



VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

FUNDAMENTINIŲ MOKSLŲ FAKULTETAS

INFORMACINIŲ TECHNOLOGIJŲ KATEDRA

Justas Grinaveckis

**KOMPIUTERINIO „BATTLE ROYALE“ ŽANRO ŽAIDIMO VEIKĖJŲ
IŠLIKIMO ANALIZĖ**

**SURVIVAL ANALYSIS OF PLAYERS IN VIDEO GAME OF GENRE BATTLE
ROYALE**

Baigiamasis magistro darbas

Informacinių technologijų studijų programa, valstybinis kodas 6211BX016

Duomenų gavybos technologijų specializacija

Informatikos inžinerijos studijų kryptis

Vilnius, 2020

Vilniaus Gedimino technikos universitetas Fundamentinių mokslų fakultetas Informacinių technologijų katedra		ISBN ISSN Egz. sk. Data-.....-.....
Antrosios pakopos studijų Informacinių technologijų programos magistro baigiamasis darbas 4		
Pavadinimas	Kompiuterinio „Battle Royale“ žanro žaidimo veikėjų išlikimo analizė	
Autorius	Justas Grinaveckis	
Vadovas	Dmitrij Šešok	
Kalba: lietuvių		
Anotacija Elektroninis sportas pastaraisiais metais tapo rimtu tradicinių sporto šakų konkurentu, surenkančiu milijonus dalyvių ir žiūrovų. Kompiuteriniai žaidimai generuoja didžiulius kiekius duomenų, kurių panaudojimas analizuoti įvairius aspektus, nors ir traukia vis didesnę susidomėjimą, tačiau išlieka gana mažas. Šiame darbe buvo atlikta žaidėjų išlikimo analizė, naudojant duomenis iš „Battle Royale“ žanro kompiuterinio žaidimo, kurio pagrindinis tikslas yra kuo ilgiau išlikti žaidime. Išlikimo analizei atlikti buvo pristatyti ir išbandyti keturi dažniausiai tyrimuose naudojami modeliai – Cox'o proporcingų rizikingumų, pagreito įvykio laiko, atsitiktinių išlikimo miškų bei neuroninių tinklų. Gerumui nustatyti naudoti atitikimo ir Brierio statistiniai įvertinimai bei prognozuojamųjų ir realiųjų įvykių vidutinė absoliutinė paklaida. Taip pat pateiktos įvertinimų kitimo laikui bėgant kreivės. Remiantis gautais rezultatais buvo nustatyta, jog tiksliausias iš pasirinktųjų buvo neuroninių tinklų modelis. Itin tikslūs rezultatai buvo gauti ir taikant Cox'o regresijos modelį. Naudojant tiksliausią modelį buvo sudaryta žaidėjų prognozuojamų rizikų histograma, kurią, suskirsčius į skirtingo rizikingumo grupes, pagal šias grupes sudarytas išlikimo tikimybių kreivių grafikas. Taip pat sudarytas svarbiausių žaidėjo atributų sąrašas, kuris rodo, jog įvairiais būdais nukeliamas atstumas bei su komandiniu žaidimu ir padarytos žalos kiekiu susiję veiksmų yra svarbiausi atributai, norint padidinti išlikimo laiką. Remiantis atributais pristatytos išlikimo kreivės, kurios nurodo žaidėjo išlikimo kreivės kitimą keičiant kiekvieną iš atributų. Darbą sudaro: įvadas, 4 pagrindiniai skyriai, rezultatai ir išvados, literatūros sąrašas. Darbo apimtis – 62 puslapiai teksto be priedų, 28 iliustracijos, 8 lentelės, 56 bibliografiniai šaltiniai. Atskirai pridedami darbo priedai.		
Prasminiai žodžiai: Elektroninis sportas, išlikimo analizė, mašininis mokymasis, Cox'o regresija		

Vilnius Gediminas Technical University Faculty of Fundamental Sciences Department of Information Technologies	ISBN ISSN Copies No. Date-.....-.....
Master Degree Studies Information Technologies study programme Master Graduation Thesis Title Survival Analysis of Players in Video Game of Genre Battle Royale Author Justas Grinaveckis Academic supervisor Dmitrij Šešok	
Thesis language: Lithuanian	
Annotation Esports, which has millions of players and spectators worldwide, became serious competitor to traditional sports in recent years. Video games generate huge amounts of data and while usage by researchers of it rises, it is still relatively low. In this paper, survival analysis of players using data from video game of genre Battle Royale was introduced. For survival analysis, four state-of-the-art models were used – Cox proportional hazards, accelerated failure time, random survival forests and neural networks. To measure goodness of model, concordance index, Brier score and average absolute error between actual and predicted number of occurrences, were introduced as well as time series curves for mentioned scores. Final results concluded that the most accurate model for given dataset of players in this paper was neural networks. Also, particularly good outcome was shown by Cox proportional hazards model. After selecting the most accurate model, three different risk groups were formed of estimated risk distribution of players and by using those groups, graph of survival curves were created for average of players in that group. Furthermore, created list of most important attributes for survivability showed that various types of traveled distance, playing in squad, causing more damage or being restored in-game attributes positively increases probability to survive longer in this video game. Lastly, survival curves for those attributes were presented to show effect on survival probability when varying selected attribute. Structure of paper: Introduction, four main chapters, results and conclusions, references. Thesis consist of: 62 p. of text without appendixes, 28 pictures, 8 tables, 56 bibliographical entries. Appendixes included separately.	
Keywords: Esports, survival analysis, machine learning, Cox regression	

Turinys

IVADAS.....	8
1. MAŠININIS MOKYMASIS IR JO TAIKYMAS	11
2. LITERATŪROS APŽVALGA	14
3. IŠLIKIMO ANALIZĖ.....	16
3.1. Išlikimo analizės metodai	19
3.1.1. Statistiniai metodai.....	20
3.1.1.1. Cox'o proporcingų rizikingumų modelis	21
3.1.1.2. Pagreitas įvykio laikas.....	23
3.1.2. Mašininio mokymosi modeliai.....	24
3.1.2.1. Atsitiktiniai išlikimo miškai	24
3.1.2.2. Dirbtiniai neuroniniai tinklai	26
3.2. Modelių gerumo vertinimas.....	28
3.2.1. Atitikimo rodiklis.....	28
3.2.2. Brierio rezultatas	30
4. TYRIMO REZULTATAI IR JŲ APTARIMAS.....	31
4.1. Žvalgomoji duomenų analizė.....	32
4.2. Modelių taikymas	38
4.3. Žaidėjų išlikimo analizė.....	52
REZULTATAI IR IŠVADOS	60
Priedai	67

Paveikslų sąrašas

1 pav. Elektroninio sporto žiūrovų skaičiaus kitimas 2012—2022 metais (Statista, 2020)	9
2 pav. Mašininio mokymosi užduočių schema	12
3 pav. Cenzūruotų duomenų pavyzdinė diagrama	17
4 pav. Priklausomybės tarp trijų dydžių – $S(t)$, $f(t)$, $F(t)$	18
5 pav. Išlikimo analizės metodų taksonomija (Wang, Yan, & Chandan, 2017).....	20
6 pav. Atsitiktinio miško sudarymo schema	25
7 pav. Neuroninio tinklo struktūra	27
8 pav. Išlikimo laikų rikiavimo $y_1 < y_2 < y_3 < y_4 < y_5$ variantai. Kairėje pusėje nurodytas scenarijus, kuomet nei vienas įvykis nėra cenzūruotas, o dešinėje pusėje du stebiniai cenzūruoti. Juodi taškai nurodo necenzūruotus stebinius, o raudoni – cenzūruotus (Wang, Yan, & Chandan, 2017)	29
9 pav. Žaidimo nuotrauka (TheTechGame, 2018).....	32
10 pav. Žaidėjų pabaigos įvykių pasiskirstymo grafikas	35
11 pav. Kaplan-Meier išlikimo kreivė pagal mėlynosios zonos atributą.....	37
12 pav. Kaplan-Meier išlikimo kreivė pagal žaidimo tipo atributą	38
13 pav. Cox'o modelio Brierio įvertinimų kitimo kreivė	39
14 pav. Realus ir prognozuojamas įvykių skaičius tyrimo metu naudojant Cox'o regresiją.....	41
15 pav. Cox'o regresijoje didžiausią įtaką žaidėjo rizikai turintys atributai.....	42
16 pav. Duomenų kvantilių grafikai pasirinktiems pagreitinto įvykių laiko modelio skirstinių tipams	44
17 pav. Brierio rezultatas naudojant pagreitinto įvykio laiko modelį.....	45
18 pav. Realus ir prognozuojamas įvykių skaičius tyrimo metu, naudojant pagreitinto įvykio laiko modelį.....	46
19 pav. Penki didžiausią įtaką pagreitintam įvykio laikui turintys atributai.....	47

20 pav. Atsitiktinių išlikimo miškų Brierio įvertinimų grafikas	48
21 pav. Atsitiktinių išlikimo miškų modelių prognozuojamų įvykių skaičiaus grafikas	49
22 pav. Brierio rezultatų grafikas naudojant neuroninių tinklų modelį	51
23 pav. Realų ir prognozuojamų įvykių kiekio grafikas naudojant neuroninių tinklų modelį	51
24 pav. Žaidėjų prognozuojamo rizikingumo histograma	53
25 pav. Skirtingo rizikingumo žaidėjų išlikimo kreivių grafikas.....	54
26 pav. Išlikimo tikimybių kreivės kintant atributui dist_on_foot.....	55
27 pav. Išlikimo tikimybių kreivės kintant atributui dist_on_vehicle	56
28 pav. Išlikimo tikimybių kreivės kintant atributui squad	57
29 pav. Išlikimo tikimybių kreivės kintant atributui groggy	58
30 pav. Išlikimo tikimybių kreivės kintant atributui dmg_per_distance	58

Lentelių sąrašas

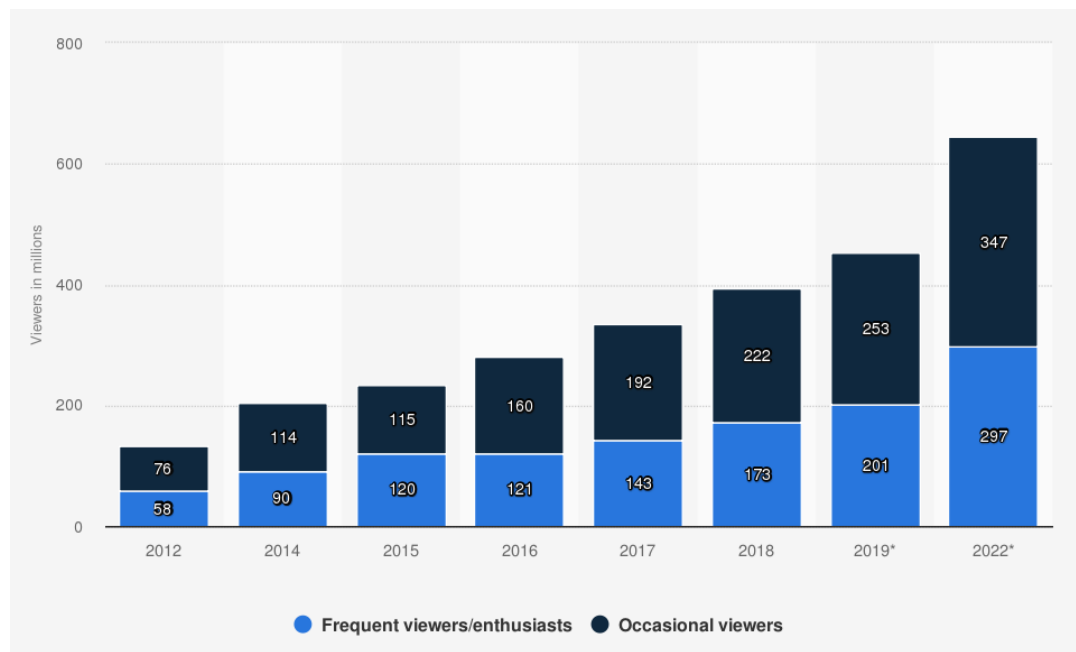
1 lentelė. Naudojamų pagrindinių duomenų atributų dalinė struktūra.....	33
2 lentelė. Svarbiausi atributai prognozavime naudojant Cox'o regresijos modelį.....	41
3 lentelė. Modelių tikėtinumų logaritmai.....	45
4 lentelė. Svarbiausi atributai prognozavime naudojant pagreitinto įvykio laiko modelį.....	46
5 lentelė. Svarbiausi atributai naudojant permutaciją Atsitiktinių išlikimo miškų modelyje	49
6 lentelė. Neuroninio tinklo parametrai	50
7 lentelė. Svarbiausi atributai naudojant neuroninius tinklus	52
8 lentelė. Išlikimo analizės modelių taikymo rezultatai.....	53

IVADAS

Išlikimo analizė (angl. *survival analysis*) yra viena seniausių statistikos atšakų, kurios pradžia yra skaičiuojama nuo septyniolikto amžiaus (Hacking, 2006). Paprasta išlikimo lentelių (angl. *life-table*) metodologija, skirta prognozuoti išlikimo vertinimui turint duomenis su kairiniu (angl. *left censoring*) arba dešiniu cenzūravimu, buvo žinoma ir tobulinama tuo pat metu. Aštuonioliktame amžiuje mokslininkai Abraomas de Muavras (*Abraham de Moivre*) ir Johanas Heinrichas Lambertas (*Johann Heinrich Lambert*) analizavo ir tobulino parametrizuotus išlikimo analizės metodus, o Danielius Bernulis (*Daniel Bernoulli*) tyrinėjo įvairių aspektų rizikas, keičiančias paskiepytų nuo raupų žmonių gyvenimo trukmę. Šie ir panašūs tyrimai bei gauti rezultatai leido susidomėjimui šia sritimi sparčiai augti (Andersen & Keiding, 1998).

Išlikimo arba laiko iki įvykio (angl. *time-to-event*) analizė yra statistikos šaka, nagrinėjanti laikotarpį iki prognozuojamo įvykio. Seniau, šio tipo analizė buvo naudojama tik medicinoje, norint prognozuoti laiką iki paciento mirties, priklausomai nuo įvairių veiksnių, arba pacientų proporcija, kuri gali išgyventi ilgiau nei tam tikras, nustatytas, laikas. Tačiau laikui bėgant kitos mokslo šakos taip pat susidomėjo išlikimo analizės metodais, o tai paskatino naujų mokslo šakų atsiradimą: sociologijos įvykių istorijos analizę (angl. *event history analysis*), ekonomikos įvykių trukmės analizę ir patikimumo (angl. *reliability*) analizę inžinerijoje. Problemos, kurias bandoma spręsti, pavyzdžiai yra likusio naudingo įrengimo darbo laiko prognozavimas, galimo klientų pasitraukimo iš įvairaus tipo kompanijų laiko prognozavimas arba likusio laiko iki paciento išgyvenimo taikant įvairius klinikinius tyrimus. Nepaisant minėtųjų, dar viena sritis, kur išlikimo analizė gali būti naudinga - elektroninis sportas (angl. *esports*).

Bendrajai prasme minint elektroninį sportą, omenyje yra turima tarpusavio konkuravimą žaidžiant vaizdo žaidimus, kurį stebi didelis skaičius žmonių (Schubert, 2016). „Statista“ (Statista, 2019) prognozuoja, jog 2022 metais elektroninio sporto entuziastų ir žiūrovų bus apie 660 milijonų.



1 pav. Elektroninio sporto žiūrovų skaičiaus kitimas 2012—2022 metais (Statista, 2020)

Dėl sparčiai augančio populiarumo ir dėl šios priežasties didėjančio turimų, su žaidimais susijusių, duomenų kiekio, elektroninis sportas tampa vis labiau populiarėjančiu tyrimų lauku, kuris pavadintas „elektroninio sporto analitika“ (angl. *esports analytics*). Tyrėjas Matas Šubertas (Schubert, 2016) jį įvardina kaip „procesą, naudojant su elektroniniu sportu susijusius duomenis įvairių reikšmingų modelių ir tendencijų radimui bei šių radinių naudojimą tolimesniame jais pagrįstų sprendimų kūrimo kelyje“. Nagrinėjamos temos išsidėsčiusios gana plačiai, pradedant marketingu (Lewis & Wu, 2018), žaidėjų įgūdžių prognozavimu (Aung, et al., 2018), svarbiausių momentų prognozavimu pagal žiūrovų reakcijas (Fu, Lee, Bansal, & Berg, 2017) baigiant komandos nugalėtojos prognozavimu (Makarov, 2017). Nepaisant jau minėtų temų, maža su elektroniniu sportu susijusių tyrimų dalis tiria žaidimo „vidinius“ duomenis – žaidimo eigos kitimą bėgant laikui. Tokie tyrimai leidžia pažvelgti į galimą didžiosios dalies žaidėjų taktiką bei padėti naujiems ar esamiems žaidėjams padidinti galimybės užimti aukštesnę vietą aptariant pagrindinius išgyvenimui darančius įtaką veiksnius.

Šiame darbe išlikimo analizė buvo taikoma nagrinėjant populiaraus visame pasaulyje žaidimo „PlayerUnknown’s Battlegrounds“, priskiriamo „Battle Royale“ žanrui, duomenų rinkinį. Jis buvo sudarytas naudojant atsitiktiniu būdu pasirinktų 200 žaistų žaidimų pakartojimų failų, kurie sudaryti iš žaidimo metu įvykusių įvairaus tipo įvykių. Analizei atlikti buvo pasirinktas vienas iš keturių, dažniausiai tokių uždavinių sprendimui naudojamų, modelių pagal tam tikrus statistinius įvertinimus.

Darbo objektas – „Battle Royale“ žanro vaizdo žaidimo „PlayerUnknown’s Battlegrounds“ atkartojimų failai.

Darbo tikslas – naudojant statistinius bei mašininio mokymosi metodus atlikti žaidime dalyvavusių žaidėjų išlikimo analizę.

Siekiant įgyvendinti šį tikslą, buvo keliami tokie **uždaviniai**:

- Pateikti išlikimo analizės tyrimų, susijusių su elektroniniu sportu, informacinių šaltinių apžvalgą bei aprašyti modeliavimui naudojamus išlikimo analizės metodus;
- Sukurti duomenų rinkinį, panaudojant žaidimų pakartojimų failus;
- Išnagrinėti gautus duomenis ir atlikti pažintinę duomenų analizę;
- Pritaikyti išlikimo analizės metodus bei pasirinkti geriausią pagal statistinius įvertinimus;
- Pasirinktu modeliu atlikti žaidėjų išlikimo analizę;
- Pateikti ir aptarti gautus rezultatus.

Šio baigiamojo darbo tyrimo dalies atlikimui pasirinkta naudoti „Python“ programavimo kalba bei jos pagrindu sukurtos bibliotekos ir paketai.

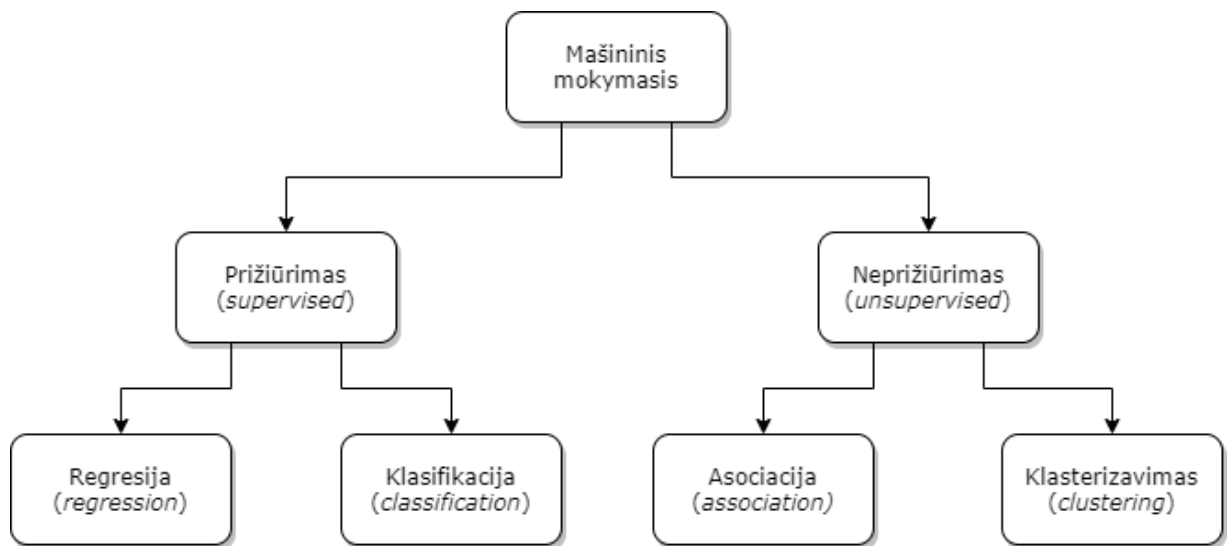
1. MAŠININIS MOKYMASIS IR JO TAIKYMAS

Dirbtinio intelekto (angl. *artificial intelligence*) įtaka šiais laikais daugelyje gyvenimo sričių yra neįtikėtinai didelė. Pradedant nuo finansų sektoriaus, kuriame siekiama automatizuoti investavimo procedūras baigiant žaidimų industrija, kurioje kuriami modeliai, gebantys atitikti realų kartu žaidžiantį žmogų. Viena pagrindinių dirbtinio intelekto užduočių yra sumažinti žmogaus darbo krūvį. Pasak Ši Zhonghzi (*Shi Zhonghzi*), savo knygoje „Advance Artificial Intelligence“ dirbtinį intelektą įvardija kaip mokslo ir inžinerijos dalį, kuri kopijuojanti, plečianti bei auginanti žmonių intelektą, naudojant dirbtines priemones ir metodiką, siekia sukurti protingas mašinas (Zhongzhi, 2011).

Pagrindiniai dirbtinio intelekto iššūkiai – žinių reprezentacija bei pagrindimas (angl. *representation and reasoning*), planavimas, mokymasis, natūralios kalbos apdorojimas (angl. *natural language processing*), suvokimas, judesys ir manipuliacija (angl. *motion and manipulation*), socialinis intelektas, bendrasis intelektas (Pannu, 2015). Kiekvienas iš jų yra savaip svarbūs, tačiau šiame darbe toliau bus nagrinėjami ir taikomi metodai, priklausantys pagrindinei mokymosi iššūkio sričiai – mašiniam mokymuisi (angl. *machine learning*).

Spartus duomenų rinkimo, apdorojimo bei didžiųjų duomenų (angl. *big data*) modelių tobulėjimas, palengvėjusi galimybė kaupti bei stebėti plataus spektro duomenis sudarė terpę vis labiau didėjančiam mašininio mokymosi srities metodų nagrinėjimui (Najafabadi, Villanustre, & Khoshgoftaar, 2015). Pats mašininis mokslas, kaip dirbtinio intelekto šaka, jungia mokymosi iš turimų duomenų problemą su bendrą gautų išvadų konceptu (Wang, Yan, & Chandan, 2017). Kiekvienas mokymosi procesas susideda iš dviejų dalių: a) sistemoje esančių nežinomų priklausomybių įvertinimas turimuose duomenyse, b) gautų priklausomybių taikymas pranašaujant norimą kintamąjį.

Mašininis mokymasis yra skaidomas į dvi skirtingo tipo užduotis – prižiūrimas (angl. *supervised*) ir neprižiūrimas (angl. *unsupervised*) mokymasis.



2 pav. Mašininio mokymosi užduočių schema

Prižiūrimas mokymasis yra taikomas kuomet turimos įėjties/išeities poros ir pagal jas gauta funkcija atranda galimas nežinomas naujų duomenų išeities vertes (Kourou, 2015). Tuomet naujiems duomenims yra prognozuojamos išvesties vertės. Šie duomenys yra lyginami su validacijos duomenyse buvusiais išvesties elementais, siekiant išsiaiškinti modelio efektyvumą. Ieškant šio tipo užduoties atitikmenų žmonių psichologijoje, ji turi paralelių su konceptų mokymosi (angl. *concept learning*) užduotimi. Šio mokymosi tipo užduotys toliau gali būti grupuojamos į regresijos ir klasifikacijos problemas.

- Regresija – šis tipas yra taikomas tuomet, kai išeities kintamasis priklauso realiųjų skaičių aibei, pavyzdžiui – „svoris“, „pinigų kiekis“ ir t.t..
- Klasifikacija – šio tipo problema iškyla tuomet, kai išeities kintamasis yra kategorija, pavyzdžiui - spalva yra „mėlyna“ arba „raudona“, skiriamas „I“, „II“ arba „III“ tipo gydymas.

Dažniausiai naudojami prižiūrimo mašininio mokymosi algoritmai: tiesinė regresija – regresijai, atsitiktiniai miškai (angl. *random forests*) – klasifikacijai ir regresijai, palaikymo vektorių mašinos (angl. *support vector machines*) – klasifikacijai (Browlee, 2016).

Neprižiūrimo tipo mokymosi pagrindinė užduotis yra pagal turimus duomenis rasti nematomus, tarp turimų duomenų esančius modelius bei priklausomybes neturint išeities duomenų. Šio tipo metodai dažnai taikomi norint klasifikuoti klientus į klasterius (Russell & Norvig, 2010), atrinkti svarbius

duomenų atributus (Malinowski, 2018). Smulkiau šis tipas yra skaidomas į dviejų rūšių problemas – klasterizavimą (angl. *clustering*) ir asociaciją.

- Klasterizavimas – šio tipo problema yra gaunama norint atrasti būdingas grupes duomenyse. Tokios grupės galėtų būti „pirkėjai, pagal skirtingas pirkimo galias“.
- Asociacija – norima surasti priklausomybes, kai tam tikra grupė nusako didžiąją dalį turimų duomenų, pavyzdžiui – pirkėjai, perkantys prekę „ X “, taip pat perka ir prekę „ I “.

Populiarūs algoritmai, sprendžiantys šio tipo problemas – k -vidurkių (angl. *k-means*) – klasterizavimui, „Apriori“ algoritmas – asociacijoms.

Per pastaruosius metus, dėka vis labiau tobulėjančių, mašiniam mokymuisi skirtų algoritmų, kurie leidžia modeliuoti netiesines priklausomybes bei gautų prognozių tikslumo, mašininis mokymasis taikomas vis didesniame kiekyje praktinių sričių. Vaizdo, garso atpažinimas, finansinės aplikacijos, sveikatos priežiūra – tai tik kelios iš jų. Šiame darbe nagrinėjamoje išlikimo arba laiko-iki-įvykio analizėje, kuri ir kilo iš klinikinių tyrimų sveikatos priežiūros srityje, mašininis mokslas taip pat yra taikomas, vis labiau didėjant duomenų kiekiui bei siekiant sudaryti kuo tikslesnes prognozes.

