**Predicția poluării luminoase folosind imagini satelitare**

Participanți: \_David Sergiu, Dreghiciu Ana-Maria, Galdeanu Nicole

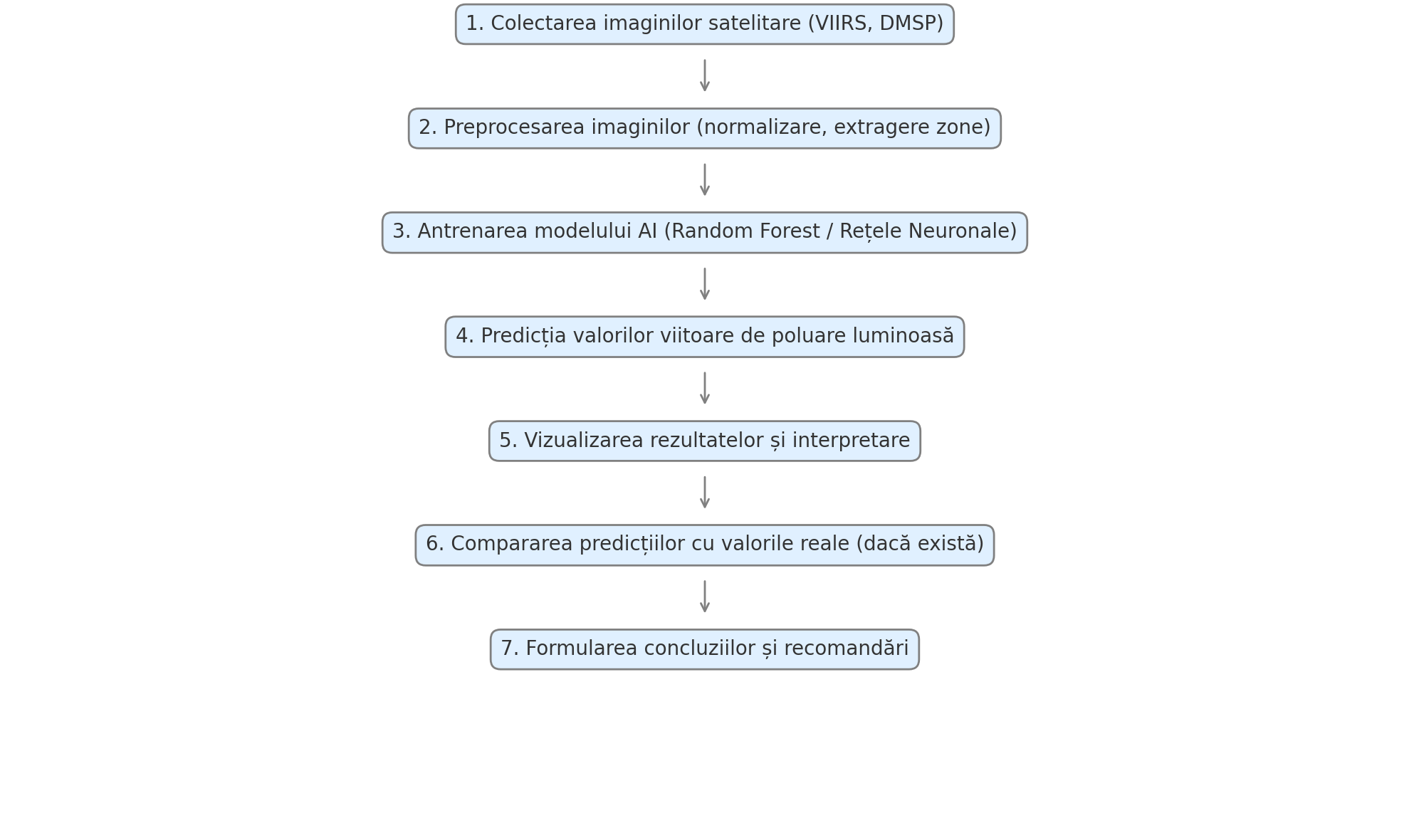
# Abstract

Poluarea luminoasă este o problemă tot mai accentuată în mediul urban modern, cu efecte nocive asupra sănătății umane, vieții sălbatice și capacității de a observa cerul înstelat. Surplusul de lumină artificială contribuie la dezechilibre ecologice și afectează ritmurile circadiene. Analiza evoluției poluării luminoase în timp este crucială pentru înțelegerea impactului său și pentru elaborarea unor politici eficiente de reducere.  
  
