learning-ului și al inteligenței artificiale sunt primele în topul tehnologiilor, unde alegerea unei baze de date adecvate joacă un rol crucial în succesul unui proiect.

Acest capitol se concentrează pe selecție și motivația din spatele alegerii bazei de date pentru proiectul meu de generare a rating-urilor la motociclete. Vom explora factorii cheie care au influențat această decizie și voi prezenta argumentele care susțin alegerea datelor utilizate.

## 1.2 MOTIVAȚIA ALEGERII BAZEI DE DATE RESPECTIVE

Proiectul a fost inițiat cu scopul de a dezvolta un model de machine learning pentru generarea rating-ului motocicletelor în funcție de anumite specificații tehnice. Alegerea unei baze de date cuprinzătoare și relevante pentru industria motocicletelor a fost esențială pentru obținerea de rezultate precise și relevante. Astfel, am explorat și evaluat mai multe baze de date, iar în final am ales "all bikes curated" ca fiind cea mai potrivită pentru proiectul nostru.

Această decizie a fost luată în urma unei analize atente a caracteristicilor și a volumului de date disponibile în baza de date "all bikes curated". Aceasta s-a dovedit a fi mult mai bogată și mai variată în informații, oferind o gamă largă de caracteristici ale motocicletelor și o diversitate extinsă de modele, față de baza de date precedentă pe care am folosit-o numită "Bike details". Prin urmare, am considerat că aceasta este alegerea potrivită pentru proiectul meu, deoarece oferă o bază solidă pentru antrenarea și evaluarea modelului meu de machine learning.

În continuare, acest proiect își propune să exploateze potențialul acestei baze de date ample pentru a dezvolta un model robust și precis de generare a rating-ului motocicletelor, contribuind astfel la înțelegerea și optimizarea procesului de evaluare a performanței acestora.

# 2 Capitolul 2

## 2.1 CONTEXTUL BAZEI DE DATE ȘI AL PROIECTULUI

În urmă cu câteva luni am achiziționat o motocicletă care se potrivea cerințelor și dorințelor mele, dar brandul nu era cel dorit. Deși eram în căutarea altei motociclete am avut norocul de a primi recomandarea altui brand și model cu

care eram sceptic la început. După studierea acestora și văzând capabilitățile acestora am decis să o achiziționez și nu am regretat nicio secundă. Probabil am fost norocos să am pe cineva care știa acea motocicletă și a văzut că este potrivită pentru mine, astfel eu am venit în ajutorul oamenilor ce nu sunt la fel de norocoși, cu un program care folosește un model de machine learning care le recomandă și alte tipuri de motociclete în funcție de rating-ul calculat pentru acestea.

În baza de date folosită pentru proiect am avut mai multe coloane cu date specifice pentru motociclete având aproximativ 38.000 de înregistrări. După curățarea bazei de date și eliminarea coloanelor de care nu aveam nevoie am rămas cu acestea, considerându-le cele mai importante pentru ce am dorit să realizez.

## Lista categoriilor de motociclete

- Brand
- Model
- Year
- Rating
- Displacement (ccm)
- Power (hp)
- Torque (Nm)
- Bore (mm)
- Stroke (mm)
- Fuel capacity (lts)
- Dry weight (kg)
- Wheelbase (mm)
- Seat height (mm)
- Category\_ATV
- Category\_Allround
- Category\_Classic
- Category\_Cross / motocross
- Category\_Custom / cruiser
- Category\_Enduro / offroad
- Category\_”Minibike, cross”
- Category\_”Minibike, sport”

## 2.2 CERINȚE

În cadrul acestui proiect, obiectivul principal a fost prezicerea rating-ului unor motociclete. Pentru a realiza acest lucru am folosit mai multe biblioteci și s-au atins câteva cerințe:

1. Importul și prelucrarea datelor: Am utilizat biblioteca Pandas pentru a încărca și manipula seturile de date.
2. Antrenarea modelelor de regresie: Am implementat modele de regresie liniară, arbori de decizie și mașini cu vectori suport (SVM) folosind scikit-learn.
3. Evaluarea modelelor: Am evaluat performanța modelelor utilizând diverse metrici precum eroarea medie pătratică (MSE), eroarea medie absolută

(MAE) și coeficientul de determinare ( $R^2$ ).

4. Vizualizarea datelor: Am folosit bibliotecile Matplotlib și Seaborn pentru a crea vizualizări relevante pentru datele noastre, cum ar fi diagramele de dispersie și diagramele de corelație.

