UNIVERSITATEA "AUREL VLAICU" DIN ARAD

FACULTATEA DE ȘTIINȚE EXACTE

DOMENIUL: INFORMATICĂ

PROGRAMUL DE STUDIU: INFORMATICĂ

FORMA DE ÎNVĂȚĂMÂNT CU FRECVENȚĂ

LUCRARE DE LICENȚĂ

ÎNDRUMĂTOR ȘTIINȚIFIC:

Lector univ. dr. – VLAD F. DRĂGOI

ABSOLVENT:

DREGHICI GH. NICOLETA

ARAD

Iunie 2021

UNIVERSITATEA "AUREL VLAICU" DIN ARAD
FACULTATEA DE ŞTIINŢE EXACTE
DOMENIUL: INFORMATICĂ
PROGRAMUL DE STUDIU: INFORMATICĂ
FORMA DE ÎNVĂŢĂMÂNT CU FRECVENŢĂ

MODELE DE REGRESIE ŞI RANDOM FOREST PENTRU DATA SCIENCE

ÎNDRUMĂTOR ŞTIINŢIFIC:
Lector univ. dr. – VLAD F. DRĂGOI

ABSOLVENT:
DREGHICI GH. NICOLETA

ARAD
Iunie 2021

Cuprins

Lista figurilor

Lista cu notaţii/abrevieri

În lucrarea de faţă au fost utilizate următoarele notaţii/abrevieri:

y = varibilă dependentă/target/aleatoare, valoarea iniţială a variabilei de explicat (y_i reprezintă un vector)

$\hat{y}$ = valoarea estimată a variabilei dependente y prin aplicarea modelelor ($\hat{y}_i$ reprezintă un vector)

$\bar{y}$ = media variabilei y

x = variabilă independentă/explicativă, predictor, atribut, variabilă care nu este aleatoare şi pe baza căreia/cărora se incearcă explicarea variabilei dependente y (x_i reprezintă un vector)

$\overline{x}$ = media variabilei x

a, b = parametrii/coeficienţii de estimat (a = constanta, b = panta/coeficientul de regresie)

ε = eroarea aleatoare neobservată sau reziduu, care conţine toată informaţia care nu este explicată de relaţia liniară între x şi y

$\overline{\varepsilon}$ = media erorilor aleatoare / reziduuri

i = numărul variabilelor cuprinse într-un vector

n = numărul de observaţii

k = numărul de variabile explicative

σ = abaterea standard

$r_{x,y}$= coeficientul de corelaţie între x şi y

SST = variaţia totală a datelor y de intrare / suma pătratelor abaterilor individuale de la medie

SSR = variaţia totală a datelor y estimate / suma pătratelor abaterilor de regresie

SSE – variaţia între datele de intrare şi cele estimate / suma pătratelor valorilor reziduale (erorilor)

R^2 = coeficientul de determinare

R^2_{adj} = coeficientul de determinare ajustat

MSE = eroarea medie patratică (mean squared error)

$RMSE$ = eroarea rădăcinii medie pătratică (root mean squared error)

MAE = media erorilor absolute (mean absolute error)

t = testul t Student

F = testul F Fisher

Introducere

Într-o lume aflată în continuă mişcare şi evoluţie permanentă, informaţia este resursa de bază care determină dezvoltarea societăţii umane. Forma brută a informaţiei se regăseşte în date. Datele reprezintă forma fizică şi suporturile de reprezentare a informaţiilor în cifre, litere, cuvinte, diverse simboluri şi alte însemne. Datele sunt ansambluri de simboluri şi nu au o semnificaţie în sine. Este necesară culegerea datelor, prelucrarea şi interpretarea lor, pentru a fi transformate în informaţii cu sens, care să poată fi utilizate în diverse scopuri şi să aducă utilitate în luarea deciziilor care ne influenţează viaţa la nivel individual şi social.

În societatea actuală, având în vedere dezvoltarea fără precedent a tehnologiei informaţiilor şi comunicaţiilor, se generează şi se vehiculează o cantitate uriaşă de date, care creşte exponenţial.

Pentru a identifica elemente importante în vastul univers al datelor şi pentru a da un sens şi un scop fluxului complex de date, s-au dezvoltat instrumente, metode şi tehnici de interpretare a datelor şi de transformare a lor în informaţii utile.

Data science a apărut ca urmare a acestor necesităţi, ca urmare a nevoilor de tratare a unor catităţi mari de date complexe şi de extragere a informaţiilor utile pe care acestea le conţin.

Data science este un domeniu interdisciplinar, care utilizează metode ştiinţifice pentru a extrage cunoştinţe şi perspective din date structurate şi nestructurate, şi aplică cunoştinţele extrase într-o gamă largă de domenii. Data science este un concept revoluţionar, capabil să abordeze cantităţile uriaşe de date care sunt generate în zilele noastre, un concept care utilizează statistica şi matematica, analiza datelor şi informatica, ca instrumente şi metode de măsurare şi interpretare obiectivă a realităţii. Data science este un concept care utilizează instrumente şi tehnici moderne pentru a identifica anumite tipare şi relaţii existente între diverse date sau seturi de date, pentru a extrage informaţii semnificative şi a descoperi noi structuri şi relaţii, în scopul construirii unor decizii fundamentate ştiinţific. Data science se bazează pe inteligenţa artificială şi pe subdomeniile ei machine learning şi deep learning, pentru a crea algoritmi capabili să extragă şi să modeleze seturi complexe de date, să facă clasificări şi grupări în funcţie de diverse criterii, să facă predicţii şi să identifice tendinţe.

Data science are întrebuinţări multiple, în diverse domenii de activitate, la nivel organizaţional şi instituţional. Rezultatele obţinute în ultimii ani prin utilizarea data science au arătat că a devenit un domeniu indispensabil pentru cunoaşterea realităţii economice şi sociale şi pentru elaborarea de decizii şi strategii de acţiune. Toţi actorii economici relevanţi, din

principalele ramuri economice, utilizează într-o formă sau alta data science pentru a se informa şi a lua decizii. Data science are beneficii majore la nivel instituţional, în business, în cercetare şi inovare, iar companiile şi statele care nu vor folosi avantajele aduse de data science riscă să rămană în urmă şi să devină incapabile să se adapteze în timp util noilor provocări ale mediului competiţional.

În prezenta lucrare este tratată analiza de regresie ca instrument de studiu în data science.

Introducerea în analiza de regresie s-a făcut printr-un studiu de caz de econometrie, în care a fost utilizată regresia liniară simplă pentru a arăta puterea unui model explicativ în identificarea unor estimatori care pot explica anumite fenomene şi relaţii economice. În acest scop a fost preluat un set de date din cartea "Econometrie – La regression lineaire simple et multiple" publicată de Ricco Rakotomalala. Pe baza unui set de date simplu, care conţine 28 de observaţii, o variabilă explicativă şi o variabilă de explicat, s-a realizat estimarea consumului de combustibil al autovehiculelor în funcţie de greutatea lor, prin modelarea cu ajutorul regresiei liniare simple prin metoda celor mai mici pătrate. Predicţiile elaborate sunt susţinute de premise ştiinţifice şi au la bază confirmarea unor ipoteze teoretice şi calcularea unor indicatori statistici care arată performanţele modelului şi susţin validitatea lui.

Pentru consolidarea analizei de regresie ca şi model explicativ s-a procedat în continuare la efectuarea unui studiu de caz mai complex, care respectă etapele şi principiile unui studiu de data science, pe baza unui set de date mai complicat. Setul de date a fost preluat de pe internet de pe site-ul Kaggle (https://www.kaggle.com/hellbuoy/car-price-prediction), care pune la dispoziţie diverse seturi de date bazate pe măsurători practice, pentru studii de data science.

Obiectivul studiului este predicţia preţului autovehiculelor pe baza unor atribute/variabile explicative preluate din viaţa reală prin măsurare. O companie din China - Geely Auto - doreşte să intre pe piaţa auto din SUA şi să înfiinţeze o unitate de producţie. A fost contactată o firmă de consultanţă care a efectuat studii de piaţă şi a obţinut un set de date despre diverse tipuri de autoturisme de pe piaţa americană. Scopul studiului este de a afla care sunt factorii principali care influenţează preţurile auto de pe piaţa americana şi cum variază preţurile în funcţie de variabilele explicative. Compania poate adapta designul produselor şi poate să stabilească o strategie de business, în funcţie de rezultatele studiului, care vor arăta dinamica preţurilor. Setul de date, conţine 205 observaţii şi are 26 de atribute (proprietăţi măsurate), care pot explica mai mult sau mai puţin nivelul preţului de vânzare.

În etapa de analiză a datelor au fost identificate principalele atribute care pot explica preţul autovehiculelor, prin implementarea unor metode de analiză. Pentru selecţia atributelor explicative relevante s-a utilizat filtrul corelaţiei slabe între predictori şi variabila target, algoritmul random forest pentru determinarea importanţei atributelor, regresia lasso (regularizarea L1) pentru selectarea atributelor. Ca urmare a aplicării acestor metode, din cele 26 de atribute au fost selectate patru, respectiv capacitatea motorului (enginesize), greutatea maşinii (curbweight), randamentul motorului/caii putere (horsepower) şi indicatorul consumului de combustibil pe autostrada (highwaympg), toate cele patru fiind foarte corelate cu preţul.

Pentru stabilirea strategiei şi a tehnicilor de modelare s-a utilizat graficul perechilor de variabile, care arată la nivel preliminar tipul de relaţie existentă între variabilele implicate şi matricea de corelaţie, care arată un nivel de corelaţie ridicat între toate cele patru variabile explicative (> 0.75). Această situaţie a corelaţiei ridicate între variabilele explicative determină multicoliniaritatea variabilelor, care duce la instabilitatea modelului, motiv pentru care se recomandă evitarea ei. Prin urmare dacă respectăm principiul multicoliniarităţii trebuie să eliminăm toate variabilele înalt corelate între ele şi să păstrăm doar o singură variabilă, cea mai bine corelată cu preţul, respectiv variabila enginesize.

Totuşi în practica economică pentru elaborarea unei strategii de business (care se bazează pe întocmirea unui buget şi a unui plan de producţie) este nevoie de cât mai multă informaţie şi trebuiesc utilizate toate variabilele semnificative. S-a pornit de la ipotezele că modelul ideal conform normelor teoretice, cu o variabilă explicativă, nu este folositor în realizarea obiectivului studiului, iar o variabilă explicativă nu poate aduce cantitatea de informaţie pe care o pot aduce mai multe variabile explicative.

