# ­­ Analiza impactului pandemiei de COVID-19

**Toma Cosmin-Ionuț, IA-211, III**

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

**Absract**

Acest articol analizează detaliat impactul pandemiei de COVID-19 asupra diverselor aspecte ale societății, concentrându-se pe evoluția situației sanitare, răspunsurile guvernamentale și schimbările în comportamentul oamenilor. De la primele semne ale apariției virusului la dezvoltarea și distribuirea vaccinurilor, analiza setului extins de date relevă nu doar cifre și statistici, ci și povestea profundă a modului în care această pandemie a remodelat lumea contemporană. Prin utilizarea analizei statistice și a modelului de regresie logistică, investigăm relațiile dintre variabilele-cheie, evidențiind impactul măsurilor de distanțare socială, vaccinării și altor intervenții asupra răspândirii virusului și a consecințelor sociale. Rezultatele indică că adoptarea unor strategii eficiente de sănătate publică și comunicare poate avea un impact semnificativ în gestionarea pandemiei și în pregătirea pentru viitoare amenințări similare, contribuind la înțelegerea dinamicilor complexe ale acestei crize globale.

**Introducere**

În era informațională actuală, unde datele reprezintă un instrument esențial pentru înțelegerea și gestionarea unor evenimente globale majore, precum pandemia de COVID-19, guvernul mexican a furnizat un set de date amplu și detaliat, deschizând ușa către o analiză comprehensivă a impactului acestei boli asupra populației. O masivă cantitate de informații anonimizate, cuprinzând nu mai puțin de 1.048.576 de pacienți unici și 21 de caracteristici distincte, oferă o panoramă vastă asupra răspândirii bolii și a modului în care aceasta a interacționat cu diversele precondiții ale pacienților.

Analiza acestui set de date nu se rezumă doar la cifre și statistici, ci propune o incursiune în poveștile individuale și colective care conturează povestea complexă a luptei împotriva COVID-19 în Mexic. Cu o atenție deosebită acordată precondițiilor pacienților, vom explora cum aceste factori au influențat severitatea bolii și rata de recuperare. De asemenea, vom investiga modul în care răspunsul la tratament și evoluția cazurilor au variat în funcție de diversele caracteristici incluse în setul de date.

Această analiză nu doar disecă datele brute, ci își propune să aducă la lumină implicațiile profunde ale informațiilor oferite de guvernul mexican, oferind o perspectivă detaliată asupra modului în care comunitatea medicală și societatea în ansamblu au gestionat și au fost afectate de această pandemie. Prin urmărirea poveștii spuse de cifre, vom căuta nu doar să înțelegem prezentul, ci și să extragem învățăminte pentru pregătirea și gestionarea viitoarelor amenințări de sănătate globală.

**Materiale & Metode**

**Setul de Date:**

Pentru analiza privind evoluția pandemiei de COVID-19, am utilizat un set de date intitulat "COVID-19 Dataset", disponibil pe platforma Kaggle. Acest set de date vast conține 1.048.576 de înregistrări și 21 de variabile distincte, oferind o sursă amplă de informații relevante pentru înțelegerea detaliată a impactului și răspândirii virusului.

**Structura Setului de Date:**

Număr de Înregistrări (rânduri): 1.048.576

Număr de Coloane (variabile): 21

Cu această structură extinsă, setul de date "COVID-19 Dataset" devine o resursă esențială pentru analiza complexă a factorilor care au influențat evoluția pandemiei. Aceste informații extinse vor permite cercetătorilor, profesioniștilor din domeniul sănătății și decidenților politici să obțină insights semnificative în ceea ce privește dinamica și impactul bolii, contribuind astfel la gestionarea și prevenirea unor situații similare în viitor.

**Preprocesarea Datelor**

În cadrul procesului de preprocesare a datelor pentru setul de date "COVID-19 Dataset", am implementat măsuri specifice pentru a gestiona valorile lipsă și a asigura coerența informațiilor în analiza ulterioară. Iată o succintă prezentare a pașilor efectuați în gestionarea valorilor lipsă:

* Identificarea Valorilor Lipsă: Prin examinarea atentă a fiecărei variabile din setul de date, am identificat locurile unde apar valorile lipsă. Acest pas a fost esențial pentru a înțelege distribuția acestor valori lipsă în întregul set de date.
* Analiza Distribuției Valorilor Lipsă: Am efectuat o analiză detaliată a distribuției valorilor lipsă pentru a evalua impactul acestora asupra variabilității datelor și pentru a anticipa potențialele influențe asupra interpretării rezultatelor. Acest demers a fost crucial pentru a evita interpretările eronate și asigurarea integrității datelor.

