# Salariu Data Science

**Ladaniuc Ion, IA-211, III**

*Universitatea Tehnică a Moldovei, Facultatea Calculatoare, Inginerie și Micro-Electronică, Chișinău, Moldova, Informatica aplicată.*

# Abstra­ct

Domeniul științei datelor este în plină expansiune, iar cererea de profesioniști calificați este în creștere. În acest context, este important să înțelegem factorii care influențează salariile în acest domeniu. Acest articol are ca scop să investigheze relația dintre salarii și factori precum anul de muncă, nivelul de experiență, tipul de angajare și titlul postului. Am analizat un set de date extins, cuprinzând informații despre salariile din domeniul științei datelor în perioada 2020-2023. Am efectuat o analiză exploratorie a datelor pentru a identifica tendințele și variațiile, iar apoi am utilizat modele de regresie liniară pentru a evalua influența factorilor de interes asupra salariilor. Rezultatele noastre arată că anul de muncă, nivelul de experiență și tipul de angajare au o influență semnificativă asupra salariilor. În mod specific, salariile cresc odată cu creșterea experienței și a numărului de ani de muncă. De asemenea, angajații care lucrează în companii mai mari și în roluri de conducere câștigă, în general, mai mult decât cei care lucrează în companii mai mici și în roluri mai puțin importante. Rezultatele noastre oferă o imagine de ansamblu asupra factorilor care influențează salariile în domeniul științei datelor. Aceste informații pot fi utile pentru profesioniștii din domeniu, angajatori și cercetători de piață.

Introducere   
 În era actuală a informației și tehnologiei, domeniul științei datelor a devenit un fenomen esențial în evoluția societății noastre. Oricare organizație sau industrie care își dorește să navigheze cu succes în peisajul complex al datelor trebuie să înțeleagă nu doar aspectele tehnice ale științei datelor, ci și implicațiile sale în ceea ce privește resursele umane. Prin analizarea detaliată a unui set de date cuprinzător privind salariile din domeniul științei datelor, ne propunem să dezvăluim nuanțe semnificative despre această industrie în continuă expansiune.

Sectorul științei datelor nu mai reprezintă doar o nișă specializată, ci a devenit o sursă de energie în economie și dezvoltare. De la experți în juniori până la lideri experimentați, analiza datelor salariale din acest domeniu oferă o perspectivă valoroasă asupra modului în care experiența, tipul de angajare și rolul ocupat influențează compensațiile financiare.

Această analiză nu reprezintă doar un exercițiu matematic sau statistic, este o incursiune în profunzimea unui domeniu care conectează talentul uman cu provocările complexe ale analizei datelor. În ciuda imposibilității de a face previziuni precise pe viitor, vom examina schimbările emergente și cerințele pieței pentru a obține o perspectivă mai completă asupra evoluției acestei industrii

Scopul principal al acestui articol este să investigheze și să ofere previziuni cu privire la date, concentrându-se în special pe variabilele precum anul de muncă, nivelul de experiență, tipul de angajare și titlul postului în relație cu salariul. Pornind de la aceste premise, vom explora conexiuni semnificative și tendințe care pot oferi perspective relevante pentru profesioniștii din domeniul științei datelor, angajatori și cercetători de piață. Setul de date utilizat, cu cele 3,756 de rânduri de date, acoperă o plajă vastă de informații și ne oferă oportunitatea de a extrage înțelesuri valoroase despre evoluția pieței forței de muncă în știința datelor în ultimii ani.

# Materiale & Metode

**Setul de date**

Pentru analiza privind salariile în domeniul Științei Datelor, am utilizat un set de date extins, cuprinzând informații relevante despre remunerațiile din acest domeniu în perioada 2020-2023. Datele au fost colectate din surse diverse și au fost centralizate pentru a oferi o imagine detaliată a evoluției salariilor în acest interval de timp.

Variabile Incluse în Setul de Date – mai multe detalii vedeți Anexa1.