Prin urmare, având în vedere complexitatea contextului, s-a stabilit utilizarea mai multor modele şi compararea performanţelor lor, pentru a se putea trage o concluzie pertinentă cu privire la cel mai eficient model explicativ. Totodată s-au analizat în profunzime influenţele multicoliniarităţii asupra modelelor, pentru a vedea dacă modelele influenţate de multicoliniaritate sunt valide.

S-a utilizat regresia liniară simplă, regresia lasso, regresia ridge şi regresia random forest. Au fost aplicate modelele simple, cu o variabilă explicativă şi modelele multiple, cu patru variabile explicative. Totodată a fost dezvoltat şi un model alternativ bazat pe patru modele de regresie random forest simplă, aplicate separat pentru fiecare variabilă explicativă. Rezultatul

final este determinat de media celor patru modele, ponderată cu coeficientul de determinare aferent fiecărui model.

Pentru determinarea celor mai performante modele au fost comparate rezultatele obţinute de acestea. Indiferent de modelul implementat, scopul este ca distanţa între valorile iniţiale ale variabilei target (preţul în cazul nostru) şi valorile estimate prin aplicarea modelului să fie cât mai mică. Întrucât valorile iniţiale şi valorile estimate se prezintă sub forma unor vectori, s-au comparat p-normele vectorilor rezultaţi din diferenţa între valorile iniţiale şi cele generate prin aplicarea modelelor.

Analiza şi modelarea datelor pentru cele două studii s-a făcut în Python, cu utilizarea Jupyter Notebook şi a pachetelor Numpy, Pandas, Matplotlib şi Sklearn.

1. Econometrie – studiu de caz

Econometria s-a dezvoltat ca o necesitate în investigarea fenomenelor şi proceselor economice, atât la nivel macroeconomic, cât şi la nivel microeconomic. „*Econometria înseamnă aplicarea metodelor statistice datelor economice, pentru a da conţinut empiric relaţiilor economice*”. (M Hashem Pesaran, 1987, „Econometrics”)

Econometria utilizează teoria statistică, statisticile matematice şi teoria probabilităţilor, aplicând modele statistice şi matematice asupra datelor din lumea reală, ca unelte pentru identificarea unor estimatori care să exprime în mod imparţial şi eficient diverse fenomene economice şi relaţiile dintre ele [1].

La nivel macroeconomic, econometria furnizează economiştilor instrumente prin care să analizeze istoria unor fenomene economice şi evoluţia lor în timp, să măsoare şi să determine relaţii de dependenţă existente între diverşi indicatori sau variabile economice, să facă previziuni aplicate în mediul economic. La nivel microeconomic, fiecare entitate economică poate utiliza econometria pentru analizarea mediului în care îşi desfasoară activitatea, pentru analizarea propriei activităţi economice, pentru determinarea unor strategii de management şi marketing care să asigure funcţionarea în condiţii de profitabilitate.

Una dintre metodele statistice fundamentale utilizate în econometrie este analiza de regresie, utilizată pentru estimarea relaţiilor între o variabilă dependentă şi una sau mai multe variabile independente. Cea mai simplă formă a analizei de regresie este regresia liniara simplă, care presupune determinarea unei linii drepte care reprezintă cel mai bine relaţia între două variabile cantitative discrete, în funcţie de un criteriu matematic specific [2].

Un model de regresie liniară simplă, bazat pe metoda celor mai mici pătrate, este reprezentat de urmatoarea ecuaţie:

$$y = a + bx + \varepsilon$$

Pentru ilustrarea modelului regresiei liniare simple, vom utiliza un exemplu, un **studiu de caz**: ***estimarea consumului de combustibil al autovehiculelor în funcţie de greutatea lor***.

Este un model simplu, pe baza unui eşantion cu 28 de observaţii, (datele fiind preluate din cartea “Econometrie La regression lineaire simple et multiple” publicată de Ricco Rakotomalala) conform Figura 1.1, în care avem:

- coloana ***greutate*** – ***variabila independentă x*** - care reprezintă greutatea în kg a vehiculelor

- coloana ***consum*** – ***variabila dependentă*** $\boldsymbol{y}$ - care reprezintă consumul de combustibil, în l/100km

n	greutate (x)	consum (y)
1	650	5.70
2	790	5.80
3	730	6.10
4	955	6.50
5	895	6.80
6	740	6.80
7	1,010	7.10
8	1,080	7.40
9	1,100	9.00
10	1,500	11.70
11	1,075	9.50
12	1,155	9.50
13	1,140	8.80
14	1,080	9.30
15	1,110	8.60
16	1,140	7.70
17	1,370	10.80
18	940	6.60
19	1,400	11.70
20	1,550	11.90
21	1,330	10.80
22	1,300	7.60
23	1,670	11.30
24	1,560	10.80
25	1,240	9.20
26	1,635	11.60
27	1,800	12.80
28	1,570	12.70

Figura 1.1: Tabel greutate-consum

Pe baza relaţiei existente între cele două variabile/atribute, care au fost măsurate în prealabil, astfel că datele sunt cunoscute, putem estima variabila y.

Graficul scatter plot din Figura 1.2 (norul de puncte), arată existenţa unei relaţii liniare între cele două variabile. Relaţia liniară este pozitivă, întrucât creşterea valorilor variabilei x (axa ox), se face concomitent cu creşterea valorilor variabilei y (axa oy).

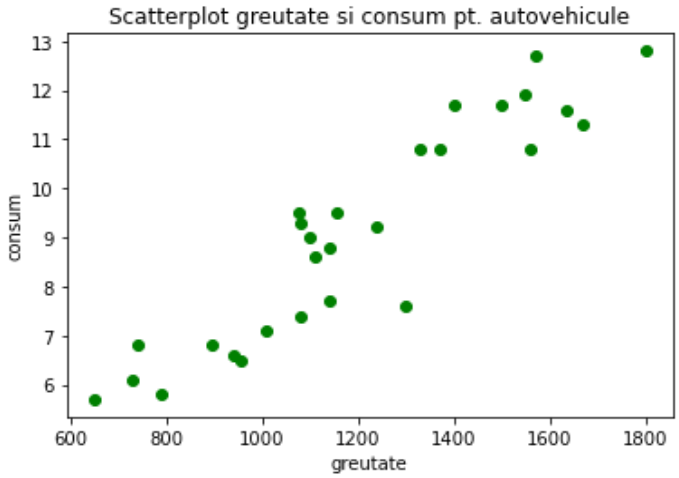

Figura 1.2: Scatterplot greutate si consum pt. autovehicule

Coeficientul de corelaţie:

Pentru determinarea corelaţiei între cele două variabile s-a calculat coeficientul de corelaţie Pearson, a cărui valoare poate fi cuprinsă între -1 şi 1.

Coeficientul de corelaţie este de 0.9263264, deci este pozitiv, ceea ce înseamnă că avem o corelaţie directă, respectiv cele două variabile variază în acelaşi sens. Deoarece corelaţia se încadrează în segmentul [0.8-1] înseamnă că avem o corelaţie foarte înaltă, ceea ce arată că poate exista o legatură statistică între variabile.

Parametrii ecuaţiei de regresie:

Utilizând metoda celor mai mici pătrate, calculăm estimatorii parametrilor din ecuaţia de regresie. Astfel, obţinem următoarele valori:

- constanta/intercept este locul de pe ordonată unde dreapta de regresie se intersectează cu axa Oy şi reprezintă o constantă care este valoarea medie a lui y atunci când $x = 0$: $\boldsymbol{a} = \mathbf{1.06269123}$
- panta de regresie/slope, care arată cu cât creşte variabila y, atunci când x creşte cu o unitate: $\boldsymbol{b} = \mathbf{0.00669386}$,

Ecuaţia funcţiei liniare de regresie este următoarea:

$$\hat{y}_i = [1.06269123] + [0.00669386]x_i$$

Vizualizarea grafică a funcţiei de regresie s-a făcut prin graficul scatter plot din Figura 1.3:

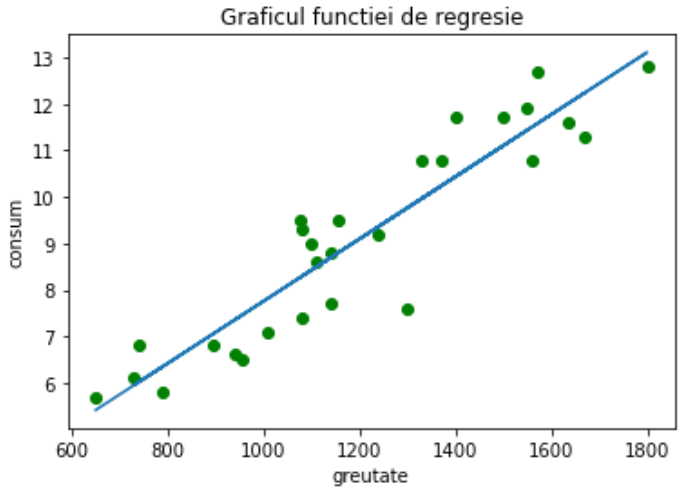

Figura 1.3: Graficul funcţiei de regresie

Se poate observa că dreapta de regresie trece aproximativ prin mijlocul norului de puncte. Evaluarea vizuală nu este însă suficientă pentru validarea modelului, astfel că avem nevoie de anumite criterii cantitative pentru evaluarea modelului.

Analiza varianţei şi coeficientul de determinare

Obiectivul regresiei liniare este acela de a minimiza suma pătratelor erorilor, respectiv suma pătratelor diferenţelor între valorile variabilei y şi valorile estimate prin ecuaţia de regresie:

$$\varepsilon = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Variaţia erorilor/reziduurilor asociată dreptei de regresie arată abaterile punctelor digramei de dispersie de la dreapta de regresie, respectiv măsura dispersiei norului de puncte faţă de dreapta de regresie.

