**Monitorizarea și predicția schimbărilor de temperatură cu efect asupra planetei**

Membrii echipei:

* Balan Silviu-Marian
* Pop Adrian
* Rotar Cristian
* Sechelaru Andrea

Denumirea echipei: **Bit by Bit**

Tema proiectului: **Monitorizarea și predicția schimbărilor de temperatură cu efect asupra planetei**

**Prezentarea problemei**

Schimbările climatice afectează și vor continua să afecteze întreaga planetă atât timp cât guvernele țărilor nu vor lua măsuri. Cu toate acestea, este posibil ca efectele schimbărilor climatice să afecteze mult mai pronunțat anumite grupuri specifice, cum ar fi comunitățile care depind de mijloacele de trai agricole sau de coastă.

Astfel, scopul acestei lucrări este implementarea unui algoritm inteligent care să prezică schimbările climatice viitoare. Prin acest mod se va putea ști evoluția climatică și se vor putea lua măsurile necesare astfel încât să se poată combate aceste modificări cu un efect negativ asupra planetei.

**Identificarea și înțelegerea setului de date**

Setul de date folosit pentru implementarea algoritmului de inteligență artificială a fost extras de pe site-ul Kaggle, <https://www.kaggle.com/sevgisarac/temperature-change>, care conține diferența de temperatură de la o lună la alta, pentru fiecare țară, din 1961 până în 2020. Acesta conține numeroase informații, printre ele enumerându-se: numele țării și codul acesteia, luna, unitatea de măsură, tipul de date, unitatea de măsură, deviația de temperatură.

După analiza setului de date, am observat că acesta conține mult prea multe detalii pentru programul pe care urmează să îl implementăm, deci din toate informațiile enumerate anterior, am ales doar țara, luna, anul și deviația de temperatură (Area, Months, Year și Value). Deci primul pas în analiza datelor a fost selectarea acestor coloane. Al doilea pas din această etapă a fost eliminarea rândurilor care conțin date statistice din întreg anotimpul sau întreg anul calendaristic, deoarece pentru prezicerea deviației de temperatură este suficientă reprezentarea datelor în funcție de lună.

Din punct de vedere al preprocesării datelor, a fost necesară completarea anumitor celule de pe coloana Values, care nu conțineau date. Pentru realizarea acestui lucru, am ales metoda „Fill missing values”, adică am completat lipsurile cu media aritmetică dintre cele mai apropiate valori existente

**Compararea cu abordări similare din literatura de specialitate**

Pentru realizarea proiectului a fost necesară documentarea cu ajutorul articolelor din literatura de specialitate, pentru a putea înțelege soluțiile care s-au abordat și pentru a putea realiza care este cu adevărat problema și cum poate fi soluționată

Așadar referințele de specialitate pe care le-am ales sunt articolele: „Forecasting Climatics Trends Using Neural Networks”, „Predicting ice flow using machine learning” și „Regression Analysis for Weather Forecasting”. În cadrul acestor articole am identificat două principale metode folosite în predicția schimbărilor climatice: regresia liniară și rețelele neuronale. Din ce am observat și analizat, ambele metode dau rezultate foarte bune și pot fi folosite în prezicerea eventualelor schimbări. Totuși, proiectul care a folosit rețelele neuronale a obținut rezultate mai bune, ajungând chiar și la o acuratețe de 98%, pe când proiectele în care a fost implementată regresia liniară au avut rezultate puțin mai slabe.

**Identificarea metodelor care se pot folosi**

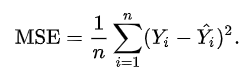
Există numeroase metode ce pot fi folosite în prezicerea schimbărilor climatice, așa cum s-a putut observa și în articolele științifice prezentate anterior. Cea mai folosită metodă este regresia, care poate fi de mai multe tipuri, deoarece este ușor de implementat și oferă rezultat extem de bune, având în vedere dificultatea metodei. Totuși, în cazul unui set de date mai mare, o alegere mai potrivită este implementarea unei rețele neuronale artificiale, aceasta reușind să ducă la bun sfârșit calcule mai complexe.

**Identificarea metricilor ce se pot calcula**

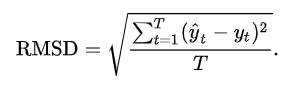
Pentru că metoda aleasă în realizarea proiectului este regresia, în continuare vom prezenta metricile ce se pot calcula pentru determinarea performanței algoritmului implementat. Nu se poate calcula acuratețea pentru un model de regresie, deci metricile ce vor fi prezentate specifică eroarea din cadrul predicțiilor.

Există numeroase metrici pentru calcularea erorii, dar dintre acestea trei dintre ele sunt mai des folosite: Mean Squared Error (MSE), Roor Mean Squared Deviation (RMSD) și Mean Absolute Error (MAE)

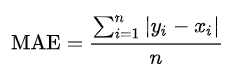
* MSE este calculată ca media aritmetică a pătratului diferențelor dintre valoarea prezisă și valoarea așteptată dintr-un set de date



* RMSD este o extensie a erorii prezentate anterior și este calculată astfel:



* MAE, spre deosebire de RMSD generează modificări liniare și astfel sunt mult mai intuitive; aceasta se calculează cu formula:



În continuare vom prezenta cele trei experimente pe care le-am realizat. Inițial, am utilizat agloritmul de gradient descendent stocastic, însă după mai multe rulări și analiza metricilor obținute am observat că nu este suficient de performant și de aceea, drept metodă de optimizare, am introdus ideea de normalizare a datelor. Rezultatele au fost semnificativ superioare, dar am continuat cercetarea și cu alt algoritm de regresie, si anume regresia logistică. Acesta din urma a fost mai simplu de implementat, nu a necesitat normalizarea datelor si valorea erorii a fost cea mai mică.

