

Predložak za nadzirano strojno učenje

Kako predvidjeti M-protein

Mile Šikić, Krešimir Križanović

Fakultet elektrotehnike i računarstva

Sadržaj

- Uvod u problematiku m-proteina
- Uvod u strojno učenje
- Opis korištenih podataka
- Inicijalna analiza
- SHAP vrijednosti i ocjena varijabli
- Izvođenje značajki
- Konačni rezultati
- Zaključak

M-protein

- Monoklonska gamapatija – posljedica benigne ili maligne bolesti
- Multipli mijelom – bol u kostima, anemija, oštećenje funkcije bubrega, povećana učestalost infekcija
 - Nespecifični simptomi
 - Laboratorijski nalazi: anemija, povećan broj leukocita, ubrzana sedimentacija eritrocita, povišena koncentracija ukupnih proteina, porast koncentracija ureja i kreatinina te porast koncentracije kalcija
 - Probit moguć specijalističkom pretragom elektroforeze serumskih proteina
- Cilj istraživanja:
 - Ispitati može li se općim laboratorijskim pretragama predvidjeti postojanje M-proteina
 - Ranija dijagnoza

Strojno učenje

- Detekcija M-proteina na temelju rezultata standardnih laboratorijskih pretraga
 - Nadzirano strojno učenje – poznati su ulazni i izlazni podaci
 - Ulazni podaci – rezultati pretraga
 - Izlazni podatak – postojanje M-proteina
 - Klasifikacija – svrstavamo svakog pacijenata u jednu od dvije klase
 - Otkriveno postojanje M-proteina
 - Nije otkriveno postojanje M-proteina
- Glavne ocjene kvalitete
 - Točnost (*engl. accuracy*)
 - Osjetljivost (*engl. sensitivity, recall, true positive rate*)
 - Pozitivna prediktivna vrijednost (*engl. positive predictive value, precision*)

Strojno učenje - klasifikacija

- Lažno negativne vrijednosti predstavljaju veći problem od lažno pozitivnih vrijednosti
 - Osjetljivost – koliki udio bolesnih je algoritam/test uspješno detektirao
- Pozitivna prediktivna vrijednost – koliki je udio stvarno bolesnih u pacijentima koje je algoritam/test proglašio bolesnima
- Površina ispod ROC krivulje – koliko dobro algoritam/test razlikuje dvije klase

Programski alati

- Programski jezik Python i Jupyter notebook – temelj za sve ostalo
- Numpy, scipy, pandas
 - Osnovne biblioteke za rad s podacima
 - Učitavanje iz datoteke
 - Manipulacija podataka
 - Matematičke i statističke funkcije
- Matplotlib – crtanje grafova
- Sklearn
 - Biblioteka za strojno učenje
 - Implementacije raznih modela i algoritama
 - Implementacija metrika
- XGBoost – brža/bolja implementacija Gradient boosting algoritma

Podaci

- Baza laboratorijskog informacijskog sustava
 - 38773 pacijenta koji imaju podatke o elektroforezi serumskih proteina (poznata ciljna vrijednost)
 - Izbačeni pacijenti koji se liječe na nekom od hematoloških odjela – smanjenje pristranosti
 - Samo jedno mjerjenje po pacijentu