Variaţia totală SST este exprimată ca suma între variaţia explicată şi variaţia neexplicată/variaţia erorilor:

$$SST = SSR + SSE$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Pornind de la ecuaţia varianţei, se poate determina coeficientul de determinare R^2, care descrie proporţia varianţei variabilei y explicată de modelul regresiei:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Cu cât valoarea lui R^2 este mai aproape de 1, cu atât modelul este mai reprezentativ, iar variabila x permite determinarea valorilor estimate ale variabilei y, întrucat valoarea SSE este mai aproape de 0, deci obiectivul minimizării sumei pătratelor erorilor este atins. Dacă valoarea lui R^2 este mai aproape de 0, atunci variaţia erorilor nu este minimizată, iar valoarea SSE se apropie de 1, ceea ce arată faptul că variabila y nu poate fi explicată prin variabila x.

În cazul modelului prezentat, **valoarea coeficientului de determinare este de 0.85808059**, ceea ce arată că modelul este reprezentativ pentru explicarea variabilei dependente.

Ipotezele asupra proprietăţilor estimatorilor

Pentru verificarea ipotezelor, au fost efectuate mai multe teste şi au fost calculaţi mai mulţi indicatori, prezentaţi în Figura 1.4:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.858
Model:	OLS	Adj. R-squared:	0.853
Method:	Least Squares	F-statistic:	157.2
Date:	Sun, 18 Apr 2021	Prob (F-statistic):	1.58e-12
Time:	16:35:46	Log-Likelihood:	-34.378
No. Observations:	28	AIC:	72.76
Df Residuals:	26	BIC:	75.42
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.0627	0.659	1.612	0.119	-0.292	2.418
x	0.0067	0.001	12.538	0.000	0.006	0.008

Omnibus:	1.054	Durbin-Watson:	1.799
Prob(Omnibus):	0.590	Jarque-Bera (JB):	0.883
Skew:	-0.409	Prob(JB):	0.643
Kurtosis:	2.705	Cond. No.	5.03e+03

Figura 1.4: Rezultate regresie liniara simpla OLS

Pentru validarea modelului au fost verificate următoarele ipoteze de lucru:

Liniaritatea modelului

Verificarea liniarităţii s-a facut grafic, prin intermediul graficului scatter plot prezentat în Figura 1.3. Întrucât norul de puncte care arată relaţia dintre variabila predictor şi cea estimată este dispus în model liniar, se poate spune că există o relaţie liniară între cele două variabile. Coeficientul de corelaţie Pearson care are valoarea 0.9263264, confirmă ipoteza existenţei unei relaţii liniare între cele două variabile.

Media erorilor este aproape de 0: $\overline{\varepsilon} = 0$

Media erorilor este 1.9032394, fiind foarte apropiată de 0, ceea ce confirmă ipoteza.

Normalitatea erorilor – variabila ε urmează o lege normală de medie zero şi variaţie σ^2: $\varepsilon_i \sim N(0, \sigma^2)$;

Normalitatea erorilor a fost verificată prin testul Jarque-Bera. A fost calculat coeficientul de asimetrie (Skewness) şi coeficientul de boltire (Kurtosis) a distribuţiei erorilor reziduale. O distribuţie perfect normală are un coeficient de asimetrie $S = 0$ si un coeficient de boltire $K = 3$.

- $S > 0$ arata o repartiţie asimetrică la dreapta, $S < 0$ arată o repartiţie asimetrică la stânga
- $K > 3$ arată o repartiţie afectată de boltire, iar $K < 3$ arată o repartiţie aplatizată

După cum se observă în Figura 1.4:

- coeficientul de asimetrie (Skew) = -0.409, deci este foarte apropiat de 0
- coeficientul de boltire (Kurtosis) = 2.705, deci este foarte apropiat de 3
- valoarea calculată a testului Jarque-Bera = 0.883, depaşeşte foarte puţin valoarea teoretică de 0.643

Conform teoremei limitei centrale, cu cât eşantionul de date creşte, cu atât distribuţia tinde sa fie normală. Dacă mărimea eşantionului depăşeşte 30 de observaţii, distribuţia mediei aritmetice a eşantioanelor va fi o distribuţie normală. Având în vedere faptul că setul de date conţine doar 28 de observaţii, putem considera că erorile reziduale sunt distribuite normal.

Homoscendasticitatea (omogenitatea varianţei erorilor) – varianţele erorilor sunt constante oricare sunt valorile variabilei predictor x: $v(\varepsilon_i) = \sigma^2$

Verificarea ipotezei homoscendasticităţii s-a făcut prin reprezentarea grafică a relaţiei dintre variabila x (axa ox) şi reziduuri (aza oy), conform Figura 1.5.

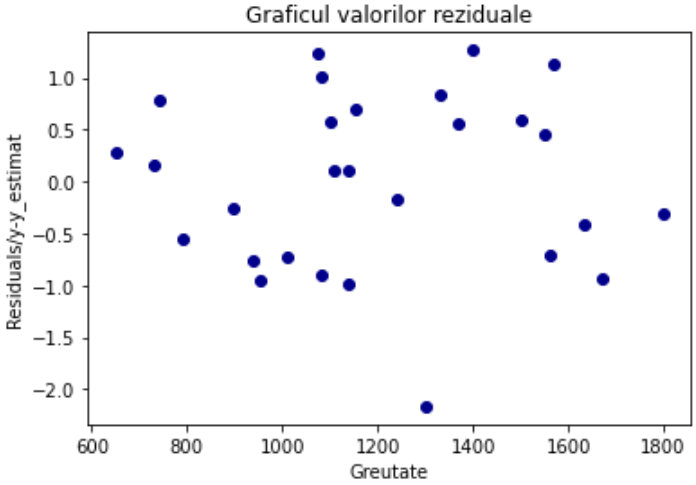

Figura 1.5: Graficul valorilor reziduale

Se poate observa că norul de puncte se află dispus normal, fără a forma un anumit model sau tipar, astfel că ipoteza homoscendasticităţii se confirmă.

Necorelarea erorilor – erorile sunt necorelate între ele: $cov(\varepsilon_i, \varepsilon_j) = 0$; erorile asociate unor valori ale variabilei y nu sunt influenţate de erorile asociate altor valori ale variabilei y.

În analiza de regresie autocorelarea erorilor intervine dacă se încearcă modelarea unei relaţii liniare asupra unor date neliniare, caz în care reziduurile vor fi autocorelate.

Ipoteza necorelării erorilor a fost verificată prin testul Durbin Watson. Valoarea testului Durbin Watson este de 1.799 şi se compară cu limita inferioară - d_L şi cu limita superioară - d_U citite din tabela Durbin şi Watson. Întrucât valoarea calculată obţinută se află în intervalul 1.325 – 1.964 obţinut din tabela cu valorile teoretice, putem considera că nu există autocorelare între valorile reziduale.

Testarea semnificativităţii pantei de regresie

Demersul porneşte de la formularea ipotezei nule H_0, conform căreia variabila y nu este influenţată de variaţia variabilei predictor x şi deci coeficientul b din ecuaţia de regresie nu este semnificativ diferit de zero.

$H_0: b = 0$

$H_1: b \neq 0$

Pentru verificarea ipotezei nule s-a folosit testul t Student pentru coeficientul de corelaţie simplă, valoarea t calculată fiind de 12.538. Valoarea calculată se compară cu valoarea teoretică din tabela t Student pentru n-2 grade de libertate şi un nivel de semnificaţie de 5%.

Întrucât $t_{calculat} = 12.538 > t_{teoretic} = 1.58$ se respinge ipoteza nulă. Coeficientul de corelaţie este semnificativ diferit de zero, iar între variabila independentă şi variabila dependentă există o legatură semnificativă.

Testarea semnificativităţii globale a modelului de regresie

Evaluarea globală a modelului s-a făcut pe baza statisticii test F. Valoarea calculată a testului este de 157.2.

Întrucât $F_{calculat} = 157.2 > F_{teoretic} = 1.58$, pentru n-2 grade de libertate şi un nivel de semnificaţie de 5%, se consideră că modelul este semnificativ.

În concluzie, întrucât modelul implementat a trecut testele de evaluare, se poate considera reprezentativ şi poate fi utilizat pentru estimarea variabilei dependente - consumul de combustibil, în funcţie de variabila independentă - greutatea autovehiculului.

Predicţia punctuală

Pe baza modelului construit se pot face previziuni punctuale ale comportamentului variabilei dependente y, în funcţie de valorile fixe ale variabilei independente x.

Pentru exemplificare s-a construit o metodă de predicţie pe baza ecuatiei de regresie. Dorim să determinăm consumul unui autovehicul cu o greutate de x = 1140 kg. Calculăm predicţia punctuală astfel:

$$y = \hat{a} + \hat{b}x = 1.062691 + 0.006694 * 1140 = 8.69368917$$

Se estimează astfel că un vehicul cu o greutate de 1140 kg va consuma 8.69 litri la 100 km.

Predicţia pe interval de predicţie

Pentru construirea unui interval de predicţie avem nevoie de un interval de încredere în cadrul căruia se găseşte cu o probabilitate semnificativă parametrul estimat, astfel că estimatorul să se gasească între valoarea inferioară şi valoarea superioară a intervalului.

Pentru exemplificare s-a construit o metodă de predicţie, pe baza intervalului de predicţie, cu un nivel de încredere de (1-α) după urmatoarea formulă:

$$\hat{y}_i \pm t_{1-\frac{\alpha}{2}} * \sigma_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Pentru calcularea intervalului de încredere luăm un risc α de 5% şi n-2 grade de libertate, şi citim din tabelul t Student valoarea 1.705618.

Pentru x = 1140 kg se calculează intervalul de predicţie după cum urmează:

[8.69368917 – 1.705618 * 0.872893 ; 8.69368917 + 1.705618 * 0.872893]

[7.204868 ; 10.182510]

Se estimează astfel că un vehicul în greutate de 1140 kg are 90% şanse să consume între 7.20 şi 10.18 litri la 100 km.

Pentru vizualizare s-a construit graficul din Figura 1.6, care arată intervalul de încredere şi intervalul de predicţie.

