**Compararea modelelor de predictie aplicate pe un singur set de date**

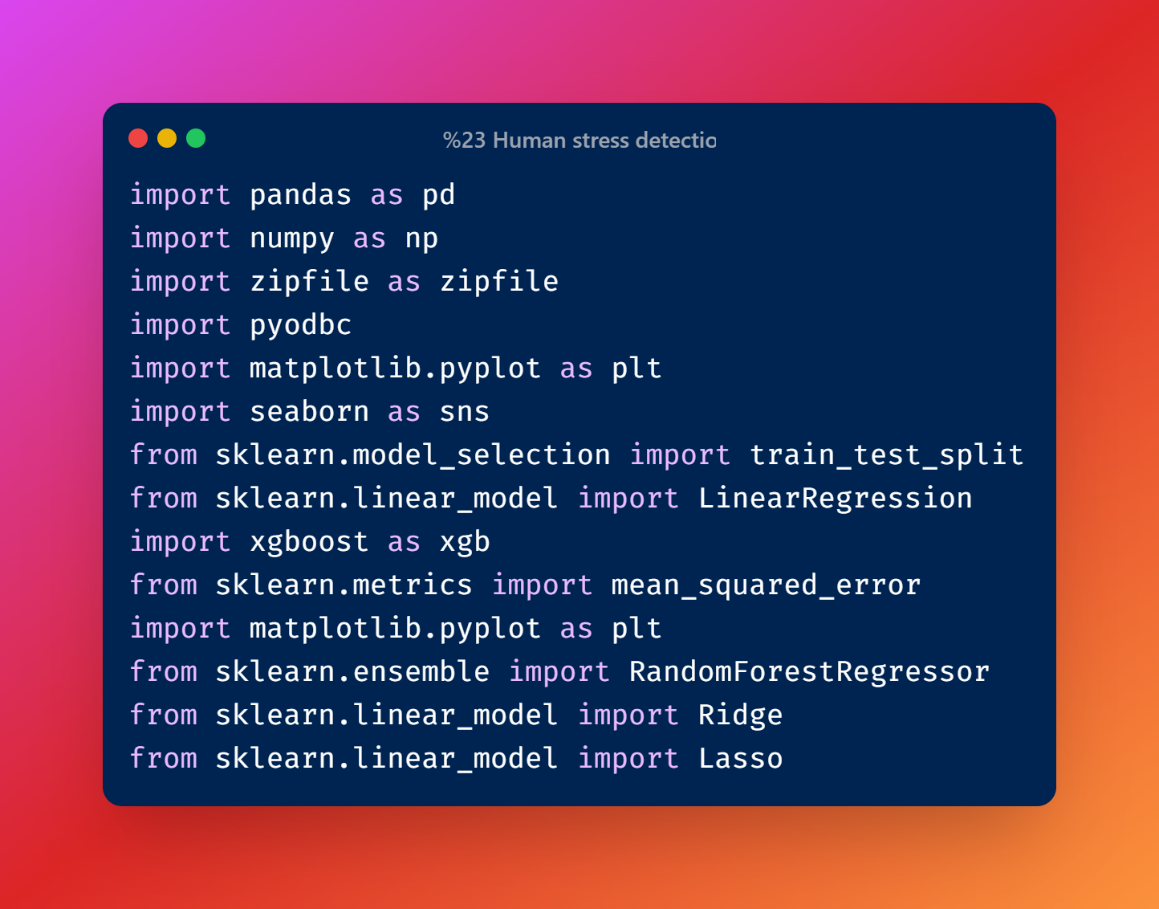
În acest proiect am vrut să arât diferența dintre acuratețea a cinci modele de predicție diferite, acestea fiind :

* **XGBoost**
* **RandomForest**
* **Regresia Ridge**
* **Regresia Lasso**
* **Regresia Lineară**

Proiectul este scris aprope în întregime între-un script de python și urmărește următoarea structură :

1. **Procesarea Inițială a Datelor**

* Importarea bibliotecilor necesare;



* Extracția datelor dintr-un fișier zip
* Transformarea valorilor temperaturii din Fahrenheit în Celsius
* Verificarea datelor pentru a asigura că nu există valori NA și tipurile de date sunt corecte
* Crearea de grafice pentru a vizualiza distribuția datelor și anomaliile
* Calcularea coeficienților de corelație Pearson și vizualizarea lor folosind un heatmap