### 2.2.1 Biblioteci utilizate:

- **Pandas:** Folosit pentru manipularea și analiza datelor tabulare.
- **NumPy:** Bibliotecă fundamentală pentru lucrul cu matrice și vectori multidimensionali.
- **Seaborn:** Utilizat pentru vizualizarea datelor statistice, oferind diagrame estetice și informative.
- **Matplotlib.pyplot:** Folosit pentru crearea de grafice și diagrame pentru vizualizarea datelor.
- **scikit-learn:** O bibliotecă de învățare automată pentru Python care oferă un set de instrumente pentru construirea și evaluarea modelelor de învățare automată.
- **joblib:** Utilizat pentru salvarea și încărcarea modelelor antrenate.

## 2.3 CE DORIM SĂ OBTÎNEM

### 2.3.1 Scopul proiectului:

Scopul proiectului a fost de a învăța cum funcționează modelele de machine learning și de a veni în ajutorul oamenilor cu puține cunoștințe în domeniul moto” pentru a putea să își achiziționeze motocicletă potrivită dorințelor sale și eventual bugetului acordat pentru o astfel de achiziție.

### 2.3.2 Obiectivele:

Principalele obiective sunt realizarea unui model funcțional și aplicabil în viața de zi cu zi și realizarea unei mici aplicații folosite de oameni.

### 2.3.3 Rezultatele așteptate:

Rezultatele așteptate sunt dezvoltarea și validarea unui model de machine learning care poate prezice rating-urile motocicletelor cu o precizie semnificativă și crearea unei aplicații sau platforme online care permite utilizatorilor să introducă preferințele lor și să primească recomandări personalizate privind achiziționarea unei motociclete.

### 2.3.4 Impactul sau beneficiile:

- Facilitarea procesului de luare a deciziilor pentru potențialii cumpărători de motociclete, reducând incertitudinea și riscul de a face o alegere nepotrivită.

- Economisirea timpului și a efortului clienților în căutarea și compararea diferitelor opțiuni de motociclete disponibile pe piață

## 3 Capitolul 3

### 3.1 Aspecte teoretice relevante, inclusiv state-of-the-art

#### 3.1.1 Machine Learning:

O cercetare realizată de Jordan et al.[4], arată progresele recente în domeniul învățării automate care au fost alimentate de dezvoltarea unor noi algoritmi și teorii. Această evoluție a fost influențată de creșterea exponențială în disponibilitatea datelor online și de reducerea costurilor de calcul. Mai mult decât atât, aceste revoluționări la nivel tehnologic au stimulat inovații semnificative într-o serie de domenii de la medicină până la industrie și educație.

Prin utilizarea metodelor de machine learning, companiile și instituțiile au reușit să obțină o mai bună înțelegere a comportamentului clienților, să îmbunătățească procesele de producție și să optimizeze deciziile financiare. De asemenea, în știință, machine learning a deschis noi perspective în analiza datelor experimentale complexe, accelerând descoperirile în domenii precum biologia, astronomia și neuroștiințele.

Astfel, în lumina progresului rapid înregistrat în domeniul machine learning, este evident că această tehnologie continuă să redefinească modul în care interacționăm cu lumea digitală și să influențeze direcția în care se îndreaptă inovația tehnologică.

#### 3.1.2 Regresia:

Regresia este o unealtă statistică pentru investigarea relațiilor dintre variabile. De obicei, investigatorul încearcă să afle efectul cauzal al unei variabile asupra alteia - de exemplu, efectul creșterii prețului asupra cererii sau efectul schimbărilor în ofertă de bani asupra ratei inflației. Pentru a explora aceste probleme, investigatorul adună date despre variabilele corelate de interes și folosește regresia pentru a estima efectul cantitativ al variabilelor cauzale asupra variabilei pe care o influențează. Investigatorul evaluează și în mod obișnuit "semnificația statistică" a relațiilor estimate, adică gradul de încredere că relația reală este aproape de relația estimată. Tehnicile de regresie au fost mult timp centrale în domeniul statisticilor economice ("econometriei"). În mod tot mai frecvent, ele au devenit importante și pentru avocați și factorii de decizie în domeniul juridic. Regresia a fost oferită ca dovadă a responsabilității în conformitate cu Legea Drepturilor Civile din 1964, ca dovadă a prejudecății rasiale în litigiile privind pedeapsa cu moartea, ca dovadă a prejudiciilor în acțiunile contractuale, ca dovadă a încălcărilor în conformitate cu Legea privind Drepturile de Vot, și ca dovadă a prejudiciilor în litigiile antitrust, printre altele.

Pentru a citi mai multe, te poți uita la lucrarea "Chicago Working Paper in Law Economics" realizată de Sykes și Alan O[10]. Este o secțiune care explică conceptele de bază ale analizei de regresie, cum funcționează și ce presupuneri are. Sper că acest fragment îți este de ajutor pentru proiectul tău!