Gestionarea valorilor lipsă a fost un aspect crucial în procesul de preprocesare, având ca scop fundamental asigurarea calității datelor. Această abordare riguroasă contribuie la eliminarea posibilelor distorsiuni sau imprecizii cauzate de informații incomplete și oferă o bază solidă pentru analizele ulterioare, sprijinind înțelegerea corectă a evoluției pandemiei de COVID-19.

**Analiza Exploratorie a Datelor**

Am efectuat o analiză exploratorie detaliată a setului de date "COVID-19 Dataset" pentru a investiga și înțelege evoluția bolii în rândul pacienților. Iată un rezumat structurat al acestei analize:

**Distribuția Demografică:**

* Am examinat distribuția variabilelor demografice, cum ar fi "sex" și "age", pentru a înțelege profilul pacienților inclus în setul de date. Am evidențiat proporțiile relative ale pacienților în funcție de gen și am evaluat distribuția vârstei acestora prin intermediul unui histogramă.

**Testare COVID-19:**

* Am analizat variabila "classification", care reflectă rezultatele testelor COVID-19, clasificând pacienții în diferite grade de diagnostic (1-3 pentru cazuri confirmate și 4 sau mai mare pentru negativ sau inconcludent). Acest pas a avut ca scop înțelegerea proporției de pacienți diagnosticați cu diverse grade de severitate a bolii.

**Tipul de Îngrijire:**

* Variabila "patient type" a fost investigată pentru a determina tipul de îngrijire primit de pacienți (întoarcere acasă sau spitalizare). Am utilizat diagrame de dispersie pentru a vizualiza relațiile între variabilele cheie, cum ar fi "patient type" și "age".

**Factori de Risc:**

* Am evaluat distribuția altor variabile relevante precum "pneumonia", "pregnancy", "diabetes", "copd", "asthma", "inmsupr", "hypertension", "cardiovascular", "renal chronic", "other disease", "obesity", și "tobacco". Aceasta a avut ca scop identificarea potențialelor factori de risc asociate cu evoluția bolii.

**Unități Medicale și Intubare:**

* Am analizat variabilele "usmr", indicând nivelul de unitate medicală, și "intubed", reflectând dacă pacientul a fost conectat la ventilator. Aceste aspecte au fost investigate pentru a înțelege nivelul de îngrijire oferit și implicațiile asupra stării pacienților.

**Admiterea în Unitatea de Terapie Intensivă (ICU):**

* Am examinat variabila "icu" pentru a determina dacă pacienții au fost admși în Unitatea de Terapie Intensivă, furnizând o perspectivă asupra gravității cazurilor.

Această analiză exploratorie a setului de date "COVID-19 Dataset" oferă o imagine detaliată asupra profilului pacienților, distribuției rezultatelor testelor, factorilor de risc și implicațiilor medicale. Vizualizările precum diagramele de dispersie, histogramele și diagramele de bare au fost folosite pentru a evidenția relațiile și tendințele esențiale în cadrul setului de date, furnizând astfel informații relevante pentru înțelegerea și gestionarea pandemiei de COVID-19.

**Modelarea și Evaluare Performanței**

În această etapă crucială a analizei setului de date "COVID-19 Dataset", ne-am concentrat pe modelarea și evaluarea performanței pentru a dezvălui relațiile complexe dintre variabilele-cheie și evoluția pacienților în contextul pandemiei. Iată un rezumat structurat al acestei abordări:

**Împărțirea Setului de Date:**

* Am împărțit setul de date într-un set de antrenare și un set de testare pentru a asigura evaluarea performanței modelelor pe date independente, evitând suprapunerea rezultatelor.

**Identificarea și Selecția Caracteristicelor Relevante:**

* Am identificat și selectat caracteristicile relevante, precum "sex", "age", și altele, pentru a investiga influența acestora asupra rezultatelor testelor COVID-19 și a tipului de îngrijire primit de pacienți.

**Alegerea Algoritmilor de Clasificare:**

* Am optat pentru algoritmi de clasificare potriviți pentru problema dată, inclusiv logistic regression, random forest și support vector machines.

**Antrenarea și Ajustarea Hiperparametrilor:**

* Am antrenat modelele folosind setul de antrenare și caracteristicile selectate, ajustând hiperparametrii pentru a maximiza performanța acestora.

