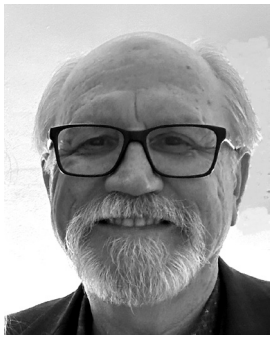
Uģis Sprūdžs

https://doi.org/10.22364/adz.59.09

SIRDS UN ASINSRITES SLIMĪBU MIRSTĪBAS RISKA PROGNOZE NĀKAMAJAM GADAM NO ANONIMIZĒTIEM LATVIJAS VESELĪBAS APRŪPES SISTĒMAS DATIEM: *XGBOOST* MAŠĪNMĀCĪŠANĀS ALGORITMA IESPĒJAMĪBAS PĀRBAUDE[1]



Uģis Sprūdžs, biznesa vadības maģistrs, ieguvis *Chartered Financial Analyst* sertifikātu, ir Čikāgas Universitātē izglītots analītikas un datu zinātnes vadības speciālists ar vairāk nekā 20 gadu darba pieredzi Ziemeļamerikas banku un apdrošināšanas sektoros, Rīgas Stradiņa universitātes Attīstības un projektu departamenta Studiju kursu izstrādes eksperts un viesdocents Biznesa, mākslas un tehnoloģiju augstskolā RISEBA, Baltijas studiju veicināšanas apvienības (*Association for the Advancement of Baltic Studies*) padomes kasieris, kopš 2021. gada ASV Fulbraita (*Fulbright*) stipendiju programmas speciālists. Pedagoģiskais darbs saistās ar datu zinātnes lietojumu uzņēmumu vadībā un sabiedrības veselības jomā. Čikāgas Universitātē ieguvis biznesa vadības maģistra grādu ekonometrijā un finansēs, ir arī ģermānistikas pedagoģijas maģistrs. *Chartered Financial Analyst Institute* biedrs. Publicējis rakstus par Imanta Ziedoņa dzeju un par privātpersonu ieguldījumu stratēģijām. Tulkojis literāros darbus no latgaliešu valodas uz angļu valodu.

Raksturvārdi: mašīnmācīšanās, mirstības riska prognoze, *XGBoost*, sabiedrības veselības dati, Latvija, sirds un asinsvadu slimību mirstības risks.

Ievads

Sabiedrības veselības aprūpes un medicīnas praksē ienāk dažādas prognožu veidošanas metodes, kuru pamatā ir mašīnmācīšanās modeļi[2]. Šajā darbā tiks pārbaudīts skaitlisks prognozēšanas algoritms *XGBoost*, izmantojot anonimizētu veselības aprūpes datu kopu, lai aprēķinātu sirds un asinsvadu slimību (turpmāk tekstā – CVD (no angļu *cardiovascular disease*)) izraisītās mirstības risku nākamo 12 mēnešu periodā, ņemot vērā tikai iepriekšējā gada datus. Dati iegūti no Latvijas Veselības aprūpes kvalitātes un efektivitātes publiskās monitorēšanas sistēmas un satur vairāk nekā 116 000 anonimizētu novērojumu personām, kuras bija dzīvas 2018. gada 1. janvārī un, iespējams, 2018. gadā izmantoja valsts apmaksātus veselības aprūpes pakalpojumus. Reģistrētie CVD izraisītie nāves gadījumi 2019. gadā bija 13 % no datu kopas. Cik autoram zināms, šis ir pirmais ar anonimizētiem Latvijas veselības aprūpes datiem pamatots pētījums par CVD mirstības riska aprēķināšanu nākamajā gadā.

[1] Pētījums sākts 2023. gada janvārī, kad autors, būdams ASV Fulbraita stipendiju programmas speciālists, viesojās Rīgas Stradiņa universitātes Sabiedrības veselības un epidemioloģijas katedrā. Autors izsaka dziļu pateicību sadarbības partnerei katedras profesorei Anitai Villerušai un katedras darbiniekiem.

[2] Sk., piemēram, https://www.wsj.com/articles/the-ai-will-see-you-now-5f8fba14?mod=hp_lista_pos1.