Acest set de date extins oferă o sursă bogată de informații pentru a înțelege dinamica salariilor în domeniul Științei Datelor și poate servi drept instrument esențial pentru profesioniștii din domeniu, cercetătorii de piață și decidenții din industrie în luarea unor decizii informate în ceea ce privește politica de salarizare și strategiile de resurse umane.

**Preprocesarea Datelor**

Pentru pregătirea datelor în vederea analizei, am utilizat limbajul de programare R, împreună cu bibliotecile relevante, inclusiv tidyverse. Am încărcat setul de date și am efectuat o analiză preliminară pentru a înțelege structura și natura datelor.

Am constatat că nu există valori lipsă (N/A) în setul de date, facilitând analiza. Tipurile de date principale sunt int (numeric) și chr (caracter), iar acest lucru a fost luat în considerare în procesul de prelucrare.

Eliminarea datelor redundante: Am eliminat coloanele care nu au fost considerate relevante pentru analiză, cum ar fi: salary, salary\_currency. Aceste coloane au fost eliminate deoarece am folosit coloana salary\_in\_usd care reprezintă salariu în dollari și respectiv această coloană înlocuiește aceste 2 coloane.

Transformarea datelor: Am aplicat transformări specifice pentru a îmbunătăți potențialul de analiză. De exemplu, am transformat unele date din format „char” în format “int” prin metoda “as.integer”, asigurând astfel o bună potrivire a unui model de regresie logică.

**Analiza Exploratorie a Datelor**

Am efectuat o analiză exploratorie a setului de date privind salariile din domeniul științei datelor, investigând relațiile dintre variabile. Am examinat distribuția salariilor în funcție de anul de lucru, nivelul de experiență și tipul de angajare, utilizând ggplot-uri pentru a ilustra tendințele și variațiile. De asemenea, am furnizat informații statistice relevante pentru fiecare variabilă prin intermediul funcției summary. Această analiză a oferit o perspectivă asupra caracteristicilor cheie ale setului de date, inclusiv medianele și variabilitatea salariilor în USD, distribuția angajaților în funcție de nivelul de experiență și proporția de angajați care lucrează de la distanță în raport cu companiile de diferite dimensiuni și locații.

**Modelarea și Evaluarea Performanței**

În această etapă esențială a analizei noastre asupra salariilor în domeniul științei datelor, ne-am concentrat pe modelarea și evaluarea performanței pentru a dezvălui relațiile complexe dintre variabilele-cheie. Scopul nostru a fost să identificăm influențele diferitelor aspecte, cum ar fi anul de muncă, nivelul de experiență, tipul de angajare și titlul postului, asupra salariilor.

În prima fază, am divizat setul de date extins în două segmente distincte: unul pentru antrenament și unul pentru testare. Această abordare ne-a permis să evaluăm performanța modelelor pe date independente, contribuind la validarea robustă a rezultatelor noastre. Selectând cu atenție variabilele relevante, precum anul de muncă, nivelul de experiență și altele menționate, am configurat un cadru analitic coerent.