Išlikimo arba laiko iki įvykio analizė skiriasi nuo tradicinio klasifikacijos ar regresijos užduočių ir išpopuliarėjo dėl atsiradusio poreikio ne tik analizuoti atsitikusių įvykių laikų pasiskirstymą ar priklausomų kintamųjų įtaką įvykio laikui, bet ir norint pagal turimus duomenis prognozuoti dar nežinomus laikus iki įvykio kitiems dalyviams. Tyrėjai naudodami bendrai tiek statistinius išlikimo analizės metodus kartu su mašininio mokymosi algoritmais, sukūrė efektyvesnius ir labiau išlikimo analizei skirtus laiko iki konkretaus įvykio prognozavimo metodus (Fadnavis, 2019).

2. LITERATŪROS APŽVALGA

Peržvelgtoje literatūroje matyti, jog bendruoju atveju pagrindinis išlikimo analizėje taikomas metodas – Cox'o proporcingų rizikingumų modelis. Populiariausi mašininio mokymosi metodai sprendžiant išlikimo analizės problemas yra neuroniniai tinklai (angl. *neural networks*), atsitiktiniai miškai (angl. *random forests*). Taip pat išlikimo analizėje vis populiarnesnis tampa bendrai populiarumo sulaukęs gilusis mokymasis (angl. *deep learning*). Populiariausioje mokslinių straipsnių duomenų bazių „Google Scholar“ atliktoje paieškoje naudojant frazes „*deep learning*“, „*survival analysis*“ nustatyta, jog per pastaruosius 4 metus (2015 – 2019 m.) parašyta 25 100 susijusių straipsnių.

Vieni pirmųjų pritaikiusių mašininį mokymąsi išlikimo analizėje – Deividas Faraggi (*David Faraggi*) ir Ričardas Simonas (*Richard Simon*) (Simon & Faraggi, 1995). Savo tyrime jie pakeitė Cox'o proporcingų rizikingumų modelyje (Cox, 1972), esantį tiesinį prognozatorių (angl. *predictor*) į neuroninio tinklo vieno iš paslėptųjų sluoksnių daugiasluoksnį perceptroną (angl. *multilayer perceptron*). Džaredas Katzmanas (*Jared Katzman*) savo darbe (Katzman, et al., 2017) pasiremdamas minėtuoju modeliu bei naudodamas gilųjų mokymąsi (angl. *deep learning*) sukūrė naują modelį pavadinimu „DeepSurv“¹, kuris pagal atitikimo (angl. *concordance*) indekso rezultatą yra pranašesnis negu ankstesni, kurie naudojo Cox'o regresijos modelį. Havardas Kvammas (*Havard Kvamme*) su kolegomis (Kvamme, Borgan, & Scheel, 2019) pratęsė Farragi ir Simon darbą ir pasiūlė neuroninių tinklų modelį, kuriame nuostolių (angl. *loss*) funkcijos rezultatai neprastėja net esant ir dideliu duomenų kiekiui bei leidžia tiesinius ir netiesinius Cox'o regresijos modelio tęstinumus (angl. *extension*). Naudojant Gleno Brierio (*Glenn Brier*) integruotą balų sistemą (Brier, 1950) palyginimui su kitais (DeepSurv, DeepHit² ir kt.) Kvamme pasiūlytas modelis rodė geriausius rezultatus – atitinkamai 0.213, 0.214 ir 0.223 (šioje sistemoje kuo arčiau skaičius yra 0 ribos, tuo tiksliau modelis prognozuoja).

Egilas Martinssonas (*Egil Martinsson*) savo tyrime (Martinsson, 2017) pristatė modelį, kuriame laikas iki įvykio yra prognozuojamas naudojant pasikartojančius (angl. *recurrent*) neuroninius tinklus. Modelyje prognozuojamas laiko iki įvykio pasiskirstymas yra panašus į tęstinį (angl. *continuous*) arba diskretųjį (angl. *discrete*) Weibullio pasiskirstymą, kurio parametrai yra pasikartojančio neuroninio

¹ <https://github.com/jaredleekatzman/DeepSurv>

² <https://github.com/chl8856/DeepHit>

tinklo išeities vertės. Šis modelis išbandytas naudojant didelį dimensijų skaičių turinčius duomenis apie reaktyvinių lėktuvų laiką iki gedimo (C-MAPSS) ir gauti tikslesni rezultatai lyginant su kitais metodais.

Peržvelgus literatūrą apie išlikimo analizės metodų taikymą elektroninio sporto žaidimų duomenų analizėje, nors ir duomenų kiekis yra didelis, daug pavyzdžių nebuvo rasta. Dažniausiai apsiribojama nugalėtojų prognozavimu (Hodge, et al., 2017), o išlikimo analizės metodai naudojant tiriant žaidėjų galimą pasitraukimą iš aktyvaus dalyvavimo žaidime (Demediuk, et al., 2018). Arčiausiai šiame darbe analizuojamos metodologijos buvo tyrėjas Adamas Katona (*Adam Katona*) su kolegomis (Katona, 2019), kurie sukūrė modelį, kuris prognozuoja žaidėjo tikimybę mirti per ateinančias 5 sekundes, naudojant populiarus žaidimo „Dota 2“ duomenis. Modelis sudarytas naudojant pasikartojančius neuroninius tinklus (angl. *recurrent neural networks*). Nepaisant to, jog mirčių įrašai duomenyse buvo pakankamai reti (1% visų duomenų), pasiekta pakankamai gera 37.7 proc. precizija (angl. *precision*) bei 72.5 procentų jautrumas (angl. *recall*).

Išanalizavus susijusią mokslinę literatūrą galima pastebėti, jog problematika yra gana aktuali šių dienų moksle, tačiau nepaisant didelio kiekio turimų duomenų elektroniniame sporte, jų išnaudojimas nėra toks didelis, ypač išlikimo analizėje. Kitame skyriuje bus supažindinama su išlikimo analizės teorine dalimi, pristatomi statistiniai bei darbe naudojami mašininio mokymosi modeliai, skirti išlikimo analizei.

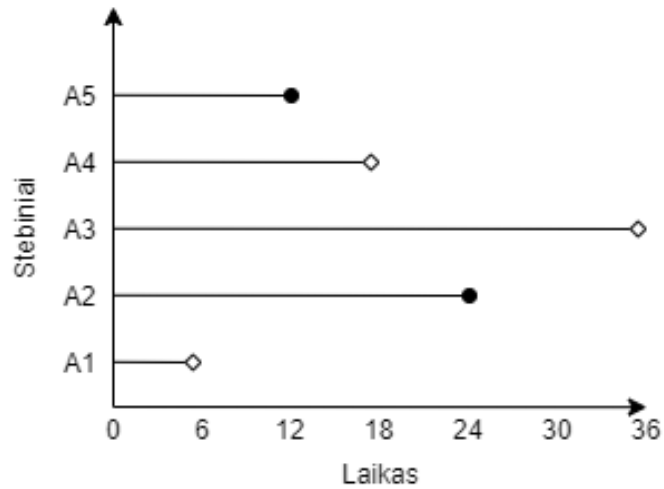
3. IŠLIKIMO ANALIZĖ

Gana ilgą laiką išlikimo analizė buvo populiari tik medicinos srityje, klinikiniuose tyrimuose, siekiant išsiaiškinti kurie pacientų gyvenimo trukmės ypatumus, pavyzdžiui, kurie gydymo tipai prailgina gyvenimo trukmę. Pastaruoju metu išlikimo analizė susilaukė populiarumo ir biostatistikoje, paslaugų sektoriuje (Stepanova & Lyn, 2002), (Andreeva, 2006), (Tang, Harvey, Dorsay, Jiang, & Rathbone, 2007), inžinerijoje, draudimo ir kitose srityse. Išlikimo analizė turi specifinių atributų, kurie išskiria šio tipo problemas nuo ankščiau minėtų regresijos ir klasifikacijos problemų. Vienas iš tokių – cenzūruoti duomenys.

Atliekant išlikimo analizę dažnai yra susiduria su situacija, kuomet nėra žinomas kiekvieno iš stebimų įvykių atsitikimo laikas visu stebėjimo laikotarpiu. Tai paprastai atsitinka esant palyginti trumpam stebėjimo laikotarpiui (pavyzdžiui gyvybės draudimo klientai atlieka mokėjimus dešimtmečiais, kuomet analizė yra vykdoma trumpesnę laiką) arba įvykus kitiems, nesusijusiems su tyrimu, įvykiams (tiriamas pacientų išgyvenimo laikotarpis taikant skirtingus vaistinius preparatus, tačiau stebėjimo laikotarpiu yra mirštama dėl kitų priežasčių). Toks duomenų praradimo konceptas yra vadinamas cenzūravimu (Klein, Houwelingen, Ibrahim, & Scheike, 2014).

Įvykio laikas yra žinomas tik tiems atvejams, kuriems jis įvyko būtent stebėjimų laikotarpiu. Kitiems atvejams, kurie dėl įvairiausių priežasčių neturi tikslaus įvykio laiko, yra priskiriamas cenzūravimo laikas, kuris gali būti stebėjimų pabaigos laikas arba paskutinio žinomo apsilankymo/įvertinimo laikas. Tai reiškia, jog kiekvienam stebiniui gali būti priskirtas arba įvykio arba cenzūruotas laikas, bet ne abu kartu. Tarkime, jeigu $y_i = \min(T_i, C_i)$, kur y_i – įvykio i laikas, T_i – išgyvenimo laikas, C_i – cenzūruotas laikas, toks atvejis yra vadinamas dešiniuuoju cenzūravimu, kuris yra dažniausiai pasitaikantis scenarijus nagrinėjant šio tipo problemas (Wang, Yan, & Chandan, 2017).

Norint aiškiau suprasti stebėjimų cenzūravimo esmę išlikimo analizėje, 2 pav. žemiau iliustruoja paprastą pavyzdį. Turimi penki stebiniai (A1-A5), kurių stebėjimo laikas yra 36 mėnesiai. Pilnaviduriu rutuliu gale pavaizduotiems stebiniams (A2 bei A5) tam tikras įvykis įvyko, o kitiems likusiems, kurių gale yra pavaizduotas rombas, įvykis neįvyko. Verta atkreipti dėmesį į tai, jog stebiniui A3 įvykis neįvyko per visą stebėjimo laikotarpį ir todėl nėra aišku, ar jis įvyko po to, o įvykiai A1, A4 yra cenzūruoti dar neįvykus stebėjimo pabaigai, nes jų stebėjimai buvo nutraukti dėl neaktyvumo arba kitų priežasčių.



3 pav. Cenzūruotų duomenų pavyzdinė diagrama

Nagrinėjant duoto pavyzdžio problemą stebiniui i , jis yra aprašomas trijų kintamųjų kombinacija (X_i, y_i, δ_i) , kur X_i – stebinio atributų vektorius ($X_i \in R^{1 \times P}$), δ_i – binarinis kintamasis, kuris esant necenzūruotam stebiniui yra lygus 1, kitu atveju – 0, o y_i nurodo užfiksuotą laiką. Jis lygus išlikimo laikui T_i necenzūruotam stebiniui ir C_i cenzūruotam:

$$y_i = \begin{cases} T_i, & \delta_i = 1 \\ C_i, & \delta_i = 0 \end{cases} \quad (1)$$

Reikia atkreipti dėmesį, jog išgyvenimo laikas T_i yra neaktualus cenzūruotiems stebiniams, nes kaip ir minėta anksčiau tokie stebiniai per turimą stebėjimo laiką neįvyko numatomas įvykis. Turint minėtuosius parametrus, pagrindinė išlikimo analizės užduotis yra prognozuoti laiką iki įvykio T_j naujiems stebiniams j , su stebinio turimu atributų rinkiniu X_j . Svarbu pažymėti ir tai, jog naujas prognozuojamas laikas iki įvykio turi būti neneigiamas ir tęstinis.

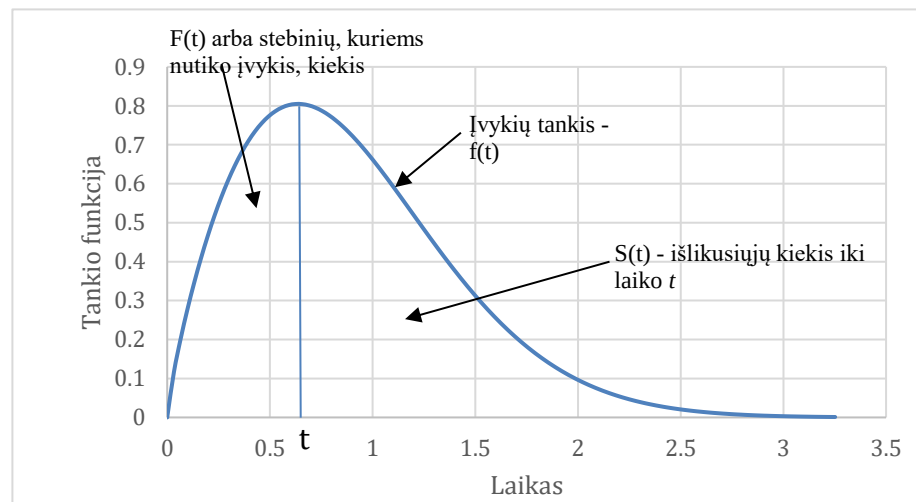
Vienas svarbiausių konceptų išlikimo analizėje yra išlikimo (angl. *survival*) ir intensyvumo (angl. *hazard*) funkcijos. Tarkime, jog T yra laikas iki kol klientas nutrauks sutartį ar pacientas pasijus geriau naudojant tam tikrą vaistą. Išlikimo funkcija yra užrašoma formule:

$$S(t) = P\{T \geq t\}, \quad \text{kur } t \geq 0 \quad (2)$$

Išlikimo funkcija yra tikimybė, kad konkrečiam stebiniui įvykis įvyks vėliau, negu nurodytas laikas t . Priešingai negu prieš tai minėtajai funkcijai, suminė pasiskirstymo funkcija $F(t)$ nurodo tikimybę, jog tam tikras įvykis įvyks prieš laiką t yra užrašoma funkcija $F(t) = 1 - S(t)$

Rizikingumo funkcija $h(t)$ parodo tikimybę įvykiui įvykti laiko momentu t , kai žinoma, jog iki to momento jis dar nėra įvykęs ir nenurodo galimybės ar tikimybės pačio įvykio. Tai yra tankio (įvykių dažnumas per laiko vienetą) funkcijos, kuri diskrečiuoju atveju užrašoma $f(t) = [F(t + \Delta t) - F(t)] / \Delta t$, o tęstiniu $f(t) = \frac{d}{dt} F(t)$ ir išlikimo funkcijos santykis. Jis užrašomas tokia lygtimi:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{F(t + \Delta t) - F(t)}{\Delta t \cdot S(t)} = \frac{f(t)}{S(t)} \quad (3)$$



4 pav. Priklausomybės tarp trijų dydžių – $S(t)$, $f(t)$, $F(t)$

Rizikingumo $h(t)$ funkcija yra neneigiama funkcija kaip ir išlikimo funkcija $S(t)$. Skirtumas tarp jų yra tas, jog išlikimo funkcija yra mažėjanti bėgant laikui, o rizikingumo funkcija gali turėti keletą skirtingų formų. Tarkime, jog funkcija $f(t)$ gali būti užrašyta lygtimi:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \ln S(t) = -\frac{d}{dt} [\ln S(t)] \quad (4)$$

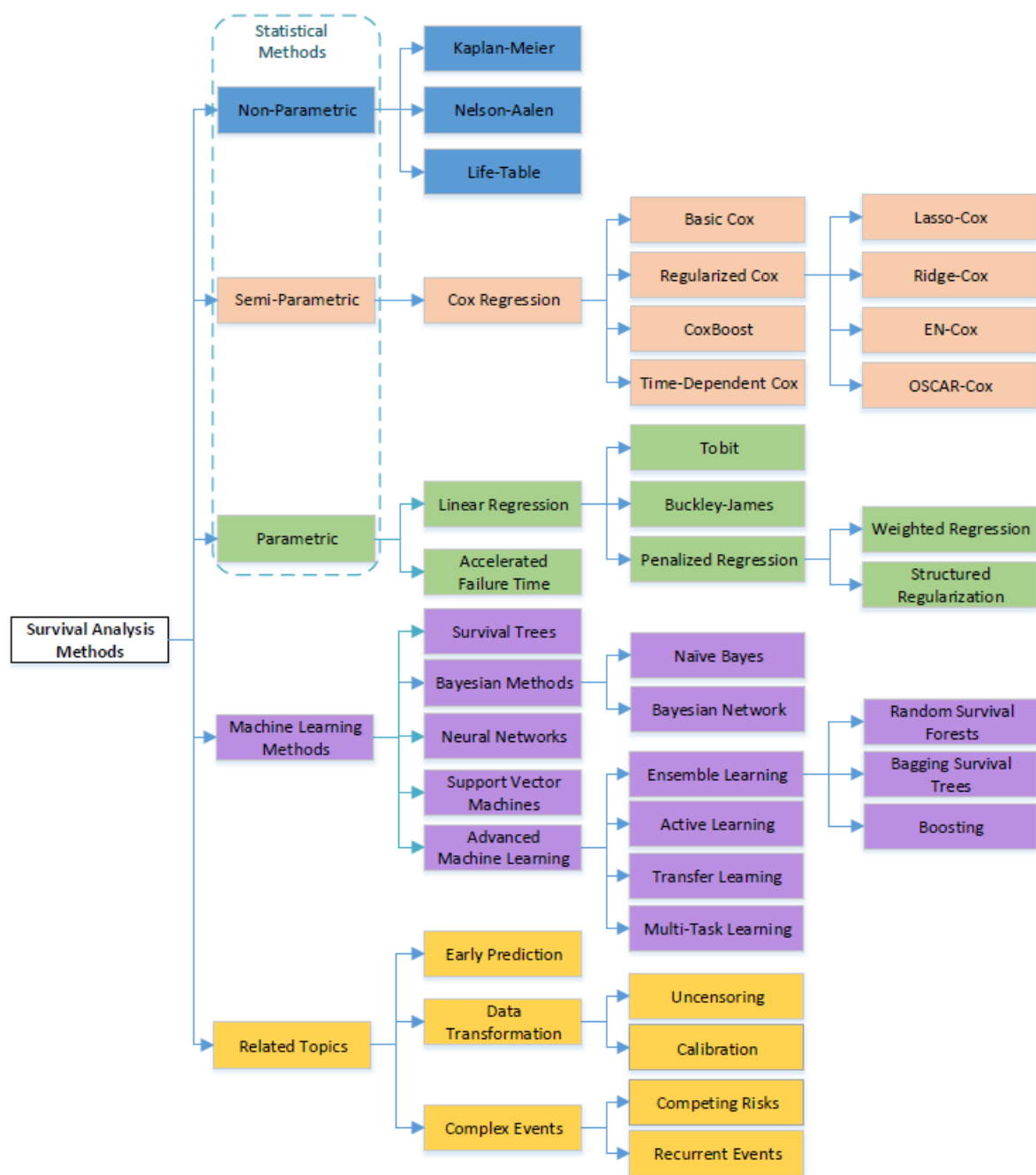
Panaudojant aukščiau aprašytą funkciją $f(t)$, lygtis (2) gali būti perrašyta į tokią:

$$S(t) = \exp(-H(t)) \quad (5)$$

Kaip matome užtenka turėti tik vieną iš funkcijų $S(t)$, $H(t)$, $h(t)$ ir visas kitas galima išvesti iš turimos. Pavyzdžiui išlikimo ir rizikingumo funkcijų rezultatai rodo kontrastingus vaizdus, todėl bendrai visos trys funkcijos yra naudingos siekiant tikslingos išlikimo analizės.

3.1. Išlikimo analizės metodai

Nagrinėjant išlikimo analizei skirtus metodus, pagrįsime, kad jie yra skirstomi į dvi grupes – statistiniai ir mašininio mokymosi metodai (Wang, Yan, & Chandan, 2017). Naudojant bet kurį iš grupėse esančių metodų, tikslas išlieka toks pat – prognozuoti išlikimo arba įvykio laiką. Tačiau skirtumų tarp jų yra, vienas iš jų – statistiniai metodai naudojami daugiau susikoncentruojant į įvykių laikų pasiskirstymą bei statistines savybes, skirtas parametrų prognozavimui. Statistinio išlikimo analizės metodo pavyzdys – Kokso regresija, apie kurią bus papasakota šiek tiek vėliau. Mašininio mokymosi metodai dažniausiai naudojami esant dideliame atributų dimensijų skaičiui, taip pat jie suteikia galimybę efektyviau nagrinėti šio tipo problemas, nes kartu jungia ir statistinius metodus bei naudojami naujausiais patobulinimais mašininio mokymosi ir optimizacijos srityse, kas leidžia skirtingais būdais analizuoti priklausomybes tarp atributų ir įvykio laikų. Dirbtiniai neuroniniai tinklai (angl. *artificial neural networks*) – vienas iš tokių metodų. Pagrindinis privalumas šio metodo išlikimo analizėje – atrasti paslėptus arba „tarpinius“ kintamuosius tarp įėjties ir išeities procesų (Clark, Bradburn, & Love, 2003). Šiuo atveju įėjties – kovariatsės, o išeities – išlikimo tikimybė. Vienas iš pavyzdžių – tyrėjai Faraggi ir Simon ankščiau minėtame savo darbe (Simon & Faraggi, 1995) patobulino Cox'o regresijos modelį, jog tiktų netiesiniam dirbtinio neuroninio tinklo prognozuotojui (angl. *predictor*) ir pasiūlė sujungti ir taikyti kartu neuroninį tinklą, kuris turi tiesinį išeities sluoksnį bei logaritminį (angl. *logistic*) paslėptą sluoksnį.



5 pav. Išlikimo analizės metodų taksonomija (Wang, Yan, & Chandan, 2017)

3.1.1. Statistiniai metodai

Statistiniai metodai dalinami į tris grupes – neparametriniai (angl. *non-parametric*), pusiau parametriniai (angl. *semi-parametric*) ir parametriniai (5 pav.). Pirmosios grupės metodai (Kaplano-Mejerio (*Kaplan-Meier*) (Kenah, 2013), Nelsono-Aaleno (*Nelson-Aalen*) ir kt.) naudojami kuomet nėra galimybės pagal įvykių laikus sudaryti jokio galimai prognozavimui reikalingo jų pasiskirstymo.

Parametriniai modeliai - pagreitinamas gedimo laikas (Swindell, 2009), Buklio-Džeimso (Buckley-James) (Lai & Zhiliang, 1991) ir kt. dažniausiai naudojami tada, kai įvykių laikai turi tam tikrą Weibullio, eksponentinį, normalųjį ir kt. pasiskirstymą. Pusiaus parametriniai metodai yra tarsi dviejų ankščiau minėtų modelių kombinacija. Šio tipo modeliai gali pasiekti nuoseklesnes prognozes naudojant platesnį intervalą sąlygų lyginant su parametriniais modeliais, taip pat gali gauti tikslesnes prognozes už neparametrinius modelius (Powell, 1994). Iš anksto žinoti įvykių laikų pasiskirstymo kreivės nereikia, tačiau yra laikoma, jog atributai, susiję su įvykio laiku, turi eksponentinę įtaką jam. Platesnis statistinių metodų ir jų kategorijų aprašymas šiame darbe nėra aktualus, tačiau norint daugiau informacijos, galima ją rasti čia (Wang, Yan, & Chandan, 2017). Kaip ir minėta ankščiau, kiekvienos kategorijos metodai turi savo pritaikomumą, tačiau vienas iš dažniausiai naudojamų metodų išlikimo analizėje, kuris bus naudojamas ir šiame darbe – Cox'o proporcingų rizikingumų modelis.

3.1.1.1. Cox'o proporcingų rizikingumų modelis

Modelis pristatytas 1972 metais (Cox, 1972) ir pavadintas jį sukūrusio tyrėjo D.R. Cox vardu. Proporcingų rizikingumų modelis paremtas rizikingumo funkcija $h(t_i)$ i-tajam dalyviui, kuriam priklauso x_i atributų vektorius. Kadangi modelis pusiau parametrinis, nereikia žinoti jokios informacijos apie įvykių laikų statistinį pasiskirstymą. Vienas svarbiausių modelio punktų yra tas, jog rizikingumas yra proporcingas viso tyrimo metu (Faruk, 2018). Pačią funkciją sudaro bazinio (angl. *baseline*) rizikingumo bei eksponentinės tiesinio modelio funkcijos sandauga, kuri yra užrašoma:

$$h(t|x_{i1}, \dots, x_{ip}) = h_0(t) \exp \left(\sum_{j=1}^p x_{ij} \beta_j \right) \quad (6)$$

kur β – koeficientai priklausantys prie kiekvieno iš p atributų ir nėra jokio pradinio atidėjimo (angl. *intercept*) koordinatų ašyje. Pagrindinė šio modelio savybė, skirianti iš kitų modelių, yra tai, jog jis yra skiriamas į dvi dalis – bazinė rizikingumo funkcija h_0 yra priklausoma tik nuo laiko t , o eksponentė yra nepriklausoma nuo laiko ir priklauso tik nuo atributų x_i .

Tinkamų modelio koeficientų paieška (angl. *model fitting*) yra vykdoma maksimizuojant dalinio tikėtumo (angl. *partial likelihood*) funkciją, kuri yra paremta tikimybe, jog i-tajam dalyviui nutiks įvykis laiko momentu t_i . Taip pat reikia pabrėžti, jog minėtuojų laiku momentu nutiks būtent tik šis

įvykis. Kaip bazinio rizikingumo funkcija šiame etape nėra reikalinga ir jos nereikia įtraukti siekiant rasti koeficientus β .