### 3.1.3 Algoritmii de Regresie:

În proiectul meu, am folosit Random Forest Regressor pentru a dezvolta modele predictive, alegând această metodă pentru a extrage informații relevante din date. Având în vedere numărul mare de tehnici de data mining disponibile, este impractic să le testăm pe toate pentru a găsi cea mai bună.

Această abordare este similară cu cea descrisă în articolul "Selecting Machine Learning Algorithms Using Regression Models" de Doan et al.[3], Jugul. Autorii demonstrează utilizarea regresiei liniare multivariate bazate pe arbori pentru a prezice performanța algoritmilor, folosind meta-cunoștințe din experiențele anterioare de învățare automată.

Am aplicat tehnici de reducere a dimensiunii spațiului de caracteristici, asigurând că datele transformate păstrează informațiile esențiale pentru predicții precise. Random Forest Regressor a fost ales și optimizat pe baza acestor meta-cunoștințe, dovedindu-și eficiența în predicțiile efectuate pe date numerice și nominale din medii reale.

### 3.1.4 Evaluarea Performanței Modelelor:

În proiectul meu, am evaluat performanța modelelor de regresie folosind o varietate de metrici, reflectând diverse aspecte ale preciziei și adecvării acestora. Aceste metrici sunt esențiale pentru a înțelege și cuantifica cât de bine se potrivesc modelele noastre datelor de testare și în ce măsură predicțiile lor corespund realității.

Am ales să utilizez metrici precum Mean Absolute Error (MAE), Mean Squared Error (MSE) și Root Mean Squared Error (RMSE), care sunt printre cele mai comune metrici folosite în evaluarea modelelor de regresie, conform secțiunii "PART I: THE MOST COMMON METRICS" din articol. Aceste metrici oferă o măsură a diferenței absolute sau pătratice medii între valorile prezise și cele reale, furnizând astfel informații importante despre cât de aproape sunt predicțiile noastre de adevăr.

În plus, am integrat și alte metrici, cum ar fi Median Absolute Error (MedAE) și Mean Absolute Percentage Error (MAPE), menționate în articolul studiat pentru evaluarea specifică a performanței modelelor în anumite domenii sau situații. Aceste metrici ne permit să evaluăm precizia mediană sau procentul mediu de eroare al predicțiilor noastre, oferind o perspectivă suplimentară asupra performanței modelelor noastre.

În lumina discuțiilor din secțiunile "Known Metrics Classifications, Their Benefits, and Drawbacks" și "Performance Metrics Typology", articol scris de Botchkarev

și Alexei [1], am recunoscut importanța evaluării multiplelor metrici pentru a obține o înțelegere cuprinzătoare a performanței modelelor noastre. Aceste discuții au subliniat diversitatea metricilor disponibile și modul în care acestea pot reflecta diferite aspecte ale erorii de predicție și adecvării modelului.

### 3.1.5 Preprocesarea Datelor:

În proiectul meu, preprocesarea datelor a fost esențială pentru a elimina variațiile nedorite și a asigura calitatea datelor utilizate în modelele alese. Am aplicat diverse tehnici pentru a gestiona și corecta artefactele prezente în datele multivariate, astfel încât să putem construi modele eficiente și precise.

Conform secțiunii "Objectives of data preprocessing" din articolul "New data preprocessing trends based on ensemble of multiple preprocessing techniques" scris de Mishra et al.[8], scopul global al preprocesării datelor este de a elimina variabilitatea sau efectele nedorite din semnal, astfel încât informațiile utile legate de proprietățile de interes să poată fi utilizate eficient pentru modelare. Specific, am eliminat coloanele și înregistrările cu valori lipsă pentru a reduce impactul acestora asupra analizei noastre. Cauzele valorilor lipsă pot varia de la erori instrumentale la erori umane, iar abordarea noastră a fost de a exclude aceste date pentru a asigura integritatea setului de date.

Astfel, preprocesarea datelor a fost un pas important în asigurarea calității datelor noastre, permițând modelelor noastre să funcționeze optim și să ofere predicții precise.

### 3.1.6 Starea actuală a domeniului:

În contextul actual, dezvoltarea tehnologiilor de învățare automată a devenit tot mai relevantă și în industria moto. Unul dintre aspectele importante abordate este dezvoltarea algoritmilor de detecție a accidentelor șoferilor de motociclete, utilizând analiza modelelor de învățare automată. Totodată, se desfășoară cercetări asupra severității leziunilor în accidente cu motociclete, iar performanța algoritmilor de învățare automată este comparată în acest context. Aceste eforturi vizează îmbunătățirea siguranței și prevenirea accidentelor în mediul moto, prin aplicarea avansată a tehnologiilor de inteligență artificială.