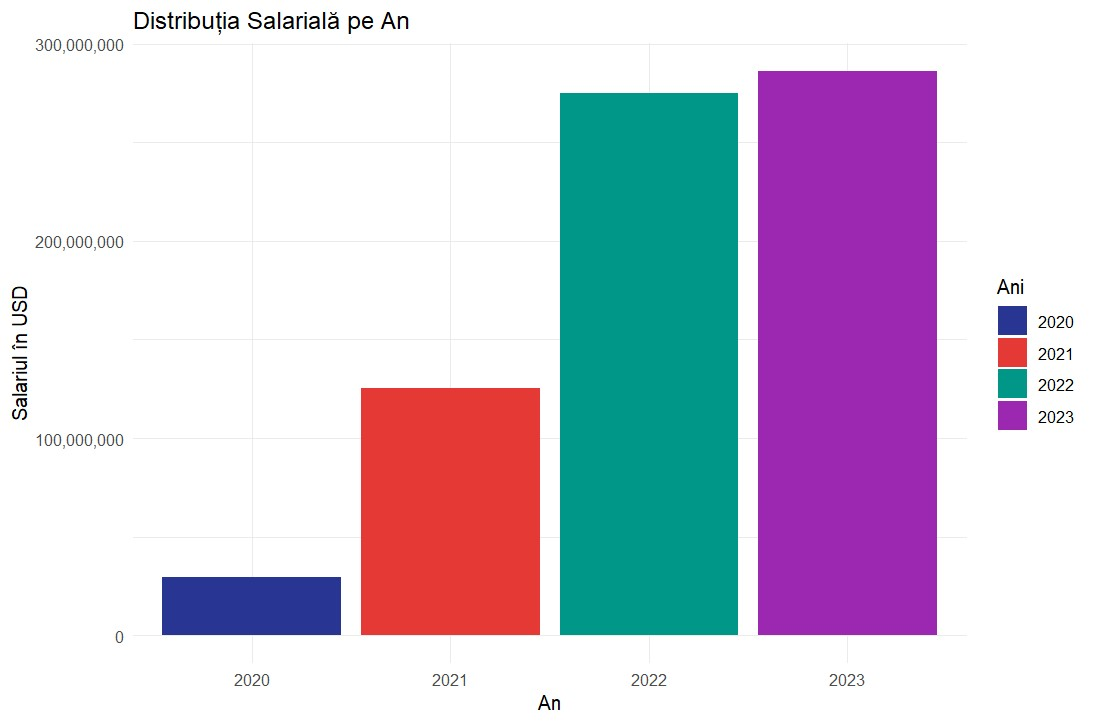
O parte foarte importantă a procesului de modelare a fost gestionarea variabilelor complexe, cum ar fi nivelul de experiență de angajare și locația angajatului. Pentru a asigura acuratețea și interpretabilitatea rezultatelor, am aplicat transformări specifice asupra datelor, adaptându-ne la particularitățile setului nostru de date. De exemplu, am explorat impactul variabilei "salary\_in\_usd" și am ajustat parametrii pentru a obține o potrivire optimă cu datele disponibile.

**Codul disponibil**

Întregul cod pentru executarea operațiilor, inclusiv crearea modelului de regresie liniară, este disponibil pe linkul: <https://github.com/11Ion/AnalizaDatelor>