Iliustruojant dalinio tikėtinumo pavyzdį, tarkime jog $R_i = \{j|y_j \geq y_i\}$ yra rizikingų įvykių grupė, kuriems dar neįvyko įvykis prieš pat laiko momentą y_i , o $I(\cdot)$ yra indikatoriaus funkcija, tuomet:

$$\begin{aligned} & P(\text{dalyvio įvykis laiko momentu } y_i | \text{tik vienas įvykis } y_i) \\ &= \frac{P(\text{dalyvio įvykis laiko momentu } y_i | \text{jokių įvykių iki } y_i)}{P(\text{vienas įvykis laiko momentu } y_i | \text{jokių įvykių iki } y_i)} \\ &= \frac{h(y_i|x_i)}{\sum_{j=1}^n I(y_j \geq y_i) h(y_j|x_j)} = \frac{h_0(y_i) e^{x_i^T \beta}}{\sum_{j=1}^n I(y_j \geq y_i) h_0(y_j) e^{x_j^T \beta}} = \frac{e^{x_i^T \beta}}{\sum_{j \in R_i} e^{x_j^T \beta}} \end{aligned} \quad (7)$$

Padauginus aukščiau užrašytą sąlyginę tikimybę (7) visiems dalyviams, kuriems jau įvyko įvykis ir ištraukus logaritmą yra gaunama dalinė tikėtinumo funkcija:

$$\log L(\beta) = \sum_{i=1}^n \delta_i \left[x_i^T \beta - \log \left(\sum_{j \in R_i} e^{x_j^T \beta} \right) \right] \quad (8)$$

kur δ_i yra binarinis kintamasis skirtas cenzūravimui. Jeigu jis yra lygus 1, tai įvykis įvyko laiko momentu i , o jeigu lygus 0 – įvykis yra cenzūruotas. Tokiu būdu į dalinio tikėtinumo skaičiavimą yra įtraukiami tik tie dalyviai, kuriems įvykis įvyko.

Dalinio tikėtinumo funkcijoje (8) laužtiniuose skliaustuose pirmasis narys $x_i^T \beta$ yra i -tojo dalyvio rizikingumas, o antrasis narys $\log \left(\sum_{j \in R_i} e^{x_j^T \beta} \right)$ yra logaritmas visų dalyvių, kurie yra rizikos srityje, tačiau jiems įvykis neįvyko tuo laiko momentu, kuomet įvyko dalyviui i , suma. Reikia taip pat pastebėti, jog abu šie nariai nepriklauso nuo laiko t , tačiau labai svarbu yra įvykių laikų rikiavimas, nes kitu atveju nėra galimybės prasmingai atlikti proporcingų rizikingumų analizės.

Šis modelis išlikimo analizėje išlieka populiariu dėl lankstaus panaudojimo ir gana paprasto interpretavimo. Jis sukuriamas kaip regresijos modelis ir gali pakankamai tiksliai nusakyti ryšius tarp subjekto atributų ir laiko iki įvykio.

3.1.1.2. Pagreitintas įvykio laikas

Šis modelis dažnai naudojamas vietoje Kokso proporcingų rizikingumų regresijos, norint išvengti statistinių problemų, su kuriomis susiduriama kuomet yra pažeidžiama proporcingų rizikingumų sąlyga (Wei, 1992). Modelio pagalba yra tiriami atributai „pagreitina“ arba „sulėtina“ išlikimo laiką, ko pasėkoje tiesiogiai daro įtaką laikams iki įvykių vietoj rizikingumo lygmens, kurį lemia atributai Cox regresijos (6 formulė) modelyje.

Tarkime, jog T yra įvykių laikų vektorius, o X yra atributų $X_1, X_2, \dots, X_p$ vektorius. Turimas modelis santykį tarp įvykio laiko funkcijos $S(t|X)$ kiekvienu laiko momentu $t \in T$ ir turimų kovariačių X

$$S(t|X) = S_0[t \exp(\beta^t X)] \quad (9)$$

S_0 – bazinė išlikimo funkcija, o $\beta^t = (\beta_1, \beta_2, \dots, \beta_p)$ – regresijos koeficientai. Lygtyje (9) esantis narys $(\beta^t X)$ ir yra vadinamasis „greitinimo“ daugiklis, kuris pakeičia išlikimo funkcijos „greitį“ kai atributai $X = 0$.

Atributų ir išlikimo laiko santykis gali būti parodytas tiesiniu santykiu tarp natūrinio išlikimo laiko logaritmo ir atributų X , kuris yra užrašomas tokia lygtimi:

$$Y = \ln T = \mu + \theta^t X + \sigma W \quad (10)$$

kur μ – nuolydis (angl. *slope*), $\sigma > 0$ – mastelis (angl. *scale*), θ^t – regresijos koeficientų vektorius, o W yra paklaidų kintamasis, kuris turi tam tikrą parametrinę pasiskirstymą. Kadangi paklaidos turi tam tikrą pasiskirstymą, galima teigti, jog ir išlikimo laikai T turi tam tikrą pasiskirstymą. Dažniausiai naudojami pasiskirstymai šio modelio taikymui yra Weibullio, eksponentinis, log-normalusis, log-logistinis ir kt. Šiame darbe, pagreitinto įvykio laiko modelyje, bus renkamasi iš aukščiau minėtųjų skirstinių, nes prieš taikant modelį nėra žinomas tikslus turimų įvykių skirstinys.

Norint geriau įsivaizduoti pagreitinto įvykio laiko modelio idėją, palyginkime dvi išlikimo funkcijas – rūkančiųjų $S_1(t)$ bei nerūkančiųjų $S_2(t)$. Tuomet turimo modelio prielaida yra užrašoma tokiu pavidalu:

$$S_2(t) = S_1(\gamma t) \quad (11)$$

kur γ ir $t \geq 0$ nurodo konstantą, kuri yra vadinama pagreičio faktoriumi tarp ankščiau minėtų rūkančiųjų ir nerūkančiųjų grupių. Šis faktorius yra naudojamas norint sužinoti atributų įtaką išgyvenimo laikui, nes tai yra pagrindinė priemonė tiriant skirtingų grupių santykius šiame modelyje (Wang, Yan, & Chandan, 2017).

3.1.2. Mašininio mokymosi modeliai

Mašininio mokymosi metodai išlikimo analizėje naudojami nuo praėjusio tūkstantmečio pabaigos. Tačiau ankstesni bandymai naudoti mašininį mokymąsi išlikimo analizės srityje nepasiteisindavo, nes lyginant jų rezultatus (Zupan, 2000), (Xiang, 2000) su tiesinio Cox regresijos modelio rezultatais, jokių pranašumų nebuvo gauta. Tai paaiškinti galima tuomet buvusių menku susidomėjimu mašininio mokymosi tobulinimu (Katzman, et al., 2017). Tačiau per paskutinius keletą metų, dėka vis labiau tobulinamų mašininio mokymosi metodų (galimybė įvertinti netiesinius (angl. *non-linear*) atributų sąryšius bei vis didesnis prognozavimo tikslumas), jų naudojimas įvairiose praktinėse srityse, įskaitant ir išlikimo analizę, tapo standartiniu (Wang, Yan, & Chandan, 2017). Atsiradę tokie modeliai kaip atsitiktiniai išlikimo miškai (angl. *Random Survival Forests*), gilusis mokymasis (angl. *deep learning*) ar atraminių vektorių mašinos (angl. *Support Vector Machines*) išsprendžia vieną didžiausių problemų, su kuriomis susiduriama išlikimo analizėje - leidžia modeliuoti ir prognozuoti laiką turint cenzūruotus duomenis. Šiuo metu mašininio mokymosi modeliai prognozių tikslumu lenkia klasikinius statistinius modelius ir dažnas pasirinkimas sprendžiant išlikimo analizės problemą.

3.1.2.1. Atsitiktiniai išlikimo miškai

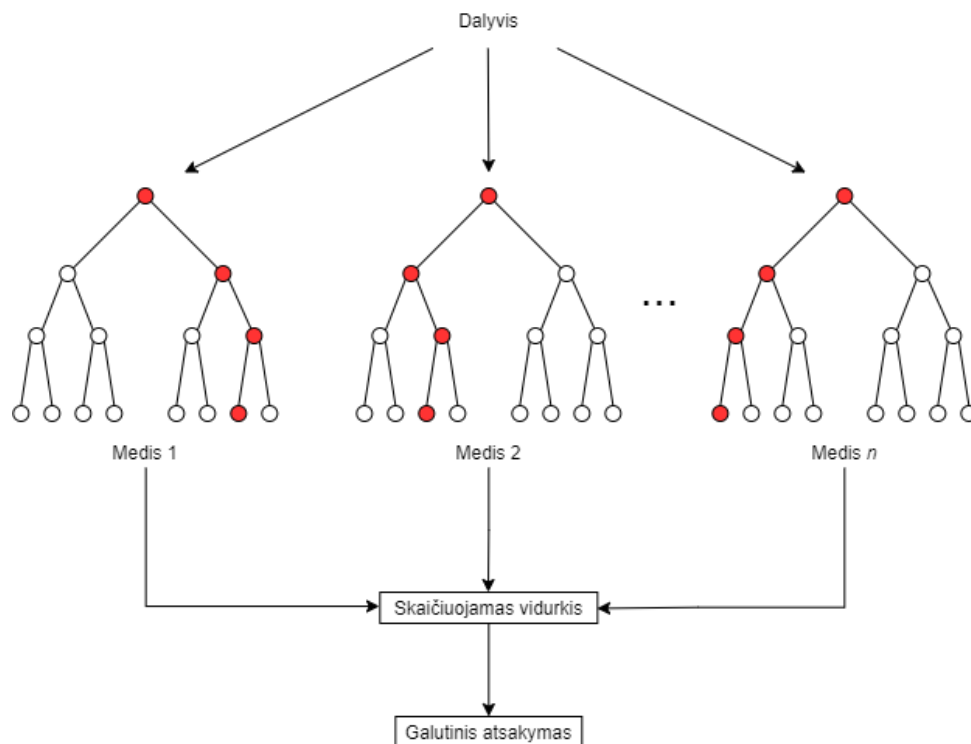
Atsitiktinis miškas (Breiman, 2001) yra nparametrinis metodas, priskiriamas ansamblinių (angl. *ensemble*) metodų klasei, kuris nereikalauja priešlaikinių prielaidų apie atributų pasiskirstymą. Šis metodas yra netiesinis ir optimizuoja prognozavimo tikslumą naudojant grupę sudarytų medžių, kurie stabilizuoja modelio prognozes. Miškas auginamas pakuojant (angl. *bagging*) kartu klasifikacijos ir regresijos medžius, o pakavimas atliekamas sudarant naujus, mokymui skirtus, duomenų rinkinius. Metode naudojamas rinkinys pasirinktas skaičius pavyzdžių ir tokiu būdu auginamas nepriklausomo medžio modelis kiekvienam populiacijos porūšiui. Kiekvienas medis yra auginamas rekursyviai dalinant populiaciją. Dalinimas yra paremtas atributų dalinimo taisyklės optimizacija. Kiekvieno dalinimo metu, poabio m ($m \leq p$) kovariatės testuojamos naudojant ankščiau minėtą optimizaciją ir kiekvienas mazgas yra dalinamas į du dukterinius mazgus. Tokia procedūra yra kartojama kol pasiekiamas norimas

sustojimo kriterijus – pasiektas norimas mazgo tyrumas (angl. *purity*) arba norimas mazgo narių skaičius, kuris nurodo imtį narių, kurią pasiekus, toliau mazgas yra nedalinamas. Būtina paminėti, jog mazgo dalinimo kokybė (mazgo tyrumas) yra matuojama dviem būdais – vidutine kvadratine paklaida (angl. *mean square error*) arba Gini indeksu. Kadangi šiame darbe yra sprendžiama regresijos problema (ieškomas laikas iki įvykio), kokybė bus matuojama naudojant vidutinę kvadratinę paklaidą – būdingas pasirinkimas esant tokio tipo problemoms (Friedman, 2000). Ji yra užrašoma tokia išraiška:

$$\frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2 \quad (12)$$

kur y_i yra subjekto i -toji klasė, N – klasių skaičius, μ – vidurkis, kuris skaičiuojamas $\frac{1}{N} \sum_{i=1}^N y_i$

6 pav. vizualiai pavaizduotas atsitiktinio miško metodo veikimas. Sudaromas nustatytas skaičius medžių - dažniausiai naudojamas 500-5000 medžių intervalas, tačiau tyrimai rodo, jog ir 50 medžių gali lemti gerus rezultatus (Cutler, et al., 2007). Priklausomai nuo turimos problemos tipo, galutinis atsakymas yra gaunamas apskaičiavus daugumą (klasifikacija) arba suskaičiavus medžių atsakymų vidurkį (regresija).



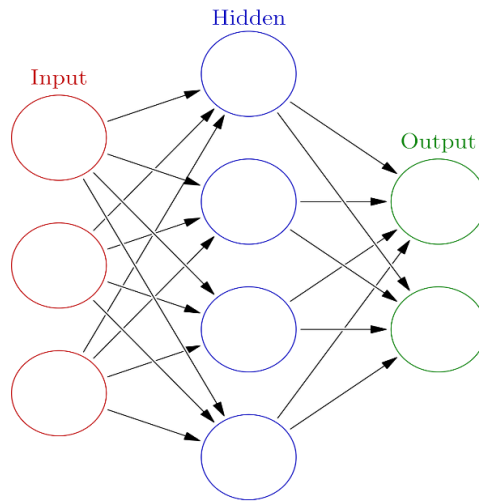
6 pav. Atsitiktinio miško sudarymo schema

Atsitiktiniai išlikimo miškai (Ishwaran, Kogalur, Blackstone, & Lauer, 2008) yra tęsinys ankščiau minėto modelio, leidžiantis naudoti iš dešinės pusės cenzūruotus (angl. *right-censored*) duomenis atliekant išlikimo analizę. Šis modelis yra gana lankstus, nes automatiškai nustato tam tikrus sąveikų tarp nepriklausomų kintamųjų tipus (Nasejje, JB.; Mwambi, H.; Dheda, K., 2017). Miškas sudaromas iš išlikimo medžių, kurių paskirtis pagal parinktų skirtingų atributų kombinacijas dalinti panašius subjektus į grupes priklausomai nuo subjektų laiko iki įvykio (priklausomo kintamojo). Tai reiškia, jog geriausias mazgo padalinimas įvyktų tuomet, jei būtų peržiūrėti visi nepriklausomi kintamieji x , jų dalinimo vertės c ir būtų pasirinktos vertės x^* , c^* , kurios turi didžiausią išgyvenimo skirtumą. Vis labiau didinant išgyvenimo skirtumą bei didėjant mazgų skaičiui, skirtingi atvejai yra vis labiau stumiami vienas nuo kito, kiekvienas mazgas tampa labiau homogeniškas ir subjektai esantys juose turi panašius išgyvenimo laikus. Pasirinktas mazgo dalinimo būdas yra labai svarbus modeliuojant išlikimo medžius. Jis atliekamas *log-rank* arba *log-rank score* būdais. Pirmojo atveju geriausias mazgo dalinimas yra randamas tame taške, kuriame apskaičiuojamas didžiausias atstumas tarp padalintų dukterinių mazgų.

3.1.2.2. Dirbtiniai neuroniniai tinklai

Šiame skyrelyje aprašoma šiuo metu daug dėmesio sulaukianti mašininio mokymosi metodologija – gilusis mokymasis. Populiarumas auga dėl spartaus giliojo mokymosi metodų tobulinimo ir galimybės pritaikyti netiesinį modelį prognozuojant logistinę rizikingumo funkciją $h(x)$, nes naudojant Cox'o regresijos modelį, viena iš prielaidų yra tokia, jog prognozuojant $h(x)$ yra naudojama tiesinė funkcija, tačiau daugybėje tyrimų duomenys neatitinka tiesinio proporcinio rizikingumo sąlygos (Katzman, et al., 2017).

Giliojo mokymosi modeliai yra sudaryti iš daugialypių dirbtinių neuroninių tinklų. Pats tinklas susideda iš dirbtinių taškų, pavadintų neuronais, kurie yra sujungti tarpusavyje tam tikrą „svorį“ turinčiais saitais.



7 pav. Neuroninio tinklo struktūra

Bendra visuma sudaro tinklą, savo architektūra atspindintį biologinį neuroninį tinklą (Wang, Yan, & Chandan, 2017). Pačią architektūrą sudaro įeities (angl. *input*), paslėptasis (angl. *hidden*) bei išvesties (angl. *output*) sluoksniai. Neuronu šiame kontekste yra laikomas skaičiavimui skirtas elementas, kuris susideda iš rinkinio adaptyvių svorių, kurie padeda suskaičiuoti tam tikrą atsakymą panaudojus kartu ir aktyvacijos funkciją, kuri padeda sušvelninti įvesties pokyčių kitimą, t. y. mažas įvesties pokytis lems taip pat ir mažą išvesties pasikeitimą. Neuroną apibrėžianti lygtis:

$$y_k = \varphi\left(\sum_{j=0}^m w_{kj}x_j\right) \quad (13)$$

kur φ – aktyvacijos funkcija, w_{kj} – k-tosios jungties j-tasis svoris, x_j – j-tasis įvesties vienetas.

Dažniausiai naudojamos aktyvacijos funkcijos – sigmoidinė (angl. *sigmoid*), hiperbolinė-tangentinė (angl. *tanh*), sureguliuoto tiesinio vieneto (angl. *rectified linear unit*). Kiekviena iš jų turi savų privalumų, tačiau šiuo metu efektyviausia ir dažniausiai naudojama yra sureguliuoto tiesinio vieneto aktyvacijos funkcija. Naudojant šią funkciją, jei įvestis yra mažesnė nei 0 – išvestis bus lygi 0, o jei didesnė nei 0 – išvestis bus lygi įvesčiai:

$$f(x) = \begin{cases} 0, & \text{jei } x \leq 0 \\ x, & \text{jei } x \geq 0 \end{cases} \quad (14)$$

Šiuolaikiniai giliojo mokymosi modeliai koncentruojasi į gilių (daug sluoksnių sudarančių) neuroninių tinklų apmokymą naudojant atgalinio dauginimo (angl. *backpropagation*) algoritmą. Populiariausi modeliai yra šie:

- Daugiasluoksniu perceptoru tinklai;
- Konvoliuciniai neuroniniai tinklai;
- Ilgieji trumpos atminties (angl. *long short-term memory*) grįžtamieji neuroniniai tinklai.

Didžioji dalis išlikimo (laiko iki įvykio) analizėje taikomų giliųjų neuroninių tinklų modelių naudoja specialią kaštų (angl. *cost*) funkciją, kuri leidžia įvertinti ir cenzūruotus duomenis. Patys modeliai sudaromi taikant sudėtingas, konvoliucinių arba pasikartojančių neuroninių tinklų architektūras (Zhao & Feng, 2019).

Šiame darbe išlikimo analizėje bus naudojamas *DeepSurv* (Katzman, et al., 2017) modelis, kuris yra paremtas Cox'o proporcinio rizikingumo modelio adaptacija, pritaikyta neuroniniams tinklams. Tinkle išeities sluoksnis atlieka Cox'o regresiją pagal paslėptojo sluoksniu aktyvacijos lygius. Modelis sudarytas iš daugiasluoksniu perceptoru, o tinklo išeitį sudaro vienas mazgas ir jis prognozuoja rizikos funkciją $\hat{h}_0(\mathbf{x})$, kuri yra parametrizuota naudojant tinklo θ svorius.

3.2. Modelių gerumo vertinimas

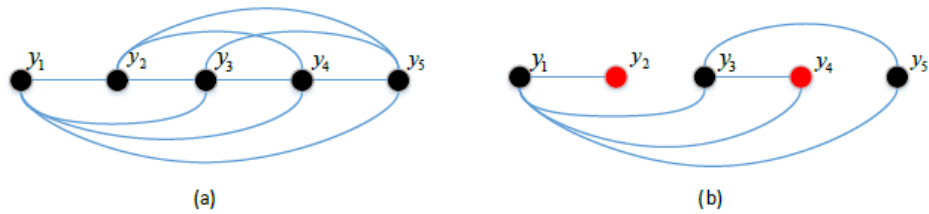
Šiame skyriuje bus apžvelgti metodai, nustatantys išlikimo laiko prognozei naudojamo modelio gerumą. Dėl išlikimo duomenyse dažnai pasitaikančio cenzūravimo, standartinės regresijoje uždaviniuose taikomos vertinimo metrikos – šaknis iš vidutinės kvadratinės paklaidos arba R^2 nėra tinkamos matuojant išlikimo analizės modelio gerumą (Heagerty & Zheng, 2005). Tai lemia, jog prognozavimo gerumas šio tipo analizėje matuojamas specializuotomis ir labiau pritaikytomis metrikomis.

3.2.1. Atitikimo rodiklis

Atitikimo (angl. *concordance*) rodiklis, nurodantis laiko iki išlikimo rūšiavimo kokybę, yra charakteristika, padedanti įvertinti modelio išskirstymo gerumą. Rodiklio naudojimą Kokso modeliuose pristatė tyrėjas Harrell (Harrell, Califf, & Pryor, 1982) ir šiuo metu tai yra vienas pagrindinių bei populiariausių matavimų vertinant išlikimo modelio gerumą turint tam tikrus duomenis (angl. *goodness-*

of-fit). Vienas iš šio vertinimo modelio privalumų, jog jis gali būti taikomas netik išlikimo analizėje, bet ir logistinėje arba tiesinėje regresijoje.

Išlikimo analizėje bendruoju atveju modelio gerumo įvertinimas yra vykdomas vietoj absoliutinio išlikimo laiko kiekvienam atvejui vertinant atskiro atvejo santykinę įvykio riziką. Išlikimo laikai dviem atvejams gali būti išrūšiuoti pagal du scenarijus: a) abu yra necenzūruoti; b) įvykusio įvykio laikas yra mažesnis nei cenzūruoto įvykio laikas (Wang, Yan, & Chandan, 2017). 3 pav. galima matyti abu variantus vizualiai.



8 pav. Išlikimo laikų rikiavimo ($y_1 < y_2 < y_3 < y_4 < y_5$) variantai. Kairėje pusėje nurodytas scenarijus, kuomet nei vienas įvykis nėra cenzūruotas, o dešinėje pusėje du stebiniai cenzūruoti. Juodi taškai nurodo necenzūruotus stebinius, o raudoni – cenzūruotus (Wang, Yan, & Chandan, 2017)

Abiejuose grafuose matomi galimi išrikiuotų stebinių palyginimai, kurie yra pavaizduoti linijomis tarp jų. Pirmuoju atveju yra 10 galimų porinių palyginimų tarp 5 necenzūruotų stebinių, o antruoju porinių palyginimų yra tik 6, dėl dviejų stebinių cenzūravimo. Reikia atkreipti dėmesį, jog cenzūruoti stebiniai gali būti lyginami su prieš tai esamu stebiniu, tik tuo atveju, jei jis yra necenzūruotas (y_2 ir y_1). Tuo pat metu, cenzūruoti stebiniai negali būti lyginami su už jais esančiais stebiniais bet kokių atveju, nes nėra žinomas stebinio įvykio laikas (y_2 ir y_3).

Įsivaizduokime dviejų stebinių tikrąsias bei prognozuojamas išlikimo laiko vertes ($y_1, \widehat{y}_1$) bei ($y_2, \widehat{y}_2$), kur y_1 – tikras, o $\widehat{y}_1$ – prognozuojamas išlikimo laikas. Atitikimo tikimybė tarp šių stebinių yra užrašoma tokia lygtimi:

$$c = \Pr(\widehat{y}_1 > \widehat{y}_2 | y_1 > y_2) \quad (15)$$

Pora taškų yra vadinami atitinkančiais, jei prognozuojami laikai iki įvykio ir tikrieji stebėjimai yra tos pačios krypties, t. y. ($y_i > y_j | \widehat{y}_i > \widehat{y}_j$) arba ($y_i < y_j | \widehat{y}_i < \widehat{y}_j$).

Atitikimo rodiklis yra apibendrintas variantas ploto po ROC - sprendimus priimančiojo ypatybių (angl. *receiver operating characteristic*) kreivę, kuris dar vadinamas AUC (angl. *area under curve*) ir jis skaičiavimui gali naudoti ankščiau minėtus cenzūruotus stebinius. Panašiai kaip ir AUC, jei atitikimo indeksas *c-index* yra lygus 1 – tai rodo geriausią įmanomą prognozavimo modelį, o jei 0.5 – modelio prognozės yra atsitiktinės.

Nepaisant to, jog šis rodiklis yra gana lengvai apskaičiuojamas bei interpretuojamas, jis turi keletą minusų:

- Esant dideliame cenzūruotų subjektų skaičiui duomenyse, gautas rezultatas yra pernelyg optimistiškas (Uno, Cai, Pencina, D'Agostino, & Wei, 2011);
- Nėra labai naudingas gerumo vertinimo matas, jeigu yra ieškomas tam tikras išgyvenimo laiko tarpas (pavyzdžiui – įvykio prognozavimas per ateinančius 3 metus) .

Šiame darbe subjektai (žaidėjai) nėra cenzūruoti - visi jie vienu ar kitu žaidimo metu baigė pasirodymą, todėl šios problemos nėra aktualios ir rodiklis tinkamas įvertinti taikomus išlikimo analizės metodus.

3.2.2. Brierio rezultatas

Šis rezultatas buvo pristatytas mokslininko Glenno Brierio (*Brier*, 1950), kuomet buvo norima nustatyti gautų oro prognozių netikslumą. Vertinimo kriterijus yra taikomas tik tiems modeliams, kurių išvestis yra tikimybinės formos, t. y. išvesties rezultatai yra intervale $[0,1]$, o suma visų galimų išvesčių yra lygi 1 (Wang, Yan, & Chandan, 2017). Išlikimo analizėje, šis vertinimas yra skirtas įvertinti tikslumą prognozuojamos išlikimo funkcijos kiekvienu laiko momentu t . Jis nurodo vidutinį kvadratinį atstumą tarp subjekto turimos bei prognozuojamos išlikimo tikimybių. Geriausia galima gauta rezultato vertė yra lygi 0, nes tai parodo idealų prognozavimo modelį ir mažiausią galimą atstumą tarp prognozuojamos ir realios išlikimo tikimybių.

Tarkime, prognozuojama dvejetainė išvestis N subjektų, kur prognozuojamas kintamasis yra $\hat{y}_i$, o realus išvesties kintamasis yra y_i . Tokiu atveju esant dešininiam subjektų cenzūravimui, rezultato skaičiavimo lygtis yra užrašoma:

$$BS(t) = \frac{1}{N} \sum_{i=1}^N \omega_i(t) [\hat{y}_i(t) - y_i(t)]^2 \quad (16)$$

Aukščiau esančioje lygtyje (16), kintamuoju $\omega_i(t)$ yra aprašomas svoris i-tajam subjektui, kuris priklauso nuo cenzūruotos informacijos. Cenzūruoti duomenys prie Brierio rezultato prisideda netiesiogiai, skaičiuojant cenzūravimo pasiskirstymą. Daugiau informacijos apie šį vertinimo kriterijų galima rasti įvairių autorių (Wang, Yan, & Chandan, 2017; Graf, Schmoor, Sauerbrei, & Schumacher, 1999) darbuose.