- În primul caz de utilizare al machine learningului în Dezvoltarea algoritmilor de detecție a accidentelor șoferilor de motociclete [2], spune că sunt puține cercetări legate de detectarea coliziunilor, dar urmează să fie explorate și dezvoltate. Gabriel Matuszczyk și Rasmus Aberg au realizat un model de machine learning care a fost implementat pe telefonul mobil fără a se folosi alte componente dedicate pentru a identifica coliziunile[7].
- În al doilea caz de utilizare al machine learningului exista destul de multe studii cum ar fi "Analysis of Motorcycle Accident Injury Severity and Performance Comparison of Machine Learning Algorithms" scris de Santos et al.[9] care a analizat pe o perioadă de 10 ani accidentele cu motoci-



clete din Portugalia. Acesta a analizat și comparat 14 modele de învățare automată și a ajuns la concluzia că modelele RF și LR au obținut cea mai bună performanță în prezicerea gravității leziunilor la motocicliști implicați în accidente. Aceste modele au evidențiat riscurile asociate cu factori precum consumul de alcool, tipul de drum și vârsta motocicletei.

## 4 Capitolul 3

### 4.1 Implementarea aspectelor teoretice în cadrul proiectului

Valori lipsă:			
Brand		0	
Model		28	
Year		0	
Category		0	
Rating		16684	
Displacement (ccm)		1011	
Power (hp)		12362	
Torque (Nm)		21838	
Engine cylinder		11	
Engine stroke		11	
Gearbox		5797	
Bore (mm)		9783	
Stroke (mm)		9783	
Fuel capacity (lts)		6768	
Fuel system		10628	
Fuel control		16464	
Cooling system		4214	
Transmission type		5611	
Dry weight (kg)		15989	
Wheelbase (mm)		12979	
Seat height (mm)		14290	
Front brakes		1583	
Rear brakes		1776	
...			
75%	NaN	NaN	NaN
max	NaN	NaN	NaN

Figure 1: Date lipsă după prelucrare

#### 4.1.1 Preprocesarea datelor

În primă fază a proiectului am utilizat *SimpleImputer* din biblioteca *sklearn* pentru a completa valorile lipsă din setul de date. Pentru variabilele numerice

am folosit media, iar pentru variabilele categorice am folosit cea mai frecventă valoare. După completarea valorilor am folosit *drop\_duplicates()* din biblioteca Pandas pentru a elimina duplicatelor.

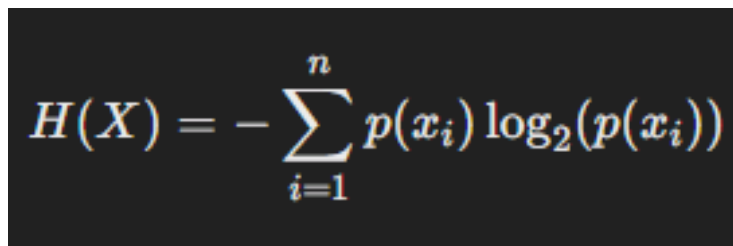
```
float64
Valori lipsă după imputare:
Brand          0
Model          0
Year           0
Category       0
Rating         0
Displacement (ccm)  0
Power (hp)      0
Torque (Nm)     0
Engine cylinder  0
Engine stroke   0
Gearbox         0
Bore (mm)       0
Stroke (mm)     1
Fuel capacity (lts)  0
Fuel system     0
Fuel control    0
Cooling system  0
Transmission type  0
Dry weight (kg)  0
Wheelbase (mm)  0
Seat height (mm)  0
Front brakes    0
Rear brakes     0
...
Front suspension  0
Rear suspension  0
Color options    0
dtype: int64
```

Figure 2: Date lipsă după de prelucrare

#### 4.1.2 Analiza exploratorie a datelor

Pentru a analiza datele am utilizat un set de indicatori cum ar fi Entropia, Indexul Gini și Cantitatea de informație pentru un set de etichete de clasă.

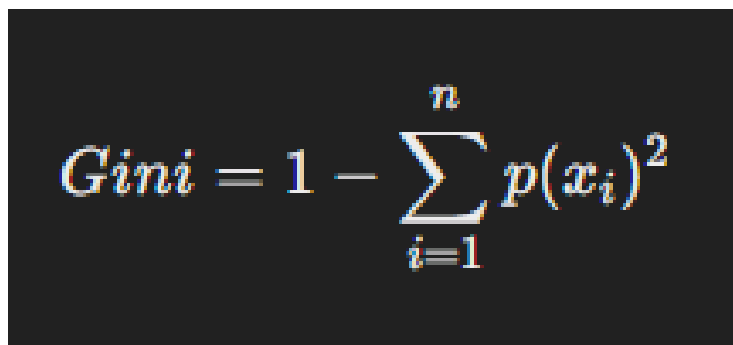
- Entropia măsoară incertitudinea unui set de date. Formula pe care am utilizat-o pentru a calcula entropia este afișat în figura 3 [ $p(x_i)$  este probabilitatea apariției clasei  $i$ ].



