Aplicatie Practica 1

Bujenita Lucian-Andrei

January 2025

Descrierea problemei

Proiectul își propune să prezică soldul energetic pentru luna decembrie 2024 pe baza unui set de date istorice care conține informații despre consumul și producția de energie pe diverse surse. Datele includ variabile precum consumul de energie, producția de energie pe diferite surse (carbune, hidrocarburi, ape, nuclear, eolian, foto, biomasa), și alte caracteristici relevante pentru prognoza soldului energetic, dar după mai multe încercari în care m-am folosit de toate coloanele, am ajuns la concluzia că mă voi folosi doar de coloana de Sold.

Scopul proiectului este de a construi un model care să ajute la prezicerea soldului pentru luna decembrie, utilizând algoritmi de învățare automată, și să comparăm performanța acestora pe baza unor măsuri statistice de evaluare, eu personal alegând să folosesc RMSE(Root Mean Squared Error), ce are formula $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$ și MAE(Mean Absolute Error), cu formula $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$.

Justificarea abordării

Pentru a rezolva această problemă de regresie, am ales inițial algoritmul **ID3**, un algoritm clasic de învățare automată bazat pe arbori de decizie. Motivul alegerii acestui algoritm a fost simplitatea și interpretabilitatea sa: ID3 creează un arbore de decizie care este ușor de înțeles și oferă informații clare despre relațiile dintre variabilele de intrare. Algoritmul a fost aplicat pentru a prezice soldul pe baza variabilelor de consum și producție de energie.

Totuși, pe parcursul experimentării, am observat că modelul ID3 nu a oferit cele mai precise predicții, în special atunci când datele sunt

complexe și conțin relații nelineare între caracteristici. Acesta a dus la concluzia că ID3 nu este suficient de robust pentru acest tip de date și că există alternative mai bune pentru îmbunătățirea acurateței predicțiilor.

În acest context, am decis să schimb algoritmul ID3 cu un alt algoritm, **Random Forest**, care este o metodă de ansamblu ce combină multiple arbori de decizie pentru a obține predicții mai stabile și mai precise. Random Forest abordează problema supraînvățării (overfitting) mult mai bine decât ID3 prin crearea unui număr mare de arbori de decizie și combinarea predicțiilor lor. Acesta este mai bine echipat pentru a gestiona complexitatea datelor și relațiile nelineare dintre variabile.

De asemenea, am utilizat **Naive Bayes**, un alt algoritm probabilistic, care funcționează bine cu datele de tip continuu și cu distribuții de date simple, oferind o abordare diferită pentru compararea performanței predicțiilor.

Prezentarea rezultatelor

În cadrul proiectului, am evaluat performanța celor trei algoritmi (ID3, Naive Bayes și Random Forest) pe un set de date de test, folosind măsurile de performanță **RMSE** (Root Mean Squared Error) și **MAE** (Mean Absolute Error). Am comparat aceste valori pentru a înțelege care model este cel mai precis în prezicerea soldului energetic pentru luna decembrie 2024.

Rezultatele arată că, în ciuda performanței decente a algoritmului ID3, Random Forest a avut un RMSE semnificativ mai mic și un MAE mai scăzut, ceea ce sugerează o performanță mai bună în prezicerea

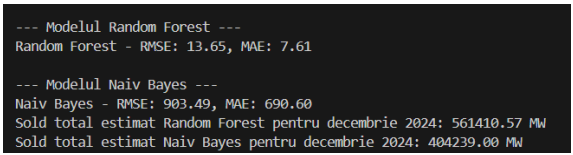

Figure 1: Analiza comparativă a performanței algoritmilor Naive Bayes și Random Forest

soldului energetic. Acest lucru este un indiciu clar că Random Forest poate modela mai bine complexitatea datelor și relațiile nelineare, în comparație cu ID3. De asemenea, Naive Bayes, deși util în anumite cazuri, nu a reușit să atingă aceeași performanță ca Random Forest.

Concluzii

În urma acestui proiect, am învățat importanța alegerii corecte a algoritmilor de învățare automată pentru a rezolva o problemă specifică. Algoritmul ID3, deși eficient pentru date discrete și probleme de clasificare simple, nu este suficient de puternic pentru a gestiona relațiile complexe din setul nostru de date. Acesta a condus la concluzia că algoritmi de ansamblu, precum **Random Forest**, sunt mult mai eficienți în astfel de contexte, datorită capacității lor de a gestiona multiple interacțiuni între variabile.

În viitor, ar putea fi explorate metode suplimentare, cum ar fi **Gradient Boosting** sau **XGBoost**, care sunt și mai sofisticate și pot oferi performanțe chiar mai bune decât Random Forest. De asemenea, am putea îmbunătăți procesul de preprocesare a datelor prin tehnici

avansate de inginerie a caracteristicilor și selecție a acestora, pentru a reduce dimensionalitatea și a spori acuratețea modelului. Experimentarea cu alți algoritmi și ajustarea parametrilor acestora ar putea duce, de asemenea, la îmbunătățiri semnificative ale performanței predicțiilor.