Clasificarea tumorilor

**Cuprins**

[**1.Introducere** 1](#_Toc187960744)

[**2.Metodologie** 1](#_Toc187960745)

[**3.Rezultate obținute** 3](#_Toc187960746)

[**4.Concluzii** 8](#_Toc187960747)

# **1.Introducere**

* Tema: Proiectul are ca scop clasificarea unor tumori localizate la nivelul creierului, și anume gliomul, meningiomul și adenomul pituitar. Avem de asemenea și un set de date cu un creier sănătos, adică lipsit de orice tumoare.
* Motivație: Inteligența artificială utilizată în domeniul medical consider că este de mare folos, și trebuie realizate cât mai multe cercetări în aceasta direcție, pentru a schimba viitorul pacienților în bine.
* Utilitate: Această aplicație vine în ajutorul medicilor, care ar primi o confirmare în privința diagnosticării pacienților.
* Algoritmii cu care am realizat acest proiect sunt K-Nearest Neighbors și Naive Bayes, a căror performanța asupra setului de date urmează să o comparăm.

# **2.Metodologie**

* Setul de date: Este împărțit în 4 directoare:
* Adenom pituitar(150 imagini)
* Gliom(150 imagini)
* Meningiom(151 imagini)
* Sănătos(150 imagini)

Imaginile au fost extrase de pe Kaggle, o platformă care oferă acces la resurse de înaltă calitate.

* Etapa de prepocesare:
* Toate imaginile au extensia **jpg**
* Imaginile au fost furnizate alb-negru, prin urmare nu a mai fost necesară convertirea la niveluri de gri
* Imaginile au fost redimensionate la 128x128 pixeli pentru reducerea complexității
* Fiecare imagine a fost aplatizată pentru a fi utilizată în clasificatori( X\_train\_reshape)
* Împărțirea setului de date: Datele au fost împărțite în trei subseturi: antrenare (80%), validare(10%) și testare(10%) cu ajutorul funcției train\_test\_split
* Algoritmi:

Algoritmii pe care i-am utilizat sunt K-Nearest Neighbors și Naive Bayes.

**K-Nearest Neighbors** este un algoritm de clasificare care atribuie o etichetă unui nou punct necunoscut pe baza etichetelor celor mai apropiați k vecini ai săi în spațiul caracteristicilor.

* Avantajele utilizării acestui algoritm sunt:
* Este ușor de ințeles și de implementat
* Singurul parametru care necesită reglare este K
* Funcționează bine pentru un set de date mare
* Dezavantajele utilizării acestui algoritm sunt:
* Pe măsură ce numărul de caracteristici crește, performanța poate scădea
* Necesită mult timp de calcul în cazul unui set mare de date

**Naive Bayes** este un algoritm de clasificare care se bazează pe probabilități, ce presupune că toate caracteristicile sunt independente.

* Avantajele utilizării acestui algoritm sunt:
* Este rapid
* Face față bine unui număr mare de caracteristici
* Dezavantajele utilizării acestui algoritm sunt:
* Clasele cu număr mare de exemple pot domina predicțiile(în cazul nostru nu se aplică)
* Optimalitatea poate scădea odată cu creșterea complexității distribuției datelor
* Bibliotecile folosite sunt:
* os
* numpy
* cv2
* sklearn.metrics
* accuracy\_score(acuratețea)
* classification\_report(rezultatele)
* confusion\_matrix(matricea de confuzie)
* sklearn.naive\_bayes
* GaussianNB
* matplotlib

# **3.Rezultate obținute**

* Acuratețea pentru KNN:
  + Validare: 85,00%
  + Testare: 86,67%

Cu parametrul K = 7.

* Acuratețea pentru Naive Bayes:
  + Validare: 91,67%
  + Testare: 83,33%
* Indicatorii de performanță:
* **Precizia(Precision):** indică ce procent din instanțele care au fost atribuite unei anumite clase prin predicție sunt într-adevăr corecte (
* **Sensibilitate(Recall):** indică ce procent din instanțele care aparțin unei anumite clase au fost prezise corect (**)**
* **F1-Score:** este media armonică între **Precizie** și **Recall** care conferă o evaluare de ansamblu a modelului (**)**
* **Support:** indică ce număr de instanțe din fiecare clasă a fost prezent în datele de testare

**Pentru KNN**:

A screenshot of a computer screen

Description automatically generated

* Matricea de confuzie:

A diagram of a number of different colored squares

Description automatically generated with medium confidence

Din cele 15 imagini cu gliom

* 15 au fost prezise corect

Din cele 15 imagini cu meningiom

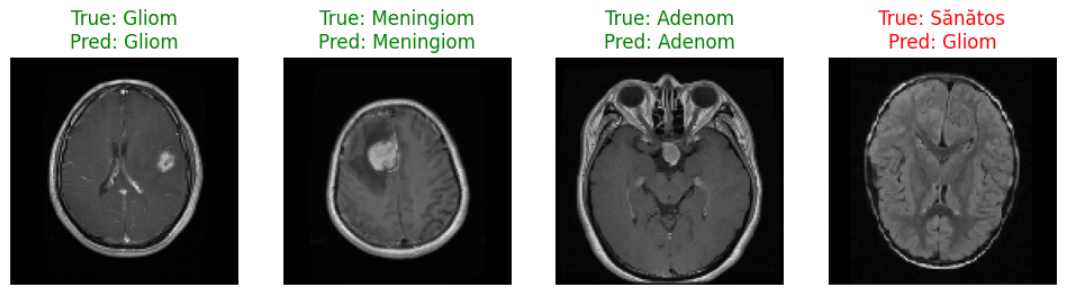
* 10 au fost prezise corect
* 5 au fost prezise ca gliom

Din cele 15 imagini cu adenom pituitar

* 15 au fost prezise corect

Din cele 15 imagini cu un creier sănătos

* 12 au fost prezise corect
* 3 au fost prezise ca gliom



**Pentru Naive Bayes**:

A screenshot of a computer screen

Description automatically generated

* Matricea de confuzie:

A blue squares with white text

Description automatically generated

Din cele 15 imagini cu gliom

* 13 au fost prezise corect
* 1 a fost prezisă ca fiind meningiom
* 1 a fost prezisă ca fiind adenom pituitar

Din cele 15 imagini cu meningiom

* 14 au fost prezise corect
* 1 a fost prezisă ca adenom pituitar

Din cele 15 imagini cu adenom pituitar

* 12 au fost prezise corect
* 3 au fost prezise ca fiind un creier sănătos

Din cele 10 imagini cu un creier sănătos

* 11 au fost prezise corect
* 3 au fost prezise ca meningiom
* 1 a fost prezisă ca adenom pituitar

A comparison of a brain scan

Description automatically generated

# **4.Concluzii**

* Pentru a mă asigura ca acuratețea va fi cât se poate de ridicată, am preluat niște imagini în care CT-ul s-a produs asemănător, adică de la aceeași distanță, deoarece dimensiunile sunt esențiale în cazul utilizării acestor algoritmi, care verifică intensitatea nivelului de gri pentru fiecare pixel. Aș fi avut posibilitatea de a prelua și imagini în care CT-ul s-a realizat din profilul persoanei, ci nu deasupra capului, mai ales în cazul adenomului pituitar, fapt care am crezut că îmi va dăuna acurateții. Avantajul a fost că localizarea adenomului are o localizare exactă, și pentru ca algoritmii de tip KNN si Naive Bayes să ofere rezultate excelente, acesta este un lucru vital
* Chiar și cu K = 7, rezultele KNN sunt excelente. Însă, rezultatele pot fi și mai bune daca setăm K = 1, și anume:
* Validare = 91,67% (cel mai bun rezultat posibil)
* Testare = 96,67% (cel mai bun rezultat posibil)

Acest lucru se datorează faptului că meningiomul are anumite localizări în creier unde se dezvoltă și are dimensiuni considerabile. Cât despre gliom, acesta este de dimensiuni mai mici, însă are o forma specifică. Creierul sănătos are o intensitate de gri constantă pe întreaga suprafață, iar adenomul este cel mai distinctiv dintre toate clasele, deci nici în aceste situații zgomotul nu a fost un factor decisiv. Acestea sunt motivele pentru care consider că algoritmul întoarce un rezultat bun chiar și cu această setare.

Pe de altă parte, atunci cand K = 15

* Validare = 80.00%
* Testare = 80.00%

Deci putem observa că atunci când mărim K, acuratețea scade. Cu toate acestea, consider că K = 7 este o alegere precaută care oferă oricum niște rezultate excelente

* Am ales ca fiecare clasă să aibă un set de date egal, tocmai pentru a nu afecta eficiența algoritmului Naive Bayes, care ia în calcul numărul de apariții al elementelor fiecărei clase din faza de antrenare pentru a realiza predicțiile, caracteristică care nu mă interesa în cadrul acestui proiect
* În ansamblu, putem observa că K-Nearest Neighbors a făcut față mai bine setului de date decât Naive Bayes, însă diferența nu este deloc mare, ținând cont și de faptul ca Naive Bayes presupune independența caracteristicilor, ceea ce poate fi nerealist pentru imagini
* Spre deosebire de primul proiect, de această dată am ales o redimensionare de 128x128 pixeli, alegere care cred că a influențat semnificativ rezultatele, timpul de rulare fiind aproape identic = 9 secunde (chiar și cu un set de date mai mare decât cel de la primul proiect)