**Funcția pentru vizualizarea coeficientului de corelație si coeficientul mai jos:**

A red squares with white text

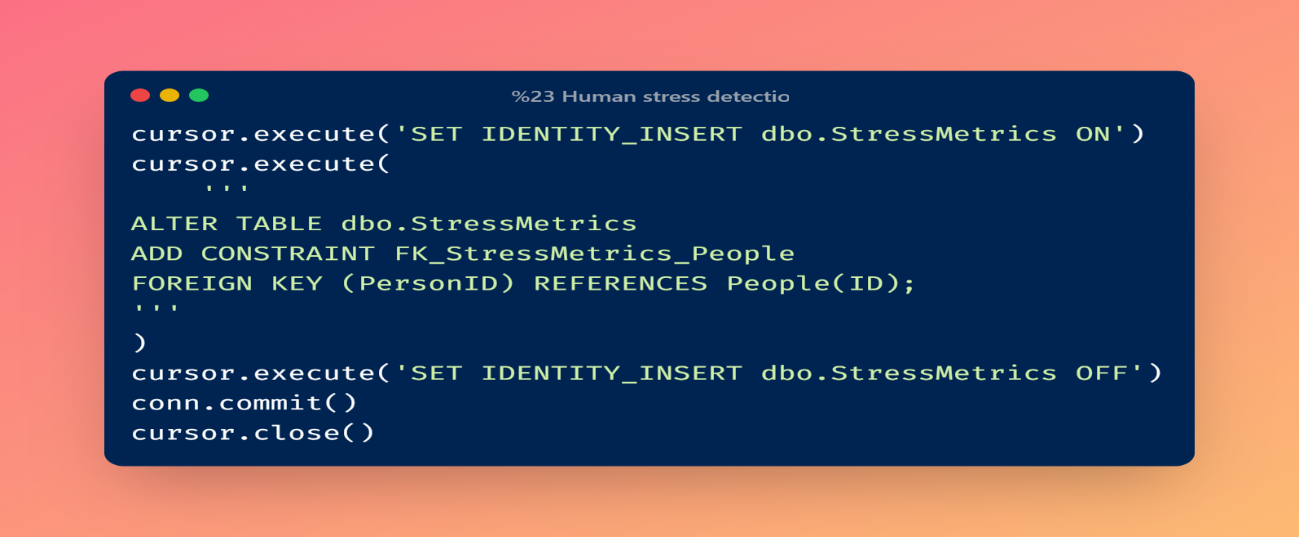
Description automatically generated**Vizualizarea Heatmap pentru corelație:**