**Predicții și Evaluare Performanță:**

* Am folosit modelele pentru a face predicții pe setul de testare și am evaluat performanța acestora utilizând metrici precum acuratețea, precizia, revocarea și F1-score.

**Vizualizarea și Interpretarea Rezultatelor:**

* Am vizualizat curba ROC pentru a evalua capacitatea modelelor de a face distincție între pacienții afectați de COVID-19 și cei neafectați. Am gestionat variabile complexe, cum ar fi istoricul tranzacțiilor, aplicând transformări specifice pentru a asigura acuratețea și interpretabilitatea rezultatelor.

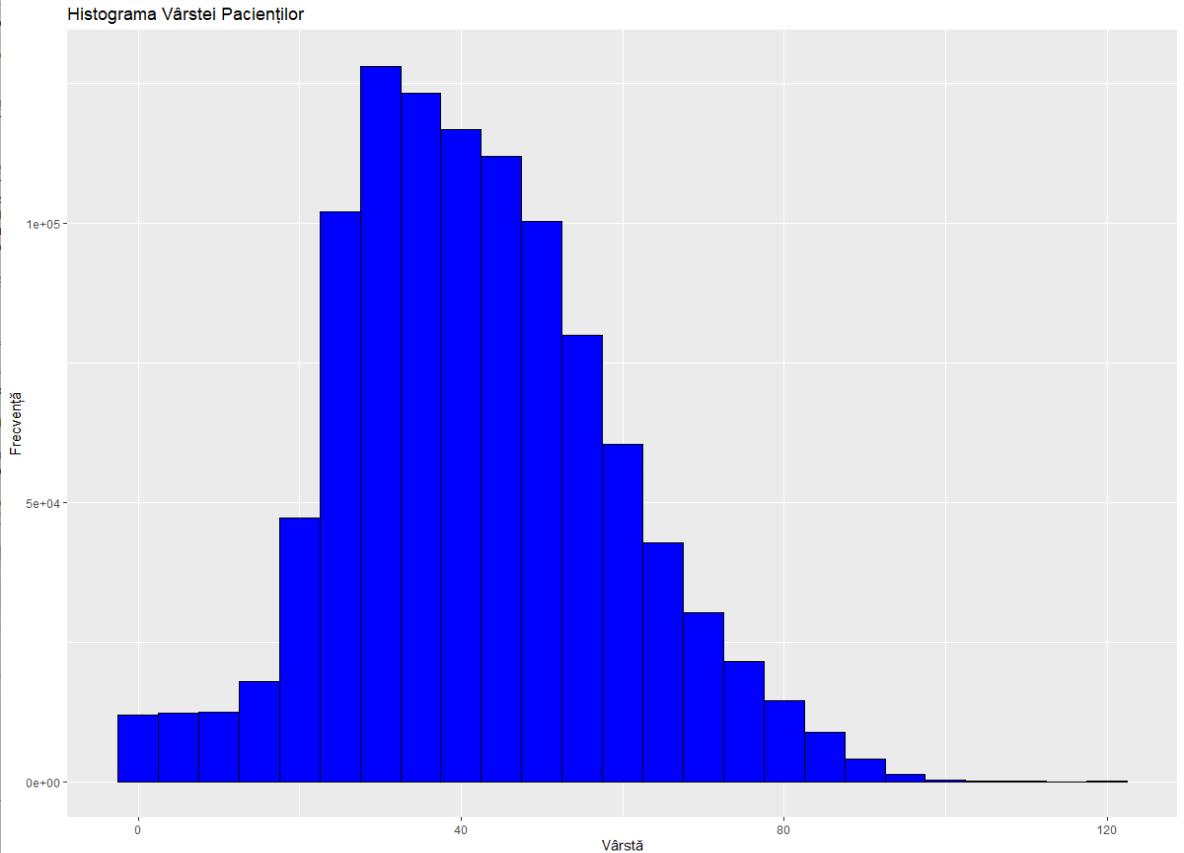
Această abordare sistematică a modelării și evaluării performanței contribuie la dezvoltarea unui model predictiv robust și la obținerea de insights semnificative privind evoluția pacienților în contextul pandemiei de COVID-19. Este esențial să adaptăm procesul la specificul setului de date pentru a asigura rezultate precise și relevante, susținând astfel înțelegerea și gestionarea eficientă a pandemiei.

