Uporaba metod strojnega učenja za analizo vpliva lipoproteinov (a) na bolezni srca in ožilja

Tajda Bogovič[1], Peter Kokol[1], Tadej Završnik[3], Jernej Završnik[2], Helena Blažun Vošner[2], David Šuran[3]

[1]*Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Maribor*
[2]*Zdravstveni dom dr. Adolfa Drolca, Maribor*
[3]*Univerzitetni klinični Center Maribor, Maribor*
E-pošta: tajda.bogovic@student.um.si peter.kokol@um.si
tadej.zavrnik@student.um.si jernej.zavrsnik@zd-mb.si
helena.blazun@zd-mb.si david.suran@student.um.s

Using machine learning methods to evaluate lipoprotein (a) and its impact on cardiovascular diseases

Cardiovascular diseases are the number one cause of death and it is expected to remain so. Machine learning has had a significant impact in many areas, including various medical applications. The aim of the study was to analyse and evaluate lipoprotein a - Lp (a) as a cardiovascular risk factor by using regression machine learning techniques. The study was performed on the Anonymous Cardiovascular Database collected in the University Hospital of Maribor. The results indicate that lipoprotein a, cholesterol, triglycerides and salt are closely connected to each other.

1 Uvod

Kardiovaskularne bolezni oz. KVB so eden izmed glavnih vzrokov za smrtnost. Pojavitev bolezni z leti narašča. [1] Najpogostejša kardiovaskularna bolezen je t. i. koronalna arterija, ki jo povzroča premajhen pretok krvi v srčno mišico. Ta nastane zaradi nabiranja maščob, holesterola in ostalih substanc na stenah arterijskih žil. [2]
Dejavniki tveganja vključujejo družinsko anamnezo, visok krvni tlak, kajenje, povišano raven holesterola LDL, sladkorno bolezen, debelost, pomanjkanje rekreacije in prekomerno uživanje alkohola. [3] Številne študije [9, 7] nakazujejo tudi vpliv povišane vrednosti lipoproteina (a) na tveganje za KVB.
Lipoprotein (a) ali na kratko Lp (a) je velik delček lipoproteina, ki ga tvorijo jetra. Visoka raven Lp (a) je bila zdaj opredeljena kot neodvisen dejavnik tveganja za srčno-žilne bolezni, ki je vzročno povezana z aterosklerozo (krčenje arterij), srčnimi napadi, možgansko kapjo, boleznijo aortne zaklopke in srčnim popuščanjem. Meritve vrednosti Lp (a) se na splošno ne izvajajo pogosto, ampak je v veliki meri odvisno od družinskih anamnez oz. stanj, saj je Lp (a) genetsko pogojen. Trenutno znano najboljše zdravljenje visoke ravni Lp (a) je zmanjšanje LDL holesterola. [5] To poteka s pomočjo intravenoznega zdravljenja, ki ima lahko do 90 % uspešnost [4]. Večina dosedanjih študij [6] obravnava zgolj uporabo metod strojnega učenja za namen klasifikacije kardiovaskularnih zapletov, pri tem pa ne vključujejo povezave z vrednostjo Lp (a), saj je merjenje le-te zahtevnejši oz. kompleksnejši proces v primerjavi s pogostejšimi laboratorijskimi krvnimi analizami. [4] Zhao, Juan in drugi [12] navajajo uporabo statistične Pearsonove analize in logistične regresije za iskanje povezav Lp (a) in KVB. Rezultati so pokazali, da tisti posamezniki, ki imajo visoko vrednost Lp (a), bodo imeli večjo verjetnost za nastanek kardiovaskularnih zapletov v primerjavi s tistimi, ki imajo normalno oz. nizko vrednost. V študiji [10] so primerjali izvedbo treh odločitvenih dreves, štirih statističnih algoritmov in dveh nevronskih mrež za napovedanje hipertenzije. Atribut Lp (a) se je izkazal za zanesljiv kazalec pri napovedovanju. Li, Huidong [8] in Xu, Yuan [11] so uporabili metodi logistično regresijsko analizo in XGboost model za diagnosticiranje ateroskleroze oz. napovedovanje možganske kapi pri bolnikih, ki so že predhodno imeli ishemično kap. Bazi, uporabljeni v študijah, sta vsebovali atribut Lp (a).