1. tabula. **Datu kopas mainīgo aprakstošā statistika, 2018. gads**

	Vecums	Ambulatorās vizītes	Atprečotās receptes	Stacionārās vizītes	NMPD izsaukumi
Aritmētiskais vidējais	53	9	12	2	2
Mediāna	53	6	8	1	1
Standartnovirze	21	9	11	2	2
N	116 393	116 393	116 393	116 393	116 393
Trūkstošo vērtību skaits	0	31 495	42 856	101 353	104 930

Sieviešu % (nav trūkstošu vērtību)	55%
2019. gada janvārī–decembrī CVD dēļ mirušo % (trūkstošu vērtību nav)	13%

Tabulā lietotais saīsinājums: NMPD – Neatliekamās medicīnas palīdzības dienests

Pētījuma mērķis

Pētījuma mērķis bija pārbaudīt, cik labi populāra datu zinātnes tehnika *XGBoost* var aplēst sirds un asinsvadu slimību izraisītās mirstības iespējamību turpmākajā 12 mēnešu periodā, izmantojot anonimizētus iepriekšējā gada veselības aprūpes datus.

Pētītā populācija un dati

Pētīto iedzīvotāju populācija ir 2018. gada 1. janvārī Latvijā dzīvojošas pieaugušas personas, kuras izmantoja vai varēja izmantot Latvijas valsts apmaksātos veselības aprūpes pakalpojumus, kuru pieejamību nodrošina Nacionālais veselības dienests. Bija pieejami anonimizēti dati par valsts apmaksātajām ambulatorajām vizītēm un uzturēšanos stacionārā, atprečotajām valsts apmaksātajām receptēm un ātrās palīdzības izsaukumiem. Datos tika iekļauts arī personu dzimšanas gads un dzimums, bet dati par ārstniecības personu apmeklējumiem bija nepilnīgi, jo ne visas vizītes apmaksā valsts. Nebija pieejami dati par privāti finansētu pakalpojumu izmantošanu, piemēram, tādu pakalpojumu izmantošanu, kurus apmaksā pacienti paši vai komerciālie veselības apdrošināšanas pakalpojumu sniedzēji.

Latvijas Veselības aprūpes kvalitātes un efektivitātes publiskās monitorēšanas sistēma šos datus uztur daudztabulu arhitektūrā, kas saistīta personas līmenī, un šim pētījumam darīja tos pieejamus, ievērojot spēkā esošos datu privātuma noteikumus (*Health Personalised Data Guidebook for Health Care Monitoring Datalink* (lumii.lv))[3]. Dati nav publiski pieejami, bet pieprasīt piekļuvi datiem var saskaņā ar minētajā tīmekļa vietnē uzdotajiem norādījumiem (http://med.oranzais.lumii.lv/index.html).

Modeļa izstrādei CVD mirstība tika izteikta kā binārs mainīgā lielums, kas norāda miršanu vai izdzīvošanu ar sirds un asinsvadu slimībām 2019. gadā. Visi prognozētāju mainīgie ņemti tikai no 2018. gada datiem. Tātad modelis tika apmācīts ar 2018. gada prognozētāju datiem, bet atkarīgā mainīgā dati, kas apzīmē CVD mirstību, ņemti no nākamā gada laikposma – no 2019. gada janvāra līdz decembrim.

Tālāk detalizētāka informācija par datu kopas izveidi.

1. Sākotnējā nejaušās izlases datu kopa: 2018. gada 1. janvārī Latvijā dzīvojošas 100 000 personas:
   1.1. vecums > 15,
   1.2. sistēmas dati par valsts apmaksātu medicīnas pakalpojumu izmantošanu no 2014. līdz 2021. gadam.

[3] Par ievada datu kopu izveidošanu pateicība Slimību profilakses un kontroles centra speciālistei Jolantai Skrulei un citiem šī centra darbiniekiem.

2. Šai nejaušajai izlasei bija maz (apmēram 160) akūtu CVD gadījumu (diagnozes kodi I21 un I22) 2019. gadā, tāpēc to papildināja ar papildu datiem:
   2.1. par personām, kuras 2019. gadā bija stacionētas ar akūtu CVD diagnozi (diagnozes kodi I21 un I22),
   2.2. ar zināmiem CVD (diagnozes kodi I00–I99) miršanas gadījumiem 2019. gadā.
3. Pēc rindu dublikātu noņemšanas datu kopā bija 116 393 personu aprakstošas datu rindas, no kurām 14 998 personas bija mirušas ar CVD (diagnozes kodi I00–I99) posmā no 2019. gada 1. janvāra līdz 2019. gada 31. decembrim.
4. Katrā rindā bija binārais CVD miršanas rādītājs un dati par personas dzimšanas gadu, dzimumu un valsts apmaksāto medicīnisko pakalpojumu izmantošanu, no kuriem četri tika izvēlēti par modeļa iespējamības pārbaudes prognozētājiem.