$$H(X) = - \sum_{i=1}^n p(x_i) \log_2(p(x_i))$$

Figure 3: Formula entropiei

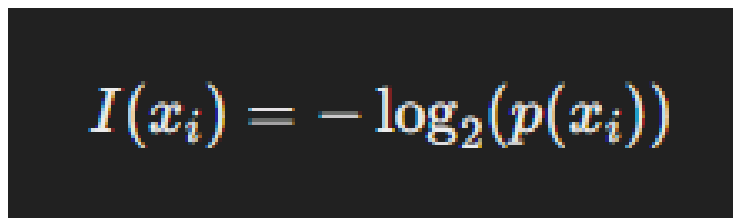
- Indexul Gini măsoară inconsistența setului de date și este utilizat frecvent în construirea arborilor de decizie. Formula folosită este afișată în figura 5.



$$Gini = 1 - \sum_{i=1}^n p(x_i)^2$$

Figure 4: Formula indexului gini

- Cantitatea de informație, cunoscută și sub numele de informație proprie, măsoară cantitatea de informație asociată cu apariția unei anumite clase. Formula pentru cantitatea de informație este reprezentată în figura de mai jos.



$$I(x_i) = - \log_2(p(x_i))$$

Figure 5: Formula cantității de informație

- Rezultatele obținute pentru toate acestea sunt:

- Entropy: 0.9709505944546686
- Gini Index: 0.48
- Information Quantity: 0: 1.3219280948873622, 1: 0.7369655941662062

După calcularea tuturor indicatorilor am realizat matricea de corelație pentru coloanele pe care le am folosit. Matricea de corelație afișată în imagine conține

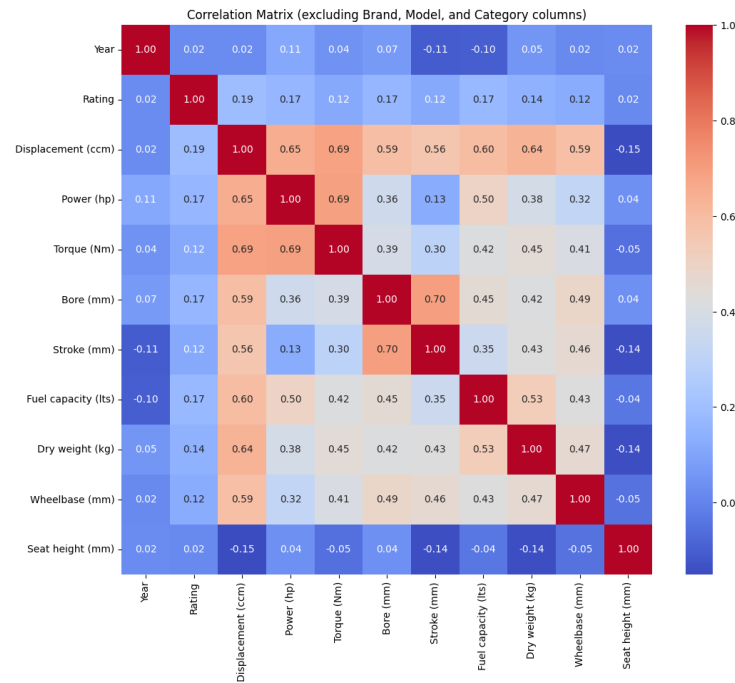


Figure 6: Matricea de corelație

corelațiile dintre diferite caracteristici ale motocicletelor, excluzând coloanele 'Brand', 'Model' și cele care încep cu 'Category'.

- Relații puternice:
  - Displacement (ccm) și Power (hp): Corelație de 0.65 indică o relație pozitivă puternică. Asta înseamnă că, în general, motocicletele cu o capacitate cilindrică mai mare tind să aibă o putere mai mare.
  - Displacement (ccm) și Torque (Nm): Corelație de 0.69, arătând o relație pozitivă puternică, similar cu cea dintre 'Displacement' și 'Power'.
  - Bore (mm) și Stroke (mm): Corelație de 0.70 indică faptul că dimensiunile cilindrilor sunt adesea proporționale în motocicletele analizate. (puteți citi mai multe despre aceste relații în următoarea lucrare)[5]

- Relații moderate:
  - Fuel capacity (lts) și Dry weight (kg): Corelație de 0.53, sugerând că motocicletele cu rezervoare mai mari tind să fie mai grele.
  - Dry weight (kg) și Wheelbase (mm): Corelație de 0.47 indică o relație pozitivă moderată, sugerând că motocicletele mai grele tind să aibă și un ampatament(=distanța dintre osii) mai mare. Pentru mai multe informații vă rog să studiați ce a scris Martins et al.[6].
- Relații slabe:
  - Year și celelalte variabile: Corelații foarte mici (aproape de 0), indicând că anul producției motocicletei nu are o relație liniară semnificativă cu celelalte caracteristici.