Norint apskaičiuoti bendrą modelio gerumą viso tyrimo metu, naudojamas integruotas Brierio rezultatas (angl. *Integrated Brier Score*). Jis užrašomas taip:

$$IBS(t_{max}) = \frac{1}{t_{max}} \int_0^{t_{max}} BS(t) dt \quad (17)$$

Kalbant apie bendro vertinimo kriterijus, svarbu paminėti, jog naudingas išlikimo analizės modelis turi turėti bendrą rezultatą mažesnę nei 0.25. Susidaryti tokias išvadas nėra sunku, nes kai prognozuojamas kintamasis $\hat{y}_i = 0.5$, tuomet $BS(t) = 0.25$. Tokiu atveju modelis prognozuoja išlikimo tikimybę tokiu pačiu tikslumu kaip ir atsitiktinai prognozuojantis modelis.

4. TYRIMO REZULTATAI IR JŲ APTARIMAS

Prieš pradedant taikyti modelius bei interpretuoti gautus rezultatus būtina susipažinti su turimais duomenimis. Šiuo atveju yra atliekama duomenų apžvalgomoji analizė (angl. *exploratory data analysis*). Prieš pradedant duomenų analizę, svarbu yra susipažinti su dalykine sritimi.

Tyrime naudojami duomenys yra gauti iš kompiuterinio žaidimo pavadinimu „PlayerUnknown’s Battleground“. Žanras šio žaidimo yra „Battle Royale“, kuriame žaidėjai kovoja vieni su kitais iki tol, kol lieka paskutinis gyvas, kas nurodo pagrindinę kiekvieno žaidėjo užduotį norint laimėti – išlikti žaidime gyvam kuo ilgiau.



9 pav. Žaidimo nuotrauka (*TheTechGame*, 2018)

Viename žemėlapyje kovoja iki šimto žaidėjų pagal vieną iš trijų skirtingų stilių, kuriuose formuojami komandų dydžiai – komandos sudarytos iš 1 žaidėjo (kiekvienas už save), 2 žaidėjų ir 2-4 žaidėjų. Žaidimo pradžioje žaidėjai neturi nė vieno daikto, galinčio padėti norint laimėti žaidimą, kuriuos jie renka žaidimo eigoje. Žaidimas šiuo metu išlieka labai populiarus pasaulyje, vidutiniškai kiekvieną dieną jį žaidžia apie 262000 unikalių žaidėjų³ ir jis išlieka trejetuke populiariausių žaidimų vienoje didžiausių žaidimų platformų „Steam“.

4.1. Žvalgomoji duomenų analizė

Šiame darbe naudojami duomenys yra gauti naudojant jau minėto žaidimo kūrėjų suteiktą programavimo sąsają (angl. *application programming interface*).⁴ Ja gauti duomenys yra telemetrijos tipo, t. y. kiekvieną laiko momentą (šiuo atveju – 1 sekundė) yra pažymimas tam tikras įvykis su jam priskirtais atributais bei jų reikšmėmis. Įvykių intervalas gana didelis - nuo žaidėjo pozicijos iki paimto tam tikro daikto arba atliktos žalos kitam žaidėjui. Analizuojama informacija yra paimta iš atsitiktinai parinktų 200 žaidimų, vykusių 2019-09-19 dieną. Pati informacija skelbiama failuose naudojant JSON (angl. *Javascript Object Notation*) formatą. Naudojant jį, failai sudaryti iš masyvų, kuriuos sudaro

³ Naudojantis platformos „STEAM“ duomenimis 2020-02-11 dieną

⁴ <https://documentation.pubg.com/en/telemetry.html>

atributų ir reikšmių poros. Šiuo atveju atskiras masyvas yra tam tikras įvykis, o atributų ir reikšmių poros – specialūs, tam įvykiui prasmingi, atributai bei atitinkamos jų reikšmės.

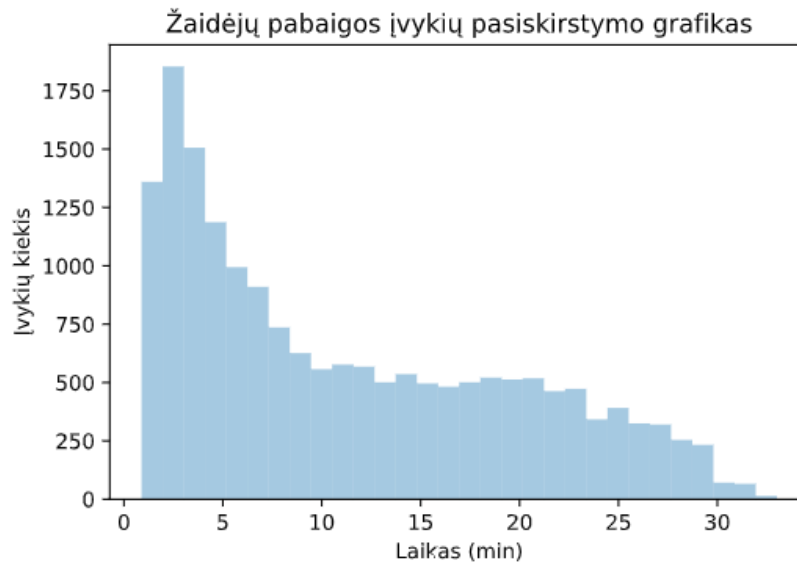
1 lentelė. Naudojamų pagrindinių duomenų atributų dalinė struktūra

Kintamasis	Tipas	Reikšmės	Aprašymas
Name	Skaitinis	<i>Player1,</i> <i>Player2</i>	Žaidėjo slapyvardis žaidime
Rank	Sveikasis skaičius	1,2,3...	Žaidėjo vieta tarp kitų žaidėjų pasibaigus žaidimui
Kills	Sveikasis skaičius	0,1,2...	Žaidėjo nužudymų skaičius
Is_in_blue_zone	Kategorinis	0,1	Ar žaidėjas pasibaigus žaidimui buvo mėlynojoje zonoje
Is_in_red_zone	Kategorinis	0,1	Ar žaidėjas pasibaigus žaidimui buvo raudonoje zonoje
Dist_on_foot	Realusis skaičius	28.97, 1487.22	Kojomis nukeliantas žaidėjo atstumas
Dist_on_swim	Realusis skaičius	33.11, 224.29	Nuplauktas žaidėjo atstumas
Dist_on_vehicle	Realusis skaičius	1.22, 504.85	Nuvažiutas transporto priemone žaidėjo atstumas
Dist_on_parachute	Realusis skaičius	2.22, 56.88	Leidžiantis parašiotu žaidėjo atstumas
Dist_on_freefall	Realusis skaičius	1422.15, 1598.55	Žaidimo pradžioje ore nuskleistas žaidėjo atstumas
Groggy	Kategorinis	0,1	Ar žaidėjas buvo prikeltas prieš pasibaigiant žaidimui
Heal_amount	Realusis skaičius	51.22, 101.22	Žaidėjo atstatytų gyvybių skaičius
Assist	Sveikasis skaičius	0,1,2	Žaidėjo asistavimų skaičius
Damage	Realusis skaičius	55.44, 221.05	Žaidėjo padarytos žalos kiekis
Playing_type	Kategorinis	1,2,3	Žaidimo tipas
Death_time	Realusis skaičius	1981.41, 521.55	Žaidėjo žaidimo pabaigos laikas

Atlikus įvairius agregavimus ir išnagrinėjus turimus failus, sudarytas duomenų rinkinys, kurio formatas yra toks – 17897 žaidėjai ir juos apibūdinantys 80 atributų. Verta paminėti, jog pradiniai duomenys (žaidimų atkartojimų failai) užėmė kiek daugiau nei 3.2 GB vietos, tačiau sudarius duomenų rinkinį jo talpa sumažėjo iki 8 MB.

1 lentelėje galima matyti dalies duomenų struktūrą. Duomenyse daugiausia yra realiųjų skaičių atributų. Verta paminėti du kategorinius atributus, apibūdinančius zonas, kuriose žaidėjai buvo įvykio pabaigos metu – mėlynoji (angl. *is_in_blue_zone*) ir raudonoji (angl. *is_in_red_zone*). Mėlynoji zona atsiranda žaidimo pradžioje ir joje esantiems žaidėjams periodiškai mažėja gyvybės iki tol, kol jie išeina iš zonos arba jiems baigiasi žaidimas. Skersmuo šios zonos laikui bėgant mažėja, kas sudaro tendenciją žaidėjams būriuotis į vis mažesnį plotą ir didinti tarpusavio aktyvumą. Raudonoji zona atsiranda žemėlapyje atsitiktiniu būdu ir daro didelę žalą žaidėjams, esantiems joje. Zonos skersmuo yra 1/64 viso žemėlapio ploto ir atsiranda vis kitoje žemėlapio vietoje 75 sekundžių intervalu. Svarbu atkreipti dėmesį, jog čia neaprašyti atributai yra realiųjų skaičių tipo ir apibūdina tam tikro žaidime esančio daikto kiekį, kurį bendrai viso žaidimo metu surinko žaidėjas. Maksimalus leidžiamas žaidėjo daiktų kiekis nėra ribojamas, todėl pateiktas maksimalus kiekis. Pilną duomenų struktūrą galima rasti skiltyje „Priedai“.

Prieš nagrinėjant kitus turimus atributus, svarbu peržiūrėti ir vieno pagrindinių atributų, kurio pagalba ir atliekama išlikimo analizė – įvykio laiko atributą *death_time*. 10 pav. matomas jo skirstinys, kuris atskleidžia, jog pagrindinis aktyvumas vyksta žaidimo pradžioje – iki ~8 minutės, toliau jis išlieka pastovus, kol galiausiai ima mažėti nuo 22 minutės.



10 pav. Žaidėjų pabaigos įvykių pasiskirstymo grafikas

Turint sudarytą duomenų struktūrą atlikta tam tikra atributų modifikacija. „*Playing_type*“ atributas, apibūdinantis žaidėjo žaidžiamo žaidimo tipą (1 – kiekvienas už save, 2 – žaidžiama poromis, 3 – žaidžiama komandomis nuo 2 iki 4 žaidėjų) keičiamas į atskirus atributus pagal žaidimo tipą. Toks atributų keitimas yra vadinamas vienu kintamuoju paremtu keitimu (angl. *OneHot encoding*). Pradinis atributas yra pašalinamas iš duomenų, o likę kintamieji binariniu principu (1 – priklauso, 0 – nepriklauso) nurodo kurį žaidimo principą pasirinko žaidėjas. Po keitimo, tarp naujų atributų, atsiranda koreliacija, todėl šio tipo keitimo metu sukurtų atributų skaičius yra lygus $n - 1$ (n – kategorijų skaičius). Šiame darbe pasirinkta panaikinti atributą, kuris nurodo, kuomet žaidžiama poromis (*playing_type_2*) ir šiems subjektams priskirta reikšmė lygi tai, kuomet žaidėjai žaidžia komandomis. Taip pat buvo pristatytas naujas kintamasis – „*dmg_per_distance*“, nurodantis žaidėjo padarytos žalos kiekį, atliktą nukeliavus vieną atstumo vienetą. Jis pristatytas remiantis žiniomis apie kintamojo „*damage*“ kolineariškumą su kitais atributais ir nenorint prarasti minėtojo atributo turimos informacijos.

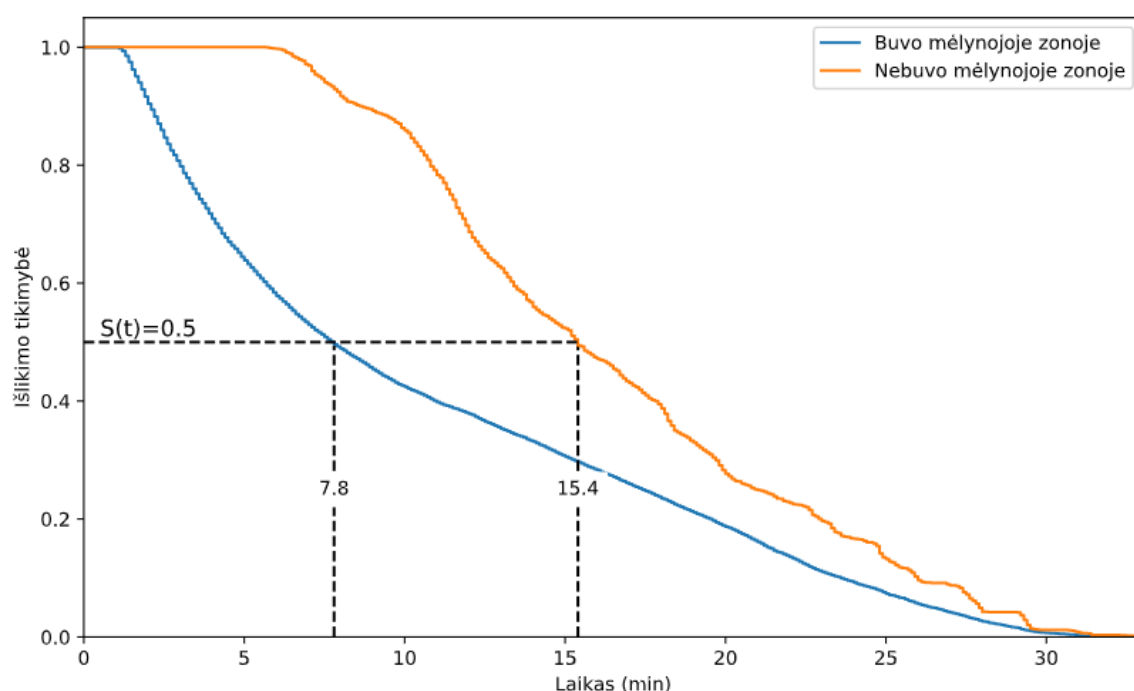
Peržvelgus turimus atributus, pašalintas sveikųjų skaičių tipo atributas „*rank*“, nurodantis pabaigoje turimą žaidėjo vietą tarp žaidime dalyvavusių žaidėjų. Šį atributą nuspręsta pašalinti dėl to, jog jis tiesiogiai nurodo žaidėjo išgyvenimo laiką, nes kuo mažesnė „*rank*“ vertė – kuo ilgiau žaidėjas gyvena, tuo aukščiau finišuoja, todėl žaidėjas daryti šiam atributui įtakos neturi galimybės ir jis neprisideda bandant prognozuoti žaidėjo išgyvenimo tikimybę, nes tik pabaigoje žaidėjai rikiuojami pagal išgyvenimo laiką ir priskiriama „*rank*“ atributo vertė.

Prieš taikant regresijos ar kitus mašininio mokymosi metodus yra peržvelgiama duomenų atributų koreliacijų matrica. Ji skirta išsiaiškinti ir pastebėti stiprias tarpusavio koreliacijas, kurioms esant regresijos metodų rezultatai gali būti nepatikimi. Tarpusavio koreliacija dažniausiai matuojama Pirsono (*Pearson*) arba Spirmeno (*Spearman*) koreliacijos koeficientais. Pirmasis matuoja atributų tarpusavio tiesinio ryšio stiprumą, kuomet skirstiniai yra normalieji, o antrasis – tarpusavio ryšio stiprumą, kuomet normalumo prielaida nėra taikoma (Janilionis, 2010). Koreliacijos koeficientų reikšmės yra visada intervale $[-1; +1]$, kur -1 reiškia stiprią atvirkštinę koreliaciją (didėjant vieno atributo vertei, mažėja kito), o $+1$ – stiprią tiesinę koreliaciją. Peržvelgus turimų duomenų atributų skirstinius pasirinktas Spirmeno koreliacijos koeficientas, nes skirstinių reikšmių normalizavimas nebuvo taikomas.

Verta pažymėti, jog sudarius koreliacijos matricą, pasirinkta pašalinti atributus, kurių koreliacijos koeficientai su kitais atributais yra tarp intervalų $[-1; -0.8)$ arba $(0.8; 1]$. Turint atributų koreliacijos koeficientus tarp šių intervalų galima vieną iš tokių atributų porų šalinti iš tolimesnės analizės, nes tokiu atveju kartu šie atributai nurodo tokią pačią informaciją apie priklausomą kintamąjį. Sudarius koreliacijos matricą (pateikta priede) ir peržvelgus gautus rezultatus, pašalinti kintamieji „*assists*“, „*distance_sum*“, „*damage*“, „*item_stack_count*“, nes koeficientai su vienu ar daugiau atributų patenka į aukščiau minėtus intervalus. Pašalinus minėtuosius atributus, papildomai atlikta VIF (angl. *variance inflation factor*) kriterijų analizė, siekiant išsiaiškinti ar esamas daugialypiškumas (angl. *multicollinearity*) tarp likusių atributų ir priklausomo kintamojo. Atliekant šią analizę yra vadovaujama taisykle, jog jeigu atributo VIF koeficientas yra didesnis negu 10 – jis yra šalinamas iš atributų sąrašo. Yra keletas būdų siekiant išspręsti šią problemą, tačiau šiame darbe atributai, turintys aukščiau minėtą arba didesnę koeficientą yra šalinami iš duomenų rinkinio. Suskaičiavus koeficientus, pagal gautus rezultatus, buvo pašalintas vienas atributas – „*dist_on_freefall*“, kurio koeficientas siekė 15.85.

Žvelgiant iš išlikimo analizės perspektyvos, prieš taikant įvairius modelius, galima sudaryti išlikimo tikimybes grafikus kategoriniams atributams, kurie skirti peržvelgti kokią tikimybę išgyventi po tam tikro laiko turi kiekvienas iš kategorinių atributų laikui bėgant. Tai atlikti leidžia sudaryta Kaplan-Meier (Kaplan & Meier, 1958) išlikimo kreivė. Ji sudaroma pagal tai pagal, kokia populiacijos (šiuo atveju – žaidėjų) dalis išgyveno iki tam tikro laiko momento (Rich, J. T.; Neely, J. G.; Paniello, R. C.; Voelker, C. C.; Nussenbaum, B.; Wang, E. W., 2010). Santykis tarp išgyvenusių ir neišgyvenusių sudaro išlikimo tikimybę tuo laiko momentu. Verta paminėti, jog kreivė sudaroma visiškai neatsižvelgiant į žaidėjų turimus atributus.

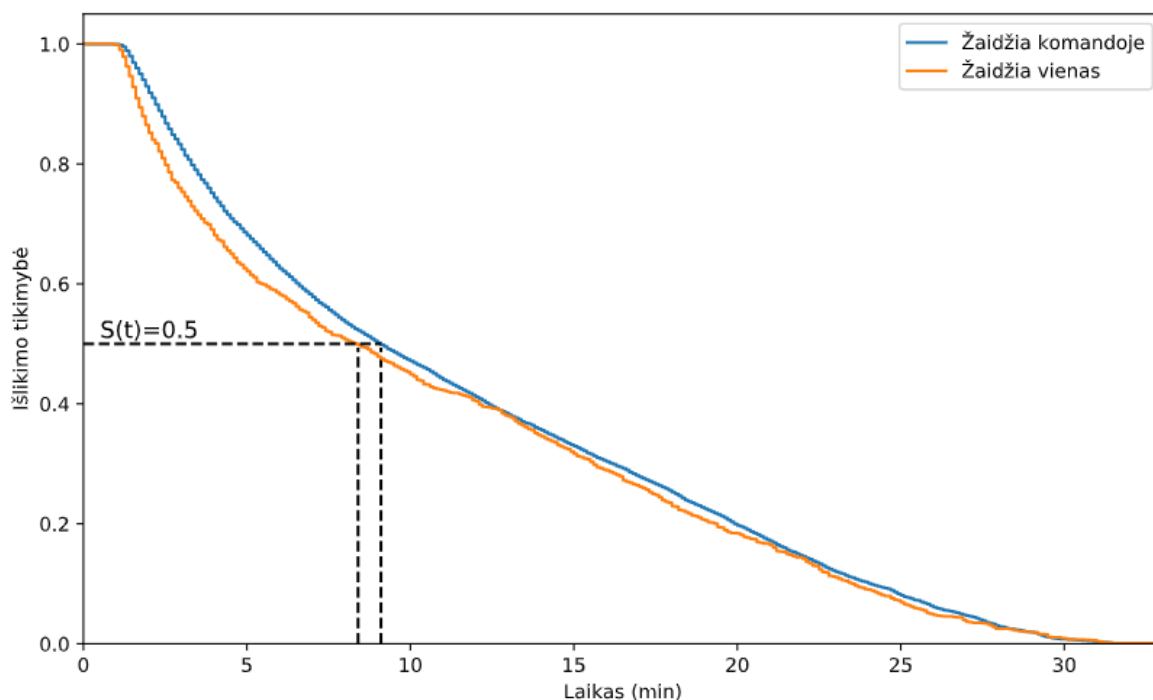
Ištyrus sudarytas išlikimo kreives, pastebėta, jog vienas iš labiausiai atskiriančių išlikimo tikimybę yra ar žaidėjas įvykio laiko momentu buvo mėlynojoje zonoje (11 pav.). Pažvelgus tik iš šio atributo perspektyvos, tai gana intuityvu, nes žaidėjui esant šioje zonoje jo gyvybės periodiškai mažėja ir vis ilgiau esant joje mažėjimo tempas didėja, kol žaidėjui baigiasi žaidimas. Zonos skersmuo žaidimo metu mažėja, todėl laikui bėgant padidėja galimybė patekti į ją. Sudarytas grafikas atskleidžia tokią tendenciją – žaidėjų, buvusių mėlynojoje zonoje išlikimo kreivė nuo pat žaidimo pradžios mažėja dideliu tempu. Vidutinis išgyvenimo laikas (išlikimo tikimybė $S(t)$ yra lygi 0.5) skiriasi 49 proc. Žaidėjų, kurie buvo mėlynojoje (mėlyna juosta) vidutinis išgyvenimo laikas – 7.8 minutės, o tų, kurie nebuvo – 15.4 minutės. Patikrinus atributą naudojant *log-rank* testą, gauta p-reikšmė lygi $3.17 \cdot 10^{-84}$ ir kadangi tai yra mažiau nei 0.05 – atmetama nulinė hipotezė, jog tarp dviejų atributo grupių nėra statistiškai reikšmingo skirtumo. Išlikimo tikimybė pasiekia nulinę ribą viename taške – (~32 minutę), nes tai yra maksimalus žaidimo laikas turimoje žaidėjų duomenų imtyje.



11 pav. Kaplan-Meier išlikimo kreivė pagal mėlynosios zonos atributą

Turint šią informaciją, taip pat patikrinta, ar skiriasi išlikimo kreivės tarp žaidėjų, žaidžiančių komandomis ir pavieniui. Sudarytame grafike (12 pav.) vizualiai skirtumą, kuris būtų reikšmingas, įžvelgti gana sunku. Išlikimo tikimybės medianos pasiekiamos 8.4 ir 9.1 minutėmis atitinkamai žaidėjų, kurie žaidžia vieni, ir tų, kurie žaidžia komandoje. Tai nurodo, kad vertinant išlikimo tikimybę tik pagal

ši atributą, žaidėjai turi 50 proc. tikimybę išlikti po pirmųjų 8-9 minučių. Tačiau panaudojus minėtąjį *log-rank* testą gauta p-reikšmė lygi 3.17^{-85} , todėl taip pat yra atmetama nulinė hipotezė, jog nėra tarp kreivių statistiškai reikšmingo skirtumo.



12 pav. Kaplan-Meier išlikimo kreivė pagal žaidimo tipo atributą

Kitų kategorinių kintamųjų Kaplan-Meier išlikimo kreivės pavaizduotos priedų skiltyje.

4.2. Modelių taikymas

Šiame skyrelyje bus aptariami skirtingų modelių, skirtų išlikimo analizei, modeliavimo ypatumai bei rezultatai, kurie bus panaudoti išrenkant geriausią modelį ir jį naudojant atlikti išlikimo prognozavimą žaidėjams. Tyrime bus naudojami šie modeliai:

- Cox'o proporcingų rizikingumų;
- Pagreitinto įvykio laikas;
- Atsitiktiniai išlikimo miškai
- Neuroniniai tinklai.

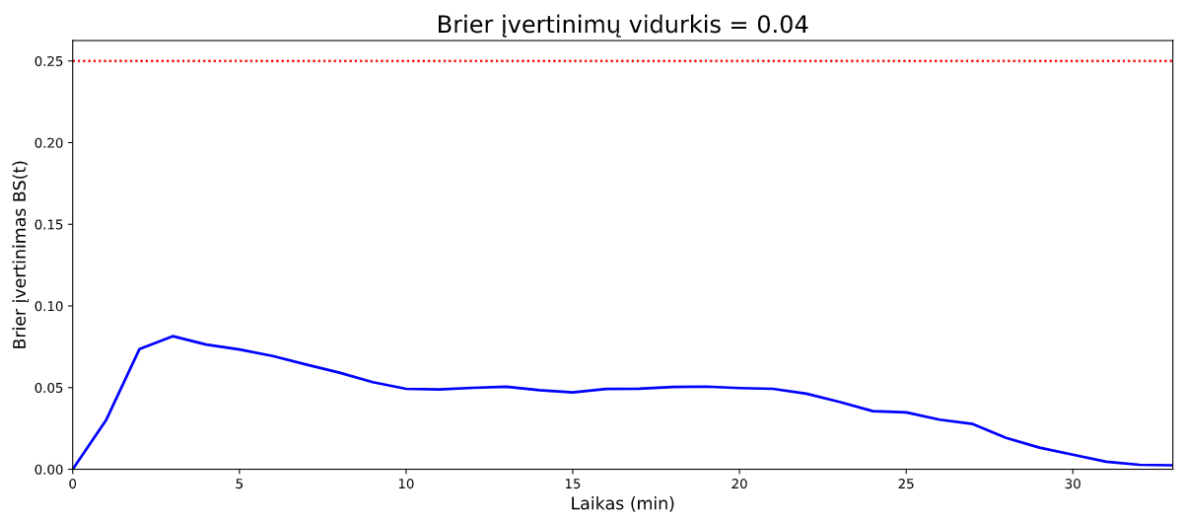
Kiekvienam iš modelių apmokymui buvo skirti 80 proc. turimų duomenų, o testavimo imtį sudarė likusieji 20 proc.

4.2.1. Cox'o regresijos modelis

Prisimenant teorinę dalį, Cox'o regresijos modelio taikymas nereikalauja jokio išankstinio žaidėjų išgyvenimo laikų skirstinio pasirinkimo. Prieš taikant modelį, turimų skaitinių atributų mastelis yra keičiamas į $[0; 10]$ intervalą. Tai daroma norint pamatyti ir palyginti atributų daromą įtaką žaidėjo įvykio rizikai, nes struktūroje yra atributų, kurių skalės yra dešimtis kartų didesnės nei visų kitų likusių.