Iespējamās datu novirzes problēmas un modeļa piemērojamība

Aplēses liecina, ka Latvijas Nacionālā veselības dienesta norēķinu sistēmās nav reģistrēti līdz pat 42 % Latvijas veselības aprūpes izdevumu.[4] Tomēr lielākā daļa hospitalizācijas izdevumu ir valsts finansēti.[5] Tā kā CVD mirstību parasti reģistrē stacionārās ārstniecības iestādēs, varam secināt, ka šī modeļa pamatā ir galvenokārt pilnīgi iznākuma dati, bet salīdzinoši nepilnīgi prognozētāju dati. Diagnostikas kodu piešķiršanā pastāv arī iespējamas kodēšanas kļūdas vai nevienlīdzība starp iestādēm, kas var radīt statistisku troksni prognozētājos vai atkarīgā mainīgā definīcijā. Neraugoties uz šiem ierobežojumiem, modeļa rezultāti bija pārsteidzoši stabili, kā tālāk aprakstīts.

Jāatzīmē, ka šajā datu kopā nebija pārstāvēti Latvijas iedzīvotāji, kuri no 2014. līdz 2021. gadam par visiem saņemtajiem medicīniskajiem pakalpojumiem norēķinājās no saviem līdzekļiem vai ar privāti iegādātu veselības apdrošināšanas polisi un kuru nāves cēlonis 2019. gadā nebija sirds un asinsvadu slimības. Tādēļ šeit aprakstītais modelis nebūtu piemērojams iedzīvotāju segmentam, kura dati par medicīnas pakalpojumu samaksu netiek glabāti Latvijas Nacionālā veselības dienesta regulārajās darījumu apstrādes sistēmās[6].

Prognožu modeļa apraksts un izstrāde

Ievērojot standarta paraugpraksi prognožu modeļu izstrādē[7], sākuma datus nejauši sadalīja apmācības datos (70 %) un validācijas datos (30 %). Pēc tam *XGBoost* algoritms tika iedarbināts apmācību datos un pēc procedūras pabeigšanas piemērots validācijas datiem. Modeļa veiktspēju novērtēja validācijas datu kopā – datos, kas no algoritma tika atturēti modeļa izstrādes posmā.

XGBoost algoritms pieder *Gradient Boosting Machine* (GBM) mašīnmācīšanās modeļu grupai, ko ātruma un precizitātes dēļ plaši lieto klasifikācijas problēmu risināšanā.[8] GBM tiek veidots kā liels daudzslāņainu individuālo lēmumu koku ansamblis, kur katrs nākamais slānis cenšas izlabot sava priekšgājēja kļūdas. Atšķirībā no loģistikas regresijas modeļa *XGBoost* nav jutīgs pret trūkstošiem datiem prognozētāju mainīgajos un tamdēļ var strādāt ar datu matricu, kas nav pilnībā aizpildīta – trūkstošās vērtības tas vienkārši uzskata par atsevišķu mainīgo vērtību kategoriju. Vēl GBM priekšrocība ir tā, ka tas nepieprasa lineāru asociāciju starp prognozēšanai lietotajiem mainīgajiem (prediktoriem) un atkarīgo mainīgo, kā tas būtu ar tradicionālās statistikas regresijas modeli.

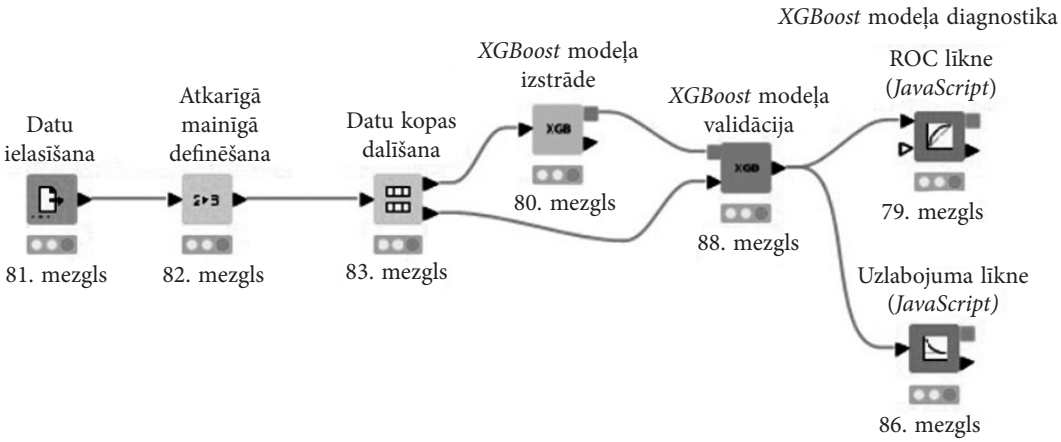
[4] OECD/ European Observatory on Health Systems and Policies 2019.