## 5 Testare și validare

### 5.1 Crearea modelelor și alegerea celui potrivit

- Împărțirea Datelor
  - Datele au fost împărțite în seturi de antrenament și testare:
    - \* Set de Antrenament: 30
    - \* Set de Testare: 70
- Modelele utilizate
  - Au fost folosite următoarele modele de regresie:
    - \* Linear Regression
    - \* Decision Tree Regressor
    - \* Support Vector Machine (SVR)
    - \* Random Forest Regressor (model adăugat ulterior pentru optimizare)
- Evaluarea Performanței
  - Performanța fiecărui model a fost evaluată folosind următorii indicatori de performanță:
    - \* Mean Absolute Error (MAE)
    - \* Mean Squared Error (MSE)
    - \* Root Mean Squared Error (RMSE)
    - \* R-squared (R2) Score
  - Rezultate:

- \* Linear Regression (MAE: 0.177, MSE: 0.067, RMSE: 0.259, R2: 0.071)
- \* Decision Tree Regressor (MAE: 0.177, MSE: 0.117, RMSE: 0.343, R2: -0.012)
- \* Support Vector Machine (MAE: 0.177, MSE: 0.066, RMSE: 0.258, R2: 0.071)
- \* Random Forest Regressor (MAE: 0.159, MSE: 0.059, RMSE: 0.244, R2: 0.175)
- Alegerea modelului final
  - \* Modelul final ales a fost Random Forest Regressor, deoarece a avut cele mai bune rezultate în evaluarea de performanță.
- Optimizarea Random Forest Regressor
  - \* Am utilizat GridSearchCV pentru a optimiza parametrii modelului Random Forest Regressor:
    - Număr de arbori (n\_estimators): 150
    - Adâncime maximă (max\_depth): 20
    - Număr minim de mostre pentru a diviza un nod (min\_samples\_split): 10
    - Număr minim de mostre pe frunză (min\_samples\_leaf): 4

## 6 Rezultate

### 6.1 Vizualizarea Rezultatelor

Graficele de mai jos ilustrează predicțiile modelelor în raport cu valorile reale. Linia roșie reprezintă linia ideală unde predicțiile ar trebui să se afle dacă modelul ar prezice perfect.

- Linear Regression și Support Vector Machine au arătat o concentrare a predicțiilor în jurul valorii medii, indicând o capacitate limitată de a capta variațiile datelor.
- Decision Tree Regressor a prezentat o dispersie mai mare a predicțiilor, indicând o supraînvățare pe setul de antrenament.
- Random Forest Regressor a demonstrat o performanță superioară, cu o mai bună aliniere a predicțiilor cu valorile reale.