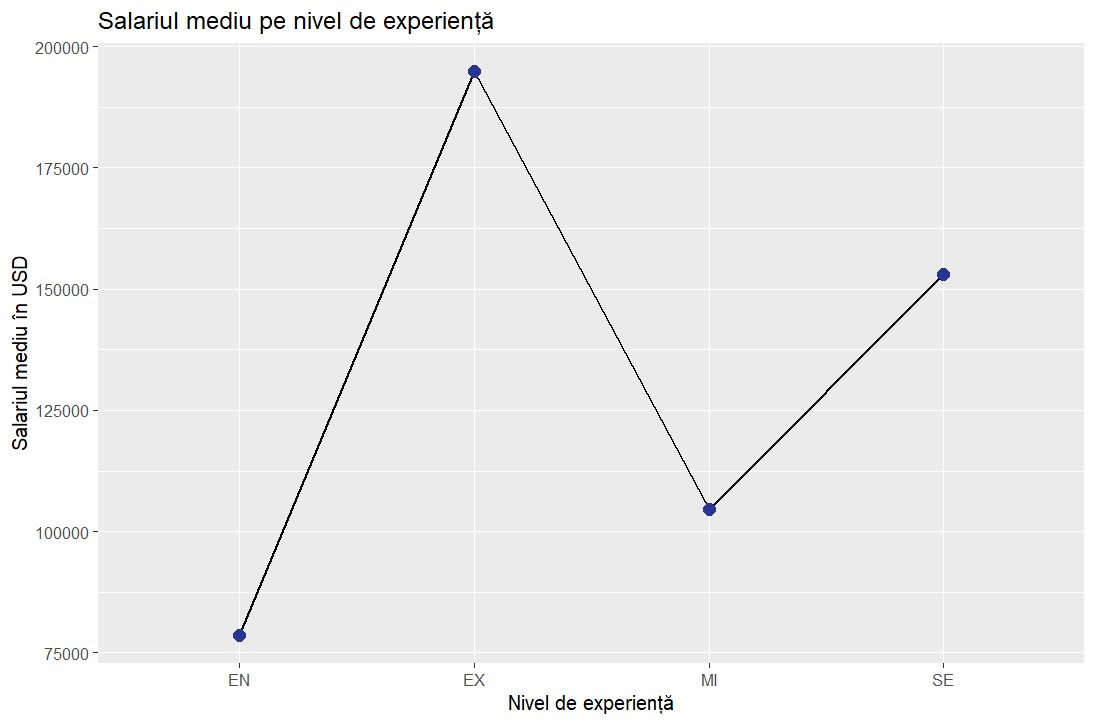
# Rezultate

După preprocesarea datelor referitoare la salariile din domeniul științei datelor, am inițiat analiza setului de date rezultat pentru a identifica tendințele și relațiile semnificative. Cu ajutorul ggplot-urilor, am explorat și evidențiat aspecte relevante privind variabilele implicate, precum anul de muncă, nivelul de experiență, tipul de angajare, etc. Această analiză vizuală este esențială pentru a înțelege dacă există corelații notabile între aceste variabile și salariile din domeniul științei datelor, contribuind astfel la obținerea unei perspective mai bogate și mai informate asupra peisajului salarial în această industrie dinamica.



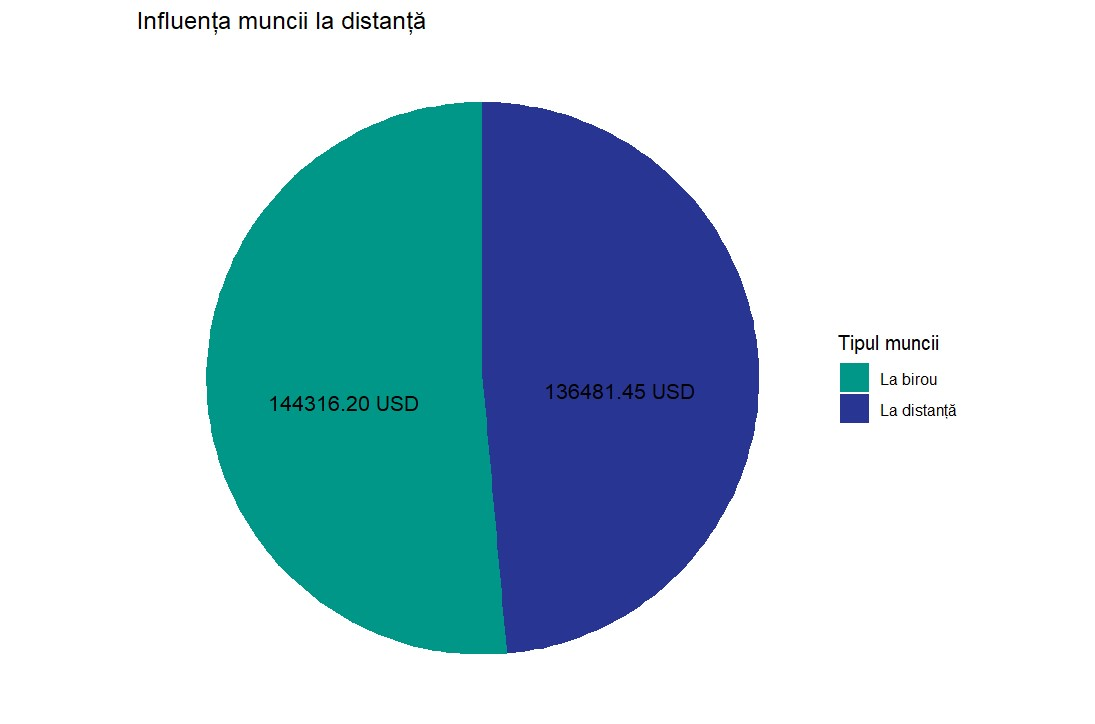
**Figura 1.1** Distribuția Salarială pe ani