Comparând rezultatele obținute de noi cu rezultatele din articolele științifice pe care le-am studiat, am observat că metricile obținute de cercetători au fost mai bune decât cele returnate de algoritmul nostru. Acest fapt se poate datora și utilizării unor metode diferite de abordare a problemei, ei folosind, de exemplu, mașină cu suport vectorial sau rețele neuronale.

În tabelul de mai jos sunt descrise experimentele, în care am specificat și setup-ul și metricile obținute:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nr.** | **Metoda implementată** | **Setup** | **Rezultate** | **Observații** |
| 1 | Gradient descendent stocastic | Learning rate = optimal  Max\_iter = 3000  Dupa mai multe incercari am observat ca acesti parametri ofera cele mai bune rezultate | Error: 3.14576854  R2 score: 0.1110161 | Metricile obținute nu au fost optime și din acest motiv am continuat căutarea altor metode de îmbunătățire |
| 2 | Gradient descendent stocastic cu datele de intrare normalizate | Learning rate = optimal  Max\_iter = 3000 | Error: 1.0311243351699  R2 score: 0.1410388862047 | Gradientul descendent stocastic oferă rezultate mult mai bune dacă datele de intrare sunt normalizate.  Am folosit normalizarea standard |
| 3 | Regresie liniară | Parametrii impliciți | Error: 1.01840365379  R2 score: 0.15163563993 | Metricile obținute în urma aplicării acestui tip de regresie au fost cele mai bune dintre toate experimentele realizate. |

Pentru clarificarea rezultatelor obținute, am realizat și graficile predicției rezultate în urma aplicării metodelor de gradient descendent stocastic și, respectiv, regresie liniara. Pentru simplificarea figurilor am ales o singură țară, Afghanistan, reprezentând pe axa Ox anul și pe axa Oy predicția. Punctele roși semnifică schimbările de temperatură reale, preluate din setul de date, iar cu albastru sunt simbolizate valorile prezise de algoritmul implementat de către noi.

Se poate observa că predicția este asemănătoare între cele două metode, ambele reușind să obțină rezultate bune, dar totuși în cazul regresiei liniare graficul este mult mai precis și cursiv.

Chart, scatter chart

Description automatically generated

Figura 1 – Prezicerea rezultată de gradientul descendent stocastic

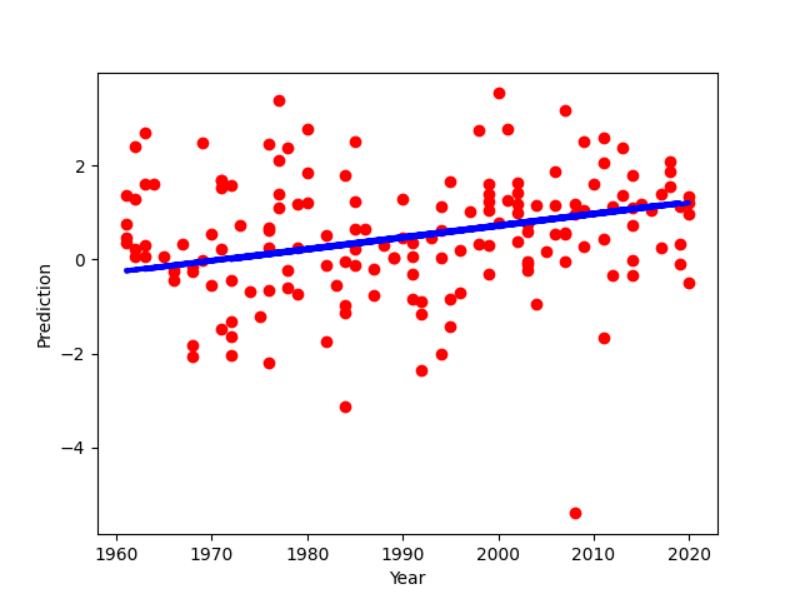


Figura 2 – Prezicerea rezultată de regresia liniară

În concluzie, rezultatele prezise de algoritmii implementați de către noi sunt eficienți, rapizi și oferă o precizie cu o eroare relativ mică. Cu toate acestea suportă îmbunătățiri, atât din punct de vedere a metodei implementate, cât și aș alegerii setului de date, având în vedere că au fost eliminate destul de multe coloane din setul inițial. Pe lângă acest lucru, considerăm că o perspectivă asupra prezicerii temperaturii, nu asupra diferenței acesteia, ar fi mult mai relevantă, doarece este mult mai ușor să conștientizezi gravitatea schimbării climatice.

**Cod sursă:** <https://github.com/CristiRotar/ClimateChangePredictor>