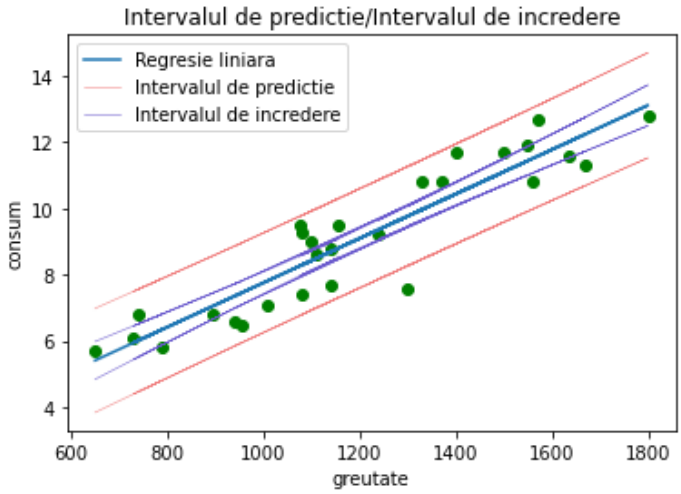

Figura 1.6: Intervalul de predicţie/Intervalul de încredere

Modelul de regresie liniară simplă pe baza metodei celor mai mici pătrate este un model puternic, care poate fi utilizat pe multiple seturi de date din lumea reala. Modelul este consolidat pe fundamente teoretice solide, iar predicţiile elaborate pe baza lui sunt susţinute de premise testate şi verificate.

2. Etapele şi conţinutul unui proiect Data Science

Un proiect Data Science este un demers complex, care trebuie să parcurgă o serie de etape şi să îndeplinească anumite condiţii pentru a fi un proiect viabil şi util. Etapele nu urmează o ordine strictă, de multe ori fiind necesară revenirea la etape anterioare, în funcţie de rezultatele obţinute pe parcursul demersului [3], care este prezentat în Figura 2.1.

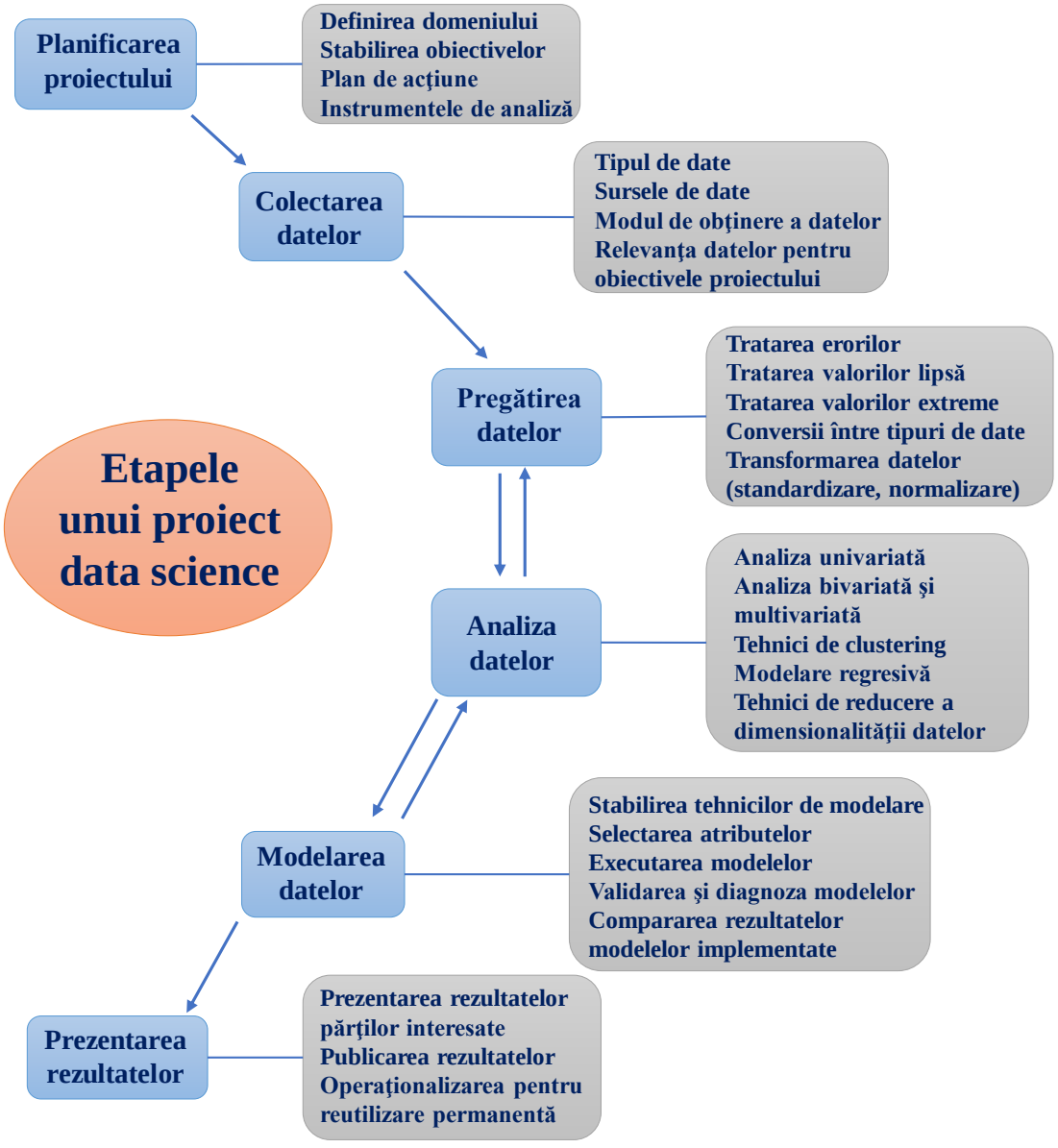

Figura 2.1: Etapele unui proiect data science

2.1. Planificarea proiectului

Un proiect de data science trebuie să înceapă prin definirea domeniului studiat, stabilirea obiectivelor proiectului, datele şi resursele necesare pentru colectarea lor, stabilirea planului de acţiune şi a etapelor de urmat.

Proiectul trebuie să aibe un scop şi obiective de cercetare foarte clar stabilite, un plan de acţiune, termene de finalizare pentru etapele proiectului, pentru a putea obţine rezultate bine justificate. Înţelegerea obiectivelor studiului, a contextului în care se desfaşoară cercetarea, stabilirea surselor de colectare a informaţiilor şi a instrumentelor de efectuare a analizei, identificarea resurselor necesare şi disponibile, sunt condiţii esenţiale pentru reuşita proiectului.

2.2. Obţinerea/colectarea datelor

Primele demersuri care trebuiesc făcute pentru colectarea datelor sunt stabilirea tipului de date necesare analizei, stabilirea surselor de date şi a modului de obţinere a datelor. Datele pot fi colectate din diverse surse şi pot fi obţinute în diverse formate.

În funcţie de domeniul de acţiune şi de obiectivele analizei, datele pot fi obţinute din mai multe surse [4]:

- prin interogarea a diverse tipuri de baze de date, dacă proiectul se desfăşoară în cadrul unei companii şi avem acces la bazele de date sau la diverse informaţii
- prin webscraping, de pe diverse website-uri
- prin intermediul Web APIs de pe website-uri; Facebook sau Twitter de exemplu permit utilizatorilor să se conecteze la serverele lor web şi să acceseze diverse date prin Web API-urile lor
- importante seturi de date pot fi descărcate direct de pe diverse site-uri (inclusiv site-uri guvernamentale), ca de exemplu Kaggle, UCI Machine Learning Repository, Data World, Dataset Search de la Google, AWS Public Data Sets de la Amazon şi multe altele

Odată obţinute datele, este necesar să se facă verificarea relevanţei lor pentru scopul şi obiectivele proiectului, respectiv trebuie verificat dacă colecţia de date asigură resursele necesare pentru efectuarea analizei şi modelării.

2.3. Pregatirea datelor

Pregătirea datelor presupune organizarea şi structurarea datelor brute pentru transformarea lor în date viabile pentru analiză şi modelare. Este o etapă care se concentrează pe asigurarea calităţii şi consistenţei datelor, astfel ca datele să devină reprezentative pentru procesele de analiza şi modelare.

Etapa pregatirii datelor are 3 subetape: curăţarea datelor, integrarea datelor, transformarea datelor [3]. Pregătirea datelor are ca scop eliminarea datelor inconsistente sau completarea datelor lipsă acolo unde este posibil, astfel încât să nu se influenţeze rezultatul.

2.3.1. Tratarea erorilor

Există diverse erori în cadrul setului de date, atât pentru datele numerice, cât şi pentru cele categoriale, care afectează analiza şi modelarea datelor. Aceste erori trebuiesc identificate şi eliminate, pentru a nu produce rezultate eronate.

În cadrul variabilelor categoriale trebuiesc găsite şi eliminate spaţiile libere prezente la începutul sau sfârşitul unei valori de tip string. De asemenea trebuie evitată capcana sensibilităţii la majuscule ale datelor de tip string. În ceea ce priveşte variabilele numerice, trebuiesc verificate valorile imposibile, de ex. vârsta de peste 100 de ani, înălţimea de peste 2 metri, etc.

2.3.2. Tratarea valorilor care lipsesc

Unul dintre primele aspecte care trebuiesc analizate la un set de date, este prezenţa spaţiilor libere, a elementelor lipsă. Există diverse metode pentru tratarea datelor lipsă, printre care, eliminarea instanţelor (linii) care conţin observaţii cu valoare zero sau a celor fără valoare (nan), eliminarea unor atribute (coloane) în cazul în care conţin prea multe valori lipsă şi nu pot fi luate în calcul la analiza şi modelarea datelor [5]. Datele lipsă pot fi înlocuite şi cu indicatori statistici adecvaţi, de ex: media, mediana sau modulul coloanei pe care se găsesc datele, acolo unde este posibil şi înlocuirea nu va influenţa rezultatul final.

În cazul în care există o relaţie liniară între 2 sau mai multe variabile, se poate utiliza şi regresia liniară pentru identificarea şi înlocuirea valorilor lipsă.

2.3.3. Conversii între tipuri de date

Datele care fac obiectul analizei pot fi numerice, categoriale, binare, imagini, etc. În scopul analizei şi modelării datelor se poate impune codificarea sau convertirea lor dintr-un tip de date în alt tip de date.

Asupra datelor numerice (cantitative) reprezentate de numere întregi sau reale, se pot face operaţii aritmetice de adunare, scădere, înmulţire, împărţire, se pot calcula indicatori statistici de tendinţă centrală, de variaţie, etc. De asemenea se poate defini o ordine între valorile atributelor numerice, un minim, un maxim, se pot încadra în anumite intervale sau cuantile. Exemple de date numerice pot fi vârsta, greutatea, preţul, temperatura, etc.