* Căutarea anomaliilor și eliminarea acestora

1. **Crearea și Inserarea Datelor în Baza de Date:**

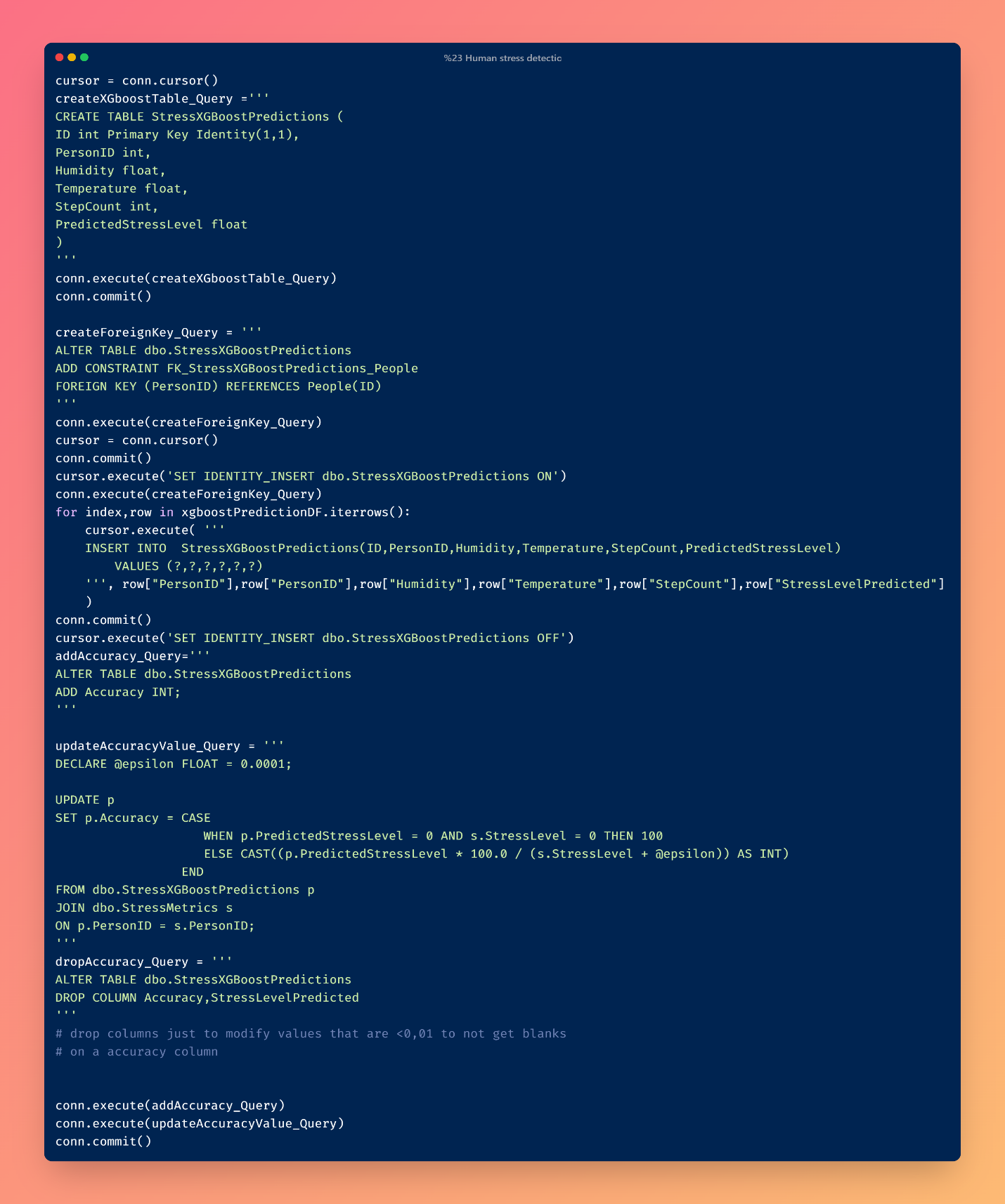
* **Conectarea la o bază de date locală**
* **Crearea tabelelor StressMetrics și People**
* **Inserarea valorilor din fișierele CSV și Excel în tabelele respective**
* **Stabilirea unei relații de cheie străină între StressMetrics și People**