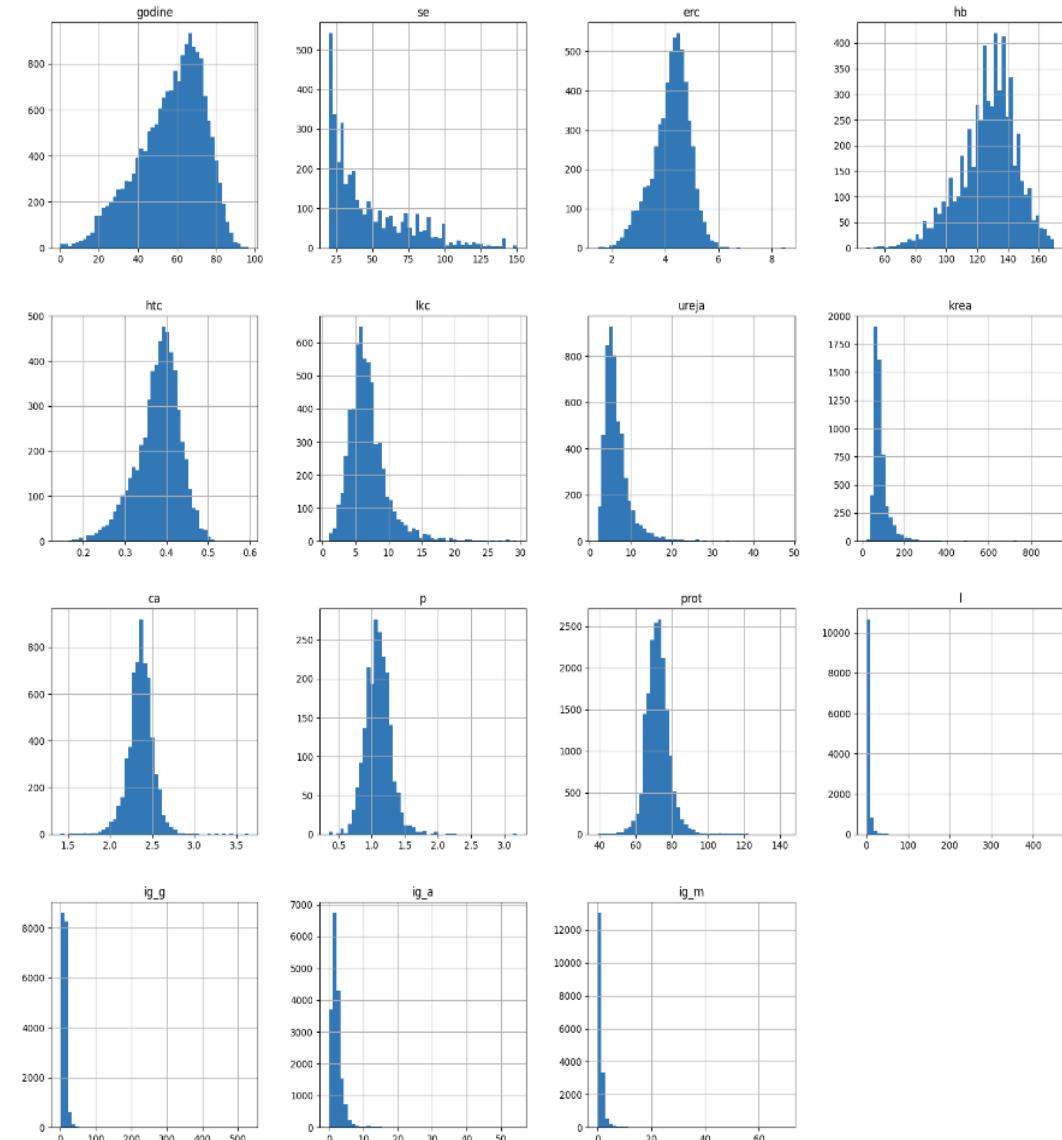
Kratica varijable	Naziv varijable	Opis varijable
spol	Spol pacijenta	Demografski podaci
godine	Dob pacijenta	
se	Sedimentacija eritrocita	
erc	Eritrociti	
hb	Hemoglobin	Rezultati općih biokemijskih i hematoloških pretraga
htc	Hematokrit	
lkc	Leukociti	
ureja	Ureja	
krea	Kreatinin	
ca	Ukupni kalcij	
p	Fosfati	
prot	Ukupni proteini	
l	Indeks lipemije	
ig_g	Imunoglobulin G	
ig_a	Imunoglobulin A	Rezultati specijalističkih pretraga
ig_m	Imunoglobulin M	
monok_it	Monoklonalni protein	Ciljna varijabla – ono što se želi predvidjeti