Datele categoriale sau discrete sunt cele care au două sau mai multe categorii, şi ele pot fi nominale sau ordinale.

Datele categoriale nominale sunt valori discrete care nu au o ordine intrinsecă şi pot fi simboluri, caractere, şiruri de caractere, etc. Exemple de date nominale sunt genul (F/M), rasa, starea civilă, etc. Asupra datelor nominale se pot efectua operaţii de calcul al frecvenţelor sau de verificare a egalităţii.

Datele categoriale ordinale pot fi sub formă de numere, simboluri, şiruri de caractere, şi sunt date între care există o ordine clară în cadrul mulţimii de valori. Exemple de date ordinale sunt temperatura (ridicată, medie, scăzută), nivele de calitate, etc. Asupra acestui tip de date nu are sens să se efectueze operaţii aritmetice.

Datele binare sunt date care acceptă doar două valori posibile (0 sau 1, adevărat sau fals) şi se utilizează pentru a codifica prezenţa sau absenţa unor caracteristici, în general notându-se cu 1 prezenţa caracteristicii şi cu 0 absenţa ei.

Datele categoriale pot fi convertite în echivalente numerice sau binare prin operaţii de codare (encoding), întrucât anumite tehnici de machine learning (de exemplu: regresia liniară, support vector machine, neural networks) pot fi aplicate doar pentru date numerice. De asemenea, operaţiunile de normalizare, standardizare, analiza în componente principale pot fi aplicate doar atributelor numerice.

Conversia unui atribut numeric în unul categorial se numeşte discretizare (de exemplu vârsta poate fi împarţită pe categorii 20-35, 36-45, etc. şi se asignează câte o valoare fiecărui subdomeniu) sau binarizare (binning) dacă este vorba despre conversia în atribute binare.

2.3.4. Tratarea valorilor extreme (outliers)

Identificarea valorilor extreme (outliers) şi studierea naturii acestora în vederea determinării modului de tratare a acestor valori, este un alt element important al pregătirii datelor. Valorile extreme pot fi identificate prin utilizarea histogramelor (cea mai întâlnită distribuţie în ştiinţele naturii este distribuţia Gaussiana) [3], a graficelor scatter plot, a tabelelor de frecvenţă (în cazul variabilelor categoriale), etc.

Valorile extreme pot fi rezultatul unor metode inadecvate de măsurare sau colectare a datelor, caz în care sunt anomalii care trebuiesc eliminate. De exemplu dacă într-un eşantion avem înălţimea unor persoane în funcţie de sexul lor şi observăm valori anormale, atunci aceste date trebuiesc tratate ca anomalii şi eliminate.

Există şi cazuri în care valorile extreme pot fi observaţii reale care ies din tiparul general al populaţiei statistice măsurate, caz în care modul de tratare a acestor valori se face în funcţie de situaţia dată. Dacă datele extreme nu modifică rezultatul, dar influenţează ipotezele, atunci ele pot fi eliminate. Dacă datele extreme afectează şi rezultatul şi ipotezele, atunci nu este indicată eliminarea lor, ci este recomandabil să se facă modelarea datelor atât în varianta cu includerea valorilor extreme, cât şi în varianta cu eliminarea lor, prezentându-se rezultatele comparativ. De asemenea, se pot testa comparativ mai multe modele pentru a vedea care este mai potrivit.

Identificarea şi tratarea valorilor extreme se poate face şi prin utilizarea modelelor statistice robuste. Media şi abaterea standard sunt sensibile la outliers, de aceea în locul mediei poate fi utilizată mediana, iar în locul deviaţiei standard se poate utiliza deviaţia cvartilă. Distribuţia normală şi abaterea standard pot fi utilizate în eliminarea datelor extreme, întrucât conform regulii celor 3 sigma pentru variabilele aleatoare normale, 68% din datele unei distribuţii normale se situează într-o deviaţie standard, 95% din date se află în două deviaţii standard, iar 99.7% se situează în 3 deviaţii standard, probabilitatea ca abaterea în valoare absolută să depăşească 3 sigma este de 0.026, adică practic 0.

În cazul modelului regresiei liniare, se poate utiliza comparativ regresia robustă pentru identificarea datelor extreme, de exemplu regresia RANSAC.

2.3.5. Transformarea datelor

Datele trebuie să fie transformate într-o formă adecvată pentru a permite măsurarea adecvată a proceselor pe care le exprimă, astfel că ele trebuiesc normalizate sau standardizate

astfel ca datele pe coloană să fie de acelaşi tip, să nu existe unităţi de măsură diferite ale observaţiilor unei variabile. Normalizarea şi standardizarea datelor sunt metode de scalare a variabilelor.

Standardizarea presupune scalarea datelor pentru a avea media 0 şi abaterea standard 1 (se utilizează de exemplu la analiza în componente principale pentru calcularea matricei de covarianţă) şi este utilă în cazul datelor care urmează o distribuţie normală, Gaussiana [6]. Standardizarea se face după urmatoarea formulă:

$$x` = \frac{x - \overline{x}}{\sigma}$$

Normalizarea (Min-Max scalling) presupune scalarea datelor astfel ca valorile să se încadreze într-un interval prestabilit, între 0 şi 1. Normalizarea este indicat să se folosească atunci când datele nu urmează o distribuţie normala, Gaussiana, cum sunt algoritmii neural networks sau algoritmii de clasificare bazaţi pe calcularea distanţelor [6].

Normalizarea se face după urmatoarea formulă:

$$x` = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Algoritmii de machine learning bazaţi pe arbori (decision trees, random forest, gradient boosting) utilizaţi pentru probleme de clasificare sau regresie, precum şi regresia liniară bazată pe modelul celor mai mici pătrate, nu necesită scalarea variabilelor.

Regresia liniară, regresia logistică, reţelele neurale, etc., care utilizează tehnica gradient descent (pasul descendent) ca şi tehnică de optimizare, necesită scalarea datelor, deoarece diferenţa de mărime între valorile variabilelor independente va determina praguri diferite pentru fiecare variabilă [6]. Utilizarea variabilelor scalate similar va ajuta algoritmul să conveargă mai repede spre valoarea minimă, în timp ce paşii vor fi actualizaţi la aceeaşi rată pentru toate variabilele independente.

Algoritmii de clasificare bazaţi pe calcularea distanţelor, cum sunt K-Nearest Neighbors (KNN), K-means sau Support Vector Machine (SVM) sunt de asemenea influenţaţi de intervalul de mărime a variabilelor, deoarece aceşti algoritmi utilizează calcularea distanţelor între puncte pentru a determina similaritatea lor [6]. De aceea, pentru creşterea performanţei algoritmului şi pentru ca algoritmul să nu fie afectat de bias, se recomandă scalarea datelor.

2.4. Analiza datelor

Datele colectate şi pregătite sunt supuse analizei exploratorii pentru identificarea caracteristicilor principale, obiectivul fiind înţelegerea datelor şi a relaţiilor dintre ele, identificarea unor rezultate statistice semnificative. Analiza exploratorie a datelor contribuie la identificarea anumitor tipare, la descoperirea eventualelor anomalii şi date extreme care au trecut de etapa de pregătire a datelor, la extragerea variabilelor semnificative şi reducerea dimensionalităţii setului de date, la formularea şi testarea unor ipoteze de lucru, la alegerea celor mai adecvate tehnici de modelare a datelor.

În sensul analizei datelor exploratorii sunt utilizate diverse tehnici şi instrumente statistice care includ [7]:

- analiza univariată, care presupune utilizarea indicatorilor statistici în vederea studierii datelor în funcţie de o singură variabilă pentru a se vedea principalele caracteristici şi tendinţe, astfel că nu este nevoie de o analiză care să arate raporturi de cauzalitate sau corelaţie;
- analiza bivariată şi multivariată care presupune utilizarea metodelor statistico-matematice pentru identificarea legăturilor de asociere existente între două sau mai multe variabile;
- tehnici de clustering pentru identificarea eventualelor tendinţe de grupare a datelor;
- tehnici de reducere a dimensionalităţii datelor care contribuie la creşterea performanţei tehnicilor de modelare şi la reducerea considerabilă a sistemului;
- modelare regresivă.

Analiza statistică univariată şi multivariată se poate face prin tehnici statistice non-grafice sau prin vizualizare grafică, întrucât tehnicile fără vizualizare grafică nu pot reda imaginea distribuirii datelor şi a relaţiilor dintre ele.

Analiza univariată si multivariată prin reprezentare grafică se poate face prin diverse grafice univariate sau multivariate ca de exemplu: histograme sau diagrame de frecvenţă, curba densităţii de probabilitate, diagrama boxplot care arată deviaţia quartilă, diagrama cu coloane, diagrame cu bare, diagrame cu linii, grafice scatter plot care arată relaţia dintre variabile, diagrama de procente (pie-chart), grafice time-series, grafice heat map, etc.

O etapă importantă a analizei exploratorii a datelor este reducerea dimensionalităţii, care presupune reducerea numărului de atribute prin care poate fi explicat modelul, în scopul eliminării informaţiilor redundante şi a eficientizării implementării tehnicilor algoritmice de modelare a datelor.

2.4.1. Reducerea dimensionalităţii

Reducerea dimensionalităţii este una dintre cele mai importante etape în analiza datelor, înainte de modelarea lor. În machine learning dimensionalitatea se referă la numărul variabilelor prezente în setul de date. Atunci când numărul variabilelor independente este mare, algoritmii pot avea probleme de performanţă, iar aceasta se numeşte „blestemul dimensionalităţii" [8].

Reducerea dimensionalităţii înseamnă transformarea datelor dintr-un spaţiu cu un număr mare de dimensiuni/atribute, într-un spaţiu cu un număr redus de dimensiuni/atribute. Transformarea trebuie să se realizeze fără a pierde prea mult din informaţia iniţială, astfel ca reprezentarea să elimine atributele irelevante, dar prin reducerea dimensiunilor să păstreze proprietăţile datelor iniţiale într-un procent semnificativ.

Reducerea dimensiunii poate fi abordată prin selectarea atributelor (feature selection) şi extragerea atributelor (feature extraction).