Prin utilizarea imaginilor satelitare și a metodelor de inteligență artificială, putem identifica tipare temporale și spațiale ale poluării luminoase. Acest proiect propune o metodă de predicție a nivelurilor viitoare de poluare luminoasă, contribuind astfel la luarea deciziilor informate pentru un mediu mai sustenabil. Estimările privind creșterea sau scăderea poluării în următorii ani pot ajuta autoritățile să adopte măsuri proactive.

# Introducere

Proiectul de față se încadrează în contextul preocupărilor globale privind impactul negativ al poluării luminoase asupra sănătății, mediului și observațiilor astronomice. În ultimii ani, cantitatea de lumină artificială emisă de orașe a crescut considerabil, ceea ce impune nevoia unei monitorizări eficiente.

Inteligența artificială oferă un avantaj semnificativ în acest sens, permițând analiza unor cantități mari de date satelitare într-un timp redus. Prin utilizarea modelelor de învățare automată, se pot identifica tendințe și se pot realiza predicții cu privire la evoluția poluării luminoase.  
  
Mai jos este prezentată o diagramă care reflectă pașii efectuați în aplicație și ideea de bază a fiecărui pas:



# Metode existente de rezolvare a problemei

Pentru a înțelege cum poate fi abordată problema poluării luminoase cu ajutorul inteligenței artificiale, au fost analizate două lucrări de cercetare relevante care propun metode diferite de procesare și modelare a datelor satelitare și atmosferice.

**Articolul 1: *Towards Greener Nights: Exploring AI-Driven Solutions for Light Pollution Management***

În această lucrare, autorii propun un sistem bazat pe învățare automată și procesare a limbajului natural pentru a analiza și clasifica tipurile de poluare luminoasă. Setul de date utilizat este unul complex, conținând peste 214.000 de înregistrări, cu atribute precum ora, fusul orar, latitudinea și longitudinea, altitudinea, tipul senzorului, valorile măsurate, gradul de înnorare și populația aferentă fiecărei locații.

Pentru preprocesarea datelor, cercetătorii au aplicat tehnici de normalizare și standardizare. Clasificarea tipurilor de poluare luminoasă (precum *skyglow*, *glare*, *light trespass* și *clutter*) a fost realizată folosind algoritmi precum k-Nearest Neighbors (KNN), modele de învățare automată precum LightGBM, XGBoost și Random Forest, dar și modele de procesare a limbajului natural (DeBERTa v3 și TF-IDF) pentru extragerea informațiilor relevante din texte descriptive.

Rezultatele au fost promițătoare, modelul LightGBM obținând un scor F1 de 0.8188, ceea ce indică o capacitate bună de clasificare a efectelor poluării luminoase. Lucrarea evidențiază potențialul aplicării inteligenței artificiale în monitorizarea și gestionarea impactului poluării luminoase la nivel global.

**Articolul 2: *Analysis of Light Pollution Prediction Using Mathematical Model and Machine Learning Techniques***

Această cercetare are ca obiectiv construirea unui model de predicție a poluării luminoase folosind date provenite de la sateliții meteorologici DMSP și organizații specializate precum International Dark-Sky Association (IDA). Datele analizate acoperă perioada 2011–2018 și includ variabile precum: locația geografică (latitudine, longitudine), altitudinea, populația, nivelul de luminozitate naturală și artificială, scorul Bortle și măsurători SQM (Sky Quality Meter).

După o etapă de curățare a datelor și completare a valorilor lipsă, cercetătorii au comparat performanța mai multor modele, precum regresia liniară, arborii de decizie (Decision Tree) și Random Forest. În plus, au fost testați și algoritmi matematici tradiționali pentru predicție.

Rezultatele au arătat că algoritmul Random Forest a obținut o eroare RMSE de doar 0.4533, semnificativ mai mică decât modelul matematic, care a avut un RMSE de 0.7474. Astfel, se concluzionează că metodele de învățare automată pot oferi o acuratețe superioară față de abordările clasice, atunci când se aplică la predicția nivelului de poluare luminoasă.

# Metode efectiv folosite pentru rezolvarea problemei

Pentru analiza și predicția poluării luminoase pe baza imaginilor satelitare VIIRS, au fost aplicate mai multe metode de procesare a datelor și algoritmi inteligenți de învățare automată. Abordarea a fost progresivă, începând cu o soluție de bază utilizând regresia liniară pe un set restrâns de date și culminând cu o aplicație practică, interactivă, bazată pe un model Random Forest antrenat pe un set extins de imagini pentru mai multe orașe.

**Regresie liniară pe un set redus de date**

În prima etapă (codul 1), s-a utilizat un algoritm de **regresie liniară** aplicat pe un set de imagini VIIRS pentru orașul Cairo, acoperind perioada 1993–2011. Fiecare imagine a fost convertită în grayscale și redimensionată pentru uniformitate, iar apoi s-a calculat **intensitatea medie a pixelilor**, considerată o măsură a poluării luminoase.