2 Metode

V naši raziskavi smo uporabili anonimno bazo kardovaskularnega oddelka Univerzitetnega kliničnega centra Maribor v obdobju 1999—2019. Baza vsebuje podatke 4477 bolnišničnih obiskov z 3940 različnimi vnosi in 39 pripadajočimi atributi, ki so podrobneje opisani v tabeli 1.

2.1 Predprocesiranje podatkov

Atribute, kateri so imeli več kot 95 % manjkajočih vrednosti (z izjemo zaporednega števila), smo odstranili, in sicer: zap. št, FT3, FT4, TSH, S-NT-pro BNP, S-magnezij (Mg), S-fosfat (P), S-apolipoprotein A1, S-AST (GOT), S-CRP, S-HDL-holesterol, S-LDH, S-alkalna fosfataza, S-bilirubin celokupni, S-bilirubin direktni, S-holesterol, S-kalij (K), S-kloridi (Cl), S-natrij (Na), S-trigliceridi, S-urea (sečnina), S-Troponin T.

Atributom Ocena GF, S-ALT (GPT), S-AST (GOT), S-CRP, S-alkalna fosfataza, S-bilirubin celokupni, S-bilirubin direktni, S-gama GT S-kalij (K), S-kloridi (Cl), S-kreatinin-E, S-natrij (Na) S-troponin I, S-urea (sečnina), S-HDL-holesterol, S-LDH, S-holesterol, S-trigliceridi, S-Lp (a), S-srčni troponin I in S-Troponin T smo manjkajoče vrednosti dopolnili s povprečno vrednostjo.

Atributa diag1 in sprejemnaDiagnoza smo spremenili v kategorični tip podatka, saj so vpisane le tiste diagnoze, ki se začnejo s šifrantom I25.

Tabela 1: Atributi baze kardovaskularnega oddelka Univerzitetnega kliničnega centra Maribor

Ime atributa	**Tip atributa**	**Opis atributa**
zap. št.	Celo število	Identifikacija pacienta
smrt	Kategorični podatek	0- Ne 1- Da
spol	Kategorični podatek	0 - Ženski 1- Moški
sprejemnaDiagnoza	Niz	Diagnoze I25.X
AH	Kategorični podatek	Prisotnost arterijske hipertenzije 0 - Ne 1 - Da
SB	Kategorični podatek	Prisotnost sladkorne bolezni 0 - Ne 1 - Da
SP	Kategorični podatek	Prisotnost srčnega popuščanja 0 - Ne 1 - Da
KBL	Kategorični podatek	Prisotnost kronične ledvične bolezeni 0 - Ne 1 - Da
D H HTg	Kategorični podatek	Prisotnost hipertrigliceridemije 0 - Ne 1 - Da
diag1	Niz	Diagnoze I25.X
star	Celo število	Starost pacienta izražena v letih
fpri	Kategorični podatek	Št. oddelka
Ocena GF	Decimalno število	Vrednosti laboratorijskih oz. krvnih analiz
S-ALT (GPT)	Decimalno število	
S-AST (GOT)	Decimalno število	
S-CRP	Decimalno število	
S-alkalna fosfataza	Decimalno število	
S-bilirubin celokupni	Decimalno število	
S-bilirubin direktni	Decimalno število	
S-gama GT	Decimalno število	
S-kalij (K)	Decimalno število	
S-kloridi (Cl)	Decimalno število	
S-kreatinin-E	Decimalno število	
S-natrij (Na)	Decimalno število	
S-troponin I	Decimalno število	
S-urea (sečnina)	Decimalno število	
S-HDL-holesterol	Decimalno število	
S-LDH	Decimalno število	
S-holesterol	Decimalno število	
S-trigliceridi	Decimalno število	
S-Lp (a)	Decimalno število	
S-srčni troponin I	Decimalno število	
S-Troponin T	Decimalno število	
FT3	Decimalno število	
FT4	Decimalno število	
TSH	Decimalno število	
S-NT-pro BNP	Decimalno število	
S-magnezij (Mg)	Decimalno število	
S-fosfat (P)	Decimalno število	
S-apolipoprotein A1	Decimalno število	

Po predpocesiranju smo tako dobili 3427 vnosov z 32 različnimi atributi.