Podaci - čišćenje

- Inicijalni skup nije prikladan za izravno korištenje
 - Isključeni pacijenti kojima spol nije definiran kao „M“ ili „Z“
 - Isključeni pacijenti s datumom rođenja 1.1.1900. i stariji od 100 godina
 - Isključeni pacijenti ako je nedostajala ciljna vrijednost zbog nesukladnosti (npr. u interpretaciju M-proteina upisana „/“)
 - U programima/bibliotekama za strojno učenje nepoznate vrijednosti se pohranjuju na poseban način (oznaka NaN), budući da neki algoritmi znaju raditi s NaN vrijednostima, te su vrijednosti ostavljene u podacima
 - U podacima postoje različite oznake za nepoznate vrijednosti (prazno, '/') – vrijednosti se ujednačavaju te u modelu imaju posebnu vrijednost NaN
 - Isključeni pacijenti s rezultatima analiza koje ne očekujemo u ciljnoj populaciji: sedimentacija eritrocita < 20 mm/h, hemoglobin > 170 g/L, broj leukocita <1 i >30 x10⁹/L, koncentracija ureje < 2 µmol/L.

Podaci - konačno

- Korišteni podaci o 17680 pacijenata
 - 2851 pozitivnih primjera (detektiran M-protein)
 - 14829 negativnih primjera
- Značajan broj nepoznatih vrijednosti
- Bez nepoznatih vrijednosti su samo spol, godine, ciljna varijabla te rezultati specijalističkih pretraga



Inicijalna analiza

- Želimo na brzinu ocijeniti što bi mogli napraviti s podacima
- Odaberemo jedan model/algoritam
- Pitajmo ChatGPT 😊



You

what is the best general machine learning classification algorithm

- XGBoost zna raditi s nepoznatim vrijednostima!



ChatGPT

It's difficult to definitively say which machine learning classification algorithm is the "best" overall, as it heavily depends on the specific dataset, problem context, computational resources, and performance metrics. However, some algorithms are commonly considered strong performers across a wide range of tasks:

1. **Logistic Regression:** Despite its name, logistic regression is a classification algorithm. It's simple, interpretable, and works well for linearly separable data.
2. **Decision Trees:** Decision trees are intuitive and can handle both numerical and categorical data. They're easy to understand and interpret, but they can overfit if not pruned properly.
3. **Random Forest:** Random forests are an ensemble learning method that builds multiple decision trees and combines their predictions. They offer high accuracy and are less prone to overfitting compared to individual decision trees.
4. **Gradient Boosting Machines (GBM):** Algorithms like XGBoost, LightGBM, and CatBoost are popular implementations of gradient boosting. They sequentially train weak learners and combine them to form a strong learner, achieving high predictive accuracy.
5. **Support Vector Machines (SVM):** SVMs are effective for high-dimensional spaces and are versatile due to different kernel functions. They work well with clear margin of separation and are effective in cases where the number of dimensions is greater than the number of samples.
6. **Neural Networks:** Particularly deep learning architectures like Convolutional Neural Networks (CNNs) for image data, Recurrent Neural Networks (RNNs) for sequential data, and Transformer models for natural language processing have shown state-of-the-art performance in various tasks. However, they often require large amounts of data and computational resources.
7. **K-Nearest Neighbors (KNN):** KNN is a simple and intuitive algorithm that classifies data points based on the majority class among their k-nearest neighbors. It's non-parametric and doesn't make strong assumptions about the underlying data distribution.

Inicijalna analiza

- Dijelimo podatke u skup za učenje i skup za testiranje/validaciju u omjeru 80:20
- Dva testa
 - Samo koristeći rezultate općih pretraga (i demografske podatke)
 - Koristeći sve podatke

Rezultati:

	Točnost	Osjetljivost	PPV	ROC AUC
Opće pretrage	87,2%	34,5%	69,1%	0,66
Sve pretrage	93,0%	64,6%	87,8%	0,81

