

Предикција рака дојке

Даница Газдић SV12/2020
Милош Обрадовић SV55/2020



20. март - национални дан борбе против рака дојке

Циљ пројекта

Досадашња истраживања у медицини показују велику потребу за применом експертских система и метода у дијагностици болести. Циљ пројекта је дијагностиковати тумор на основу познатих обележја која представљају физичке и морфолошке карактеристике.

Подаци

Коришћен је скуп података који је прикупљен у америчким здравственим установама у Висконсину. Садржи 11 колона:

• Идентификација	Величине узимају вредности
• Дебљина грудвице	из распону 1-10.
• Величина ћелија	Последња колона представља
• Облик ћелија	дијагнозу и има вредности
• Маргинална адхезија	2 - бенигни и 4 - малигни.
• Величина епителних ћелија...	Подаци су подељени на тренинг и тестни скуп у размери 70:30

Коришћене методологије

За израду пројекта коришћен је python 3.9. Коришћена је библиотека pandas за припрему података, а за моделе машинског учења искоришћена је библиотека scikit-learn, из које смо користили и могућност оцене прецизности. Такође, коришћена је и сопствена имплементацију неуронске мреже.

У претпроцесирању су дефинисане две технике рада са подацима који су имали непотпуне вредности:

- одбацити неисправне вредности
- вршити израчунавање непознате вредности на основу сличних случајева података.

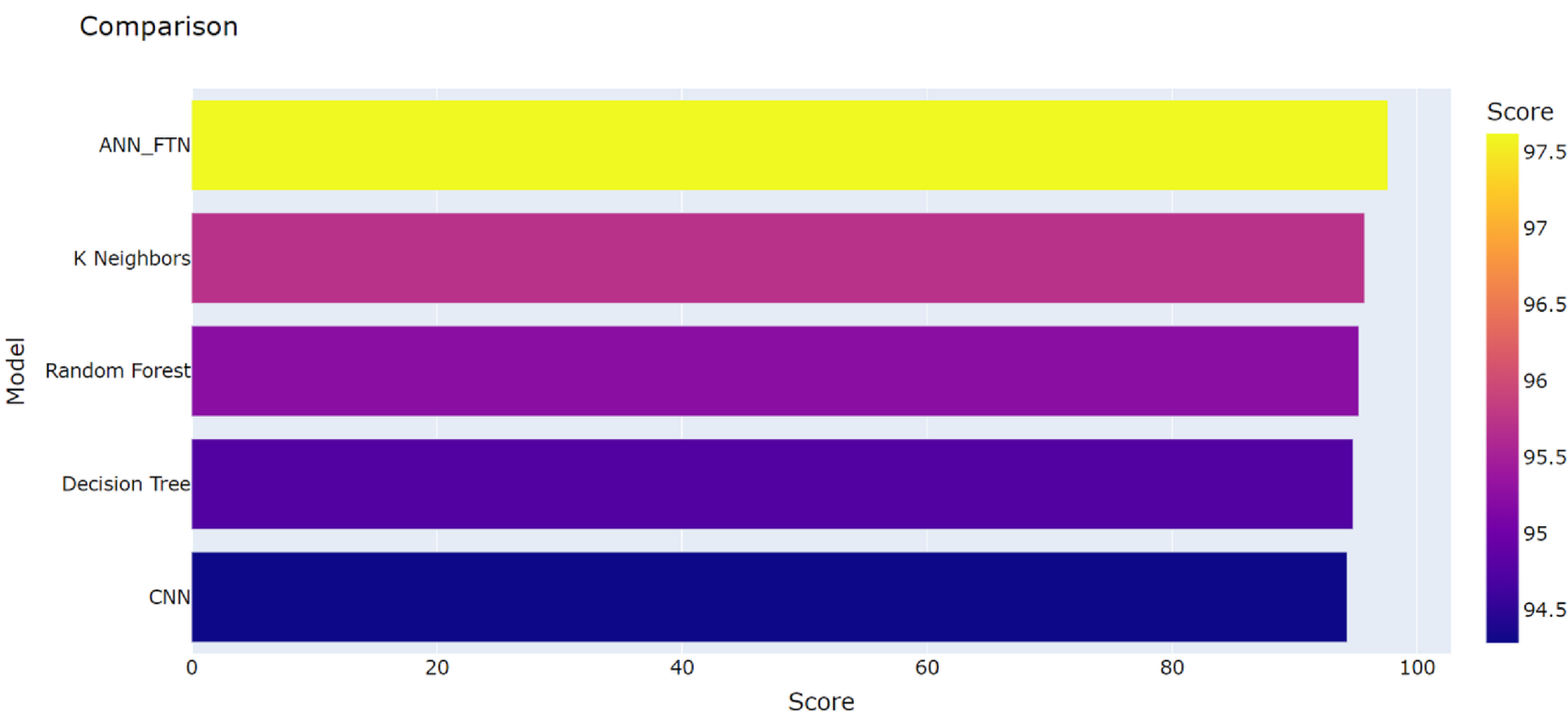
Укупно је коришћено 5 алгоритама, и то:

1. Decision Tree
2. Random Forest
3. K Neighbors
4. SciKit NN
5. FTN NN

Прецизност је израчуната по формули:

$$\frac{\text{бр. тачних предикција}}{\text{укупан бр. предикција}}$$

Евалуација решења



На графикону се види да се сопствена имплементација неуронске мреже показала као најпрецизнија за постављене вредности броја епоха и величине групе.

Такође, прецизност свих модела је преко 90%, што је задовољавајући резултат.

Даљи рад

За даљи рад неопходно је сакупити више података и поново покренути тестове над њима, као и истражити напредније технике руковања подацима који недостају.