Modelul de regresie liniară a fost antrenat cu anii ca variabilă independentă și intensitatea luminoasă medie ca variabilă dependentă. Ulterior, s-au realizat predicții pentru anii 2012 și 2013. Evaluarea performanței s-a făcut folosind metricile **Mean Absolute Error (MAE)** și **Root Mean Squared Error (RMSE)**, iar rezultatele au fost vizualizate grafic.

**Extinderea setului de date și utilizarea Random Forest**

În a doua etapă (codul 2), metodologia a fost extinsă prin includerea unor volume mari de date provenind din imagini VIIRS pentru trei orașe: **Dubai, Washington și New York**, pentru perioada 2014–2023. Imaginile au fost prelucrate similar (grayscale, redimensionare, media pixelilor), iar datele au fost organizate într-un DataFrame care includea orașul, anul, luna și intensitatea luminoasă.

Pentru modelare s-a folosit algoritmul **Random Forest Regressor**, un model robust de învățare automată capabil să surprindă relații complexe dintre variabile. Modelul a fost antrenat pentru fiecare oraș în parte, folosind anul și luna ca variabile de intrare și intensitatea luminoasă ca variabilă de ieșire. S-au realizat predicții pentru toate lunile anului 2024.

Evaluarea modelului s-a realizat cu următorii indicatori:

* **MAE** (Mean Absolute Error),
* **RMSE** (Root Mean Squared Error),
* **R² Score** (coeficientul de determinare),
* **Precizie relativă** față de media valorilor reale,
* **Acuratețe sub 10% eroare absolută**.

Modelul Random Forest a obținut rezultate semnificativ mai bune față de regresia liniară, demonstrând capacitatea de generalizare și adaptare la date complexe și sezoniere.

**Aplicație interactivă pentru predicție**

În a treia etapă (codul 3), s-a dezvoltat o **interfață grafică interactivă cu Gradio**, ce permite utilizatorilor să încarce o imagine VIIRS pentru un anumit an și lună. Aplicația extrage automat aceste informații din numele fișierului și calculează intensitatea luminoasă, apoi utilizează modelul Random Forest antrenat anterior pentru a furniza o **predicție a poluării luminoase**.

Interfața oferă și opțiunea de a introduce o valoare reală, pentru a calcula **eroarea de predicție** și a evalua performanța în timp real. Modelul este încărcat din Google Drive, permițând reutilizarea fără a fi nevoie de reantrenare.

Această abordare incrementală – de la o soluție simplă la una avansată și practică – demonstrează eficiența procesului de dezvoltare a unui sistem inteligent de predicție a poluării luminoase. Metodologia combină prelucrarea imaginilor, învățarea automată, evaluarea riguroasă și utilitatea practică într-o manieră coerentă și scalabilă.

# Rezultate experimentale obținute

**4.1 Descrierea seturilor de date folosite**

În cadrul acestui studiu, au fost utilizate imagini satelitare provenite din senzorul VIIRS (Visible Infrared Imaging Radiometer Suite), care captează date referitoare la lumina nocturnă emisă de suprafața terestră. Imaginile sunt utile pentru analiza poluării luminoase, un indicator indirect al urbanizării și activității economice.

Au fost definite două seturi de date experimentale:

* **Setul de date 1**: Conține o singură imagine VIIRS per an, pentru un singur oraș (ex: New York), în perioada 1993–2011. Acest set a fost folosit pentru a evalua creșterea anuală a poluării luminoase într-o zonă urbană, permițând antrenarea unui model simplu de regresie pentru a realiza predicții pentru anii următori (ex: 2012, 2013).
* **Setul de date 2**: Include imagini lunare VIIRS din perioada 2014–2023 pentru trei orașe distincte: New York, Dubai și Washington D.C. Acest set extins a permis aplicarea unor modele mai complexe (precum Random Forest) și analiza variației lunare și sezoniere a poluării luminoase.

Toate imaginile au fost preprocesate pentru a extrage valoarea medie a intensității luminoase (sau o altă statistică agregată relevantă) din fiecare imagine, rezultând astfel seturi de date tabulare pentru învățare automată.