Inicijalna analiza

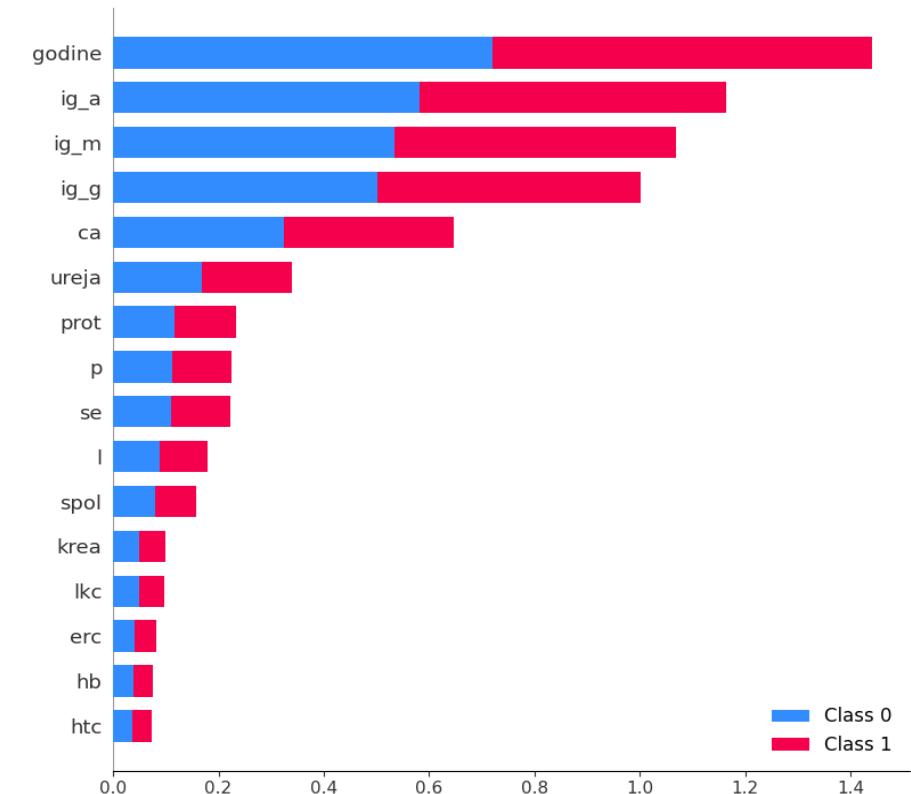
Rezultati:

	Točnost	Osjetljivost	PPV	ROC AUC
Opće pretrage	87,2%	34,5%	69,1%	0,66
Sve pretrage	93,0%	64,6%	87,8%	0,81

- Inicijalni rezultati nisu baš dobri
- Specijalističke pretrage značajno poboljšavaju rezultate
- Točnost je varljivo visoka – u našem slučaju nije dobra mjeru

SHAP vrijednosti

- SHAP vrijednosti mjere izolirani utjecaj pojedine varijable na rezultat modela
 - (Model se sastoji od svih korištenih značajki)
- Najveći utjecaj imaju
 - Godine pacijenta
 - Specijalističke pretrage
 - Ukupni kalcij