Modelio gerumui įvertinti naudojami teorinėje dalyje minėti atitikimo (angl. *concordance*) bei Brier įvertinimai. Pritaikius modelį gautas atitikimo įvertinimas – 0.921. Toks rezultatas rodo labai gerą modelio rizikingumą žaidėjams prognozavimą, nes taikant šį modelį, net 92.1 procento žaidėjų rizikingumas yra prognozuojamas teisingai atsižvelgiant į jų įvykių laikus.

Gautas Brier rezultatas (0.044) taip pat indikuoja apie modelio gerumą prognozuojant išlikimo žaidėjų išlikimo tikimybę kiekvienu laiko momentu. Kuo rezultatas arčiau 0, tuo modelis geriau prognozuoja, nes įvertinimas yra skaičiuojamas kaip subjekto išlikimo tikimybės ir išlikimo statuso kiekvieno laiko momentu t skirtumas. Bendras įvertinimas yra apskaičiuojamas integralu visų rezultatų nuo pirmo iki paskutinio laiko momento.

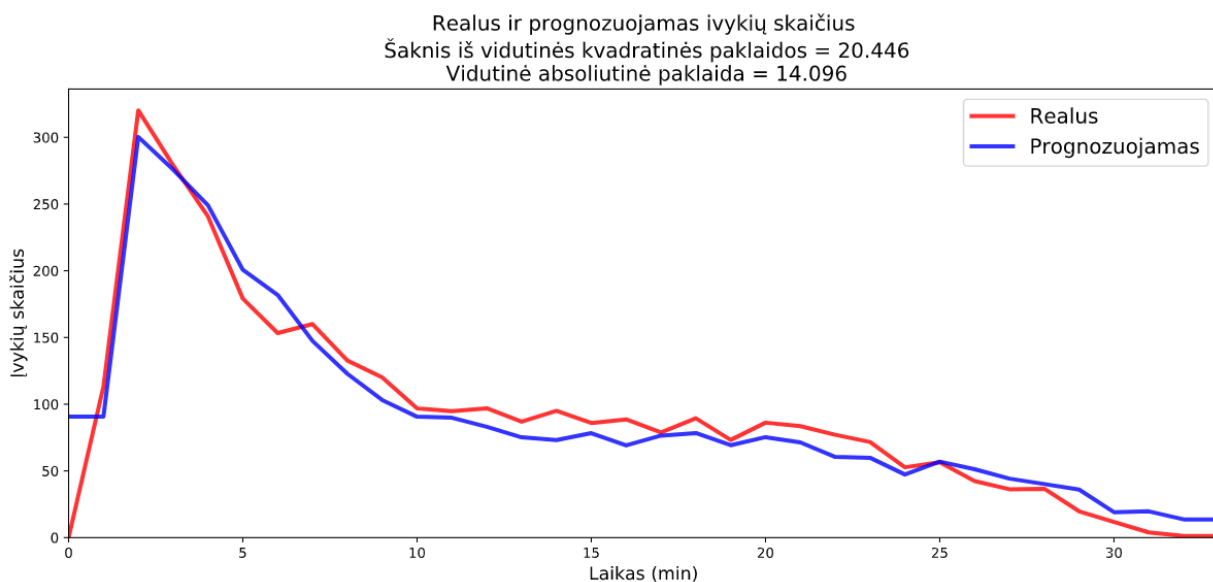


13 pav. Cox'o modelio Brierio įvertinimų kitimo kreivė

Aukščiau pavaizduotoje kreivėje (13 pav.) matome, jog modelis prasčiausius rezultatus rodo apie 2-3 minutes, po kurios seka vis tikslesnės prognozės, kol galiausiai įvertinimo reikšmė pasiekia artimą 0 ribą pačioje laiko pabaigoje. Tačiau apskaičiavus bendrą rezultatą, integruojant visais laiko t momentais, rezultatas (0.04) yra reikšmingai žemiau nustatytos ribos (0.25), kas nurodo, jog pagal turimus duomenis sudarytas modelis pakankamai tiksliai nustato žaidėjų išlikimo tikimybę.

Tiriant sudarytą modelį, viena iš gerųjų praktikų yra palyginti prognozuojamą ir realų įvykių skaičių kiekvienu laiko momentu t . Tai atliekama skaičiuojant išlikimo funkciją turimiems duomenims naudojant Kaplan-Meier statistinį įvertinimą įvykių tankiui nustatyti ir lyginant jį su prognozuotų įvykių tankiu kiekvienu laiko momentu.

14 pav. pavaizduotame grafike matyti, jog modelio prognozuojama įvykių tendencija (mėlyna linija) atitinka pagal realius duomenis paskaičiuotąją. Prognozuojamų įvykių grafikas sudarytas iš įvykusių įvykių skaičiaus kiekvieną žaidimo minutę. Pagal vidutinės absoliutinės paklaidos vertę (14.1) turint įvykių intervalą $[0; 330]$ galima teigti, jog modelis pakankamai tiksliai prognozuoja įvykių kiekį kiekvienu laiko momentu, kas sudaro vidutiniškai 14 žaidėjų prognozės skirtumą. Didžiausias nuokrypis pastebimas 5-6 minutėmis, kuomet modelis prognozuoja didesnę žaidėjų aktyvumą lyginant su realiu aktyvumu, o kitais laiko momentais, iki 24 minutės, modelis prognozuoja mažesnę aktyvumą. Taip pat pagal grafiką galima pastebėti bendrą žaidėjų aktyvumą. Pats aktyviausias laikotarpis žaidime yra pačioje pradžioje – pirmosios penkios minutes, tuomet jis eksponentiškai mažėja iki 10 minutes bei išlaiko pastovumą iki pat žaidimo pabaigos.



14 pav. Realus ir prognozuojamas įvykių skaičius tyrimo metu naudojant Cox'o regresiją

Lyginant realių ir prognozuojamų įvykių kreives, buvo naudojami visi turimi atributai. Tačiau svarbu palyginti ir atskirų atributų gerumą prognozuojant žaidėjų išlikimą žaidime, siekiant nustatyti, kurie atributai yra svarbiausi prognozuojant žaidėjų išlikimą. 2 lentelėje nurodyti 5 didžiausią įtaką prognozei turintys atributai. Atributo svarba matuojama pagal atitikimo indekso didumą analizėje naudojant tik vieną atributą.

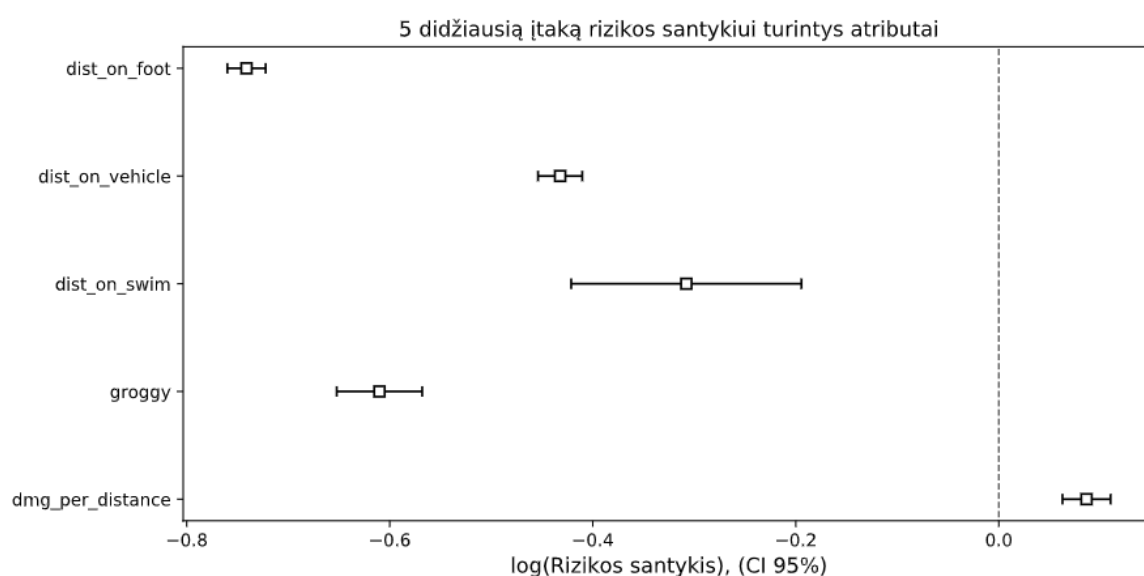
2 lentelė. Svarbiausi atributai prognozavime naudojant Cox'o regresijos modelį

Atributas	Atitikimo indeksas
<i>Dist_on_foot</i> (Nubėgtas atstumas pėsčiomis)	0.906355
<i>First Aid Kit</i> (Pirmosios medicininės pagalbos rinkinių skaičius)	0.755447
<i>Energy Drink</i> (Energetinių gėrimų, skirtų pagydyti, skaičius)	0.754864
<i>Heal_amount</i> (Pagydytų gyvybių skaičius)	0.753378
<i>5.56mm Ammo</i> (Šoviniai, skirti UMP9 ginklui)	0.728024

Lentelėje matoma, jog geriausią rezultatą rodo pėsčiomis nukeliauto atstumo atributas – „*dist_on_foot*“. Toliau seka su gyvybių didinimu susiję atributai bei šovinių, skirtų tam tikram ginklui, skaičius. Verta paminėti, jog atributai, skirti nurodyti ar žaidėjas žaidžia komandoje (angl. *squad*) ar

pavieniui (angl. *solo*), neturi jokios svarbos prognozavimui – atitikimo įvertinimai yra lygūs 0.504 ir tai nurodo atsitiktinį žaidėjų prognozavimą. Lentelę su visais atributais galima rasti priedų skiltyje.

Tuo pačiu galima peržvelgti didžiausią įtaką žaidėjo rizikai išgyventi darančius atributus. 15 paveikslėlyje matomas grafikas, su pavaizduotais 5 atributais, kurių įtaka žaidėjų rizikai buvo didžiausia – atstumas nukeliautas pėsčiomis (angl. *dist_on_foot*), transporto priemone (angl. *dist_on_vehicle*), plaukiant (angl. *dist_on_swim*), ar žaidėjas buvo pagydytas nuo sunkaus sužeidimo žaidimo metu (angl. *groggy*) ir žaidėjo atliktos žalos ir nukeliauto atstumo santykis (angl. *dmg_per_distance*). Grafike reikšmės pavaizduotos su 95 proc. pasitikėjimo intervalais, pavaizduotais linijomis iš šonų.



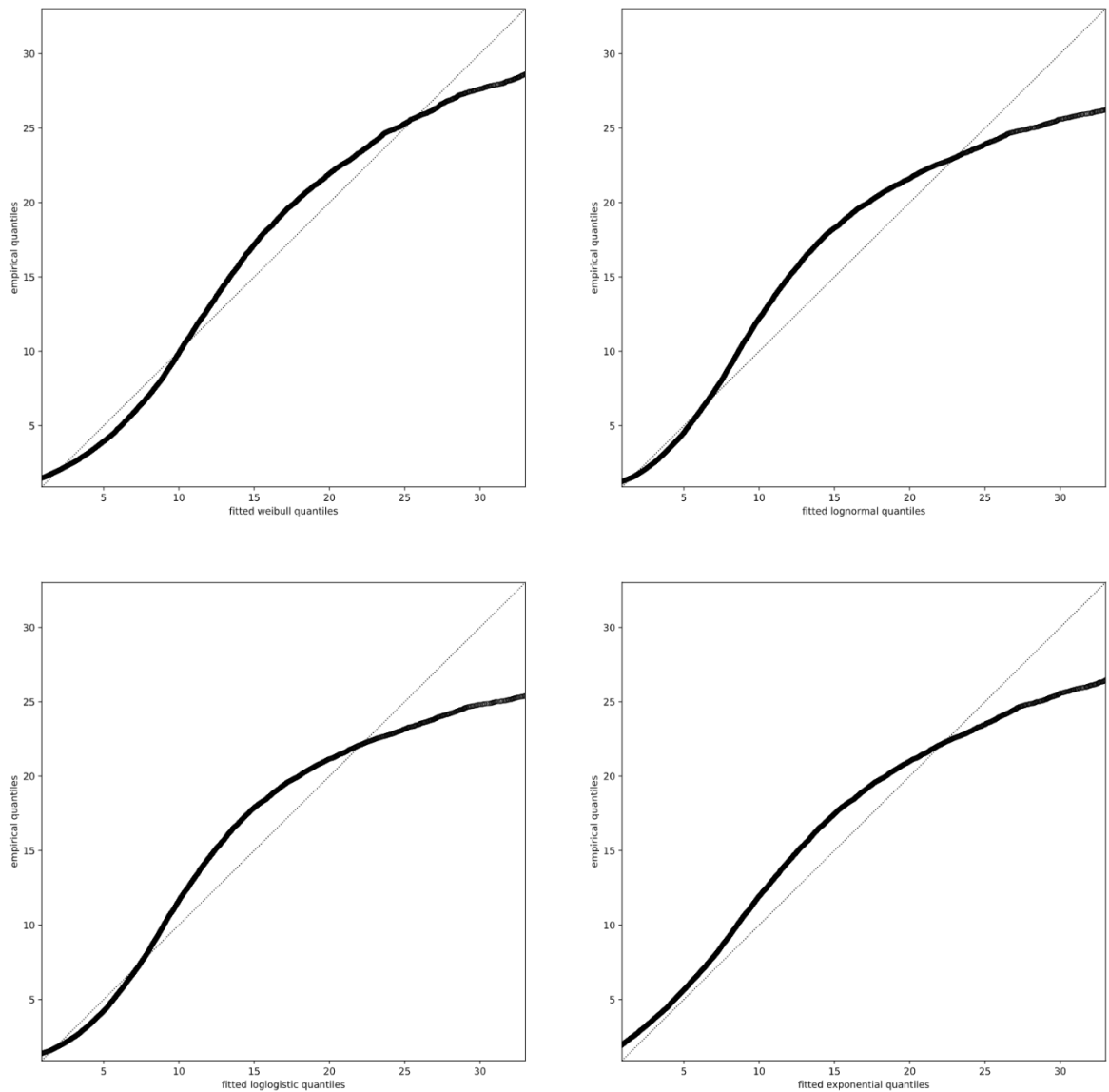
15 pav. Cox'o regresijoje didžiausią įtaką žaidėjo rizikai turintys atributai

Svarbu prisiminti, jog Cox'o regresijoje didesnė rizika reiškia didesnę galimybę žaidimui artimiausiu metu pasibaigti. Grafike matoma neigiama atributo koeficiento reikšmė reiškia atributui didėjant rizika mažėja, o teigiama – atributui didėjant rizika taip pat didėja. Pavyzdžiui, pasirinkus atributą *dist_on_foot*, galima apskaičiuoti, jog didėjant šiam atributui 1 dydžiu – žaidėjo rizikos santykis yra lygus $\exp(-0.74) = 0.48$. Tai reiškia, jog rizika mažėja 48 procentais. Grafike pavaizduotų atributų p-reikšmė yra mažesnė nei 0.005, todėl galima teigti, jog jie yra statistiškai reikšmingi rizikingumų modelyje. Pilną visų atributų grafiką galima rasti priedų skiltyje.

4.2.2. Pagreitintas įvykio laikas

Prieš taikant šį modelį, reikia pasirinkti tinkamą turimų įvykių laikų skirstinio tipą ir dažnu atveju prieš pasirinkimą nėra galimybės tiksliai pasakyti, kurio skirstinio tipo modelį naudoti. Vienas iš būdų tai sužinoti – vizualiai palyginti Q-Q (angl. *quantile-quantile*) grafikus (18 pav.) tarp skirtingų skirstinių modelių. Dažniausiai išlikimo analizėje pasitaikantys skirstinių tipai – Weibullio, log-normalusis, log-logistinis, eksponentinis. Būtent šia tvarka pradedant kairėje pusėje ir išrikiuoti grafikai žemiau esančiame paveiksle. Peržiūrėjus grafikus, sunku susidaryti išvadą, kuris iš pasirinktų skirstinių labiausiai atitinka įvykių laikus, nes nei viename iš jų nėra pastebimas santykinai didelis duomenų pasiskirstymas apie įstrižą liniją, kuri atspindi pasirinktą skirstinį. Tokiu atveju yra naudojamas kitas būdas siekiant pasirinkti vieną iš naudojamų skirstinių tipų – tikėtumo logaritmo (angl. *log-likelihood*) skaičiavimas.

Bendruoju atveju, naudojant tą patį duomenų rinkinį (toks pat subjektų ir atributų skaičius), tinkamesnis taikymui modelis yra tas, kurio tikėtumo logaritmas yra didesnis. Taigi, šiuo atveju buvo apskaičiuoti (3 lentelė) logaritmai ir tolimesniam taikymui pasirinktas log-logistinis skirstinys, kurio rezultatas buvo didžiausias ir lygus -115059.7.



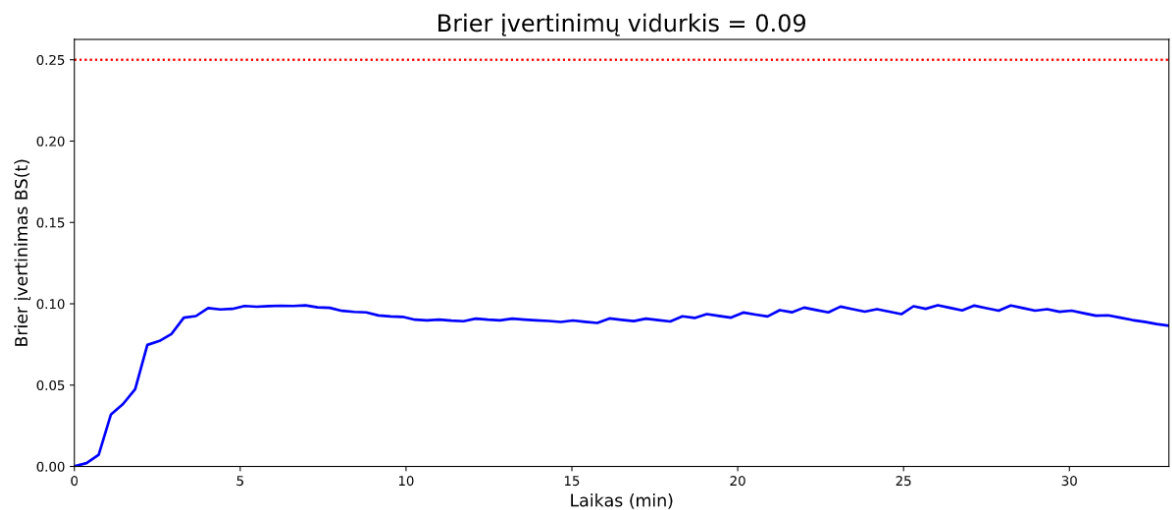
16 pav. Duomenų kvantilių grafikai pasirinktiems pagreitinto įvykių laiko modelio skirstinių tipams

Pritaikius pasirinktą modelį, gautas atitikimo statistinis įvertinimas (*c-index*) lygus 0.912. Šis rezultatas gana panašus (mažesnis 1.2 proc.) į gautąjį naudojant Cox'o regresiją ir taip pat įrodo šio modelio gerumą (subjektų poros surikiuotos 91.2 proc. tikslumu) skaičiuojant subjektų rizikingumą bei lyginant jį su turimu įvykio laiku.

3 lentelė. Modelių tikėtinumų logaritmai

Skirstinys	Rezultatas
Eksponentinis	-117241.2
Log-logistinis	-115059.7
Log-normalusis	-115104.1
Weibullio	-116207.0

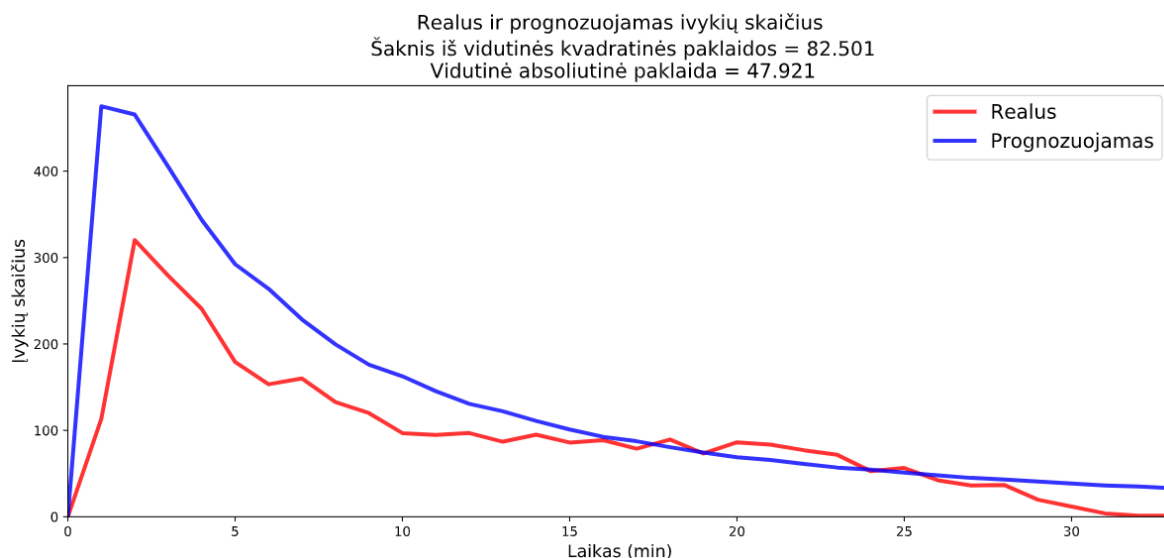
Atlikus Brierio testą ir nubraižius įvėčio kitimo laikui bėgant grafiką (17 pav.) galima teigti, jog sudarytas modelis pasirinkus log-logistinį skirstinį pakankamai sėkmingai (bendras vidurkis $IBS(t) = 0.088$) prognozuoja subjektų išlikimo tikimybes kiekvieno laiko momentu t .



17 pav. Brierio rezultatas naudojant pagreitinoto įvykio laiko modelį

Išnagrinėjus gautą grafiką, galima daryti išvadą, jog modelis kiekvienu laiko momentu gana tiksliai prognozuoja išlikimo tikimybes, nes grafikas nepasiekia atsitiktinio prognozavimo ribos, kuri yra pažymėta grafike punktyrine linija ir lygi 0.25. Skirtingai, nei nuo prieš tai taikyto modelio, šio rezultatai išlaiko panašią trajektoriją didžiąją dalį laiko ir nemažėja artėjant laikui prie pabaigos – laikosi apie 0.10 ribą. Dėl šios priežasties minėtas bendras rezultatas yra 2.1 karto prastesnis negu prieš tai taikyto Cox'o regresijos modelio. Prastesnį rezultatą patvirtina ir realaus ir prognozuojamo įvykių skaičius kiekvienu laiko momentu t grafikas (18 pav.). Jame matome, jog prognozuojama įvykių kreivė atitinka pasirinktą skirstinio tipą (log-logistinis), tačiau pats grafikas yra pasislinkęs į viršų ir sudaro didelį prognozuojamų įvykių skaičiaus skirtumą žaidimo pradžioje. Vidutinė kvadratinė modelio paklaida yra lygi 82.50. Bendrai šie rezultatai patvirtina, jog pasirinktas log-logistinis skirstinys neviseškai atitinka turimą įvykių skaičiaus skirstinį, ko pasekoje prognozės turi reikšmingą skirtumą. Nurodyta kvadratinė paklaida ypač

išryškina didelį prognozuojamų įvykių skaičiaus skirtumą žaidimo pradžioje. Vidutinė absoliutinė modelio paklaida kiekvienu laiko momentu yra lygi ~ 48 įvykiams ir patvirtina bendrą didelę prognozių paklaidą, bet ir nelygų paklaidų pasiskirstymą viso žaidimo metu, nes yra mažesnė nei šaknis iš vidutinės kvadratinės paklaidos.



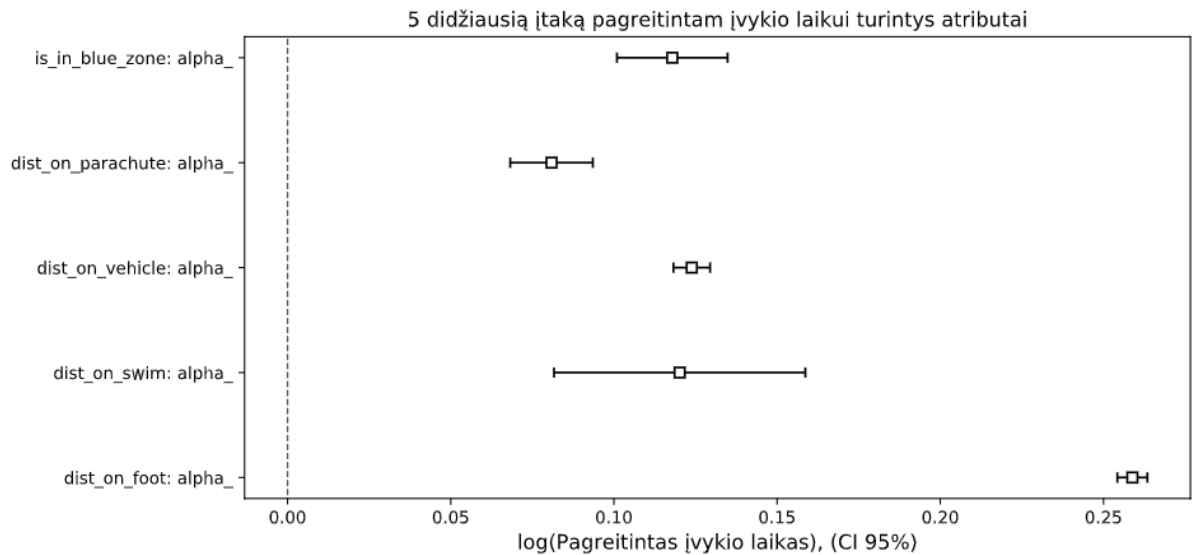
18 pav. Realus ir prognozuojamas įvykių skaičius tyrimo metu, naudojant pagreitinoto įvykio laiko modelį

Tikrinant atskirų atributų prognozavimo galimybes, gauta lygiai toks pat geriausiai prognozuojančių atributų penketukas, pavaizduotas 4 lentelėje. Bendroju atveju patys įvertinimai yra mažesni, turint omenyje bendrą modelio atitikimo įvertinimą, tačiau tiek išsidėstymas, tiek tarp atributų verčių esančios proporcijos išlieka panašios.