[5] Turpat.

[6] Datu kopā ir 29 020 personas, kurām zināms vecums un dzimums, bet nav informācijas, vai tās 2018. gadā izmantojušas valsts apmaksātos medicīnas pakalpojumus. 2019. gadā CVD mirstības incidence šajā grupā bija 7%. Iespējams, ka viņu vidū bija personas, kas pirms hospitalizēšanas ar CVD bija privātu veselības aprūpes iestāžu klienti.

[7] Wiemken, Kelley 2020.

[8] Chen, Guestrin 2016.



1. attēls. ***XGBoost*** **modeļa izstrādes darba plūsma**

GBM formulējums arī neprasa iepriekš norādīt, starp kuriem mainīgajiem aprēķināt mijiedarbības lielumu, kā tas būtu tradicionālās statistikas modelī. GBM algoritms konstruē lielu skaitu lēmumu koku, no kuriem katrs atsevišķi uztver kādu mijiedarbības kombināciju. Kad algoritms ir izpildījis savu darbu, katra datu rinda saņem iznākuma notikuma varbūtības aprēķinu, kas atbilst visu modeļa lēmumu koku galapunktu vidējai prognozei un ietver visas nozīmīgās mainīgo mijiedarbības.[9]

KNIME (*Konstanz Information Miner*) bezmaksas vizuālās datu analītikas platformā *XGBoost* modeļa izstrādes darba plūsma vienkāršākā formā vizualizējama tā, kā parādīts 1. attēlā.[10]

1. attēlā redzamo modeli parasti izmanto situācijās, kad mērķis ir prognozēt iespējamos nākotnes rezultātus, maksimāli pastiprinot pašreizējā stāvoklī esošo informāciju. GBM modelis savas sarežģītības dēļ nav piemērots, lai identificētu individuālos faktorus vai iemeslus, kāpēc dotā datu rinda no modeļa saņēma savu konkrēto iznākuma varbūtību. Tamdēļ sabiedrības veselības kontekstā šādu modeli ieteicams izmantot tikai kā konsultatīvu instrumentu – papildinājumu visai citai pacienta informācijai, ko ārsts izmanto savā parastajā praksē.[11]

Šajā priekšizpētes testā *XGBoost* algoritmam par katru novērojumu datu kopā tika doti seši prognozētāju mainīgie:

1) vecums 2018. gadā,
2) dzimums,
3) ambulatoro apmeklējumu skaits 2018. gadā,
4) 2018. gadā atprečoto recepšu skaits,
5) hospitalizāciju skaits 2018. gadā,
6) neatliekamās medicīniskās palīdzības izsaukumu skaits 2018. gadā.