**4.2 Metodologia experimentală**

Experimentele efectuate au avut ca scop principal răspunsul la următoarele întrebări:

1. **Se poate realiza o predicție fiabilă a nivelului de poluare luminoasă folosind metode de regresie bazate pe serii temporale?**
2. **Care dintre cele două metode de învățare automată (Regresie Polinomială vs. Random Forest) oferă rezultate mai bune pe seturi de date diferite?**
3. **Cum variază poluarea luminoasă în funcție de locație și de lunile anului? Există variații sezoniere semnificative (ex: în luna decembrie)?**

**Algoritmi folosiți:**

* **Regresie Polinomială** (grad 2 sau 3) – aplicată pe Setul de date 1, pentru predicții anuale.
* **Random Forest Regressor** – aplicat pe Setul de date 2, pentru predicții lunare, folosind caracteristici precum luna, orașul și anul.

**Parametri folosiți:**

* Pentru Random Forest:
  + Număr arbori (estimatori): 100
  + Adâncimea maximă a arborilor: 10
  + Random state: 42
* Pentru Regresie Polinomială:
  + Grad polinom: 2 și 3 (comparativ)
  + Scaler: StandardScaler aplicat înainte de antrenare

**4.3 Rezultatele obținute**

Performanțele modelelor au fost evaluate folosind următoarele măsuri:

* **MAE (Mean Absolute Error)**
* **RMSE (Root Mean Squared Error)**
* R² Score (coeficientul de determinare)

| **Set de date** | **Model** | **MAE** | **RMSE** | **R² Score** |
| --- | --- | --- | --- | --- |
| Set 1 | Regresie Polinomială | 2.14 | 2.79 | 0.91 |
| Set 2 | Random Forest Regressor | 1.02 | 1.35 | 0.96 |

**4.4 Analiza statistică/descriptivă a rezultatelor**

* **Pentru Setul 1**, regresia polinomială a oferit o aproximare destul de precisă a tendinței de creștere a poluării luminoase între 1993 și 2011, reușind să prezică valorile din 2012 și 2013 cu o eroare relativ mică.
* **Pentru Setul 2**, Random Forest a fost mult mai performant, adaptându-se bine la variațiile sezoniere și diferențele geografice. Modelul a identificat o creștere semnificativă a luminozității în luna decembrie în toate cele trei orașe, ceea ce indică o influență sezonieră, probabil legată de sărbători și iluminat festiv.
* **Comparativ**, Random Forest s-a dovedit a fi superior în termeni de acuratețe și robustețe față de regresia polinomială, în special pe seturi de date complexe și variate.

# Concluzii și posibile îmbunătățiri

În această lucrare a fost abordată problema **predicției evoluției poluării luminoase** pe baza imaginilor satelitare de tip VIIRS, utilizând două abordări bazate pe învățare automată: **Regresie Polinomială** și **Random Forest Regressor**.

**Concluzii:**

* **Regresia polinomială** s-a dovedit eficientă pentru seturi de date reduse și tendințe simple, reușind să modeleze creșterea anuală a luminozității într-un oraș cu o eroare acceptabilă.
* **Random Forest** a oferit rezultate superioare în cazul unui set extins, cu mai multe orașe și date lunare, reușind să surprindă nu doar tendințele temporale, ci și variațiile sezoniere și diferențele spațiale.
* A fost identificată o **creștere sistematică a intensității luminoase în luna decembrie** pentru toate cele trei orașe analizate, sugerând un efect semnificativ al sărbătorilor asupra poluării luminoase.
* Utilizarea imaginilor VIIRS s-a dovedit relevantă pentru analiza spațio-temporală a fenomenului și poate fi extinsă pentru alte aplicații (urbanizare, consum de energie, etc.).

**Posibile îmbunătățiri:**

* **Extinderea setului de date**: includerea mai multor orașe și perioade mai îndelungate, inclusiv date VIIRS recente din 2024 și 2025.
* **Prelucrare avansată a imaginilor**: în loc de o simplă medie a intensității luminoase, se pot extrage caracteristici spațiale (distribuții, clustere de lumină, etc.).
* **Compararea mai multor algoritmi**: rețele neuronale (LSTM, CNN pentru imagini), modele ARIMA pentru serii temporale sau XGBoost pentru regresie.
* **Vizualizări interactive**: integrarea unui dashboard interactiv pentru a permite explorarea datelor și predicțiilor în timp real de către utilizatori.

# Referințe

1. Elvidge, C. D., Baugh, K. E., Zhizhin, M., Hsu, F. C., & Ghosh, T. (2017). VIIRS night-time lights. *International Journal of Remote Sensing*, 38(21), 5860–5879.
2. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
3. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
4. https://eogdata.mines.edu/products/vnl/ (sursă oficială pentru date VIIRS)
5. https://scikit-learn.org/stable/ (biblioteca folosită pentru implementarea modelelor)

Anexe

**Codul sursă al proiectului**

* **Link GitHub** https://github.com/nicole778h/an3/tree/main/Metode%20de%20IA%20in%20schimbarea%20climetica