4 lentelė. Svarbiausi atributai prognozavime naudojant pagreitinoto įvykio laiko modelį

Atributas	Atitikimo įvertinimas
<i>Dist_on_foot</i> (Nubėgtas atstumas pėsčiomis)	0.904580
<i>First Aid Kit</i> (Pirmosios medicininės pagalbos rinkinių skaičius)	0.753465
<i>Energy Drink</i> (Energetinių gėrimų, skirtų pagydyti, skaičius)	0.753350
<i>Heal_amount</i> (Pagydytų gyvybių skaičius)	0.752700
<i>5.56mm Ammo</i> (Šoviniai, skirti UMP9 ginklui)	0.725513

Tuo pat metu galima pažvelgti ir į daugiausia įtakos pagreitintam įvykio laikui turinčius atributus. 21 paveiksle matomas penkių didžiausių įtaką turinčių atributų grafikas, kuriame pavaizduoti jų koeficientai kartu su 95 proc. pasitikėjimo intervalais.



19 pav. Penki didžiausių įtaką pagreitintam įvykio laikui turintys atributai

Didžiąją dalį atributų užima įvairiais būdais nukeliauto atstumo atributai – pėsčiomis (angl. *dist_on_foot*), parašiotu (angl. *dist_on_parachute*), transporto priemone (angl. *dist_on_vehicle*), vandeniui (angl. *dist_on_swim*) ir taip pat atributas – ar įvykis įvyko esant žaidėjui mėlynojoje zonoje (angl. *is_in_blue_zone*). Pavyzdžiui pasirinkus atributą *dist_on_vehicle*, eksponentę pakėlus gautu rezultatu galima apskaičiuoti, jog padidėjus šiam atributui vienu vienetu, vidutinis išgyvenimo laikas padidėja 13 procentų, nes $\exp(0.12) = 1.13$. Verta atkreipti dėmesį, jog šiame modelyje prie koeficientų esantys ženklai turi atvirkščią reikšmę, nei Cox'o regresijos modelyje – teigiamas koeficientas lemia prailgintą išgyvenimo laiką, kaip ir pateikta pavyzdyje aukščiau. Visų koeficientų grafikas taip pat pateiktas priedų skiltyje.

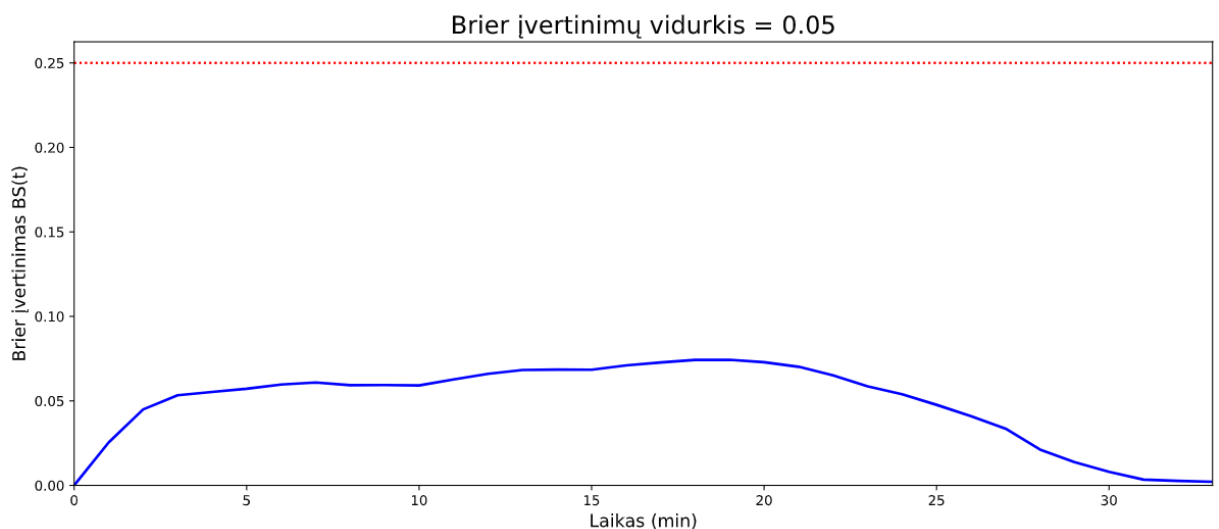
4.2.3. Atsitiktiniai išlikimo miškai

Naudojant modelį buvo pasirinkti šie parametrai, atsižvelgiant į miško formavimo laiką, kompiuterinės įrangos pajėgumą bei rezultatų gerumą:

- Medžių kiekis – 100;
- Atributų kiekis, renkantis geriausią mazgo dalinimą - $\sqrt{\text{atributų skaičius}} = \sqrt{72} = 8.48$;
- Mažiausias mazge esančių subjektų skaičius – 15;
- Miško gylis – 20 mazgų.

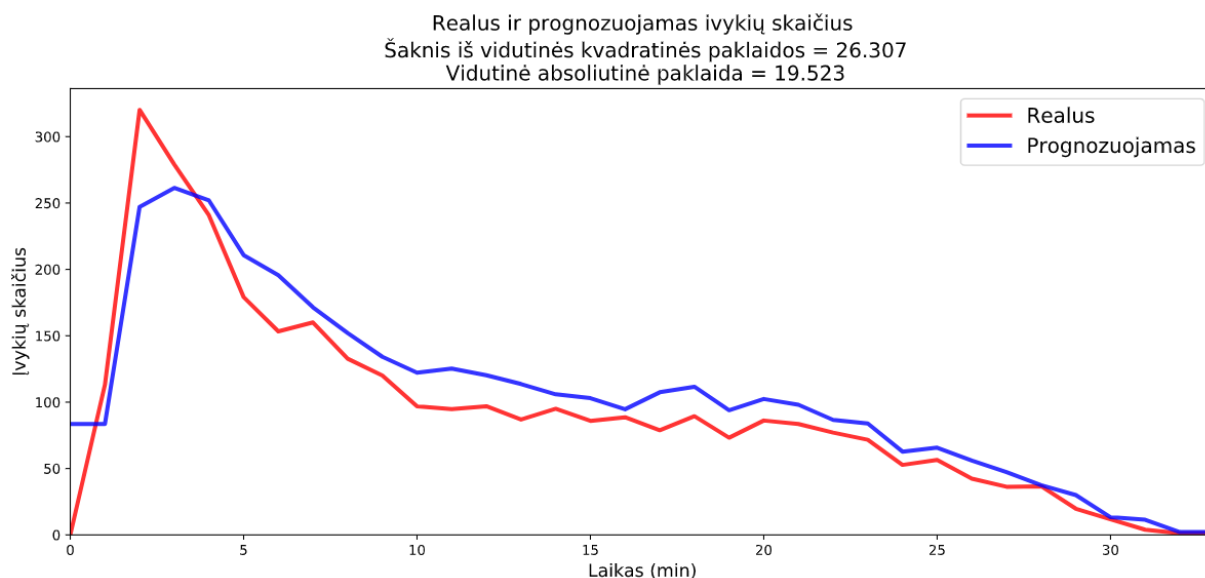
Sudarius modelį pagal turimus parametrus, gautas atitikimo statistinis įvertinimas lygus 0.911. Pagal šį įvertinimą, šis mašininio mokymosi metodas subjektus pagal rizikingumo ir įvykio laiko poras išdėsto didesniu tikslumu negu anksčiau taikyti statistiniai išlikimo analizės metodai.

Gautas Brierio bendras rezultatas ($IBS(30) = 0.050$), nurodo aukštą modelio tikslumą. Žvelgiant į rezultatų grafiką pavaizduotą 20 pav., matoma, jog kitaip negu prieš tai taikytuose modeliuose, paklaidos didėja iki pat antros žaidimo pusės ir pasiekus didžiausią vertę 20 minutę ima mažėti kol pasiekia minimalią vertę paskutinėmis tyrimo minutėmis.



20 pav. Atsitiktinių išlikimo miškų Brierio įvertinimų grafikas

Lyginant modelio prognozuojamų ir realių įvykių skaičiaus grafikus, galima teigti, jog modelis pakankamai tiksliai prognozuoja įvykių skaičius kiekvienu laiko momentu (vidutinė absoliutinė paklaida lygi ~19 įvykių). Didžiausios paklaidos matomos pirmojoje žaidimo dalyje, kuomet žaidėjų skaičius ir aktyvumas tarp jų yra didžiausias, ypač aktyvumo piko metu – 2–6 minutėmis. Toliau modelis išlaiko panašią prognozių paklaidą iki pat pabaigos, kuomet grafikai susilygina.



21 pav. Atsitiktinių išlikimo miškų modelių prognozuojamų įvykių skaičiaus grafikas

Galiausiai, pritaikius modelį, galima sudaryti svarbiausių kintamųjų sąrašą, skaičiuojant kintamųjų svarbą (*angl. feature importance*). Svarba nustatoma apskaičiuojant priemaišų (*angl. impurity*) dydį dalinant mazgus kiekvienu atributu (Breiman, 2001). Kuo mažesnis priemaišų dydis dukteriniuose mazguose – tuo atributas yra svarbesnis. Tačiau išlikimo analizėje, dėka galimo subjektų cenzūravimo, norint sužinoti svarbiausius atributus, skaičiuojama perstatymo (*angl. permutation*) svarba. Ji nurodo, kiek pagrindinis rezultatas (šiuo atveju – atitikimo įvertinimas) sumažėja, jeigu atributo nėra duomenų rinkinyje. Gautas svarbiausių atributų sąrašas pavaizduotas 3 lentelėje.

5 lentelė. Svarbiausi atributai naudojant permutaciją Atsitiktinių išlikimo miškų modelyje

Atributas	Skirtumo dydis
dist_on_foot	0.0696
dist_on_vehicle	0.0177
heal_amount	0.0159
5.56mm Ammo	0.0051
Military Helmet (Level 2)	0.0033

Pagal aukščiau nurodytus rezultatus, galima teigti, jog svarbiausias atributas išlikimo analizėje yra žaidimo metu nukeliamas atstumas kojomis (*angl. dist_on_foot*). Panaikinus jo sąryšį su išlikimo laiku (atsitiktinio perrinkimo metu), atitikimo įvertinimas sumažėja vidutiniškai 0.0696 taško (7.6 proc.). Toliau seka atstumo, nukeliamas naudojant transporto priemonę, atributas (*angl. dist_on_vehicle*), papildytų gyvybių kiekio (*angl. heal_amount*) atributas, surinktų 5.56mm kalibro šovinių (*angl. 5.56mm*

Ammo) ir žaidimo metu turėtų apsauginių šalmų (angl. *Military Helmet (Level 2)*) kiekio atributai. Svarbu atkreipti dėmesį, jog skirtumas tarp pirmos ir antros vietos atributų svarbos yra 4 kartai, o tai įrodo koks šiame modelyje svarbus yra „*dist_on_foot*“ atributas šiame modelyje.

4.2.4. Neuroniniai tinklai

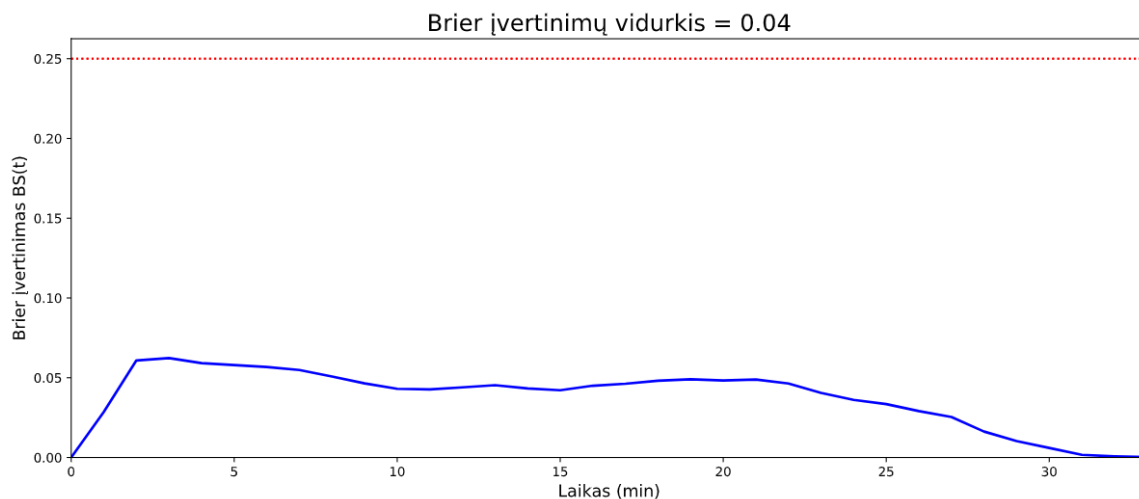
Šiame tyrime buvo naudojamas pilnai sujungtas daugiasluoksnis perceptronas, turintis 2 vidinius sluoksnius. Tinklo išorinį sluoksnį sudaro 1 neuronas, kuriame prognozuojama rizikos funkcija $\widehat{h}_{\theta}(x)$, kuri parametrizuota tinklo svoriais θ . Apmokymui naudojami parametrai pradžioje buvo gauti optimizacijos būdu pagal turimus duomenis, pasinaudojus autorių (Katzman, et al., 2017), kurių giliojo mokymosi modelis naudojamas šiame tyrime, optimizacijos funkcija. Gavus parametrus ir juos panaudojus modeliuojant, pagal gautus rezultatus buvo nuspręsta, pradedant nuo turimų parametrų, klaidų ir bandymų metodu ieškoti optimalresnių verčių, kurios yra pavaizduotos 4 lentelėje.

6 lentelė. Neuroninio tinklo parametrai

Parametras	Vertė
Sluoksnių skaičius	2
Aktyvacijos funkcija	SELU (angl. <i>Scaled Exponential Linear Unit</i>)
Neuronų skaičius vidiniame sluoksnyje	129
Mokymosi greitis (angl. <i>learning rate</i>)	$1 \cdot 10^{-3}$
Išmetimas (angl. <i>dropout</i>)	0.469
L2-regulatorius	4.692
Gradiento mažėjimo optimizacijos algoritmas	ADAM (angl. <i>Adaptive Moment Estimation</i>)

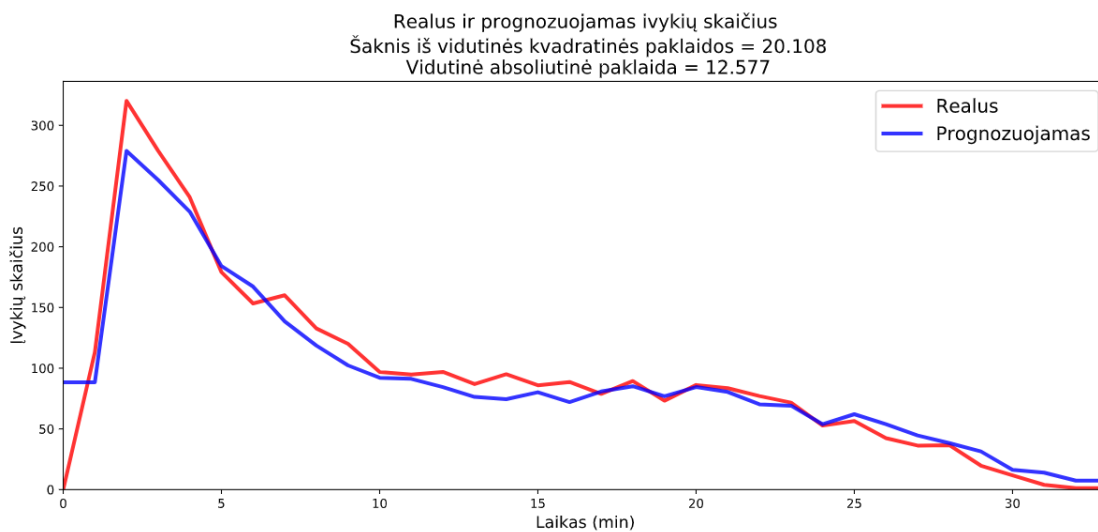
Pritaikius sudarytą neuroninį tinklą pagal nurodytus parametrus, gautas atitikimo rezultatas pranoksta prieš tai taikytų modelių ir yra lygus 0.926. Šis rezultatas parodo, tai yra tiksliausias naudojamas modelis ir tarp turimų atributų ir išgyvenimo laiko yra tam tikrų netiesinių priklausomybių, kurių įprastiems tiesinės regresijos modeliams nepavyksta aptikti.

Modelio prognozavimo savybes patvirtina ir gautas Brierio rezultatų grafikas. Bendras rezultatas siekia 0.038, o ir grafike (22 pav.) matoma, kad aukštas tikslumas yra išlaikomas visais laiko momentais. Santykinai prasčiausias prognozavimas vyksta žaidimo pradžioje, kaip ir minėta, esant didžiausiam žaidėjų aktyvumui.



22 pav. Brierio rezultatų grafikas naudojant neuroninių tinklų modelį

Lyginant prognozuojamą skaičių įvykių su realiu skaičiumi (23 pav.) kiekvienu laiko momentu (kiekis skaičiuojamas kiekvieną minutę), žaidimo pradžioje matomas panašus prognozavimas, kuris nurodomas mėlyna linija grafike, kaip ir naudojant Atsitiktinių Išlikimo Miškų modelį.



23 pav. Realio ir prognozuojamo įvykių kiekio grafikas naudojant neuroninių tinklų modelį

Didžiausias skirtumas fiksuojamas 2 minutę, kuomet jis yra lygus 41 žaidėjui – tai yra 3.3 karto daugiau negu vidutinė paklaida viso žaidimo metu. Tačiau bendra vidutinė paklaida (~12 žaidėjų) fiksuoja tiksliausią naudotą, siekiant nustatyti subjektų įvykių kiekį kiekvienu momentu, modelį šiame tyrime. Grafike galima pastebėti, jog pradedant 10 minute modelio tikslumas yra labai aukštas, o nuo 17 iki 24 minutės jis yra praktiškai maksimalus – prognozių grafikas atitinką realių įvykių grafiką. Tai indikuoja, jog modelis yra naudingesnis prognozuojant įvykių kiekį antroje žaidimo dalyje.

Galiausiai aptariami ir pasirinkto modelio svarbiausi atributai prognozuojant žaidėjų rizikingumą. Informacinė atributų svarba pateikiama atsižvelgiant į atitikimo įvertinimo skirtumą – kuo jis didesnis, tuo atributas yra svarbesnis modelyje.

7 lentelė. Svarbiausi atributai naudojant neuroninius tinklus

Atributas	Skirtumo dydis
<i>Dist_on_foot</i>	0.0382
<i>Dist_on_vehicle</i>	0.0085
<i>squad</i>	0.0043
<i>groggy</i>	0.0036
<i>Dmg_per_distance</i>	0.0035

Šio, kaip ir kitų, modelio svarbiausių atributų sąrašė pagrindinis atributas yra nueitas atstumas pėsčiomis (angl. *dist_on_foot*). Skirtumas taip pat išlieka panašus kaip ir kituose modeliuose – tarp pirmos ir antros vietų 4 kartai, o toliau atributai išsiskirsto tolygiai.

Pritaikius ir išnagrinėjus visus turimus modelius, kitame skyrelyje bus pasirinktas geriausias modelis, kurio pagalba bus atlikta išlikimo analizė atsitiktinai pasirinktiems subjektams iš skirtingų rizikingumo zonų bei išnagrinėta svarbiausių atributų daroma įtaka žaidėjų išlikimo tikimybei.

4.3. Žaidėjų išlikimo analizė

Žemiau pateikiama lentelė su turimais modeliavimo rezultatais (8 lentelė). Būtina paminėti, jog modeliams skirtų apmokymo ir testavimo imtys siekė atitinkamai 80 proc. ir 20 proc.

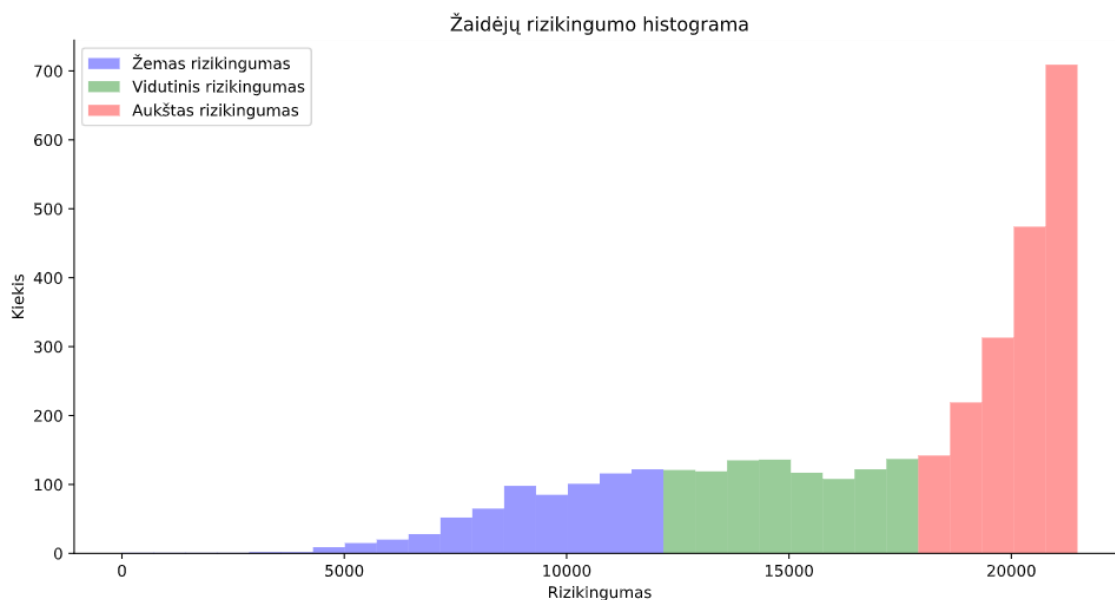
Kaip matoma 8 lentelėje, geriausius rezultatus pagal visus tikslumo parametrus fiksuoja neuroninių tinklų modelis. Išskirtinai gerus rezultatus parodė pusiau parametrinis Cox'o tiesinės regresijos modelis. Pagal visus gerumo vertinimo kriterijus jis fiksuoja gana panašias vertes. Taip pat mokymo laiko grafoje,

kuri parodo, jog apmokyti šį modelį užtrunka vos 23 sekundes. Šis privalumas bei nereikalingas išankstinis įvykių laikų skirstinio žinojimas daro Cox'o regresijos modelį ypač patrauklų sprendžiant išlikimo analizės uždavinius. Tačiau išlikimo analizei šiame darbe pasirenkamas neuroninių tinklų modelis.

8 lentelė. Išlikimo analizės modelių taikymo rezultatai

Modeliai	Statistiniai modeliai		Mašininio mokymosi modeliai	
	Cox'o regresija	Pagreitinto Įvykio Laikas	Atsitiktiniai Išlikimo Miškai	Neuroniniai tinklai
Atitikimo rodiklis (<i>c-index</i>)	0.921	0.911	0.908	0.926
Brierio rezultatas	0.044	0.091	0.050	0.038
Vidutinė absoliutinė paklaida	14.09	51.49	19.52	12.57
Mokymo laikas (min)	0:23	1:09	35:14	11:46

Pasirinkus geriausią modelį, sudaroma žaidėjų rizikingumų histograma (24 pav.), kurioje matomas prognozavimui skirtų žaidėjų rizikingumų skirstinys. Primenama, jog kuo didesnis žaidėjo rizikingumas – tuo mažesnė žaidėjo tikimybė išgyventi žaidime laikui bėgant. Žaidėjo rizikingumui skaičiuoti naudojamos turimų atributų vertės pagal sudaryto modelio koeficientus.

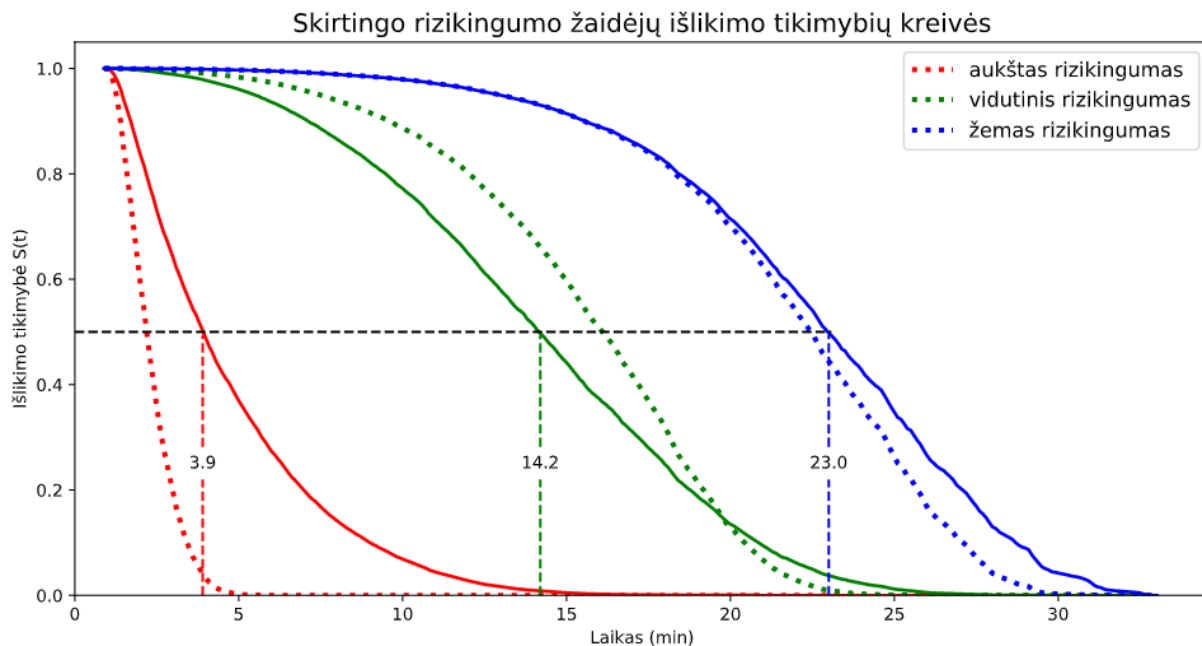


24 pav. Žaidėjų prognozuojamo rizikingumo histograma

Iš pateiktos histogramos matoma, jog didžioji dalis žaidėjų turi didelį rizikingumą, kuris eksponentiškai kyla pabaigoje, kas patvirtina turimas žinias, apie tai, jog didžioji dalis žaidėjų užbaigia

žaidimą jo pradžioje. Pati skaičių skalė grafike įtakos neturi, nes x koordinatinių ašies dydžiai yra sutartiniai ir svarbu jų išsidėstymas. Pavyzdžiui, rizikingumo intervalas $[0; 5]$ būtų toks pat kaip ir esantis pavaizduotame grafike kol išlaikomos tokios pačios skirstinio proporcijos. Rizikingumai padalinti į 3 (žemas, vidutinis, aukštas) dalis, siekiant sužinoti kaip gerai modelis prognozuoja skirtingo rizikingumo žaidėjų išlikimo tikimybes. Rizikingumai pasirinkti atsižvelgiant į rizikingumo dydį ir kitimo pobūdį. Padalinus rizikos sritis, iš kiekvienos pasirenkamas vienas žaidėjas ir naudojant turimą modelį apskaičiuojama išlikimo tikimybės funkcija bėgant laikui.

