Tema Practica

Ghorbani Darius 3A2, Burghianu Bianca 3A2

January 10, 2024

Pentru setul de date lingspam, vom folosi primele 9 parti pentru antrenare si part10 pentru testare si vom aplica unul dintre algoritmii invatati la cursuri, parcurgandfiecare fisier (bare, lemm, lemm stop, stop). Scopul este de a obtine o accuratete cat mai buna.In setul de date spam ling, se regasesc doua tipuri de fisiere, mesaje spam care au prefixul "spm" in titlu si mesaje normale.

Am considerat AdaBoost ca fiind cel mai optim algoritm pentru tipul de date Ling Spam. AdaBoost se adapteaza nu numai la greselile clasificatorului anterior, dar poate si sa ajusteze limitele decizionale intr-un mod care imbunatateste acuratetea generala a modelului. De asemeni, este un algoritm cu o performanta puternica care produce clasificatori cu performante bune. Am testat pe 50 de estimari.

AdaBoost rezultate:
Accuracy $lemm = 0.96$
Accuracy $lemmstop = 0.96$
Accuracy $stop = 0.96$
Accuracy $bare = 0.96$

Folosind Bayes Naiv am obitnut urmatoarele rezultate:

Accuracy $lemmstop = 0.88$
Accuracy $stop = 0.87$
Accuracy $bare = 0.83$
Accuracy $lemm = 0.83$

Rezultatele obtinute de Bayes Naiv sunt mai mici decat cele de la AdaBoost, totusi Bayes Naiv este mai rapid. Am testat si algoritmul ID3 pe Ling Spam si am obtinut rezultate asemanatoare ca AdaBoost, dar am considerat ca este mai potrvit AdaBoost-ul pentru acest set in mod particular.

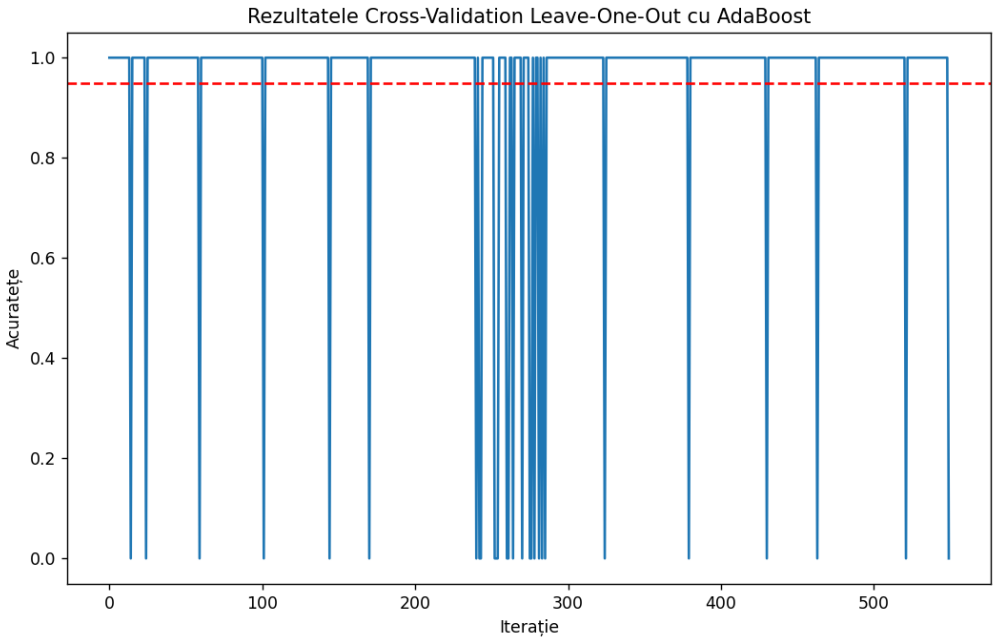

Figure 1: Rezultatele Cross-Validation Leave-One-Out cu AdaBoost

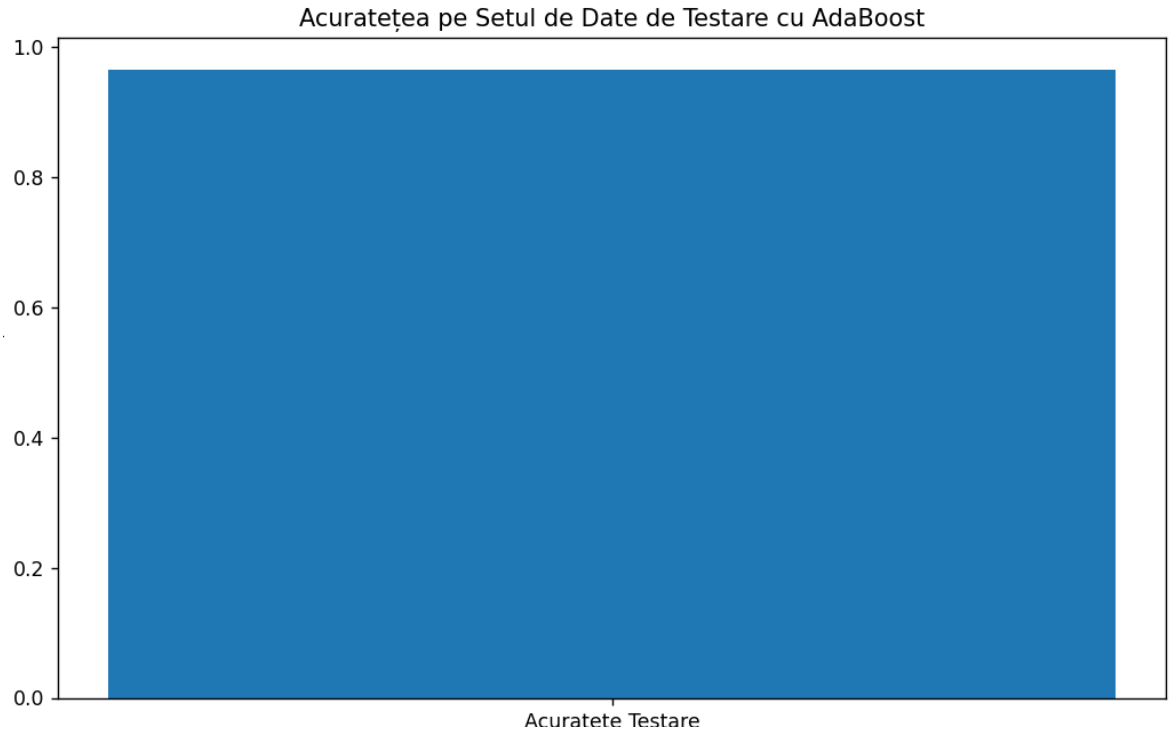

Figure 2: Acuratetea pe Setul de Date de Testare cu AdaBoost