Modelele bazate pe extragerea atributelor determină un set de date cu un număr mai mic de atribute decat setul original, dar nu păstrează caracteristicile originale şi crează caracteristici noi din combinarea matematică a atributelor iniţiale. Pentru acest motiv modelele feature extraction nu sunt fiabile în cazul în care asupra datelor urmează sa fie aplicaţi algoritmi de machine learning care presupun predicţii sau previziuni ale unor variabile pe baza altor variabile, cum sunt modelele de regresie. Unii algoritmi au încorporate modele de extragerea atributelor, cel mai bun exemplu fiind deep learning. De asemenea, extragerea atributelor poate fi nesupervizată prin analiza în componente principale (PCA) sau supervizată prin analiza discriminantă liniara (LCA).

Modelele de reducere a dimensionalităţii prin selectarea atributelor păstrează un subset al caracteristicilor iniţiale, motiv pentru care sunt fiabile pentru datele asupra cărora se vor aplica modele de regresie. Unii algoritmi de învăţare supervizată cum sunt modelele de regresie regularizate sau random forests, au încorporate modele de reducere a dimensionalităţii prin selectarea atributelor. Astfel regresia ridge regularizează coeficienţii dar păstrează toate atributele, regresia lasso regularizează coeficienţii pentru a efectua selecţia atributelor, iar elastic nets combină aceste două tipuri de regresie regularizată.

Reducerea dimensionalităţii are mai multe avantaje [9]:

- reducerea numărului atributelor înseamnă reducerea spaţiului necesar pentru stocarea datelor;

- evitarea multicoliniarităţii prin eliminarea atributelor redundante;
- algoritmii de machine learning lucrează mai eficient pe seturi de date cu dimensiuni reduse, iar modelele rezultate sunt mai simple şi datele rezultate sunt mai uşor de vizualizat.

Există mai multe metode de reducere a dimensionalităţii datelor, unele mai simple, altele mai complexe, printre care:

Filtrul corelaţiei scăzute între variabilele independente şi variabila dependentă

Corelaţia exprimă interdependenţa sau legătura dintre două variabile, fără a exista o relaţie de cauzalitate între ele. Astfel, o variabilă dependentă care nu este corelată cu variabila target, nu poate avea o pondere importantă în explicarea comportamentului variabilei target, motiv pentru care poate fi eliminată din model [9].

Filtrul corelaţiei ridicate între variabilele independente

Un indicator mare al corelaţiei între două variabile dependente arată faptul că cele doua variabile au tendinţe similare şi este posibil să aducă în model informaţii similare, redundante, determinând multicoliniaritatea modelului.

Pentru evitarea acestor probleme se calculează corelaţia între variabilele independente (se poate utiliza matricea de corelaţie), iar dacă două variabile sunt foarte corelate (coeficientul de corelaţie depaseste 0,5-0,6), atunci una poate fi eliminată fără a pierde foarte mult din informaţie [10]. Vom păstra variabila care este mai corelată cu variabila dependentă.

Eliminarea progresivă (backward feature elimination)

Eliminarea progresivă se face în mai mulţi paşi. Se începe prin antrenarea modelului utilizând toate variabilele prezente în setul de date, iar apoi se calculează performanţa modelului (pe baza coeficientului de determinare ajustat). Ulterior se elimină pe rând cate o variabilă şi se aplică modelul la toate seturile de variabile n-1, calculându-se performanţa pentru fiecare variantă. Se identifică variabila a cărei eliminare a produs cea mai mică variaţie a performanţei modelului şi se elimină din setul de date. Se repetă procedura, până când nu se mai pot elimina variabile [9].

Selecţia progresivă (forward feature selection/construction)

Este inversa procedurii eliminării progresive, întrucât în loc să se elimine pe rând variabilele, se încearcă găsirea celei mai utile variabile care îmbunătăţeşte performanţa modelului.

Se începe prin aplicarea modelului pe rând, la câte o singură variabilă independentă, până epuizăm toate cele n variabile independente. Se selectează ca şi variabilă de pornire, variabila care determină cea mai mare performanţă a algoritmului. Se repetă procesul prin adaugarea a câte o variabilă, pe rând, din cele n-1 variabile independente rămase, iar variabila care determină cea mai bună performanţă se reţine pentru iteraţiile viitoare. Se repetă procesul până când nu se mai identifică o îmbunătăţire a performanţei modelului [9].

Eliminarea progresivă şi selecţia progresivă sunt tehnici care se aplică modelelor de regresie, dar întrucât sunt metode consumatoare de timp şi resurse, aceste tehnici sunt utilizate atunci când avem un set de date cu un număr redus de atribute. Variabilele se pot elimina prin compararea performanţelor modelului aplicat la diversele variante ale setului de date pe baza coeficientului de determinare ajustat sau selecţia variabilelor se opreşte când valorile t calculate devin mai mici decât valoarea critică citită din tabela Student.

Random forest regressor – importanţa atributelor

Random forest este unul dintre cei mai utilizaţi algoritmi pentru feature selection, întrucât are încorporat un model de determinare a semnificativităţii atributelor în explicarea modelului [9].

Prin random forest, care este un model de tip ansamblu (ensemble learning), se generează un set de arbori de decizie cu privire la o variabilă target şi apoi se utilizează statistica de performanţă a fiecărui atribut predictor pentru a identifica subsetul de atribute care furnizează cea mai mare cantitate de informaţie cu privire la model.

Algoritmul generează un set de arbori, fiecare dintre aceştia fiind antrenat cu o parte mică din totalul atributelor predictor, luate în diverse combinaţii. Dacă un atribut este selectat de multe ori ca şi criteriu de ramificare (best split), atunci este foarte probabil că acesta este un atribut semnificativ al modelului.

După generarea ansamblului de arbori, se calculează scorul fiecărui atribut, în funcţie de câte ori a fost ales ca şi criteriu de ramificare şi la ce nivel, în cadrul atributelor candidate.

Scor = ramificaţii(niv.0)/candidaţi(niv.0) + ramificaţii(niv.1)/candidaţi(niv.1) + ...

Scorul arată care sunt atributele cele mai importante în explicarea modelului. Se vor reţine doar atributele care înregistrează un scor mai mare decât un prag stabilit.

Regresia lasso pentru selectarea atributelor

Regresia Lasso este un algoritm care crează un model regularizat al regresiei liniare, prin adăugarea unui parametru de regularizare. Prin modificarea acestui parametru în sensul creşterii

sau descreşterii lui, coeficienţii ecuaţiei de regresie liniară scad către valoarea 0. Variabilele a căror coeficient atinge valoarea 0 sunt variabilele care pot fi eliminate, întrucât contribuţia lor în cadrul modelului este nesemnificativă [10].

2.5. Modelarea datelor

După finalizarea operaţiunilor de pregătire şi analiză a datelor, se poate trece la modelarea lor prin tehnici de statistică, data mining sau machine learning. Tehnicile de modelare trebuiesc alese în funcţie de tipul de date care sunt supuse analizei şi de scopul analizei.

Etapele procesului de modelare cuprind [3]:

- stabilirea tehnicilor de modelare şi selectarea atributelor/variabilelor care se modelează
- executarea şi validarea modelelor
- diagnoza modelelor şi compararea rezultatelor obţinute în urma modelării

În urma rezultatelor analizei exploratorii a datelor se pot stabili variabilele care explică cel mai bine modelul şi care vor constitui baza pentru construirea unor modele performante. În alegerea modelelor trebuie să se ţină cont de scopul analizei, de resursele necesare pentru implementare, pe ce perioadă de timp vor rămâne relevante rezultatele analizei şi dacă va fi nevoie de întreţinerea ulterioară a modelului pentru păstrarea relevanţei, care este performanţa modelului.

Baza procesului de modelare a datelor o constituie tehnicile de machine learning (învăţarea automată), care sunt algoritmi care se pot autoperfecţiona prin experienţă. Algoritmii de machine learning construiesc modele bazate pe serii de date cu scopul de a elabora predicţii pe baza cărora să se poată lua decizii, prin descoperirea relaţiilor de cauzalitate sau de corelaţie între diverse variabile. Tehnicile de machine learning pot contribui la structurarea şi identificarea relaţiilor existente între anumite elemente ale unui sistem de informaţii şi la descoperirea de noi structuri, noi relaţii şi semnificaţii.

Alegerea şi implementarea algoritmilor de machine learning utilizaţi pentru modelarea unui set de date trebuie să se facă în funcţie de sursa, dimensiunea şi tipul datelor utilizate, de sarcinile pe care trebuie să le îndeplinească şi de performanţa avută în vedere în scopul autoperfecţionării prin experienţă.

Algoritmii de machine learning utilizaţi în Data Science pot fi împărţiţi în două tipuri, în funcţie de modul de învăţare, în algoritmi supervizaţi şi nesupervizaţi.

Algoritmii de învăţare supervizată (supervised learning) sunt algoritmi care construiesc modele matematice pe seturi de date în care rezultatele sunt cunoscute, iar prin instruirea pe

datele de antrenare, algoritmii învaţă să facă predicţii pe seturi noi de date, care nu au făcut parte din datele de antrenament [11].

Exemple: algoritmi de regresie, de clasificare, K-nearest neighbors, decision tree, random forest, support vector machine, naïve Bayes, neural networks (multilayer perceptron), etc.

Algoritmii de învăţare nesupervizată (unsupervised learning) sunt algoritmi care au la dispoziţie doar date neetichetate, iar rezultatele nu sunt cunoscute [11]. Aceşti algoritmi încearcă în mod iterativ să extragă anumite reguli din datele disponibile şi să organizeze datele pe baza identificării unor similitudini între date şi pot avea ca obiectiv gruparea unor eşantioane cu comportament similar sau identificarea unor caracteristici similare în cadrul datelor.

Exemple: algoritmi de grupare (clustering), reducerea dimensionalităţii datelor (analiza în componente principale), învăţarea regulilor de asociere, algoritmi de detectare a anomaliilor, etc.

Algoritmii de machine learning pot fi utilizaţi şi în alte etape ale unui proiect, nu doar în partea de modelare. Tehnicile de modelare pot fi utilizate în pregătirea şi analiza datelor, în completarea valorilor lipsă, în detectarea valorilor extreme, reducerea dimensionalităţii datelor, etc.

În procesul de modelare, tehnicile de machine learning pot fi utilizate pentru explicarea cauzelor unui fenomen (root cause analysis), pentru clasificarea unor fenomene sau pentru predicţia unor fenomene.