****

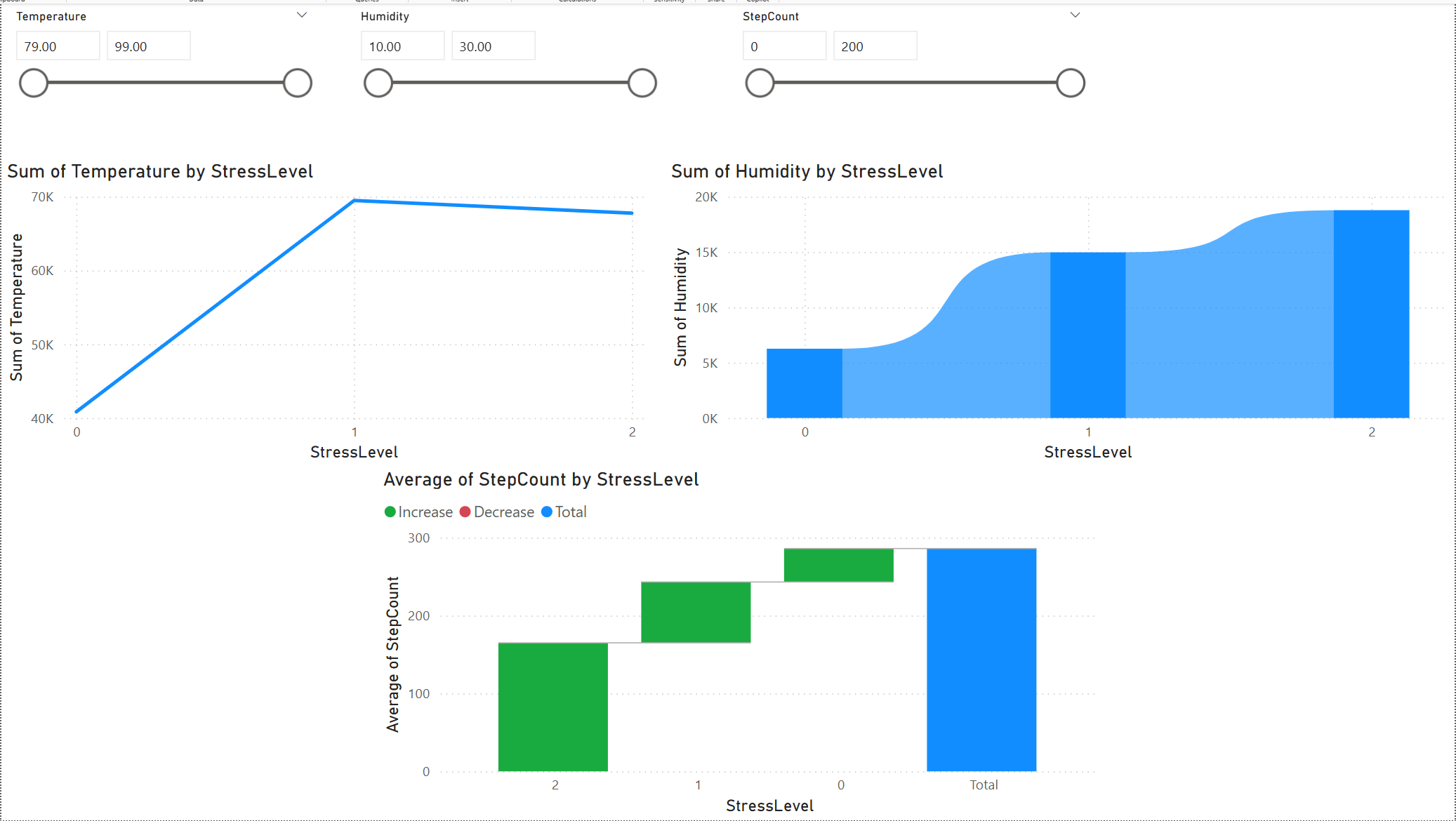
1. **Modele Predictive:**

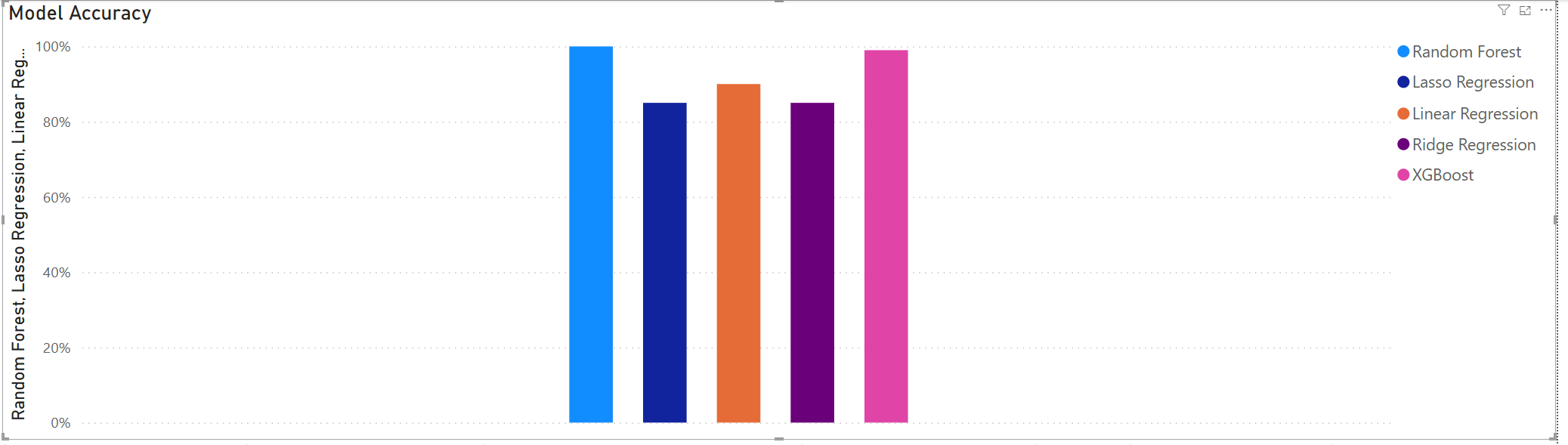
* **Pregatirea seturilor de date pentru antrenarea modelelor și alegem variabilele independente cu ajutorul cărora facem predicția ;**