2.2 Uporaba metod

V študiji smo uporabili različne regresijske metode za odkrivanje povezav med izbranim atributom Lp (a) in ostalimi atributi. Izbrali smo modele odločitvenih dreves, naključnih gozdov, GradientBoost in AdaBoost, ki smo jih implementirali s pomočjo odprto-kodne Python knjižnice Scikit-learn.

S pomočjo omenjenih regresijskih metod smo ugotovili, kateri atributi imajo največjo korelacijo z atributom Lp (a) in kakšna je bila povprečna kvadratna napaka pri posameznem modelu.

Odločitveno drevo je imelo parameter maksimalne globine enak 4. Model naključnih gozdov je vključeval 500 iteracij, maksimalno globino 5 in minimalno razdelitev 2. GradientBoost in Adaboost sta imela 500 iteracij, stopnjo učenja 0.01. Funkcija izgube je v primeru GradientBoost bila regresija najmanjših kvadratov (ang. least squares regression), medtem ko je pri Adaboost bila linearna.

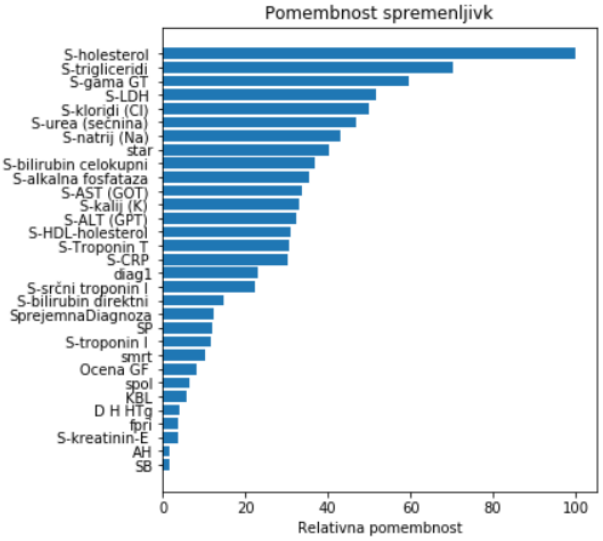
.

3 Rezultati in diskusija

Slika 1: Pomembnost spremenljivk v primeru regresijskega odločitvenega drevesa

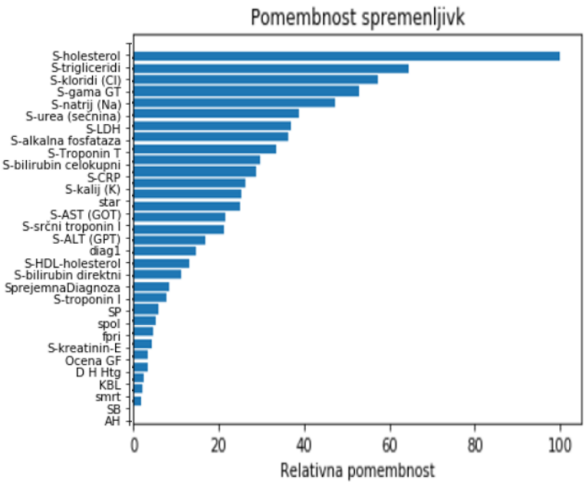
Iz slike 1 lahko razberemo pomembnost posameznih atributov z atributom Lp (a) pri uporabi regresijskega odločitvenega drevesa. Izkaže se, da je atribut S-holesterol eden izmed najpomembnejših, saj znaša njegova relativna pomembnost skoraj 100 %. Sledita atributa SP (67 %) in S-Trigliceridi (57 %). Prav tako ne gre zanemariti atributov spol, KBL in S-Troponin T, ki so dosegli več kot 30 % relativno pomembnost.