Prin analiza graficului, se distinge clar că anii 2022 și 2023 se evidențiază ca având cele mai înalte valori de salariu în comparație cu ceilalți ani din setul de date. Această observație sugerează că în acei ani a avut loc o posibilă creștere semnificativă a salariilor și numărului de specialiști în acest domeniu.



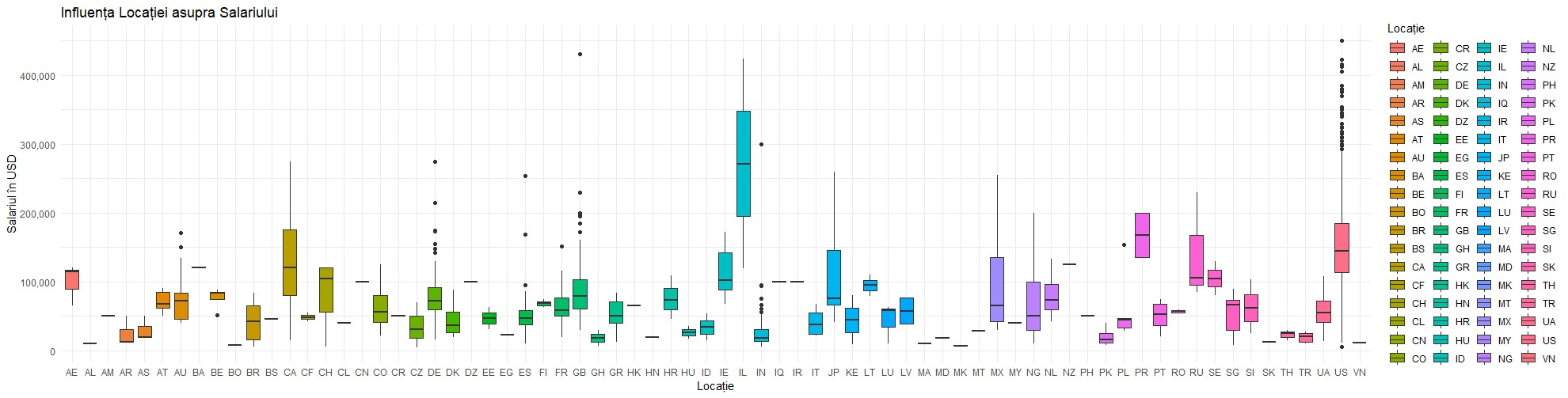
**Figura 1.2** Salariul mediu pe nevel de experință

În figura 1.2, putem observa modul în care salariul mediu variază în funcție de nivelul de experiență al angajaților. Pe axa X, avem diferitele niveluri de experiență, iar pe axa Y, avem salariul mediu corespunzător fiecărui nivel. Linia neagră conectează punctele, oferind o imagine vizuală a modului în care salariul mediu evoluează în funcție de experiența acumulată. Punctele albastre de-a lungul liniei reprezintă valorile individuale ale salariilor medii pentru fiecare nivel de experiență.