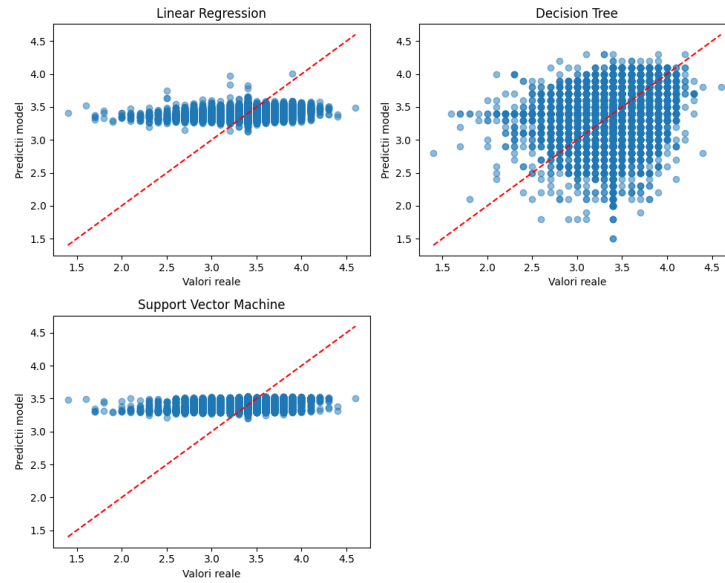


Figure 7: Linear Regression, Support Vector Machine & Decision Tree Regressor

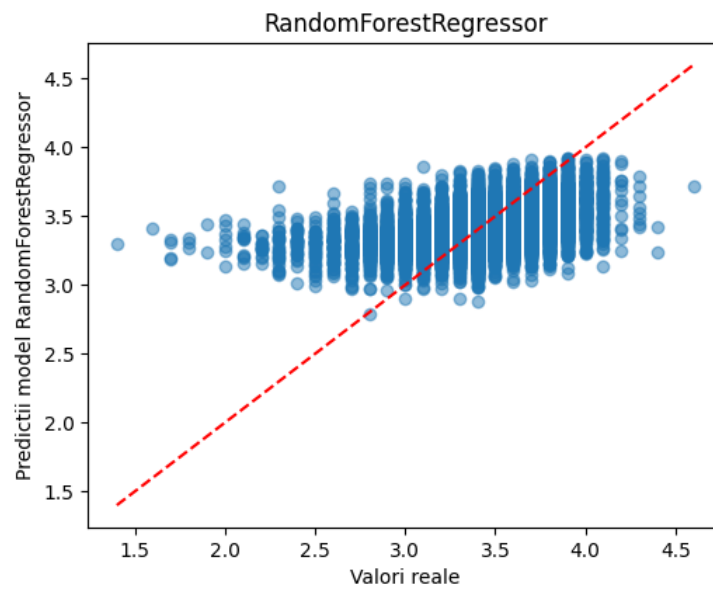


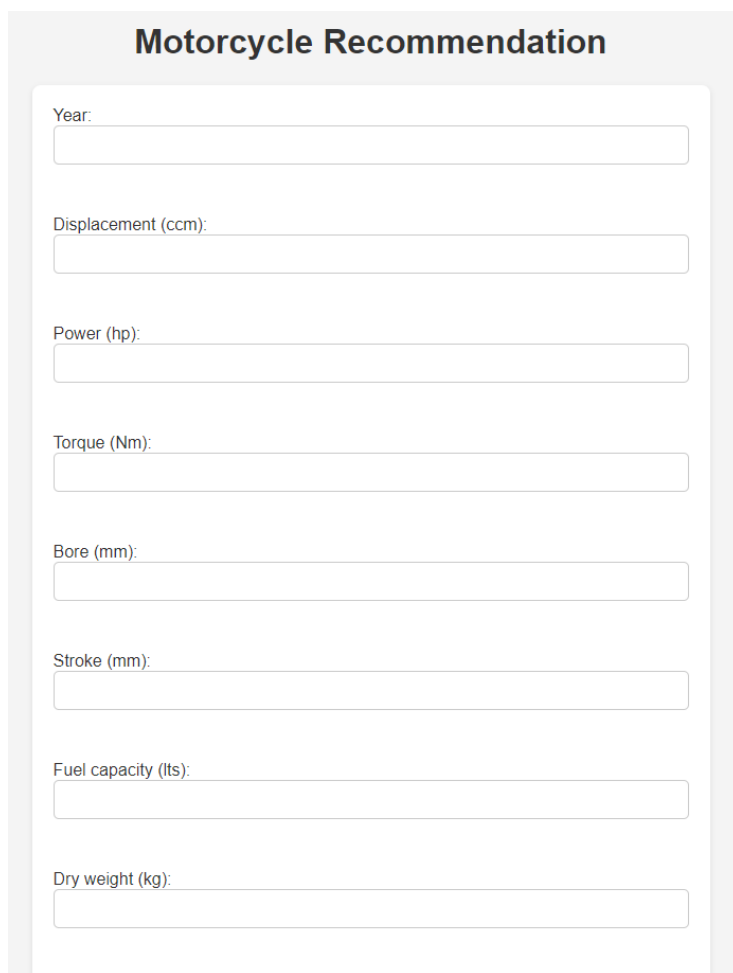
Figure 8: Random forest

## 7 Utilizarea modelului într-un proiect practic

### 7.1 Salvarea modelului

Am salvat modelul creat cu toate că nu are o acuratețe excepțională într-un fișier numit "trained\_model.pkl", pe care l-am folosit ulterior în predicția ratingului motocicletelor într-o pagină web.

### 7.2 Folosirea modelului



The image shows a web form titled "Motorcycle Recommendation". It contains eight input fields, each with a label above it: "Year:", "Displacement (ccm):", "Power (hp):", "Torque (Nm):", "Bore (mm):", "Stroke (mm):", "Fuel capacity (lts):", and "Dry weight (kg):". Each label is followed by a rectangular text input box.