Slika 2 prikazuje rezultat pomembnosti atributov pri uporabi regresijskega modela GradientBoost. Opazimo lahko, da je največja povezava med vrednostima Lp (a)



Slika 2: Pomembnost spremenljivk v primeru regresijskega modela Gradientboost

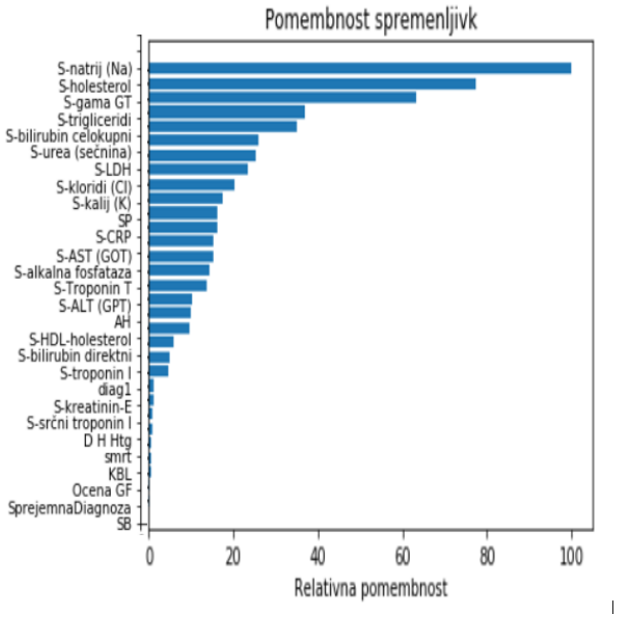
in S-holesterolom. Sledijo atributi S-trigliceridi, S-gama GT, S-LDH in S-kloridi, ki imajo 50 ali več procentno relativno pomembnost. Med 30 in 50 % pomembnost so dosegli S-urea (sečnina), S-natrij (Na), starost, S-bilirubin celokupni, S-lokalna fosfataza, S-AST, S-kalij, S-ALT, S-HDL-holesterol, S-Troponin T in S-CRP. Pri ostalih znaša manj kot 30 %.



Slika 3: Pomembnost spremenljivk v primeru regresijskega modela naključnih dreves

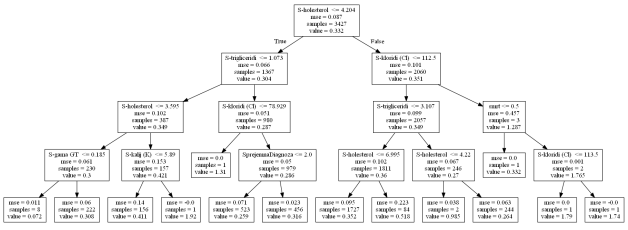
Izris pomembnosti atributov modela naključnih gozdov je viden na sliki 3. Ponovno je S-holesterol dosegel najvišjo vrednost relativne pomembnosti z Lp (a), saj je ta znašala 100 %. Visok delež so dosegli tudi atributi S-trigliceridi, S-kloridi in S-gama GT.

Za razliko od modelov odločitvenega drevesa, GradientBoost in naključnih dreves je AdaBoost dosegel najvišji procent (100 %) relativne pomembnosti z atributom S-natrij, kar lahko razberemo tudi na sliki 4. Sledita S-



Slika 4: Pomembnost spremenljivk v primeru regresijskega modela Adaboost

holesterol in S-gama GT, oba z več kot 50% relativno pomembnostjo.



Slika 5: Izris regresijskega odločitvenega drevesa

Odločitveno drevo, ki smo ga dobili kot rezultat uporabe regresijskega odločitvenega drevesa, vidnega na sliki 5, potrjuje pomembnost atributov S-holesterola, S-trigliceridov in soli, kot sta S-natrij in S-kalij. Listi dreves so vrednosti atributa Lp (a).