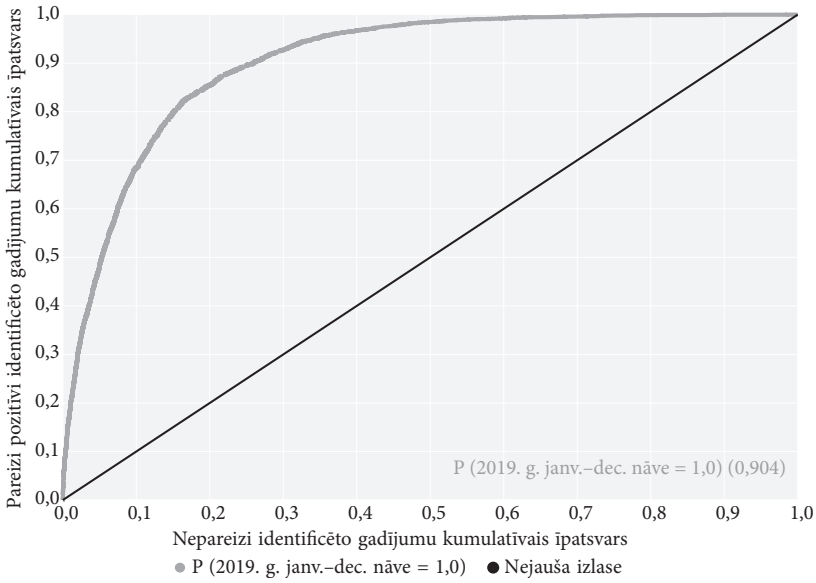
Modeļa rezultāti

Modeļa vispārējā precizitāte bija ļoti laba, ar validācijas datos konkordances (C) rādītāju 0,90 (datu zinātnes terminoloģijā to arī dēvē par AUC (*Area Under the Curve*) jeb "telpu zem līknes"). Salīdzinājumam: nesen pārkalibrējot tradicionālo uz kohortām balstīto SCORE CVD riska vērtēšanas modeli Eiropas

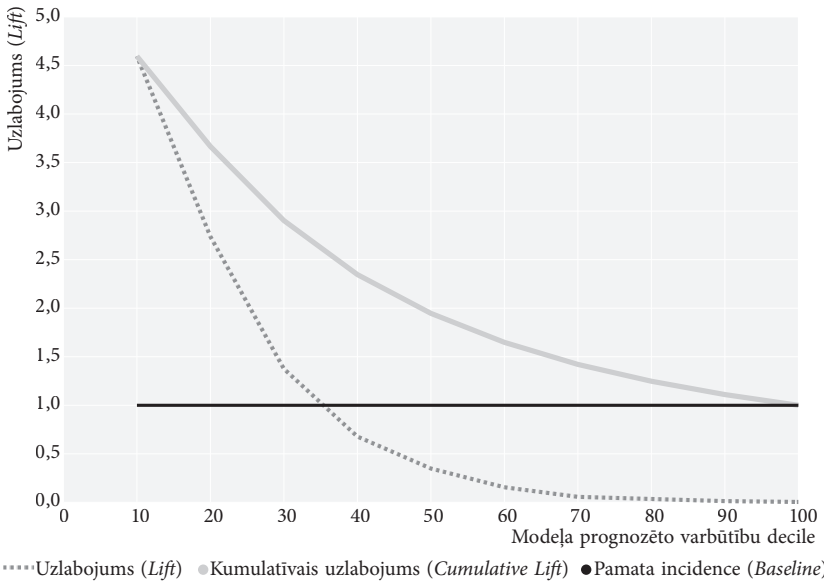
[9] *XGBoost* algoritmam ir vairāki uzstādījumi, dēvēti par hiperparametriem. Te aprakstīto modeli izstrādāja ar KNIME bezmaksas vizuālās datu analītikas platformas uzdotajiem hiperparametru sākuma lielumiem, piemēram, modeļa mācīšanās temps (*learning rate*) bija 0,3 un lēmumu koku maksimālais pakāpenisko sazarojumu skaits (*tree depth*) bija 6. Citus tehniskos modeļa algoritma parametrus, ieskaitot mainīgo svarīgumu, var saņemt, sazinoties ar autoru.

[10] Sk. https://www.knime.com/knime-analytics-platform. KNIME platformai nav latviešu valodas versijas.

[11] Rossello et al. 2019.



2. attēls. **Modeļa veiktspējas mērījuma ar ROC līkni rezultāti**
Piezīme: vertikālā ass – pareizi pozitīvi identificēto gadījumu kumulatīvais īpatsvars (*True Positive Rate*); horizontālā ass – nepareizi identificēto gadījumu kumulatīvais īpatsvars (*False Positive Rate*); pelēkā līkne rāda modeļa rezultātu, melnā līnija rāda hipotētisku nederīga modeļa iznākumu, kura prognozēm nav korelācijas ar gadījumu faktisko incidenci, respektīvi, (nederīgā modeļa) prognozēto varbūtības deciles līdzinās nejaušām (*Random*) izlasēm.