Pasirinktų žaidėjų išlikimo tikimybių kreivės (25 pav.) nurodo, jog modelis sėkmingai identifikavo skirtingo rizikingumo žaidėjus. Pasirinktiems subjektams apskaičiuotos išlikimo kreivės pagal turimus atributus įrodo teorinę išlikimo analizės dalį, jog didesnis rizikingumo lygmuo nurodo mažesnę tikimybę išgyventi laikui bėgant. Vientisomis linijomis pavaizduotos išlikimo tikimybių kreivės, kurias sudaro visų, į tam tikrą rizikingumo zoną patekusių, žaidėjų išlikimo kreivių vidurkis. Žaidėjo išgyvenimo laiko medianos (išlikimo tikimybė - $S(t) = 0.5$), esant tam tikroje rizikingumo zonoje, pavaizduotos punktyrinėmis vertikaliomis linijomis. Didžiausio rizikingumo išgyvenimo laiko mediana yra lygi 3.9 minutės, vidutinio rizikingumo – 14.2, o mažiausio – 23.0.

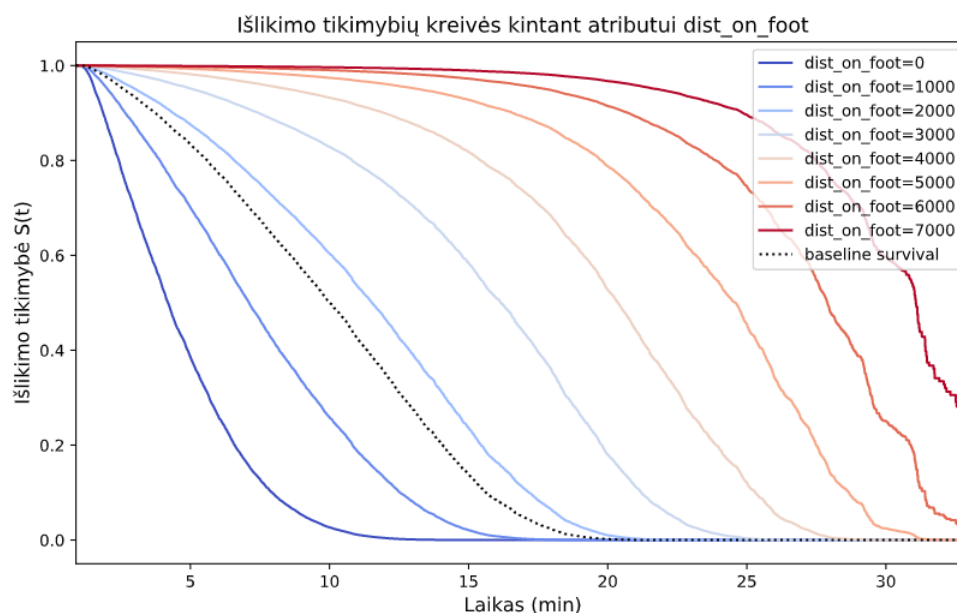


25 pav. Skirtingo rizikingumo žaidėjų išlikimo kreivių grafikas

Taškinėmis linijomis pavaizduotos atsitiktiniu būdu pasirinktų žaidėjų iš kiekvienos rizikingumo zonos išlikimo tikimybių kreivės, idant būtų galima palyginti zonos vidurkį su konkreto žaidėjo atveju.

Grafike kreivių kitimo tendencijos tarp vidutinio bei pasirinkto žaidėjo išlieka proporcingai panašios ir tuo pačiu, žvelgiant į kreives, galima teigti, jog kiekvienos grupės grafikas kinta skirtingai. Aukščiausio rizikingumo žaidėjų išlikimas nuo pat pradžios mažėja eksponentiškai, vidutinio rizikingumo išlaiko tolydų mažėjimą visu žaidimo laikotarpiu, o mažo rizikingumo grafike tikimybės išgyventi ryškus mažėjimas pastebimas tik žaidimui perkopus 15 minutę ir pasiekia žemiausią tašką tik pačioje žaidimo pabaigoje. Taigi, galima daryti išvadą, jog skirtingo rizikingumo išlikimo kreivės nurodo netik skirtingą išlikimo laiko medianą, bet ir skirtingas išlikimo tikimybių kitimo tendencijas bėgant laikui.

Pasirinkus atributus pagal jų svarbą neuroninių tinklų modelyje, galima palyginti, kaip kintant pasirinktam atributui keičiasi bazinės subjektų išlikimo tikimybės kreivė, kol visi kiti atributai nekinta. Minėta bazinė išlikimo kreivė gaunama turint išlikimo modelį, kuomet visi parametrai yra statiški (Breslow, 1972). Žemiau pateiktame grafike (26 pav.) matomas išlikimo kreivių kitimas, kintant pėsčiomis nueitam atstumui.

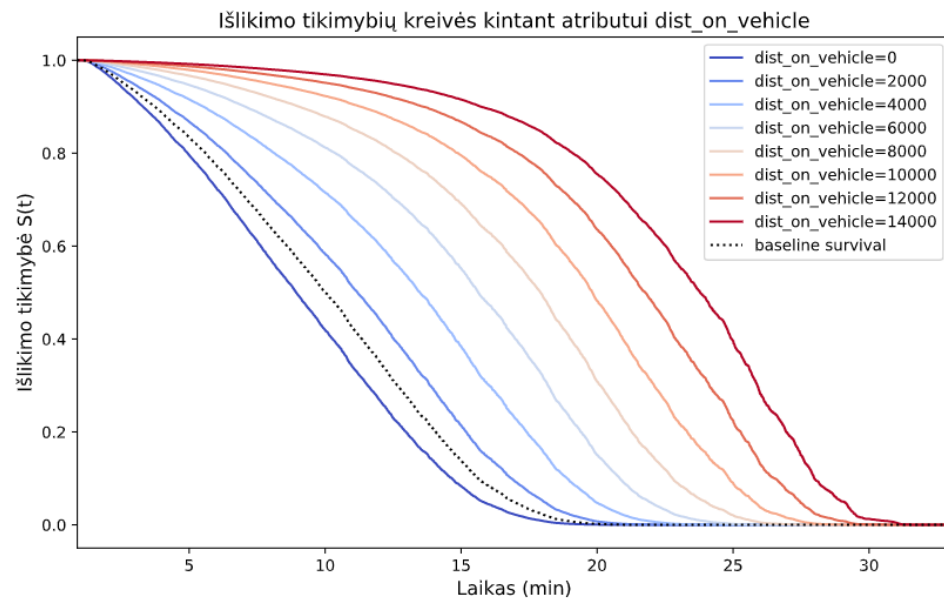


26 pav. Išlikimo tikimybių kreivės kintant atributui *dist_on_foot*

Pavaizduotame grafike taškine linija pavaizduota anksčiau minėta bazinė išlikimo kreivė, kuri įgauna vis kitokią formą, priklausomai nuo atributo vertės. Pats atributas padalintas į 8 dalis, kurios kinta nuo 0 iki 7000 metrų (maksimalus nueitas atstumas sudarytame duomenų rinkinyje). Pagal sudarytas grupės matoma, jeigu žaidėjo nueitas atstumas yra mažesnis už 3000 metrų – išlikimo tikimybė yra žemesnė nei bazinė, todėl taktika slapstytis kuo mažiau judant žemėlapyje nėra veiksminga. Didėjant

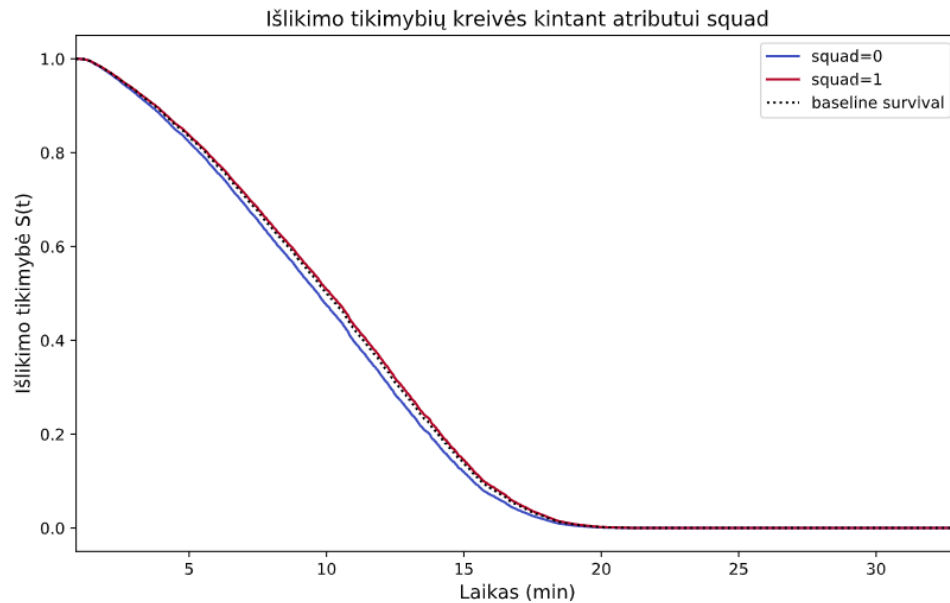
nukeliautam atstumui pėsčiomis – didėja išlikimo tikimybė kiekvienu laiko momentu. Tuo pat metu matomas ir gana ryškus pokytis tarp atstumų grupių – esant išlikimo tikimybei lygiai 0.5, vidutinis laiko tarpas tarp grupių yra lygus 3.9 minutės. Tai rodo, jog didėjantis nukeliautas pėsčiomis atstumas reikšmingai didina išlikimo tikimybes.

Toliau pavaizduotas grafikas (27 pav.), kuriame parodytas išlikimo kreivių kitimas kintant transporto priemone nukeliautam atstumui. Pateiktame paveiksle galima pastebėti tokią pačią kitimo tendenciją kaip ir prieš tai nagrinėto atributo – didėjant nuvažiuotam atstumui didėja ir išlikimo tikimybė. Nenaudoti transporto priemonių žaidimo metu yra taip pat nenaudinga, nes tai lemia žemesnę nei mažesnę nei bazinį išlikimą visais laiko momentais.



27 pav. Išlikimo tikimybių kreivės kintant atributui *dist_on_vehicle*

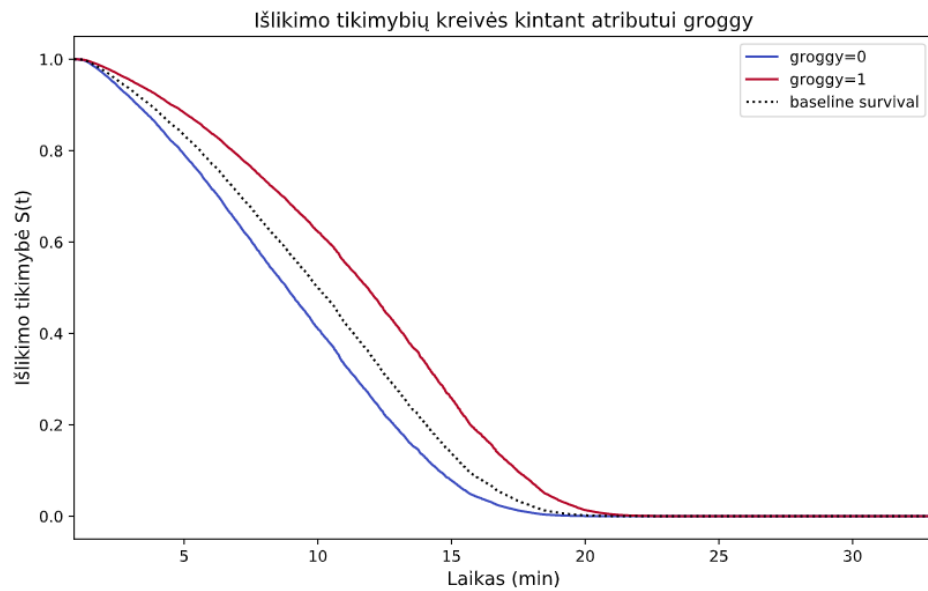
Tačiau naudojant šį grafiką galime teigti, jog šis atributas mažiau daro įtakos išlikimui – vidutiniškai tarp kreivių, esant $S(t) = 0.5$, yra 2.2 minučių skirtumas. Lyginant su prieš tai nagrinėtu atributu – skirtumas 44 proc.



28 pav. Išlikimo tikimybių kreivės kintant atributui *squad*

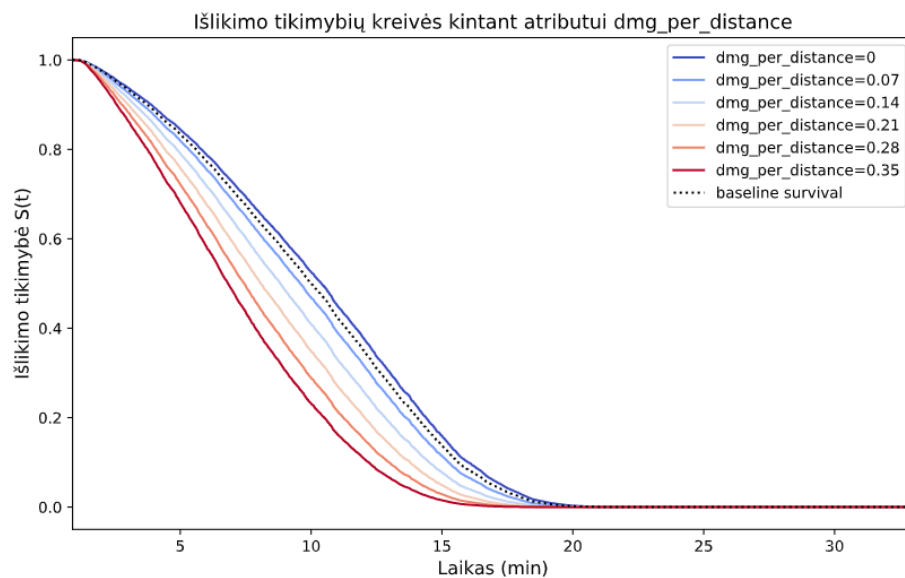
Pasirinkus atributą, nurodantį ar žaidėjas žaidė kartu su komanda (*squad*), matoma, jog didelės įtakos išlikimo tikimybei jis neturi (28 pav.). Esant komandoje, kuomet atributas lygus 1, tikimybės kreivė persidengia su bazine, tačiau žaidžiant pavieniui, tikimybė ilgiau išgyventi sumažėja. Vizualiai pastebėti skirtumą gana sunku, tačiau atlikus *logrank* testą, kuris tikrina hipotezę, ar dvi kreivės yra skirtingos, gauta p reikšmė lygi $3.84 \cdot 10^{-14}$ ir kadangi tai yra mažiau nei 0.05, tai leidžia atmesti nulinę hipotezę, jog tarp dviejų kreivių nėra statistiškai reikšmingo skirtumo.

Toliau pavaizduotas atributo, nurodančio ar žaidėjas buvo pagydytas nuo sunkaus sužeidimo žaidimo metu (angl. *groggy*) atributo tikimybių kreivių grafikas. Jame, matomas gana ryškus skirtumas tarp bazinės kreivės ir skirtingų atributo verčių. Buvus pagydytam žaidėjui (atributas lygus 1), vidutinė išgyvenimo tikimybė pasiekama 3.6 minutės vėliau, negu nebus pagydymo.



29 pav. Išlikimo tikimybių kreivės kintant atributui *groggy*

Paskutinis atributų sąrašė esantis žalos kiekio ir nueito atstumo santykio atributas, nurodo atvirkščią priklausomybę – didėjant šiam atributui trumpėja žaidėjo išgyvenimo laikas. Tai parodo, jog net ir didėjant padarytam žalos kiekiui, bet esant pastoviam nukeliautam atstumui išlikimo tikimybė mažėja.



30 pav. Išlikimo tikimybių kreivės kintant atributui *dmg_per_distance*

Galiausiai išnagrinėjus atributų įtaką išlikimo kreivei, galima apibūdinti žaidėją, kuris turi didelę tikimybę ilgai išlikti žaidime:

- Didelį atstumą keliauja pėsčiomis. Žaidėjų, nužingsniavusių mažiau nei 2000 metrų žaidimo metu, vidutinė išlikimo tikimybė yra mažesnė nei bazinė, todėl svarbu keliauti žaidimo metu kuo daugiau.
- Žaidėjai, keliaujantys įvairiomis transporto priemonėmis, taip pat didina savo išlikimo tikimybę. Galima pastebėti, jog minimalus nukeliautas nėra ypač svarbus, tačiau nukelivus transportu nemažiau nei 10000 metrų, išlikimo laiko mediana (išlikimo tikimybė, $S(t) = 0.5$), pasiekama ~10 minučių vėliau negu juo nesinaudojant.
- Žaidimas komandoje. Žaidėjai, žaidžiantys komandoje vidutiniškai išlieka 2 minutėmis ilgiau, negu pavieniui. Tam turi įtakos įvairios komandinės taktikos bei galimybė būti pagydytam.
- Žaidėjai, pagydyti po sunkaus sužeidimo, turi didesnę tikimybę išgyventi ilgiau. Atributas nenurodo, jog yra svarbu būti smarkiai sužeistam žaidime, tačiau tai: a) leidžia toliau tęsti žaidimą ir b) žaidėjas galimai ima žaisti atsargiau taip padidindamas savo tikimybę išlikti žaidime ilgesnį laiką.
- Žaidėjai, kurie koncentruojasi į slapstymąsi, o ne kovą su kitais žaidėjais. Atributas, nurodantis žalos kiekį per nukeliautą atstumą nurodo, jog didėjant žalos kiekiui mažėja išlikimo tikimybė, jeigu atstumas nekinta, todėl patartina daugiau laiko skirti keliavimui po žemėlapi, norint išlikti žaidime ilgesnį laiką.

REZULTATAI IR IŠVADOS

Atlikus uždavinius, iškeltus darbo pradžioje, buvo sudarytos šios išvados:

1. Informaciniuose šaltiniuose („Google Scholar“, „Researchgate“), galima rasti daugiau nei 25 tūkst. per pastaruosius 5 metus publikuotų straipsnių, kuriuose nagrinėjamos su išlikimo analize susijusios problemos. Atlikta literatūros analizė parodė, jog tyrimai dažniausiai apima žaidėjų pasitraukimo (angl. *churn*) arba žaidėjų elgsenos paties žaidimo metu ypatumus. Dažniausiai taikomi metodai – neuroniniai tinklai ir Cox'o regresija. Atlikus analizę, buvo pasirinkti ir išnagrinėti keturi išlikimo analizei skirti modeliai – du statistiniai ir du mašininio mokymosi. Statistinius sudaro Cox'o regresija bei pagreitintas įvykio laikas, o mašininio mokymosi – atsitiktiniai išlikimo miškai bei neuroniniai tinklai.
2. Sudarytas analizei skirtas duomenų rinkinys iš turimų, atsitiktiniu būdu pasirinktų, žaidimų pakartojimų. Rinkinį sudaro 17895 žaidėjai bei 80 atributų. Surinkus ir apdorojus reikalingus duomenis, rinkinio talpa sumažėjo iki 8 MB, iš turėtų 3.2 GB, kuriuos užėmė pakartojimų failai.
3. Atlikus pradinę duomenų analizę, nustatyta, jog didžiausias žaidėjų aktyvumas fiksuojamas žaidimo pradžioje – 3–5 žaidimo minutėmis. Sudarius Kaplan-Meier išlikimo tikimybių kreives kategoriniams kintamiesiems, pastebėta, jog didžiausią skirtumą išlikimui, nepaisant kitų atributų įtakos, sudaro atributas, nurodantis ar žaidėjas buvo mėlynojoje zonoje. Laiko momentai, esant išlikimo tikimybei lygiai 0.5, skiriasi 49 proc..
4. Pritaikius pasirinktus modelius ir lyginant pagrindinį gerumo matą – atitikimo įvertinimą, Cox'o regresija fiksuoja vertę lygiai 0.921, pagreitinto įvykio laiko metodas – 0.911, atsitiktinių išlikimo miškų metodas – 0.908, o neuroniniai tinklai – 0.926.
5. Bendrai įvertinus visus metodų gerumo matus, geriausiu iš pasirinktų modelių išrinktas neuroninių tinklų modelis. Jį naudojant buvo atlikta žaidėjų išlikimo analizė - sudarytos ir palygintos prognozuotų skirtingų rizikingumų žaidėjų išlikimo tikimybių kreivės. Jos atskleidžia, jog didėjant rizikingumui mažėja prognozuojamas išlikimo žaidime laikas. Išlikimo tikimybės mediana esant dideliame rizikingume pasiekiamas 3.9 minutę, vidutiniame – 14.2, o mažame – 23 minutę. Pateiktas pagrindinių atributų pagal svarbą sąrašas nurodo, jog žaidėjai, nukeliaujantys didesnę atstumą įvairiais būdais (pėsčiomis, transportu) yra linkę ilgiau išgyventi žaidime. Taip pat teigiamą įtaką išlikimo trukmei žaidime turi, tai ar žaidėjas žaidžia komandoje, ar buvo sužeistas žaidimo metu ir didėjantis žaidėjo atliktos žalos per nukeliautą atstumą santykis.

Taigi, pagal iškeltus uždavinius ir gautus rezultatus galima teigti, jog pasirinktas modelis padeda sėkmingai prognozuoti žaidėjų išlikimo tikimybes bei identifikuoti pagrindinius žaidėjų atributus.

LITERATŪROS SĄRAŠAS

- Andersen, P. K. & Keiding, N. (1998). Survival Analysis, Overview. *Encyclopedia of Biostatistics* 6, 4452–4461.
- Andreeva, G. (2006 m. October). European Generic Scoring Models Using Survival Analysis. *The Journal of the Operational Research Society*, 57(10), 1180–1187.
- Aung, M., Cowling, P., Bonometti, V., Drachen, A., Kokkinakis, A., Yoder, C. & Wade, A. (2018). Predicting Skill Learning in a Large, Longitudinal MOBA Dataset. *2018 IEEE Conference on Computational Intelligence and Games (CIG)* (p. 1–7). Maastricht: IEEE.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Breslow, N. E. (1972). Discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society*, 34, 216–217.
- Brier, G. (1950). *Verification Of Forecasts Expressed In Terms Of Probability*. doi:10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2
- Browlee, J. (2016). *Supervised and unsupervised machine learning algorithms*.
- Ching, T., Zhu, X. & Garmire, L. X. (2018). Coxnnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLOS Computational Biology*, 14(4).
- Clark, T., Bradburn, M. & Love, S. (2003). Survival Analysis Part IV: Further concepts and methods in survival analysis. *Br J Cancer*(89), 781–786.
- Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society*, 34, 187–202.
- Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J. & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88(11), 2783–2792.

- Demediuk, S., Murrin, A., Bulger, D., Hitchens, M., Drachen, A., Raffae, W. L. & Tamassia, M. (2018). Player retention in league of legends: a study using survival analysis. *Australasian Computer Science Week Multiconference*. New York.
- Fadnavis, R. A. (2019 May). Application of Machine Learning For Survival Analysis. A Review. *Journal of Engineering*, 09(5), 56–60.
- Faruk, A. (2018). The comparison of proportional hazards and accelerated failure time models in analyzing the first birth interval survival data. *Journal of Physics: Conference Series*, 974.
- Friedman, J. H. (2000). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29, 1189–1232.
- Fu, C.-Y., Lee, J., Bansal, M. & Berg, A. C. (2017). Video Highlight Prediction Using Audience Chat Reactions. *arXiv:1707.08559*.
- Graf, E., Schmoor, C., Sauerbrei, W. & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statist. Med.*, 18, 2529–2545. doi:doi:10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5
- Hacking, I. (2006). The Emergence of Probability: A Philosophical Study of Early Ideas About Probability Induction and Statistical Inference. *Cambridge University Press*, 2nd edition, 11–12.
- Harrell, F. E., Califf, R. M. & Pryor, D. B. (1982). Evaluating the Yield of Medical Tests. *JAMA*, 2543–2546.
- Heagerty, P. & Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1), 92–105.
- Hodge, V., Devlin, S., Sephton, N., Block, F., Drachen, A. & Cowling, P. (2017). Win Prediction in Esports: Mixed-Rank Match Prediction in Multi-player Online Battle Arena Games. *Digital Creativity Labs*.
- Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. (2008). Random Survival Forests. *The Annals of Applied Statistics*, 841–860. doi:10.1214/08-AOAS169

- Janilionis, V. (2010). Koreliacinės ir regresinės analizės pagrindai. *Mokymai apie kiekybinių ir kokybinių HSM tyrimų duomenų analizės metodus*, 6–7.
- Kaplan, E. L. & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282), 457–481. doi:10.1080/01621459.1958.10501452
- Katona, A. (2019). Time to Die: Death Prediction in DotA 2 using Deep Learning. *IEEE*.
- Katzman, J., Shaham, U., Bates, J., Cloninger, A., Jiang, T. & Kluger, Y. (2017). *DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network*. arXiv:1606.00931. doi:10.1186/s12874-018-0482-1
- Kenah, E. (2013). Nonparametric survival analysis of infections disease data. *Journal of the Royal Statistical Society.*, 75(2), 277–303.
- Klein, J., Houwelingen, H., Ibrahim, J., & Scheike, T. (2014). *Handbook of Survival Analysis*. Chapman & Hall.
- Kourou, K. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, 8–17.
- Kvamme, H., Borgan, O. & Scheel, I. (2019). *Time-to-Event Prediction with Neural Networks and Cox Regression*. University of Oslo, Oslo.
- Lai, T. & Zhiliang, Y. (1991). Large Sample Theory of a Modified Buckley-James Estimator for Regression Analysis with Censored Data. *The Annals of Statistics*, 19(3), 1370–1402.
- Lambert, J. & Chevret, S. (2014). Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent ROC curves. *Statistical Methods in Medical Research*.
- Lewis, M., & Wu, J. (2018 December). Esports Viewership and Customer Engagement: The SMITE Case Study. *Emory Marketing Analytics*.
- Makarov, I. (2017). Predicting Winning Team and Probabilistic Ratings in “Dota 2” and “Counter-Strike: Global Offensive” Video Games. *Analysis of Images, Social Networks and Texts*.

