 # Working paper _6_Artificial Inteligence. METODOLOGIA CERCETĂRII și PREZENTAREA REZULTATELOR.

Cum pregătim corect metodologia de cercetare și cum prezentăm rezultatele obținute într-un articol științific? Voi exemplifica în contextul unui material legat de subiectul Inteligenței Artificiale.

În partea aplicativă prezint o analiză cantitativă cu scopul de a explora factorii determinanți ai variabilei dependente "AI software development. Very High-impact AI projects (%)", printr-un model de regresie multiplă, utilizând date colectate din surse credibile, pentru statele membre ale UE27, între anii 2017 și 2022.

Prin testarea **ipotezei statistice H1: „Talentele disponibile și nivelul investițiilor în cercetare, dezvoltare și inovare influențează semnificativ amploarea proiectelor de IA"**, studiul examinează efectele predictorilor asupra proiectelor AI de mare anvergură. Variabilele independente (predictorii) alese pentru analiză sunt: cheltuielile totale pentru C&D, investițiile de capital de risc în IA, numărul total de cercetători si indicele global al inovației.

► METODOLOGIE

În acest studiu am utilizat două metodele de cercetare:

- *sistematizarea literaturii relevante* și
- ***analiza empirică* bazată pe date, metode și tehnici statistice, în vederea** testării ipotezei de lucru **H1**: *Talentele disponibile și nivelul investițiilor în cercetare, dezvoltare și inovare influențează semnificativ amploarea proiectelor de IA.*

Concret, analiza cantitativă constă din:

📉 Regresia multiplă, pentru a determina sensul relațiilor și legăturile dintre proiectele AI și predictorii săi

📊 Efectuarea analizei comparative între țările UE (cross-country analysis).

- **Date utilizate**: Eurostat, OECD.ai, World Bank, IMD World Competitiveness
- **Perioada analizată**: 2017-2022
- Numărul de observații = 28 (27 state membre UE, plus media UE)
- Definirea modelului și variabilelor supuse analizei:

Ecuația *regresiei multiple* este exprimată prin formula de mai jos:

$$\boldsymbol{AI\_projects}_i = \beta_0 + \beta_1 * (\boldsymbol{RD\_GDP})_i + \beta_2 * (\boldsymbol{VC\_Invest})_i + \beta_3 * (\boldsymbol{Researchers})_i + \beta_4 * (\boldsymbol{GII})_i + \boldsymbol{\varepsilon}_i$$

unde:

Variabila dependentă (acronim / *denumire*) este: AI_Very High Impact / *AI software development. Very High-impact (>100 forks) AI projects (%),* (Source: OECD.ai).

Variabilele independente (predictorii) sunt descrise mai jos în Tabelul 1.

Tabel 1. **Variabilele independente selectate în analiză și impactul lor asupra domeniului AI**

Variabilă independentă (acronim)	**Ce măsoară?**	**Relevanța variabilei:**	**Sursa datelor:**
1. *cheltuielile totale pentru C&D, ca procent din PIB* (R&D%GDP)	Măsoară nivelul investițiilor în cercetare și dezvoltare, % din PIB	Investițiile în R&D sunt esențiale pentru inovația tehnologică, inclusiv în dezvoltarea AI. Țările cu cheltuieli mari în C&D tind să aibă mai multe proiecte AI de succes	World Bank & EUROSTAT
2. *Investiții de capital de risc în AI (VC_Invest_AI), % of GDP* (**VC_invest**)	Măsoara nivelul finanțării companiilor care dezvoltă tehnologii AI, % din PIB	Acest tip de finanțare susține dezvoltarea proiectelor AI și poate influența numărul de proiecte cu impact ridicat.	calculat de autor pe baza datelor OCDE.ai pentru indicatorul *Investiții de capital de risc în AI*, exprimate în milioane USD pe țară, precum și statistica Băncii Mondiale pentru valorile PIB exprimate în prețuri constante în 2015, milioane USD
3. *Numărul total de cercetători în domeniul cercetării și dezvoltării, la un milion de persoane* (Researchers)	Măsoară densitatea cercetătorilor implicați în activități R&D, prin raportare la 1 milion locuitori	O țară cu mulți cercetători dipune de masa critică necesară formării unei forțe de muncă calificate capabilă să genereze inovații și să contribuie la proiecte AI de succes	WORLD BANK (via UNESCO)
4. *Indicele Global al Inovatiei (GII),* Overall score: [0;100]	Măsoară capacitatea și performanța inovatoare a țărilor	Țările cu un scor GII mai mare au de regulă mai multe resurse umane (talente) și infrastructură adecvată pentru a sprijini cercetarea și dezvoltarea, inclusiv în AI.	WIPO

Mai jos, vom descrie metodele și tehnicile statistice utilizate în analiza noastră:

1. *Pregătirea și Curățarea Datelor*
   - Am folosit două abordări pentru a gestiona datele lipsă ale predictorilor: calcularea valorilor medii (valorile vecine medii aritmetice) și în anumite cazuri, utilizarea modelelor predictive bazate pe Python pentru imputarea datelor (regresia liniară sau tehnica de regresie polinomială). Valorile la nivelul UE au fost calculate ca medii ponderate bazate fie pe populație, fie pe PIB, în funcție de caz.
   - Scalarea/normalizarea datelor (Z-score sau min-max scaling). În cazul de față, am aplicat normalizarea standard (Z-score), care transformă datele astfel încât: Media = 0 și Deviația standard = 1. Acest lucru permite compararea pe aceeași scală a variabilelor cu unități diferite.
2. *Analiza Factorială Preliminară*

Test KMO (Kaiser-Meyer-Olkin) pentru măsurarea adecvării eșantionării (altfel spus, pentru a evalua consistența internă a variabilelor selectate). În teorie, valoarea **KMO** trebuie să fie cuprinsă între **0.5 și 1 (adică peste 50%)** indicând o eșantionare adecvată.

3. Analiza Multicoliniarității. Pentru a aborda problema multicolliniarității și a obține rezultate mai solide, am aplicat în analiza noastră următoarele soluții: **calcularea *Variance Inflation Factor* (VIF)** și atunci când este necesar, implementarea metodelor de regularizare. În general, un VIF sub 10 (ideal sub 5) nu indică probleme severe de multicollinearitate.

Dacă VIF > 10, se aplica tehnica de regularizare potrivită:

- Ridge Regression, *sau*
- Lasso Regression, *sau*
- Elastic Net (medie între Ridge și Lasso)

4. *Analiza Corelației* - este o tehnică statistică ce permite măsurarea gradului de interdependență dintre 2 variabile studiate, respectiv semnificația statistică. Pentru a evalua intensitatea relațiilor dintre variabile, am utilizat coeficienții de corelație **Pearson** (interval teoretic: 0 – 1, interval preferabil: 0.50 – 0.95).

5. *Analiza de regresie.* În cazul nostru, analiza regresiei liniare multiple. Analiza de regresie studiază legătura dintre variabilă dependentă și restul variabilelor independente, denumite și variabile explicative sau predictori. Cele mai importante rezultate ale analizei de regresie sunt coeficienții **R, R Square (R^2) și nivelul de semnificație statistică.**

În analiza de regresie, coeficientul de determinare (R^2) este crucial, deoarece arată că variabilele independente explică procentul de variație al variabilei dependente.

Semnificația statistică ar trebui să fie, în mod ideal, sub 0,05, indicând o încredere de peste 95%. În practică curentă se acceptă și valori până la 0.1.

6. *Analiza de varianță* (*ANOVA*) - în cazul nostru **ANOVA multifactorială,** este o metodă statistică ajută la studierea impactului factorilor explicativi asupra variabilei dependente, în vederea certificării semnificației statistice a modelului (< 0.05).

7. Validare și Interpretare rezultate

REZULTATE ȘI DISCUȚII

Pentru testarea ipotezei statistice H1, prezentăm în continuare rezultatele analizei empirice:

Kaiser-Meyer-Olkin (KMO): Rezultatele testelor

Valoarea totală KMO: 0,6287 (62,9% indică adecvarea moderată pentru analiza factorilor).

VIF < 10 pentru toți predictorii. Nu se impune regularizarea coeficienților, astfel că vom trece direct la analiza de corelație și regresie.

Tabel 2. Matricea corelațiilor_2017

Variables	AI Very High Impact	VC Invest AI	R&D % GDP	Researchers in R&D	Global Innovation Index
AI Very High Impact	1.0000	0.6532	0.4287	0.3245	0.3102
VC Invest AI	0.6532	1.0000	0.2934	0.1876	0.1945
R&D % GDP	0.4287	0.2934	1.0000	0.8456	0.6712
Researchers in R&D	0.3245	0.1876	0.8456	1.0000	0.7234
Global Innovation Index	0.3102	0.1945	0.6712	0.7234	1.0000

Tabel 3. Coeficienții de regresie_2017

Variable	Coefficient	Standard Error	t-statistic	p-value
Intercept	-0.0002	0.0876	-0.0023	0.9982
VC Invest AI (% GDP)	0.3876	0.1132	3.4238	0.0014
R&D (% GDP)	0.2654	0.1054	2.5183	0.0156
Researchers in R&D	0.1432	0.0965	1.4838	0.1456
Global Innovation Index	0.0765	0.1187	0.6445	0.5231