3. attēls. **Modeļa precizitātes mērījums ar uzlabojuma līknes (*Lift Chart*) metodi**
Piezīme: vertikālā ass – novēroto gadījumu incidences attiecība decilē pret vidējo visā datu kopā; horizontālā ass – modeļa prognozēto varbūtību deciles; pelēkā līkne rāda novēroto gadījumu incidences attiecību kumulatīvā izteiksmē, punktētā līkne rāda novēroto gadījumu incidences attiecību katrā atsevišķā decilē, melnā līnija ir gadījumu incidence visā datu kopā, dalīta pati ar sevi, tātad 1,0.

populācijās, tika iegūti C rādītāji no 0,67 līdz 0,81[12]. *XGBoost* modelis šajā formulējumā un šajā datu kopā uzrāda aptuveni 13 % uzlabojumu salīdzinājumā ar labāko publicēto C rādītāju no 2021. gada SCORE-2 modeļiem.[13]

2. attēlā redzama ROC (no angļu *Receiver Operating Characteristic*) līkne, kurā attēlota faktiskā kumulatīvā mirstības incidence dilstošā secībā pēc prognozētās varbūtības. Šis grafiks veidots no validācijas datiem.

Kā redzams 2. attēlā, aptuveni 70 % pareizi pozitīvi identificēto mirstības gadījumu konstatēti prognozēto varbūtību augstākajā decilē. Pirmajās divās decilēs to bija kumulatīvi 85 %. Datu zinātnieki šādu precizitātes līmeni uzskata par spēcīga modeļa indikatoru.[14]

Cits modeļa precizitātes skatījums ir "uzlabojumu līkne" (angļu *Lift Chart)*, kurā attēlota novērotā mirstības incidences attiecība dotajā decilē pret vidējo visā izlasē, deciles sarindojot dilstošā varbūtības secībā (sk. 3. attēlu). Šis attēls arī iegūts no validācijas datiem.

3. attēlā mēs redzam, ka personām validācijas datos prognozētās mirstības augstākajā decilē novērotā mirstības incidence (*Lift*) bija 4,5 reizes lielāka nekā datiem kopumā (*Baseline*), personām divās augstākajās decilēs kumulatīvi novērotais mirstības rādītājs (*Cumulative Lift*) bija 3,6 reizes lielāks nekā vidēji, utt.

Secinājumi un iespējamais lietojums

Rezultāti liecina, ka sabiedrības veselības sistēmas dati var sniegt patiešām noderīgas aplēses par īstermiņa CVD mirstības risku. Tradicionālās CVD riska aprēķina metodes, piemēram, SCORE, prognozē akūtu CVD notikumu iestāšanos (diagnosticēšanu) ilgtermiņā daudzu gadu laikā[15], savukārt šis pētījums norāda uz *XGBoost* algoritma spēju kvantitatīvi prognozēt CVD mirstības risku nākamajiem 12 mēnešiem.

Valsts veselības iestādes, kuras ir atbildīgas par veselības aprūpei paredzēto valsts budžeta līdzekļu sadalīšanu, varētu saskatīt šī modeļa izmantošanas iespējas, lai noteiktu, kurp novirzīt ierobežotos budžeta līdzekļus, lai tie sasniegtu visaugstākā riska iedzīvotāju segmentu. Veselības politikas plānotājiem šis modelis varētu noderēt slimnīcu un citu veselības aprūpes iestāžu kapacitātes plānošanā un slodzes līdzsvarošanā. Arī ģimenes ārsti varētu gūt labumu no modeļa, kas viņu pacientu sarakstu sarindotu pēc CVD mirstības riska, potenciāli identificējot tos, kuriem draud šis risks kādu ārsta nepamanītu faktoru dēļ. Ieguvēji varētu būt arī privātie veselības aprūpes pakalpojumu sniedzēji un apdrošināšanas sabiedrības, ja to datus varētu iekļaut modeļu izstrādes datu kopās.

Šī algoritma ieviešanai vismaz Latvijā nebūtu vajadzīgi papildu izdevumi, lai savāktu un apkopotu datus, un modeļu izstrāde varētu notikt atvērtā koda datu analītikas platformās. Lai gan modeļa īstenošana nekad nav bezmaksas, šie faktori varētu pozitīvi ietekmēt izmaksu un ieguvumu aprēķinus budžetu izstrādātājiem un lēmumu pieņēmējiem. Algoritma ieviešanai Latvijā būtu nepieciešami politiski, administratīvi, finansiāli un datortehnikas risinājumi.