Figure 9: Formular web

Am creat o pagină HTML cu un formular care conține câmpurile din tabela noastră exceptând coloana Rating". După completarea formularului și apăsarea butonului de la finalul formularului trimitea omul la o nouă pagină web unde



se recomandă și alte tipuri și modele de motociclete în funcție de ratingul din baza de date pe care am folosit-o.(vezi figurile de mai jos)

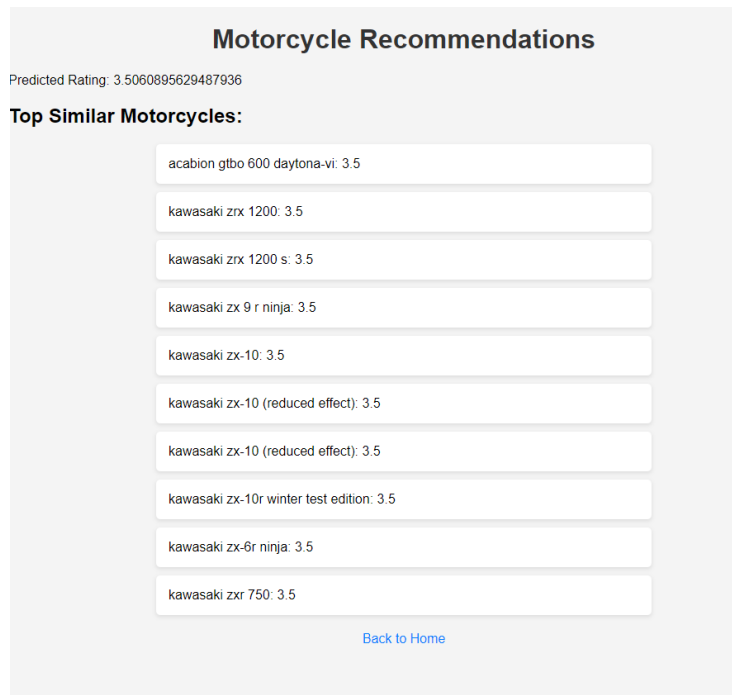


Figure 10: Predicții

## 8 Concluzii

În concluzie, am abordat problema predicției unui anumit set de date folosind mai multe tehnici, cu scopul de a identifica modelul care oferă cele mai precise predicții și cum se folosește un model după crearea acestuia. Am început prin prelucrarea datelor, includerea eliminării coloanelor inutile și împărțirea datelor în seturi de antrenare și testare. Apoi, am antrenat mai multe modele de regresie, cu accent pe RandomForestRegressor, și am evaluat performanța acestora folosind diverse metrice precum MAE, MSE, RMSE, R-squared, MedAE și MAPE. Prin optimizarea modelului ales, cu ajutorul căutării pe grilă, am identificat cei mai buni parametri. În cele din urmă, am interpretat rezultatele și am evidențiat importanța caracteristicilor în cadrul setului de date. În continuare, modelul a fost integrat într-un formular web, unde utilizatorii introduc caracteristicile unei motociclete pentru a primi o predicție a ratingului. După completarea formularului, utilizatorii sunt redirecționați către o pagină cu recomandări de motociclete în funcție de ratingul prezis. Această integrare oferă o soluție practică și interactivă pentru utilizatorii care caută motociclete potrivite.

## References

- [1] Alexei Botchkarev. A new typology design of performance metrics to measure errors in machine learning regression algorithms. *Interdisciplinary Journal of Information, Knowledge, and Management*, 14:045–076, 2019.
- [2] Lingyun Deng. Development of a crash detection algorithm for motorcycle drivers using machine learning. 2021.
- [3] Tri Doan and Jugal Kalita. Selecting machine learning algorithms using regression models. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1498–1505. IEEE, 2015.
- [4] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.
- [5] Adnan Katijan, Mohd Faruq Abdul Latif, Qamar Fairuz Zahmani, Shahid Zaman, Khairuldean Abdul Kadir, and Ibham Veza. An experimental study for emission of four stroke carbureted and fuel injection motorcycle engine. *Journal of Advanced Research in Fluid Mechanics and Thermal Sciences*, 62(2):256–264, 2019.
- [6] Guilherme Nobrega Martins, Mauro Speranza Neto, and Marco Antonio Meggiolaro. Dynamic models of bicycles and motorcycles using power flow approach. *Cep*, 22451:900, 2016.
- [7] Gabriel Matuszczyk and Rasmus Åberg. Smartphone based automatic incident detection algorithm and crash notification system for all-terrain vehicle drivers. 2016.
- [8] Puneet Mishra, Alessandra Biancolillo, Jean Michel Roger, Federico Marini, and Douglas N Rutledge. New data preprocessing trends based on ensemble of multiple preprocessing techniques. *TrAC Trends in Analytical Chemistry*, 132:116045, 2020.
- [9] Kenny Santos, Bernardo Firme, João P Dias, and Conceição Amado. Analysis of motorcycle accident injury severity and performance comparison of machine learning algorithms. *Transportation research record*, 2678(1):736–748, 2024.
- [10] Alan O Sykes. An introduction to regression analysis. 1993.