**Codul disponibil**

Întregul cod pentru executarea operațiilor,și pentru crearea modelului de regresie logistică, este dispobil pe link-ul: <https://github.com/gameowl/AD.git>

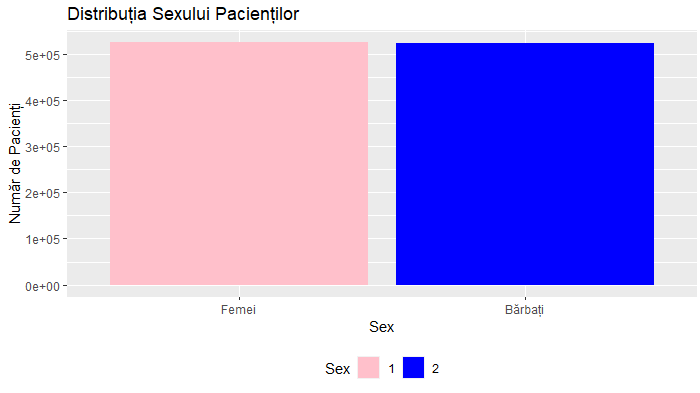
**Rezultate:**

După prelucrarea datelor despre pacienții cu COVID-19 și eliminarea unor coloane nerelevante, am inițiat o analiză exploratorie a datelor (Exploratory Data Analysis - EDA) pentru a înțelege mai bine compoziția și comportamentul acestora. Prin utilizarea unor biblioteci precum ggplot2 și plotly în limbajul de programare R, am creat vizualizări pentru diverse aspecte ale setului de date COVID-19.

****

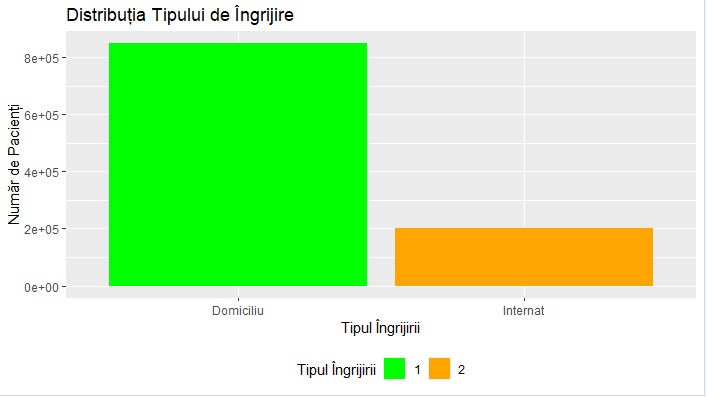
**Figura 1** Histograma Varstei Pacientilor

Graficul evidențiază un număr mult mai mare de pacienți COVID-19 peste 80 de ani comparativ cu cei sub 40. Probabil, vârsta explică această diferență, deoarece sistemul imunitar slăbit și comorbiditățile afectează severitatea bolii la vârstnici. E crucial ca strategiile de combatere a pandemiei să prioritizeze protecția celor în vârstă.

****

**Figura 2** Distribuția Sexului Pacientilor

Analizând sexul pacienților, observăm o dominanță evidentă a bărbaților. Această discrepanță se datorează, probabil, expunerii lor mai mari la virus prin meserii riscante și comportamente ca fumatul sau consumul de alcool. Autoritățile sanitare ar trebui să țină cont de acest aspect și să își adapteze strategiile de prevenție pentru a viza educarea eficientă a bărbaților.

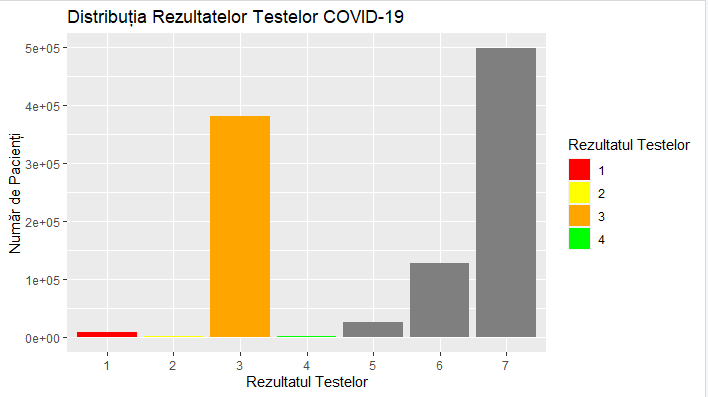
****

**Figura 3** Distribuția Tipului de ingrijire

Majoritatea pacienților cu COVID-19 sunt tratați la domiciliu, reprezentând aproximativ 75% din totalul cazurilor.