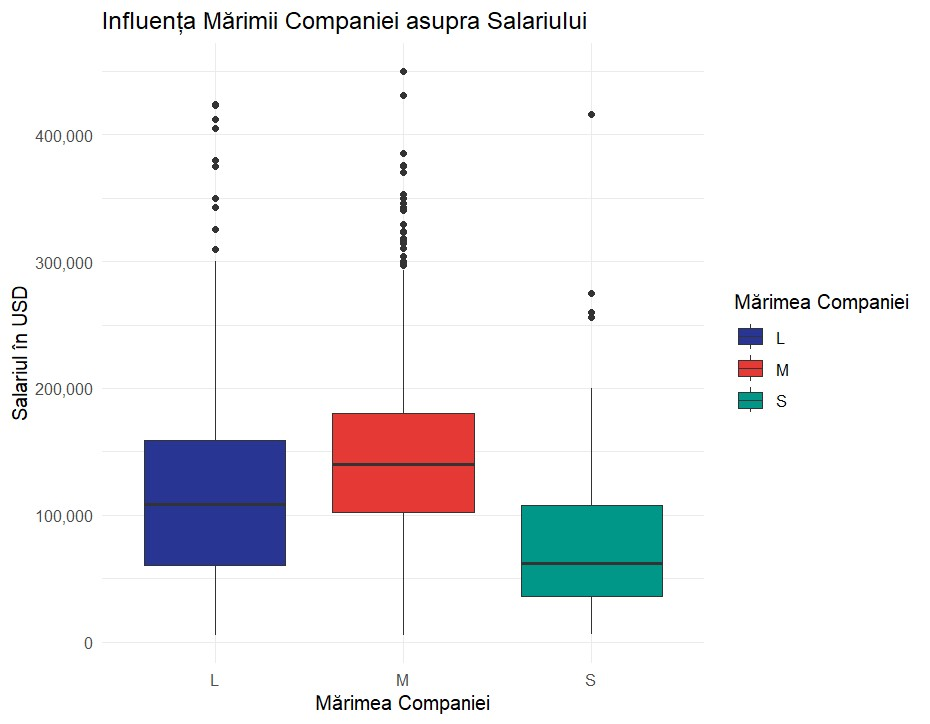
**Figura 1.3** Impactul muncii la distanță

Figura 1.3 prezintă un grafic circular (pie chart) care evidențiază informații cu privire la influența locului de muncă asupra salariilor medii. Graficul sugerează că există o diferență relativ mică între salariile medii pentru muncă "La birou" și "La distanță". Cu toate acestea, permite o comparare rapidă și vizuală între aceste două tipuri de muncă în ceea ce privește remunerația medie.



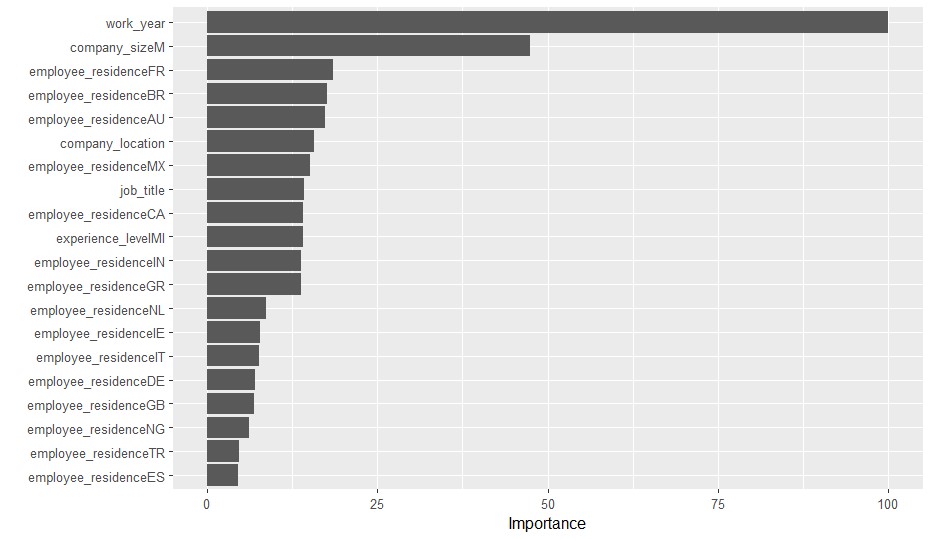
**Figura 1.4** Influența locației asupra salariului

Figura 1.4 este un grafic care reprezintă Influența locației asupra salariului . Din grafic se poate observa că salariul mediu este mai mare în țările dezvoltate, cum ar fi Statele Unite, Franța, Germania etc. În schimb, salariul mediu este mai mic în țările în curs de dezvoltare.



