PREDIKCIJA ŠUTEVA KOBIJA BRAJANTA

Katarina Branković

CILJ PROJEKTA

Cilj projekta je da uporedi različite algoritme mašinskog učenja za binarnu klasifikaciju. Konkretno, projekat ima za cilj da proceni njihovu tačnost u predviđanju da li je Kobe Bryant u određenim okolnostima dao koš.

PODACI

- Za dataset je korišćen skup podataka sa Kaggle takmičenja, koji obuhvata oko 25.500 instanci, pri čemu svaka instanca sadrži 25 atributa.

- Skup podataka je prilično uravnotežen, sa oko 11.500 instanci jedne klase i 14.000 instanci druge.

PRIPREMA PODATAKA

- Izbačene su kolone koje su procenjene kao irelevantne, dok su dodate one koje su smatrane značajnim za modele.

- Veliki broj atributa bio je kategorično — za one sa većom kardinalnosti primenjeno je target kodiranje, dok je za atribute sa manjom kardinalnošću korišćen dummy encoding.

PODELA PODATAKA

- Podaci su podeljeni na treniranje i testiranje, a za svaki model korišćen je pipeline koji uključuje target enkodiranje, skaliranje i optimalne parametre iz GridSearchCV-a. Time je osigurano da podaci za testiranje ostanu neiskorišćeni do finalnog testiranja.

- Target enkodiranje je korišćeno u svakom pipeline-u kako bi se izbeglo "curenje podataka", osiguravajući da se target varijable enkodiraju isključivo na trening podacima tokom svake fold validacije.

GRID SEARCH CROSS VALIDATION

- GridSearchCV pretražuje optimalne hiperparametre za model tako što isprobava različite kombinacije na osnovu zadatih vrednosti i meri performanse modela kroz K-fold validaciju.(cv=5)

- Zatim fituje model na celokupnom trening skupu sa optimalnim hiperparametrima.

REZULTATI KLASIČNIH MODELA

- Trenirani su različiti modeli, uključujući Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees, Random Forests, Support Vector Classification (SVC), XGBoost, AdaBoost i Naive Byes, a zatim je posmatrana njihova tačnost.

- Model sa najvećom tačnošću bio je XGBoost, dok je K-Nearest Neighbors (KNN) pokazao najmanju tačnost.

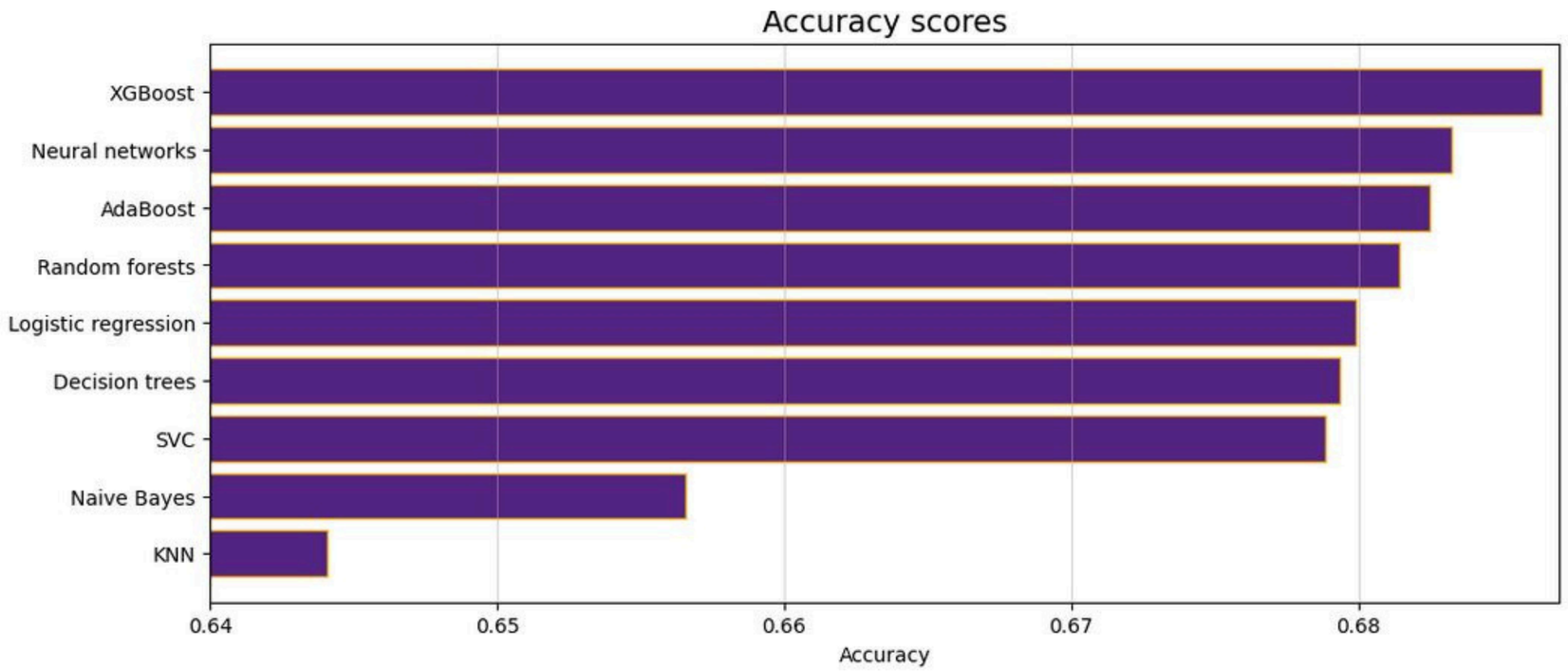
- (KNN) se pokazao kao najsporiji model jer zahteva izračunavanje udaljenosti između svih tačaka u skupu podataka, što postaje vremenski zahtevno s povećanjem broja instanci.

NEURONSKA MREŽA

- Sastoji se od tri potpuno povezana sloja. Svaki skriveni sloj koristi ReLU aktivaciju za nelinearnost i Dropout za regularizaciju Na izlazu, mreža koristi sigmoidnu funkciju koja pretvara rezultat u verovatnoću.
- U toku treniranja mreže praćene su promene loss-a i accuracy-ja kroz epohe.
- Korišćen je early stopping kako bi prekinulo treniranje ako se validacioni gubitak ne poboljša nakon određenog broja epoha.

OCENA TACNOSTI

- Kao metrika za ocenu kvaliteta modela izabrana je tačnost.



HVALA NA PAŽNJI!