Tabela 2: Napaka regresijskih modelov

Regresijski model	Povprečna kvadratna napaka
Odločitveno drevo	0.0898
GradientBoost	0.0869
Naključna drevesa	0.0860
AdaBoost	0.1344

Kot je razvidno iz tabele 2 so vsi modeli imeli majhno povprečno kvadratno napako. Najmanjšo smo zabeležili pri naključnih drevesih, sledita GradientBoost in odločitveno drevo. Najvišjo napako je imel model AdaBoost.

Vsi modeli so pokazali, da je prisotna korelacija med S-holesterolom in Lp (a) vrednostjo. Prav tako se je pokazala pomembnost S-trigliceridov z Lp (a). V treh primerih je slednja znašala (skoraj) 100 %. Le v primeru modela AdaBoost lahko opazimo nekoliko drugačne rezultate, kar pa se tudi kaže v izmerjeni povprečni kvadratni napaki. Ta je bila v primerjavi z ostalimi modeli nekoliko višja.

Glede na to, da je meritev vrednosti trigliceridov in holesterola pogostejša in enostavnejša, lahko povzamemo, da bi v primeru visoke vrednosti teh atributev bilo smiselno izmeriti oz. pridobiti tudi gensko vrednost Lp (a).

4 Zaključek

Prisotnost kardivaskularnih bolezni predstavlja večjo umrljivost in večjo možnost pridružitve zdravstvenih zapletov oz. novih diagnoz. V tej študiji smo podrobneje pregledali in analizirali povezanost vrednosti Lp (a) z ostalimi dejavniki tveganja KVB. Pri tem smo uporabili različne regresijske tehnike strojnega učenja, in sicer odločitvena drevesa, naključne gozdove, GradientBoost ter AdaBoost model. Rezultat uporabe tehnik v skoraj vseh primerih nakazuje povezanost izbranega atributa z atributi S-holesterol, S-trigliceridi in soli.

Literatura

[1] Benjamin, Emelia J., et al. Heart disease and stroke Statistics-2019 update a report from the American Heart Association. Circulation, 2019.

[2] Diseases and Conditions - Mayo Clinic. https://www.mayoclinic.org/diseases-conditions (pridobljeno 19. 6. 2020)

[3] Hajar, Rachel. Risk factors for coronary artery disease: historical perspectives. Heart views: the official journal of the Gulf Heart Association, 2017, 18.3: 109.

[4] HEART, U. K. The cholesterol charity. Bridging the gap. UK: Heart the cholesterol charity, 2013.

[5] High Lipoprotein(a). https://www.heartuk.org.uk/genetic-conditions/high-lipoproteina (pridobljeno 27. 8. 2020)

[6] Kadi, Ilham, Ali Idri, and José Luis Fernandez-Aleman. Systematic mapping study of data mining–based empirical studies in cardiology. Health Informatics Journal 25.3 (2019): 741-770.

[7] Kassner, Ursula, et al. Lipoprotein (a)–an independent causal risk factor for cardiovascular disease and current therapeutic options. Atherosclerosis Supplements, 2015, 18: 263-267.

[8] Li, Huidong, et al. Modeling analysis of the relationship between atherosclerosis and related inflammatory factors. Saudi journal of biological sciences, 2017, 24.8: 1803-1809.

[9] Nordestgaard, Børge G., et al. Lipoprotein (a) as a cardiovascular risk factor: current status. European heart journal, 2010, 31.23: 2844-2853.

[10] Ture, Mevlut, et al. Comparing classification techniques for predicting essential hypertension. Expert Systems with Applications, 2005, 29.3: 583-588.

[11] Xu, Yuan, et al. Extreme gradient boosting model has a better performance in predicting the risk of 90-day readmissions in patients with ischaemic stroke. Journal of Stroke and Cerebrovascular Diseases, 2019, 28.12: 104441.

[12] Zhao, Juan, et al. Using topic modeling via non-negative matrix factorization to identify relationships between genetic variants and disease phenotypes: A case study of Lipoprotein (a)(LPA). PloS one, 2019, 14.2: e0212112.