**Figura 1.5** Influența mărimii companiei asupra salariului

Din graficul din figura 1.5 se poate observa că salariul mediu este mai mare în companiile mai mari. De exemplu, salariul mediu în companiile cu peste 500 de angajați este de 100.000 USD, în timp ce salariul mediu în companiile cu mai puțin de 50 de angajați este de 50.000 USD. Există mai multe motive pentru această diferență. În primul rând, companiile mai mari tind să aibă o bază de clienți mai mare și să genereze mai multe venituri. Acest lucru le permite să plătească salarii mai mari angajaților lor.



**Figura 1.6** Imporntanța variabilelor ce influențează munca la distanta

După procesul de logoritmare noi am primit o frecvență bună pentru a putea face un model de regresie logica.

Am dezvoltat un model de regresie logistică utilizând formula remote\_ratio ~ work\_year + company\_size pentru a prezice variabila remote\_ratio în funcție de variabilele explicative work\_year și company\_size. Am ales metoda de modelare 'glm' (generalized linear model) și am implementat validarea cu 10 folduri pentru a evalua performanța modelului.

Procesul de validare evediențiază implicare a împărțirii setului de date în 10 părți egale, utilizând fiecare parte ca set de testare în mod iterativ, în timp ce celelalte 9 părți sunt folosite pentru antrenament. Acest procedeu este repetat de 10 ori, asigurând o evaluare robustă a modelului.

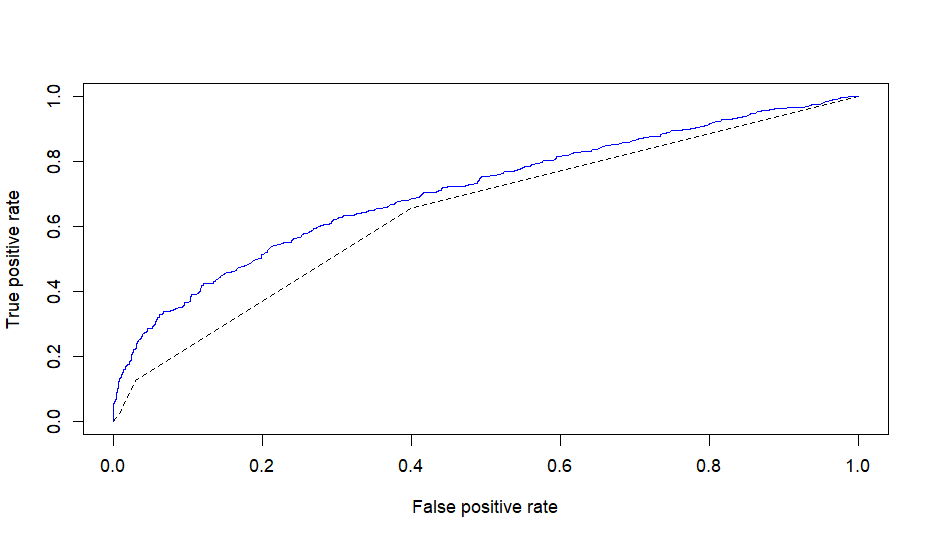
Pentru a evalua corectitudinea modelului, am efectuat interpretarea coeficienților obținuți din modelele de regresie logistică (model1 și model2). Expunerea acestora prin funcțiile exp(coef(model1)) și exp(coef(model2)) ne furnizează rapoartele de șanse pentru fiecare variabilă independentă. Aceste rapoarte indică în ce măsură șansele de a avea remote\_ratio = yes cresc sau scad pentru o unitate de creștere în variabila corespunzătoare.