Turpmāka izpēte

Turpmākos pētījumos varētu salīdzināt *XGBoost* algoritma rezultātus ar citiem mašīnmācīšanās algoritmiem, piemēram, mākslīgo neironu tīkliem. Būtu jāizvērtē šo algoritmu prognozējošā veiktspēja attiecībā uz atkarīgiem mainīgajiem, kas apzīmē citus slimības stāvokļus vai veselības rezultātus. Modelēšanas pieeja būtu jāapstiprina ar datiem pēc Covid-19 pandēmijas. Būtu jātestē arī detalizētāki prognozētāju mainīgie, piemēram, ambulatorās vizītes pēc diagnozes koda vai receptes pēc medikamentu veida. Algoritma prognozējošo veiktspēju varētu pētīt arī īsākos vai garākos laika intervālos. Daudz darba vēl priekšā.

[12] Hageman et al. 2021.

[13] Turpat.

[14] Sk. sīkāk: C-Statistic: Definition, Examples, Weighting and Significance. Pieejams: https://www.statisticshowto.com/c-statistic/.

[15] Hageman et al. 2021.

VĒRES

Chen, T.; Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August, 785–794. https://doi.org/10.1145/2939672.2939785.

C-Statistic: Definition, Examples, Weighting and Significance (bez datuma). Pieejams: https://www.statisticshowto.com/c-statistic/ (03.10.2023.).

Hageman, S.; Pennells, L.; Ojeda, F.; Kaptoge, S.; Kuulasmaa, K.; de Vries, T.; Xu, Z.; Kee, F.; Chung, R.; Wood, A.; McEvoy, J. W.; Veronesi, G.; Bolton, T.; Dendale, P.; Ference, B. A.; Halle, M.; Timmis, A.; Vardas, P.; Danesh, J.; Graham, I. ... Zhou, B. (2021). SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *European Heart Journal*, 42 (25), 2439–2454. https://doi.org/10.1093/eurheartj/ehab309.

OECD / European Observatory on Health Systems and Policies (2019) *Latvija: Valsts veselības profils 2019, State of Health in the EU*. Paris : OECD Publishing; Brussels : European Observatory on Health Systems and Policies. Pieejams: https://health.ec.europa.eu/system/files/2019-11/2019_chp_lv_latvian_0.pdf (29.09.2023.).

Rossello, X.; Dorresteijn, J. A. N.; Janssen, A.; Lambrinou, E.; Scherrenberg, M.; Bonnefoy-Cudraz, E.; Cobain, M.; Piepoli, M., F.; Visseren, F. L. J.; Dendale P. (2019) Risk prediction tools in cardiovascular disease prevention: A report from the ESC Prevention of CVD Programme led by the European Association of Preventive Cardiology (EAPC) in collaboration with the Acute Cardiovascular Care Association (ACCA) and the Association of Cardiovascular Nursing and Allied Professions (ACNAP). *European Journal of Cardiovascular Nursin*, 18 (7), 534–544. https://doi.org/10.1177/1474515119856207.

Wiemken, T. L.; Kelley, R. R. (2020) Machine Learning in Epidemiology and Health Outcomes Research. *Annual Review of Public Health,* 41 (1), 21–36. Pieejams: https://www.annualreviews.org/doi/10.1146/annurev-publhealth-040119-094437.

Summary

One year-ahead CVD mortality risk prediction from anonymized Latvian health care data records: testing feasibility of the XGBoost machine learning algorithm

Various predictive modeling methodologies utilizing machine learning models are making their way into the practice of public health and medicine. In this article a numerical forecasting method called the XGBoost algorithm is tested on a dataset of anonymized health care data records for the purpose of assessing cardiovascular disease (CVD) mortality risk in a future 12-month period, using only data from the previous year. The data originates from Latvia's Centre for Disease Prevention and Control and contains over 116 000 anonymized observations from persons who were living as of January 1, 2018 and may have utilized government-subsidized medical services in 2018. CVD case fatality in 2019 occurred among 13% of the sample. The machine learning mortality risk model was based on six input variables (age, sex, the number of out-patient visits and in-patient hospital stays, prescriptions filled, and ambulance calls) and produced an out-of-sample Concordance (AUC) measure of 0.90, which is comparable with the accuracy of cohort-based SCORE2 models recommended for use in Europe today. Potential model applications could be in public health risk quantification, future capacity and load balancing estimates for medical facilities, as well as patient stratification for individual physician practices. An earlier English language version of this article is available upon request from the author: usprudzs@chicagobooth.edu.

Keywords: machine learning, mortality risk prediction, *XGBoost*, public health data, Latvia, cardiovascular disease mortality risk.

Redakcijā saņemts: 10.04.2023.