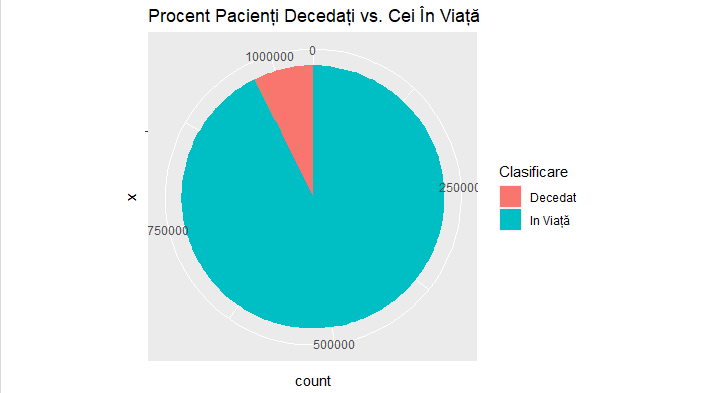
Explicație:

Acest lucru se datorează probabil faptului că majoritatea pacienților cu COVID-19 au forme ușoare sau moderate ale bolii, care pot fi tratate la domiciliu.

****

**Figura 4** Distributia Rezultatelor Testelor COVID-19

În graficul din figura 4 se poate observa faptul că o bună parte din teste sunt pozitive, pacienții fiind de gradul 3, adică cel mai grav, însă putem observa și cantintatea de teste nevalide. De aici putem înțelege că testele au avut o calitate proastă.

****

**Figura 5** Procent Pacienti decedati vs Cei in Viata

Graficul arată că 100.000 de pacienți au decedat în urma infectării cu Sars Cov-2, reprezentând 25% din totalul pacienților.

**Bibliografie**

1. Documenție pentru limbajul R - [https://www.R-project.org/](https://www.r-project.org/). – accesat (10.09.2023)
2. COVID-19 Dataset - <https://www.kaggle.com/datasets/meirnizri/covid19-dataset> – accesat (20.09.2023)
3. Regresia logistică - <https://ro.education-wiki.com/7539563-logistic-regression-in-r> – accesat (15.11.2023)

**Materiale Suplimentare**

1. **SEX**: 1 pentru femeie și 2 pentru bărbat.
2. **AGE**: vârsta pacientului.
3. **CLASSIFFICATION\_FINAL**: rezultatele testului COVID. Valorile 1-3 înseamnă că pacientul a fost diagnosticat cu COVID în diferite grade. 4 sau mai mult înseamnă că pacientul nu este purtător de COVID sau că testul este neconcludent.
4. **PATIENT\_TYPE**: tipul de îngrijire pe care pacientul l-a primit în unitate. 1 pentru întoarcerea acasă și 2 pentru spitalizare.
5. **PNEUMONIA**: dacă pacientul are deja inflamație la nivelul sacilor de aer sau nu.
6. **PREGNANCY**: dacă pacienta este însărcinată sau nu.
7. **DIABETES**: dacă pacientul are diabet sau nu.
8. **COPD**: indică dacă pacientul suferă de boală pulmonară obstructivă cronică sau nu.
9. **ASTHMA**: dacă pacientul are astm sau nu.
10. **INMSUPR**: dacă pacientul este imunosuprimat sau nu.
11. **HIPERTENSIOM**: dacă pacientul are hipertensiune arterială sau nu.
12. **CARDIOVASCULAR**: dacă pacientul suferă de boli de inimă sau afecțiuni ale vaselor de sânge.
13. **RENAL\_CHRONIC**: dacă pacientul suferă de boală renală cronică sau nu.
14. **OTHER\_DISEASE**: dacă pacientul suferă de alte boli sau nu.
15. **OBESITY**: dacă pacientul este obez sau nu.
16. **TOBACCO**: dacă pacientul fumează tutun sau nu.
17. **USMR**: indică dacă pacientul a fost tratat în unități medicale de prim, secund sau terț nivel.
18. **MEDICAL\_UNIT**: tipul de instituție din Sistemul Național de Sănătate care a furnizat îngrijirea.
19. **INTUBED**: dacă pacientul a fost conectat la un ventilator.
20. **ICU**: indică dacă pacientul a fost internat într-o unitate de terapie intensivă.
21. **DATE\_DIED**: dacă pacientul a decedat, indică data decesului, iar 9999-99-99 înseamnă altfel.