Data science are la dispoziţie diverse tehnici de modelare a datelor, de aceea este foarte important să se stabilească care sunt modelele adecvate pentru analiză, în funcţie de tipul datelor şi de scopul analizei. Asupra unui set de date pot fi aplicate în paralel mai multe tehnici de modelare, a căror performanţe trebuiesc evaluate şi comparate, pentru a determina care dintre modele explică cel mai bine fenomenele studiate.

Un model eficient trebuie să îndeplinească două condiţii de bază:

- să aibe putere predictivă asupra fenomenelor studiate
- să poată fi generalizat şi extrapolat pe seturi noi de date

Pentru îndeplinirea acestor condiţii este necesară implementarea următoarelor strategii:

- ***evaluarea modelelor*** prin definirea unor instrumente de măsurare a erorilor pe care pot să le genereze modelele, ca de exemplu rata erorii de clasificare, matricea de confuzie pentru clasificatori, sau coeficientul de determinare R^2, MSE, RMSE, MAE, pentru modelele de regresie;

- stabilirea unor ***strategii de validare a modelului***, ca de exemplu divizarea setului de date în date de antrenare şi date de test, utilizarea validării încrucişate, implementarea unor metode de regularizare a modelelor pentru tehnicile de regresie, în scopul obţinerii unui compromis între bias şi varianţă, etc.

Validarea modelului este extrem de importantă, deoarece extrapolarea rezultatelor pe seturi noi de date determină dacă modelul poate funcţiona în condiţii reale şi dacă este o reprezentare reală a fenomenelor studiate.

2.6. Prezentarea rezultatelor

În urma modelării datelor şi a obţinerii rezultatelor cu privire la fenomenele analizate, rezultatele trebuiesc prezentate părţilor interesate şi/sau pot constitui baza implementării unor aplicaţii.

Rezultatele şi concluziile studiului pot fi prezentate ca atare, prin utilizarea unor instrumente de prezentare sau pot fi publicate în medii ştiinţifice de specialitate. Rezultatele pot fi automatizate şi utilizate prin implementarea unor aplicaţii, iar întreg demersul de analiză poate fi operaţionalizat pentru reutilizare permanentă şi pentru integrarea cu alte instrumente [3].

3. Aspecte teoretice şi practice ale unui proiect data science bazat pe analiza de regresie

Un proiect data science este un demers complex prin care date reale culese din diverse domenii sunt modelate prin tehnici de statistică, matematică, machine learning sau data mining. În funcţie de tipul de date şi de obiectivul studiului, datele pot fi analizate şi modelate pentru regresie, clasificare, clustering, etc.

Analiza de regresie este un instrument prin care se pot analiza legăturile, conexiunile sau corelaţiile existente între două sau mai multe fenomene. Pe baza rezultatelor analizei de regresie se pot lua decizii obiective, bazate pe date reale, observate prin măsurarea fenomenelor sociale, economice şi din oricare alte domenii de activitate.

Analiza de regresie cuprinde tehinci statis

tice şi matematice, precum şi strategii de evaluare şi validare a modelelor. Principalele concepte care stau la baza analizei de regresie şi a strategiilor de validare sunt prezentate în prezentul capitol.

3.1. Consideratii statistice

Statistica „*este ştiinţa care studiază fenomenele şi procesele de tip colectiv din societate, natură, etc., din punct de vedere cantitativ, în scopul descrierii acestora şi al descoperirii legilor care guvernează manifestarea lor*" [12].

Statistica este utilizată în data science în scopul pregătirii şi analizării datelor, pentru a identifica anumite tendinţe sau relaţii între datele analizate, pentru a înţelege datele analizate (statistica descriptivă) sau pentru a descoperi noi structuri, noi relaţii şi semnificaţii despre problemele analizate (statistica inferentială). Datele sunt analizate prin intermediul unor indicatori statistici care arată tendinţa centrală sau variabilitatea lor.

Indicatorii tendinţei centrale arată care este tendinţa evenimentelor analizate de a se grupa în jurul unui punct central. Principalii indicatori statistici care arată tendinţa centrală a unui set de date sunt media, mediana, modul [13].

Media unei serii de date statistice este raportul dintre suma valorilor individuale şi numărul lor, şi reprezintă măsura valorii centrale a setului de date în jurul căreia fluctuează datele. Se utilizează în general atunci cand fenomenele cercetate prezintă o tendinţa liniară.

Acest indicator nu este reprezentativ dacă termenii seriei de date sunt prea dispersaţi, întrucât este sensibil faţă de valorile extreme [14].

$$\bar{x} = \frac{\Sigma x_i}{n}$$

Suma diferenţelor dintre toate valorile individuale ale seriei şi media setului de date este 0.

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Mediana este valoarea centrală din setul de date ordonat crescător sau descrescător, respectiv valoarea care împarte seria în două părţi egale. Mediana nu este afectată de valorile extreme, dar poate fi nereprezentativă dacă valorile individuale nu se grupează în jurul valorii centrale [13] [14].

- pentru o serie cu număr impar de termeni mediana se calculează astfel:

$$Me = x_{\frac{n+1}{2}}$$

- pentru o serie cu număr par de termeni mediana se calculează astfel:

$$Me = \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$$

Modul este valoarea care are cea mai mare frecvenţă în seria de date, fiind un indicator specific seriilor de distribuţie unimodală (cu un singur maxim) şi multimodală (cu mai multe maxime) [14].

Indicatorii variaţiei sau variabilităţii arată pe ce domeniu se întinde setul de date, respectiv cât de dispersate sau de extinse sunt datele, având rol în inferenţa statistică (estimarea parametrilor pentru întreaga populaţie pornind de la rezultatele obţinute la nivelul eşantionului). Se utilizează deoarece indicatorii tendinţei centrale nu dau nicio indicaţie asupra împrăştierii datelor. Principalii indicatori ai variaţiei sunt coeficientul de variaţie, deviaţia quartilă, deviaţia standard (abaterea standard), dispersia (varianţa), covarianţa [13].

Coeficientul de variaţie propus de Pearson este raportul dintre deviaţia standard şi media seriei de date, fiind un indicator al împrăştierii datelor în raport cu tendinţa centrală.

$$Cv = \frac{\sigma}{\bar{x}} * 100$$

Acest indicator statistic se utilizează în special în studiul omogenităţii unei populaţii şi pentru compararea a două serii de date atunci când au mediile mult diferite şi abaterea standard nu lămureşte diferenţele. Coeficientul de variaţie se exprimă în procente şi are valorile cuprinse între 0 şi 100 %. Cu cât valoarea lui este mai apropiată de 0 cu atât seria statistică este mai omogenă şi media este mai reprezentativă, iar cu cât valoarea se apropie de 100 cu atât seria este mai eterogenă iar media calculată este mai puţin reprezentativă. Pragul de la care se consideră că seria de date nu mai este omogenă este de 30-35% [13].

Deviaţia quartilă măsoară dispersia datelor aflate în zona de mijloc a distribuţiei şi determină un interval centrat pe mediană.

$$DQ = \frac{(Q3 - Q1)}{2}$$

Se ordonează datele, după care se calculează valoarea mediană a jumătăţii superioare Q3 = UQ (upper quartile) şi valoarea mediană a jumătăţii inferioare Q1 = LQ (lower quartile) [14].

Indicatorii de localizare se prezintă astfel:

- Q1 ≥ 25% din datele seriei
- Q2 = mediana (50% din date)
- Q3 ≥ 75% din datele seriei

***Deviaţia standard (abaterea standard*)** este un indicator al împrăştierii seriei de date în jurul valorii medii şi se calculează ca medie pătratică sau ponderată a abaterilor valorilor seriei faţă de media lor, fiind de fapt radicalul dispersiei. O serie de date este omogenă dacă abaterea standard are o valoare mică [13].

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Dispersia (varianţa) este tot un indicator al gradului de împrăştiere a datelor şi se calculează ca medie a pătratelor abaterilor termenilor seriei de la tendinţa lor centrală. Dispersia este abatarea standard ridicată la pătrat. O valoare mare a dispersiei arată o împrăştiere mare a valorilor seriei şi o serie eterogenă, iar o valoare mică arată o împrăştiere mică a datelor, respectiv o serie omogenă.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Dacă comparăm două sau mai multe serii de date, la medii aproximativ egale, este mai împrăştiată seria cu dispersia mai mare, iar la dispersii aproximativ egale, este mai împrăştiată seria cu media mai mică [15].

Covarianţa este măsura de variaţie comună a două variabile aleatorii şi se calculează după următoarea formulă:

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})$$

Dispersia (varianţa) este un caz special al covarianţei, în care cele două variabile sunt identice.

$$cov(x,x) = var(x)$$

Dacă valorile mari ale unei variabile corespund cu valorile mari ale altei variabile şi în mod corespunzător acest fapt este valabil şi pentru valorile mici, atunci avem două variabile cu comportamente similare, ceea ce înseamnă că avem covarianţă pozitivă. Dacă acest lucru nu este valabil şi avem valori mari la o variabilă şi valori corespunzătoare mici la cealaltă variabilă, atunci cele două variabile au comportamente opuse şi covarianţa este negativă. Semnul covarianţei arată practic direcţia relaţiei liniare care există între două variabile [16].

Totuşi covarianţa este greu de interpretat, deoarece nu este normalizată şi depinde de interpretarea celor două variabile. Din acest motiv, se utilizează covarianţa aplicată la date normalizate în prealabil sau varianta normalizată a covarianţei, respectiv corelaţia, care poate arăta puterea relaţiei liniare între două variabile.