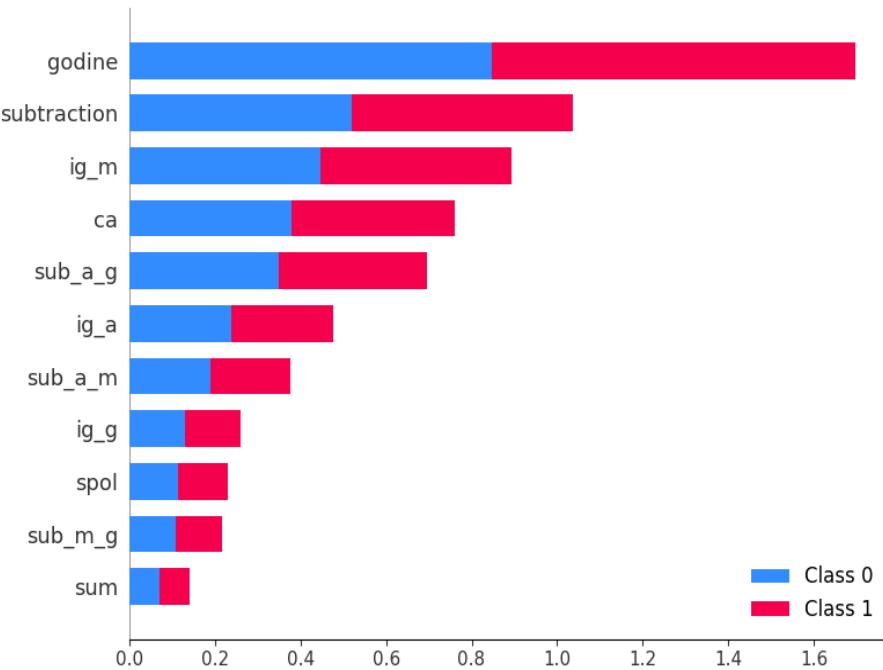
Izvođenje značajki

- Kombinacija vrijednosti različitih varijabli može imati veći utjecaj na ciljnu varijablu od početnih varijabli
 - Isproban veći broj kombinacija
 - Odabrane su one koje imaju najveći utjecaj

Naziv nove značajke	Formula za računanje
sum	$ig_g + ig_a + ig_m$
subtraction	$ig_g - ig_a - ig_m$
sum_a_m	$ig_a + ig_m$
sum_a_g	$ig_g + ig_a$
sum_m_g	$ig_g + ig_m$
sub_a_m	$ig_a - ig_m$
sub_a_g	$ig_g - ig_a$
sub_m_g	$ig_g - ig_m$

Izvođenje značajki

- U usporedbi s kalcijem, nekim varijablama je smanjena SHAP vrijednost. Nove varijable su možda bolje od početnih, ali nose „iste informacije”, pa kad ih zajedno koristimo u modelu, svaka pojedinačno ima manji utjecaj



Naziv nove značajke	Formula za računanje
sum	$ig_g + ig_a + ig_m$
subtraction	$ig_g - ig_a - ig_m$
sum_a_m	$ig_a + ig_m$
sum_a_g	$ig_g + ig_a$
sum_m_g	$ig_g + ig_m$
sub_a_m	$ig_a - ig_m$
sub_a_g	$ig_g - ig_a$
sub_m_g	$ig_g - ig_m$

Konačni rezultati

- Odabir najutjecajnijih značajki
- Testiran veći broj različitih modela pomoću biblioteke *LazyPredict*
- Kriterij za usporedbu modela: ROC AUC
- Mnogi od korištenih algoritama ne mogu raditi s nepoznatim vrijednostima varijabli – automatski se zamjenjuju srednjom vrijednosti

Konačni rezultati

- Najbolji rezultati: XGBoost, LightGBM, RandomForest
- Minimalan pomak u odnosu na početne rezultate
 - Smanjena točnost
 - Povećan ROC AUC
 - Osjetljivost i PPV nepromijenjeni

Model	Točnost	Osjetljivost	PPV	ROC AUC
XGBoost	92%	64%	88%	0,82
LightGBM	92%	64%	88%	0,82
RandomForest	92%	62%	89%	0,80
DecisionTree	88%	65%	65%	0,79
KNeighbors	91%	60%	80%	0,77
SVC	91%	51%	90%	0,74
Perceptron (plitka neuronska mreža)	82%	45%	40%	0,66
Logistička regresija	86%	22%	73%	0,61

Zaključak

- Strojno učenje pruža brojne mogućnosti za predikciju na temelju dostupnih podataka iz LIS-a
- Važna je kvaliteta podataka (čišćenje podataka)
- Na primjeru je demonstrirana predikcija M-proteina na temelju laboratorijskih pretraga
- Model je u stanju prepoznati pacijente koje je potrebno usmjeriti na daljnju specijalističku obradu
- Specijalističke pretrage poboljšavaju rezultate
- Relativno mala osjetljivost
- Može poslužiti kao dodatan alat za usmjeravanje dijagnostike