* **Calculăm eroarea medie patratică și coeficientul de determinare pentru a observa eficacitatea modelului înainte de a-l aplica pe setul de date;**
* **Vom crea un tabel nou pentru fiecare dintre predicțiile noastre la care vom adăuga un field nou de acuratețe prin care calculam acuratețea modelului raportat la valorile din setul initial**

1. **Crearea de vizualizări în PowerBI**

****

****

**Rezultate :**

**Random forest – 99.2%**

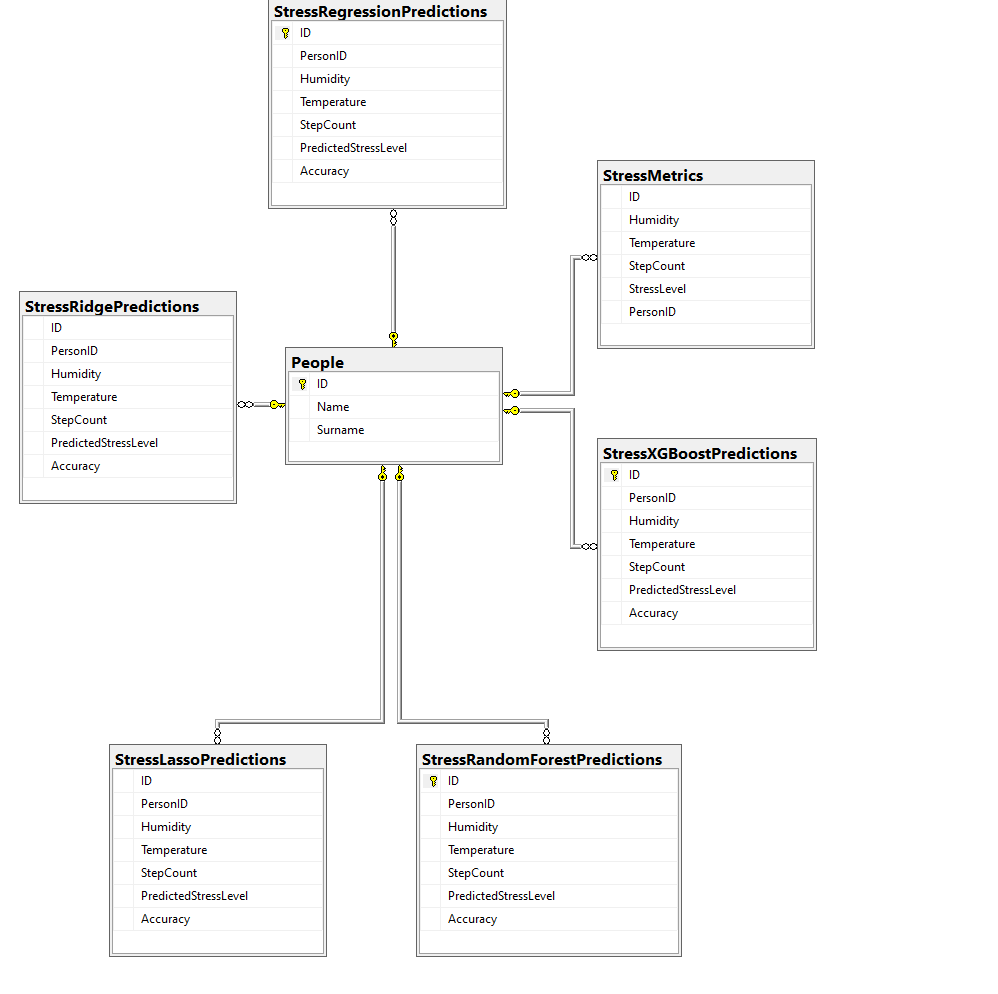
**XGBoost – 99.0%**

**Regresia Lasso – 85%**

**Regresia Liniara – 90%**

**Regresia Ridge – 86%**

**Diagramă Entitate-Relație**

****