Corelaţia (coeficientul de corelaţie) se utilizează pentru a defini interdependenţa sau legătura dintre două variabile. Acest indicator nu descrie o relaţie cauzală între două variabile şi nu permite predicţia unei variabile pe baza altei variabile. Coeficientul de corelaţie propus de Pearson se calculează astfel:

$$r_{x,y} = \frac{cov(x,y)}{\sqrt{var(x)} * \sqrt{var(y)}}$$

$$r_{x,y} = \frac{\sum((x_i - \bar{x}) * (y_i - \bar{y}))}{\sqrt{\sum(x_i - \bar{x})^2} * \sqrt{\sum(y_i - \bar{y})^2}}$$

Semnul valorii corelaţiei indică direcţia relaţiei între două variabile:

- coeficient de corelaţie pozitiv = corelaţie directă, înseamnă că cele două variabile variază în acelaşi sens (dacă x creşte - creşte şi y, dacă x scade – scade şi y)

- coeficient de corelaţie negativ = corelaţie inversă, înseamnă că cele două variabile variază în sens contrar (daca x creşte – y scade, daca x scade – y creşte) [15]

Coeficientul de corelaţie $r_{x,y}$ ia valori între [-1,1], iar valoarea absolută a coeficientului constituie un indiciu al intensităţii corelaţiei sau legăturii între cele două variabile x şi y. Corelaţia este foarte strânsă când $r_{x,y} \to 1$ şi foarte redusă când $r_{x,y} \to 0$.

Astfel, atunci când coeficientul de corelaţie ia valoarea -1 avem o corelaţie perfectă, inversă sau negativă, când ia valoarea 0 avem o corelaţie nulă sau inexistentă (asociere aleatoare) şi când ia valoarea 1 avem o corelaţie perfectă, directă sau pozitivă.

3.2. Regresia liniară

Este un instrument de modelare statistică, utilizat pentru a determina existenţa unor legături sau corelaţii între date numerice. Modelul este reprezentat grafic printr-o linie care uneşte cel mai bine datele, care sunt reprezentate prin puncte [2]. Analiza de regresie se face pe baza unei ecuaţii de regresie, care exprimă evoluţia variabilelor analizate şi printr-un calcul al semnificativităţii utilizării acestei tehnici, care arată cât de adecvat este modelul pentru analizarea datelor. Pe baza datelor de antrenare, reprezentate de variabile numerice, se pot face predicţii sau previziuni ale unor variabile pe baza altor variabile.

Regresia liniară identifică linia dreaptă care reprezintă cea mai bună aproximaţie dintr-o diagramă de dispersie a datelor [2]. Regresia liniară simplă utilizează două variabile pentru analiză, iar regresia multiplă utilizează mai multe variabile independente care afectează o variabilă dependentă.

Regresia liniară simplă face predicţii ale unei variabile dependente y, pe baza unei variabile independente care este cunoscută x, prin utilizarea metodei celor mai mici pătrate.

Ecuaţia funcţiei liniare de regresie este următoarea:

$$y = a + bx + \varepsilon$$

Pentru regresia liniară multiplă ecuaţia funcţiei de regresie este următoarea:

$$y = a + b_1 * x_1 + b_2 * x_2 + \cdots + b_i * x_i + \varepsilon$$

Interceptul şi coeficientul de regresie sunt estimaţi prin rezolvarea următoarei probleme de optimizare [17] [18]:

$$\min_{\{a,b\}} F(a,b), \quad \text{unde} \quad F(a,b) = \sum_{i=1}^{n} \{y_i - (a + bx_i)\}^2$$

Minimul se obţine prin calcularea sistemului de derivate parţiale:

$$\begin{cases} \frac{\partial F(a,b)}{\partial a}|_{a=\hat{a},b=\hat{b}} = 0, \\ \frac{\partial F(a,b)}{\partial b}|_{a=\hat{a},b=\hat{b}} = 0, \end{cases}$$

care se transformă în:

$$\begin{cases} -2\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i) = 0 \\ -2\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)x_i = 0 \end{cases}$$

Ca urmare a rezolvării sistemului de ecuaţii, se obţine formula de calcul a estimatorului interceptului şi formula de calcul a estimatorului coeficientului de regresie.

Interceptul este locul de pe ordonată unde dreapta de regresie se intersectează cu axa Oy şi reprezintă o constantă care este valoarea medie a lui y pentru $x = 0$, respectiv nivelul variabilei dependente y care nu este determinat de variabila independentă x, ci de alte variabile sau alţi factori. Interceptul se calculează după urmatoarea formulă:

$$\hat{a} = \bar{y} - \hat{b} * \bar{x}$$

Panta de regresie arată variaţia medie a variabilei y, atunci cand x creşte cu o unitate, indicând faptul că între cele două variabile x şi y există o legatură directă. Astfel la creşterea cu o unitate a variabilei independente x, variabila dependentă y creşte cu valoarea reprezentată de coeficientul b. Dacă valoarea coeficientului de regresie este 0, înseamnă că nu există nicio relaţie între cele două variabile.

Întrucât regresia liniară este strâns legată de corelaţia dintre variabilele analizate, panta de regresie se calculează dupa formula:

$$\hat{b} = \frac{\sum_{i=1}^{n}((x_i - \bar{x}) * (y_i - \bar{y}))}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r_{x,y} * \frac{\sigma_y}{\sigma_x}$$

În realitate variabila y nu depinde de o singura variabilă, ci de un ansamblu de variabile, exprimate în cadrul modelului prin variabila eroare sau reziduu ε, care însumează influenţa altor variabile, pe care modelul regresiei liniare nu le ia în calcul, asupra variabilei dependente y. Eroarea exprimă abaterile între valorile observate şi valorile estimate prin aplicarea modelului.

Pentru estimarea celor doi parametri, constanta şi panta de regresie, s-a utilizat metoda celor mai mici pătrate [19] [17] [20], care minimizează patratele abaterilor dintre valorile date (datele observate) – y şi cele calculate - $\hat{y}$, şi care presupune parcurgerea mai multor etape:

- calcularea tuturor erorilor (diferenţelor) între datele de antrenare (datele observate) şi valorile rezultate (estimate) în urma aplicării modelului, care sunt reprezentate liniar

- valoarea fiecarei erori se ridică la pătrat şi se calculează suma
- se găseşte linia unde suma pătratelor erorilor sau reziduurilor ia valoarea cea mai mică

$$\text{SSE} = \sum (y_i - \hat{a} - \hat{\text{b}}\text{x}_\text{i})^2 = \sum (y_i - \hat{y}_i)^2 = \text{minim}$$

Cu cât coeficientul de corelaţie între variabilele analizate este mai mare sau mai apropiat de valoarea absolută 1, cu atat valoarea erorii de estimare - ε va fi mai mică.

3.2.1. Ipoteze statistice asupra modelului regresiei liniare

Modelul regresiei liniare prin metoda celor mai mici pătrate trebuie să respecte anumite ipoteze de lucru cu privire la proprietăţile estimatorilor. Prin validarea sau invalidarea ipotezelor, se poate determina dacă modelul liniar este adecvat pentru explicarea fenomenelor studiate şi se poate evalua precizia estimatorilor.

Liniaritatea modelului – între variabila dependentă şi variabila independentă există o relaţie liniară.

- este esenţială pentru validarea coeficienţilor estimaţi ai ecuaţiei de regresie
- verificarea liniarităţii se poate face grafic prin intermediul scatterplots, respectiv dacă norul de puncte care arată relaţia dintre variabila predictor şi cea estimată este dispus în model liniar, atunci există o relaţie liniară între cele două variabile
- dacă relaţia între variabile nu este liniară, atunci se pot face transformări de liniarizare asupra variabilelor (modelul exponenţial, modelul de tip putere) [20]
- în cazul în care relaţia dintre variabile nu se poate liniariza, se folosesc modele de regresie neliniare, de exemplu modelele polinomiale, care au la baza functii neliniare

Inexistenţa multicoliniarităţii – variabilele predictor din model nu sunt corelate între ele (testarea se face doar pentru regresia multiplă)

- pentru identificarea existenţei corelaţiilor între predictori/variabilele independente se afişează matricea de corelaţie, calculată pe baza coeficientului de corelaţie Pearson; dacă între două variabile predictor există o corelaţie mare, atunci una dintre ele trebuie eliminată din model; se păstrează variabila care are o corelaţie mai mare cu variabila dependentă y [21]

Normalitatea erorilor – variabila ε urmează o lege normală de medie 0 şi variaţie σ^2: $\varepsilon_i \sim N(0, \sigma^2)$; implică şi verificarea ipotezei că ***media erorilor este 0***: $\overline{\varepsilon} = 0$

- dacă ipoteza e confirmată, atunci înseamnă că şi estimatorii parametrilor modelului de regresie urmează o lege normală
- daca ipoteza e încălcată, înseamnă că estimatorii parametrilor nu sunt eficienţi
- testarea ipotezei de normalitate a erorilor se poate face grafic (histograma/diagrama reziduurilor) sau prin procedee numerice (testul Kolmogorov-Smirnov, testul Jarque-Bera); prin intermediul diagramei reziduurilor se poate observa foarte uşor dacă erorile au o distribuţie normală [20] [22]

Homoscendasticitatea (omogenitatea varianţei erorilor) – varianţele erorilor sunt constante oricare sunt valorile variabilei predictor x: $v(\varepsilon_i) = \sigma^2$;

- ipoteza se poate confirma prin metode grafice sau non-grafice
- ***metode grafice***: prin reprezentarea grafică (scatterplot) a relaţiei dintre reziduuri şi variabila x, sau reziduri şi variabila estimată $\hat{y}$, în cazul regresiei multiple, se poate observa dacă norul de puncte se află dispus normal în jurul liniei de regresie, fără a forma un anumit model; dacă se identifică un model definit (liniar sau pătratic sau în formă de pâlnie), atunci avem heteroscedasticitate
- ***metode non-grafice***: testul Park, testul Glejser, testul White, testul Goldfield-Quandt, testul t Student pentru coeficientul de corelaţie neparametrică Spearman [20] [22]

Necorelarea erorilor – erorile sunt necorelate între ele: $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$; erorile asociate unor valori ale variabilei y nu sunt influenţate de erorile asociate altor valori ale variabilei y

- este aplicabilă în special în ***analiza time-series***, între valorile aceleiaşi variabile observată la diferite momente în timp; cea mai simplă metodă de verificare a ipotezei este prin intermediul unui scatterplot a valorilor reziduale vs. evoluţia în timp a variabilei observate (mai întâi se aranjează datele în ordine temporală) [20]
- în analiza de regresie autocorelarea erorilor intervine dacă modelul e incorect specificat, respectiv dacă se încearcă modelarea unei relaţii liniare asupra unor date neliniare, atunci reziduurile vor fi autocorelate
- pentru testarea ipotezei se pot utiliza metode grafice (scatterplot) sau non-grafice: testul Durbin Watson, Runs test, testul Ljungbox [20] [23]