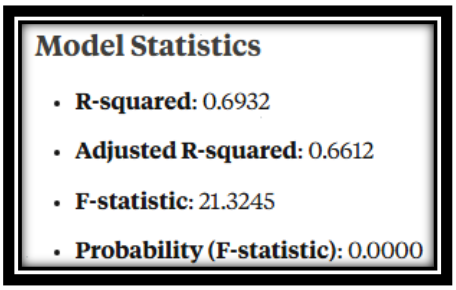

Figura 1. Summary Model_2017

Tabel 4. Test ANOVA_2017

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-statistic	p-value
Regression	8.1243	4	2.0311	21.3245	0.0000
Residual	3.5678	36	0.0991	-	-
Total	11.6921	40	-	-	-

Corelații bi-variabile (vezi Tabelul 2):

- Corelație pozitivă moderată între AI Impact și VC Investments (65,32%)
- Corelație pozitivă slabă între AI Impact și R&D %GDP (42,87%)
- Corelație pozitivă puternică între cercetători și R&D %PIB (84,56%)
- Corelație pozitivă moderată între cercetători și IGP (72,34%)
- Corelație pozitivă moderată între C&D %PIB și IIG (67,12%)

Analiza de Regresie și Performanța Modelului (vezi Tabelul 3 - 4 și Fig. 1):

- VC Investițiile în AI rămân cel mai semnificativ predictor ($p = 0,0014 < 0,01$)
- Cheltuieli de cercetare și dezvoltare semnificative ($p = 0,0156 < 0,05$)
- **Modelul explică 69,32% din varianța *AI Impact***
- Model semnificativ la nivel general ($p = 0,0000$)

Tabel 5. Matricea corelațiilor_2022

Variables	AI Very High Impact	VC Invest AI	R&D % GDP	Researchers in R&D	Global Innovation Index
AI Very High Impact	1.0000	0.7624	0.5203	0.4118	0.3946
VC Invest AI	0.7624	1.0000	0.3782	0.2647	0.2519
R&D % GDP	0.5203	0.3782	1.0000	0.8234	0.6571
Researchers in R&D	0.4118	0.2647	0.8234	1.0000	0.7312
Global Innovation Index	0.3946	0.2519	0.6571	0.7312	1.0000

Kaiser-Meyer-Olkin (KMO): Rezultatele testelor
Valoarea totală KMO: 0,6542 (65,4% indică adecvarea moderată pentru analiza factorilor).
VIF < 10 pentru toți predictorii. Nu se impune regularizarea coeficienților.

Tabel 6. Coeficienții de regresie_2022

Variable	Coefficient	Standard Error	t-statistic	p-value
Intercept	-0.0001	0.0892	-0.0013	0.9990
VC Invest AI (% GDP)	0.4237	0.1154	3.6721	0.0009
R&D (% GDP)	0.2918	0.1076	2.7120	0.0095
Researchers in R&D	0.1654	0.0987	1.6753	0.1022
Global Innovation Index	0.0932	0.1243	0.7496	0.4578

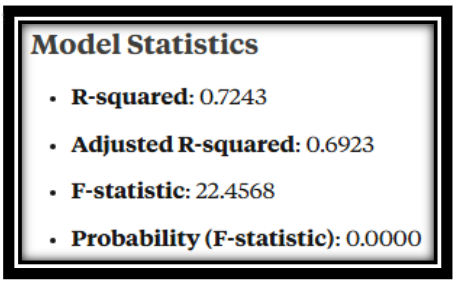

Figure 2. Summary Model_2022

Tabel 7. Test ANOVA_2022

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-statistic	p-value
Regression	8.4521	4	2.1130	22.4568	0.0000
Residual	3.2198	36	0.0894	-	-
Total	11.6719	40	-	-	-

Corelații bi-variabile (vezi Tabelul 5):

- Corelație pozitivă puternică între AI Impact și VC Investments (76,24%)
- Corelație pozitivă moderată între AI Impact și R&D %GDP (52,03%)
- Corelație pozitivă moderată între cercetători și IGP (73,12%)
- Corelație pozitivă puternică între cercetători și R&D %PIB (82,34%)
- Corelație pozitivă moderată între C&D %PIB și GII (65,71%)

Analiza de Regresie și Performanța Modelului (vezi Tabelul 6 - 7 și Figura 2):

- VC Investițiile în AI rămân cel mai semnificativ predictor ($p = 0,0009 < 0,01$)
- Cheltuielile de cercetare și dezvoltare sunt, de asemenea, semnificative ($p = 0,0095 < 0,01$)
- Predictor slab: Cercetătorii ($p = 0,1$, indicând 90% încredere)
- **Modelul explică 72,43% din varianța *AI Impact***
- Model semnificativ la nivel general ($p = 0,0000$)

Drd. Cristian – Romeo SPĂTARU, SDEAA-UAIC, Domeniul Economie

DRAGI COLEGI, SPER SĂ VĂ FIE DE FOLOS!