Am creat o matrice de confuzie pentru a evalua performanța modelelor pe setul de date de antrenament. Această matrice oferă o perspectivă detaliată asupra acurateței modelului în ceea ce privește clasificarea corectă a instanțelor dintre categoriile yes și no ale variabilei de răspuns.

|  |  |  |
| --- | --- | --- |
| Confusion Matrix and Statistics | | |
| Reference | | |
| **Prediction** | **Yes** | **No** |
| **Yes** | 825 | 458 |
| **No** | 457 | 888 |

|  |  |
| --- | --- |
| **Acuracy** | 0.6518 |
| **95%CI** | (0.63333, 0.6701) |
| **No Information Rate** | 0.51222 |
| **P-Value [Acc> NIR]** | <2e-16 |
| **Kappa** | 0.3033 |
| **Mcnemar’s Test P-Value** | 1 |
| **Sensitivity** | 0.6435 |
| **Specificity** | 0.6597 |
| **Pos Pred Value** | 0.6430 |
| **Neg Pred Value** | 0.6602 |
| **Prevalence** | 0.4878 |
| **Detection Rate** | 0.3139 |
| **Detection Prevalence** | 0.4882 |
| **Balanced Accuracy** | 0.6516 |
| **Positive Class** | yes |

Informația dată arată că modelul de clasificare a avut o acuratețe de 65,18%. Acest lucru înseamnă că modelul a clasificat corect 65,18% din instanțe. Acuratețea este o măsură generală a performanței modelului, dar nu oferă o imagine completă a performanței modelului în a clasifica atât instanțe pozitive, cât și negative.



**Figura 1.8** Regresia Logistică

Analiza de regresie logistică arată că variabilele work\_year și company\_size joacă un rol semnificativ în determinarea deciziei de a permite sau nu munca la distanță (remote\_ratio). Aceste două variabile sunt identificate drept factori importanți care influențează rezultatul variabilei de răspuns, evidențiind importanța experienței de lucru și a dimensiunii companiei în luarea deciziilor privind munca la distanță.

# Bibliografie

**Documenție pentru limbajul R** - [https://www.R-project.org/](https://www.r-project.org/). – accesat (10.09.2023)

**Dataset Data Sience Salaries** - <https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023> – accesat (17.09.2023)

**Regresia Logistică în limbajul R** - <https://www.datacamp.com/tutorial/logistic-regression-R> – accesat (12.11.2023)

**Regresia liniară în limbajul R -** <https://www.codecademy.com/learn/learn-linear-regression-in-r/modules/linear-regression-in-r/cheatsheet> – accesat (12.11.2023)

**Tutorial despre ggplot -** <https://ggplot2.tidyverse.org/reference/>– accesat (17.09.2023)

**Materiale Suplimentare**  
ANEXA 1.

1. **work\_year:** Anul de lucru în care s-a înregistrat informația.
2. **experience\_level:** Nivelul de experiență al angajatului
3. **employment\_type:** Tipul de angajare, adică dacă angajatul este angajat cu normă întreagă, cu normă parțială sau pe proiect.
4. **job\_title:** Titlul sau poziția angajatului în companie.
5. **salary:** Salariul angajatului.
6. **salary\_currency:** Valuta în care este exprimat salariul.
7. **salary\_in\_usd:** Salariul convertit în dolari americani (USD)
8. **employee\_residence**: Locul de reședință al angajatului.
9. **remote\_ratio:** Procentul de muncă efectuată la distanță în raport cu munca de la birou.
10. **company\_location:** Locația companiei în care lucrează angajatul.
11. **company\_size:** Mărimea companiei.