- Malinowski, M. (2018). *An Unsupervised Cluster: Learning Water Customer Behavior Using Variation of Information on a Reconstructed Phase Space*. Marquette University, Electrical Engineering and Computer Science.
- Martinsson, E. (2017). *WTTE-RNN: Weibull Time To Event Recurrent Neural Network*. University of Gothenburg, Gothenburg.
- Najafabadi, M. M., Villanustre, F. & Khoshgoftaar, T. M. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data* 2.
- Nasejje, JB., Mwambi, H., Dheda, K. (2017). A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC Med Res Methodol*, 17.
- Pannu, A. (2015 April). Artificial Intelligence and its application in different areas. *International Journal of Engineering and Innovative Technology*, 4(10).
- Powell, J. L. (1994). Estimation of semiparametric models. *Handbook of econometrics*(4), 2443–2521.
- Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C., Nussenbaum, B., Wang, E. W. (2010). A practical guide to understanding Kaplan-Meier curves. Otolaryngology- head and neck surgery. *Official journal of American Academy Of Otolaryngology-Head and Neck Surgery*, 143(3), 331–336.
- Russell, S. J. & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (T. III). Pretence Hall.
- Schubert, M. (2016). Esports analytics through encounter detection. *MIT Sloan Sports Analytics Conference*. MIT.
- Simon, R. & Faraggi, S. (1995). A neural network model for survival data. *Statistical Medical Journal*(14), 73–82.
- Statista. (2019). Nuskaityta iš Esports audience size worldwide from 2012 to 2022: www.statista.com/statistics/490480/global-esports-audience-size-viewer-type/
- Statista. (2020). Nuskaityta iš eSports audience size worldwide from 2012 to 2022, by type of viewers(in millions): <https://www.statista.com/statistics/490480/global-esports-audience-size-viewer-type/>

- Stepanova, M. & Lyn, T. (2002). Survival Analysis Methods for Personal Loan Data. *Operations Research*, 50(2), 277–289.
- Swindell, W. R. (2009). Accelerated failure time models provide a useful statistical framework for aging research. *Experimental gerontology*, 44(3), 190–200.
doi:<https://doi.org/10.1016/j.exger.2008.10.005>
- Tang, V., Harvey, D., Dorsay, J., Jiang, S. & Rathbone, M. (2007). Prognostic indicators in metastatic spinal cord compression: using functional independence measure and Tokuhashi scale to optimize rehabilitation planning. *Spinal Cord*, 45(10), 671–677.
- TheTechGame. (2018 September). NEWS. Nuskaityta iš 'PUBG' PS4 RATING DISCOVERED:
<https://www.thetechgame.com/News/sid=16774/pubg-ps4-rating-discovered.html>
- Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. & Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statist. Med.*, 30, 1105–1117.
- Wang, P., Yan, L. & Chandan, K. R. (2017 m. March). Machine Learning for Survival Analysis: A Survey. *ACM Computing Surveys*, 1(1).
- Wei, L. J. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Stat. Med.*, 11, 1871–1879.
- Xiang, A. (2000). Comparisson of the performance of neural network methods and cox regression for censored survival data. *Computational Statistical Data Analysis*, 34(2), 243–257.
- Zhao, L., & Feng, D. (2019). DNNSurv: Deep Neural Networks for Survival Analysis Using Pseudo Values. *arXiv:1908.0233*.
- Zhongzhi, S. (2011). *Advanced Artificial Intelligence*. China: Chinese Academy of Sciences.
doi:doi.org/10.1142/7547
- Zupan, B. (2000). Machine learning for survival analysis: a casestudy on recurrence of prostate cancer. *Artifical Intelligence in Medicine*, 20, 59–75.

Priedai

A priedas

Pilna naudojamų duomenų atributų struktūra

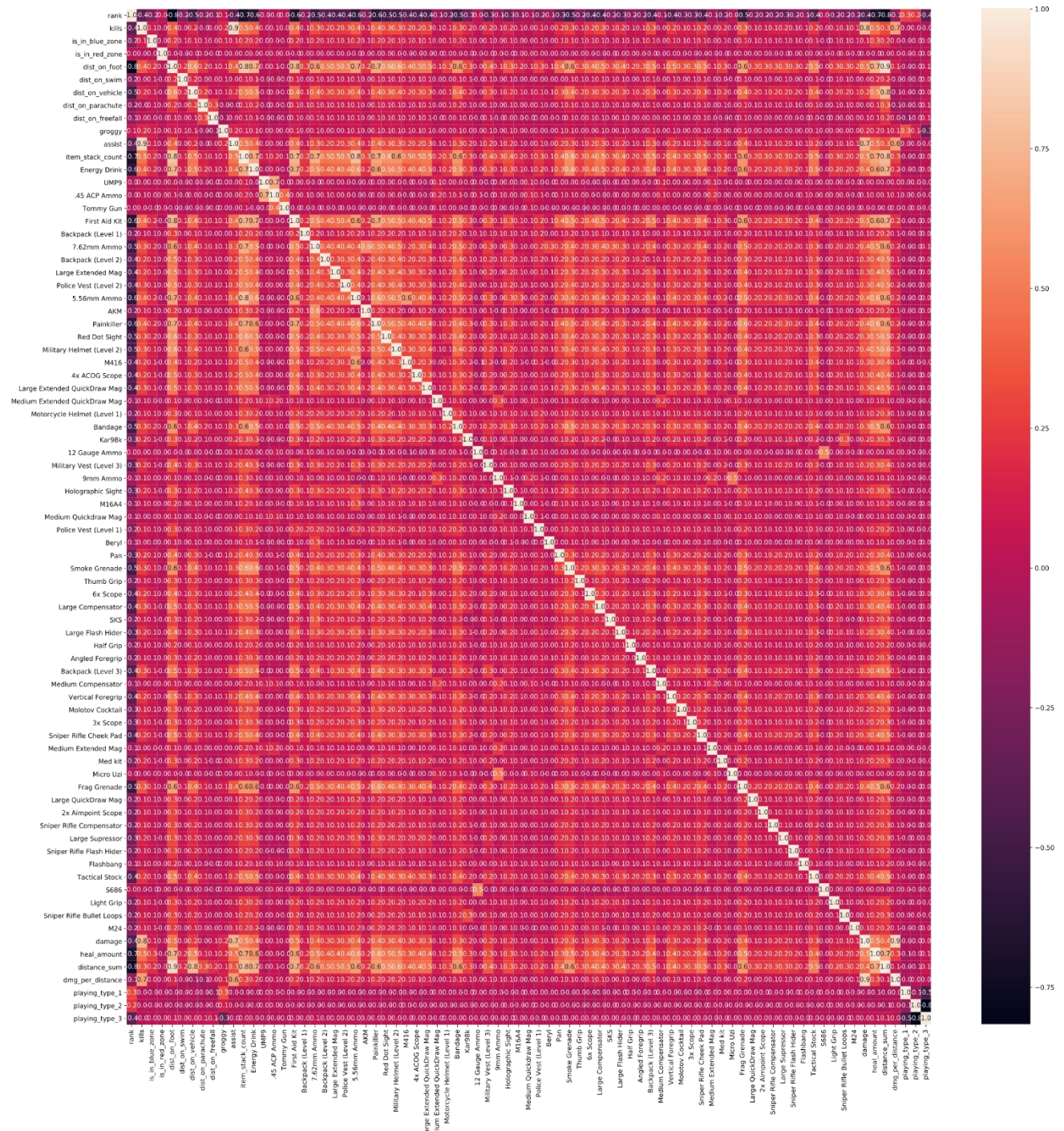
Kintamasis	Tipas	Reikšmės	Aprašymas
Name	Skaitinis	<i>Player1,</i> <i>Player2</i>	Žaidėjo slapyvardis žaidime
Rank	Sveikasis skaičius	<i>1,2,3...</i>	Žaidėjo vieta tarp kitų žaidėjų pasibaigus žaidimui
Kills	Sveikasis skaičius	<i>0,1,2...</i>	Žaidėjo nužudymų skaičius
Is_in_blue_zone	Kategorinis	<i>0,1</i>	Ar žaidėjas pasibaigus žaidimui buvo mėlynojoje zonoje
Is_in_red_zone	Kategorinis	<i>0,1</i>	Ar žaidėjas pasibaigus žaidimui buvo raudonoje zonoje
Dist_on_foot	Realusis skaičius	<i>28.97,</i> <i>1487.22</i>	Kojomis nukeliantas žaidėjo atstumas
Dist_on_swim	Realusis skaičius	<i>33.11,</i> <i>224.29</i>	Nuplauktas žaidėjo atstumas
Dist_on_vehicle	Realusis skaičius	<i>1.22,</i> <i>504.85</i>	Nuvažiutas transporto priemone žaidėjo atstumas
Dist_on_parachute	Realusis skaičius	<i>2.22,</i> <i>56.88</i>	Leidžiantis parašiotu žaidėjo atstumas
Dist_on_freefall	Realusis skaičius	<i>1422.15,</i> <i>1598.55</i>	Žaidimo pradžioje ore nusklestas žaidėjo atstumas
Groggy	Kategorinis	<i>0,1</i>	Ar žaidėjas buvo prikeltas prieš pasibaigiant žaidimui
Heal_amount	Realusis skaičius	<i>51.22,</i> <i>101.22</i>	Žaidėjo atstatytų gyvybių skaičius
Assist	Sveikasis skaičius	<i>0,1,2</i>	Žaidėjo asistavimų skaičius
Damage	Realusis skaičius	<i>55.44,</i> <i>221.05</i>	Žaidėjo padarytos žalos kiekis
Playing_type	Kategorinis	<i>1,2,3</i>	Žaidimo tipas
Death_time	Realusis skaičius	<i>1981.41,</i> <i>521.55</i>	Žaidėjo žaidimo pabaigos laikas
Energy Drink	Sveikasis skaičius	<i>0,1,2</i>	Gyvybes papildantis gėrimas
UMP9	Sveikasis skaičius	<i>0,1,2</i>	Ginklas <i>UMP9</i>
.45 ACP Ammo	Sveikasis skaičius	<i>0,1,2</i>	.45 kalibro ACP šoviniai

Tommy Gun	Sveikasis skaičius	0,1,2	Ginklas
First Aid Kit	Sveikasis skaičius	0,1,2	Medicininės pagalbos įrankis
Backpack (Level 1)	Sveikasis skaičius	0,1,2	Kuprinė (1 lygis)
7.62mm Ammo	Sveikasis skaičius	0,1,2	7.62 kalibro šoviniai
Backpack (Level 2)	Sveikasis skaičius	0,1,2	Kuprinė (2 lygis)
Large Extended Mag	Sveikasis skaičius	0,1,2	Šovinių pakuotė
Police Vest (Level 2)	Sveikasis skaičius	0,1,2	Policijos liemenė (2 lygis)
5.56mm Ammo	Sveikasis skaičius	0,1,2	5.56 kalibro šoviniai
AKM	Sveikasis skaičius	0,1,2	Ginklas
Painkiller	Sveikasis skaičius	0,1,2	Nuskausminamieji
Red Dot Sight	Sveikasis skaičius	0,1,2	Taikiklis ginklui
Military Helmet (Level 2)	Sveikasis skaičius	0,1,2	Armijos šalmas (2 lygis)
M416	Sveikasis skaičius	0,1,2	Ginklas
4x ACOG Scope	Sveikasis skaičius	0,1,2	Šoviniai
Large Extended QuickDraw Mag	Sveikasis skaičius	0,1,2	Šovinių pakuotė
Medium Extended QuickDraw Mag	Sveikasis skaičius	0,1,2	Šovinių pakuotė
Bandage	Sveikasis skaičius	0,1,2	Bintas
12 Gauge Ammo	Sveikasis skaičius	0,1,2	12 šovinių pakuotė
Military Vest (Level 3)	Sveikasis skaičius	0,1,2	Armijos liemenė (3 lygmuo)
9mm Ammo	Sveikasis skaičius	0,1,2	9 kalibro šoviniai
Holographic Sight	Sveikasis skaičius	0,1,2	Holografinis taikiklis
M16A4	Sveikasis skaičius	0,1,2	Ginklas
Medium QuickDraw Mag	Sveikasis skaičius	0,1,2	Šovinių dėklas
Police Vest (Level 1)	Sveikasis skaičius	0,1,2	Policijos liemenė (1 lygmuo)
Beryl	Sveikasis skaičius	0,1,2	Ginklas
Pan	Sveikasis skaičius	0,1,2	Ginklas
Smoke grenade	Sveikasis skaičius	0,1,2	Dūminė granata
Thumb grip	Sveikasis skaičius	0,1,2	Ginklo porankis
6x scope	Sveikasis skaičius	0,1,2	Taikiklis
Large Compensator	Sveikasis skaičius	0,1,2	Ginklo kompensatorius
SKS	Sveikasis skaičius	0,1,2	Ginklas
Large Flash Hider	Sveikasis skaičius	0,1,2	Ginklo kompensatorius
Half grip	Sveikasis skaičius	0,1,2	Ginklo porankis
Angled Foregrip	Sveikasis skaičius	0,1,2	Laikiklis
Backpack (Level 3)	Sveikasis skaičius	0,1,2	Kuprinė (3 lygmuo)
Medium compensator	Sveikasis skaičius	0,1,2	Ginklo kompensatorius
Vertical Foregrip	Sveikasis skaičius	0,1,2	Vertikalus porankis
Molotov Cocktail	Sveikasis skaičius	0,1,2	Degaus skysčio granata
3x scope	Sveikasis skaičius	0,1,2	Taikiklis
Sniper Rifle Cheek Pad	Sveikasis skaičius	0,1,2	Ginklo porankis
Medium Extended Mag	Sveikasis skaičius	0,1,2	Šovinių dėklas

Med kit	Sveikasis skaičius	0,1,2	Medicininų priemonių rinkinys
Micro Uzi	Sveikasis skaičius	0,1,2	Ginklas
Frag grenade	Sveikasis skaičius	0,1,2	Granata
Large QuickDraw Mag	Sveikasis skaičius	0,1,2	Šovinių dėklas
2x Aimpoint Scope	Sveikasis skaičius	0,1,2	Taikiklis
Sniper Rifle Compensator	Sveikasis skaičius	0,1,2	Ginklo taikymosi kompensatorius
Large Supressor	Sveikasis skaičius	0,1,2	Atatrankos mažintojas
Sniper Rifle Flash Hider	Sveikasis skaičius	0,1,2	Atatrankos mažintojas
Flashbang	Sveikasis skaičius	0,1,2	Smūginė granata
Tactical Stock	Sveikasis skaičius	0,1,2	Įrankių dėklas
S686	Sveikasis skaičius	0,1,2	Ginklas
Light Grip	Sveikasis skaičius	0,1,2	Ginklo porankis
Sniper Rifle Bullet Loops	Sveikasis skaičius	0,1,2	Šovinių dėklas
M24	Sveikasis skaičius	0,1,2	Ginklas

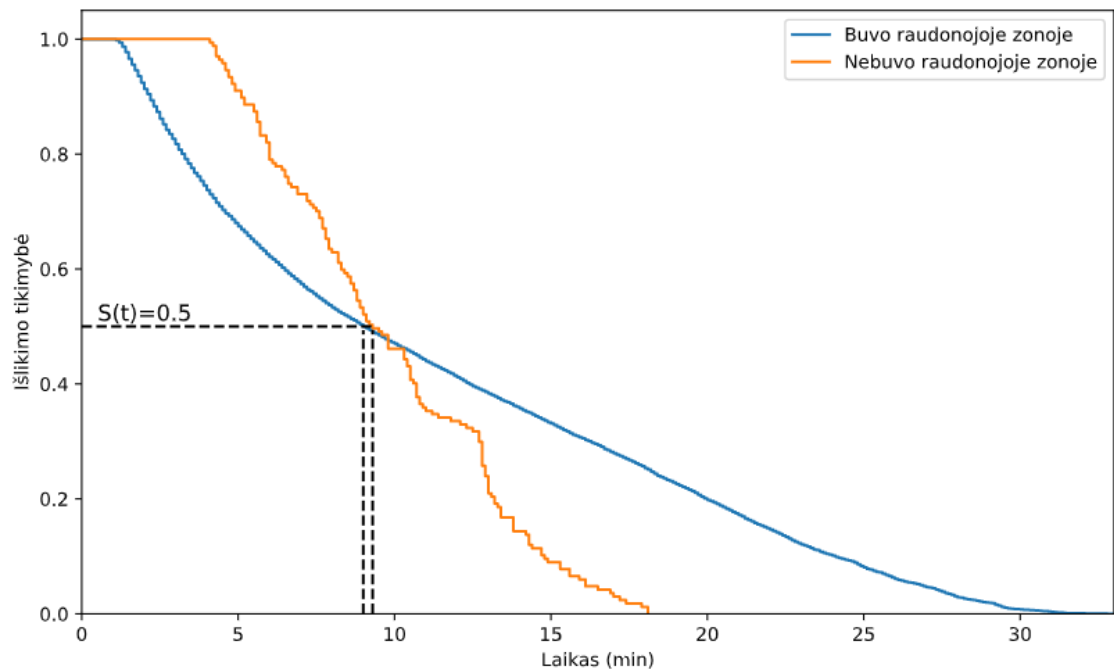
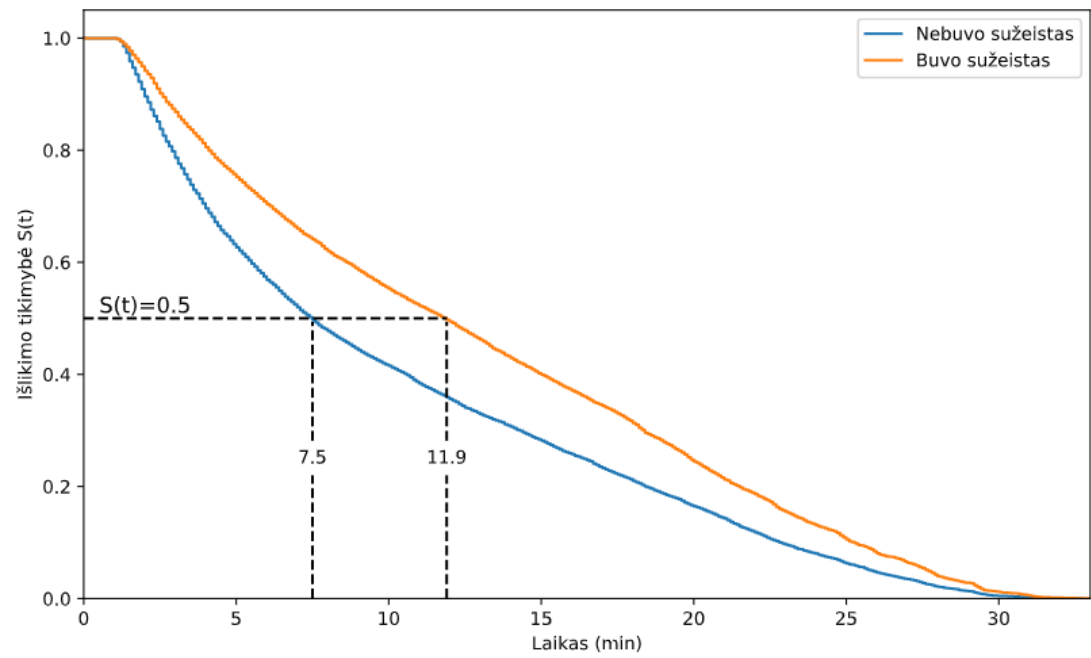
B priedas

Duomenų struktūros atributų koreliacijos matrica



C priedas

Kategorinių kintamųjų Kaplan-Meier išlikimo tikimybių kreivės



D priedas

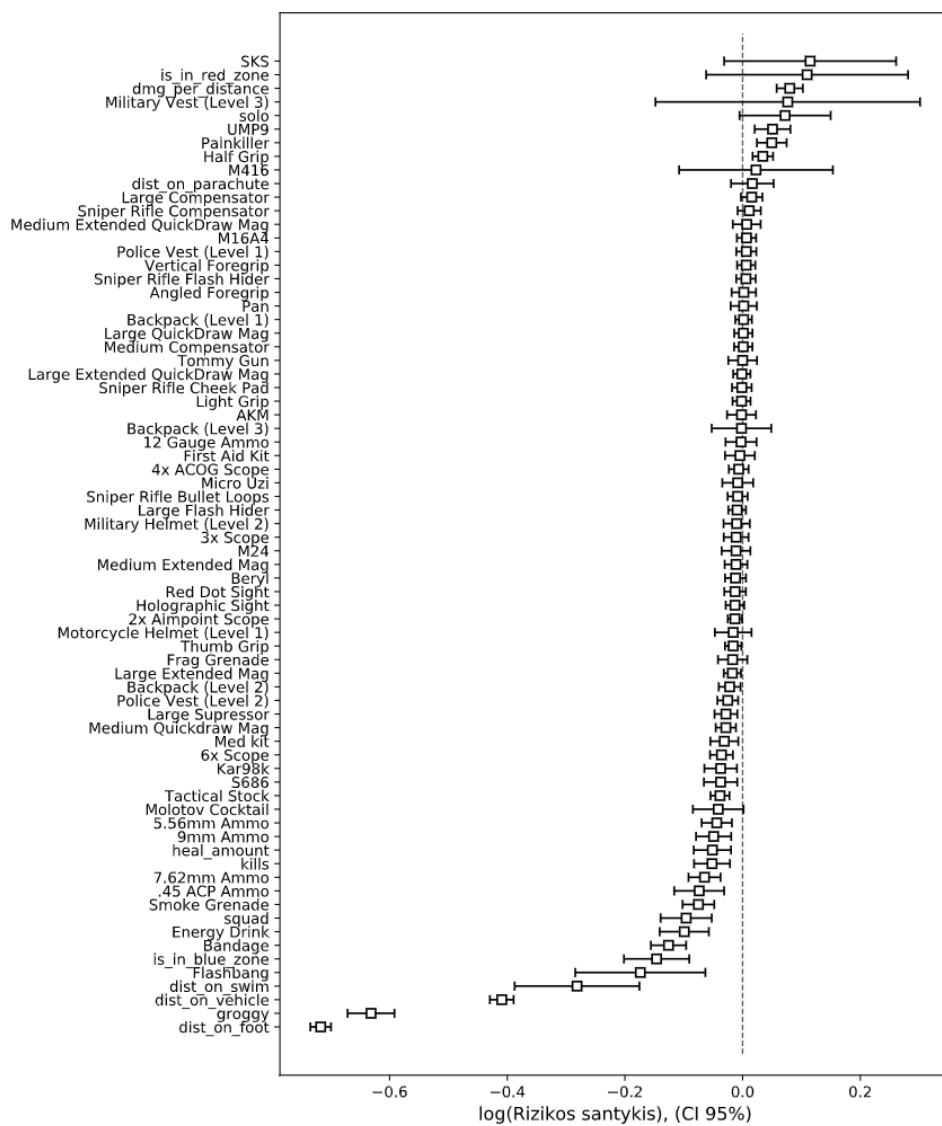
Atributų svarba prognozavime naudojant Cox modelį

dist_on_foot	0.904462
First Aid Kit	0.754237
Energy Drink	0.753653
heal_amount	0.752172
5.56mm Ammo	0.726943
Painkiller	0.719749
7.62mm Ammo	0.710484
Frag Grenade	0.709552
Bandage	0.706904
Smoke Grenade	0.700715
dist_on_vehicle	0.687935
Military Helmet (Level 2)	0.681147
Red Dot Sight	0.672667
Police Vest (Level 2)	0.653129
Large Extended Mag	0.642932
Tactical Stock	0.642196
Backpack (Level 2)	0.636327
Large Extended QuickDraw Mag	0.633467
Backpack (Level 3)	0.629849
Sniper Rifle Cheek Pad	0.627685
Large Compensator	0.627681
4x ACOG Scope	0.627451
Vertical Foregrip	0.621108
kills	0.612813
M416	0.612762
Large Flash Hider	0.606056
Pan	0.60112
6x Scope	0.598944
Holographic Sight	0.594641
Military Vest (Level 3)	0.587861
Motorcycle Helmet (Level 1)	0.580438
3x Scope	0.578722
dist_on_parachute	0.57756
Molotov Cocktail	0.575775
Police Vest (Level 1)	0.571377
Backpack (Level 1)	0.569088
Kar98k	0.567447

2x Aimpoint Scope	0.563317
Thumb Grip	0.562355
Large Supressor	0.562317
AKM	0.562215
Angled Foregrip	0.560782
Large QuickDraw Mag	0.560502
SKS	0.557058
Sniper Rifle Bullet Loops	0.55333
Med kit	0.553234
Light Grip	0.550647
Sniper Rifle Compensator	0.548843
Sniper Rifle Flash Hider	0.545774
groggy	0.544799
is_in_blue_zone	0.542454
M24	0.54153
M16A4	0.540031
dmg_per_distance	0.538723
9mm Ammo	0.535285
Half Grip	0.535247
Flashbang	0.530148
dist_on_swim	0.524152
Medium Extended QuickDraw Mag	0.523204
Medium Extended Mag	0.521206
Beryl	0.520347
Medium Compensator	0.519196
.45 ACP Ammo	0.51659
squad	0.516034
Medium Quickdraw Mag	0.51483
12 Gauge Ammo	0.509541
Micro Uzi	0.508088
S686	0.506261
solo	0.504237
UMP9	0.503752
Tommy Gun	0.500842
is_in_red_zone	0.500014

E priedas

Atributų įtaka žaidėjo rizikingumui naudojant Cox'o regresiją



F priedas

Atributų svarba prognozavime (pagal atitikimo rodiklį) naudojant pagreitinto įvykio laiko modelį

dist_on_foot	0.904554
First Aid Kit	0.753449
Energy Drink	0.753334
heal_amount	0.752684
5.56mm Ammo	0.725499
Painkiller	0.719645
7.62mm Ammo	0.710637
Frag Grenade	0.710001
Bandage	0.706486
Smoke Grenade	0.699985
dist_on_vehicle	0.687095
Military Helmet (Level 2)	0.680304
Red Dot Sight	0.671536
Police Vest (Level 2)	0.653261
Large Extended Mag	0.64286
Tactical Stock	0.641546
Backpack (Level 2)	0.635847
Large Extended QuickDraw Mag	0.633373
Backpack (Level 3)	0.630063
Large Compensator	0.628225
4x ACOG Scope	0.628149
Sniper Rifle Cheek Pad	0.627347
Vertical Foregrip	0.620864
kills	0.613309
M416	0.612025
Large Flash Hider	0.604293
Pan	0.600576
6x Scope	0.598813
Holographic Sight	0.595129
Military Vest (Level 3)	0.588429
Motorcycle Helmet (Level 1)	0.580805
3x Scope	0.578723
dist_on_parachute	0.57641
Molotov Cocktail	0.575197
Police Vest (Level 1)	0.571056
Backpack (Level 1)	0.568435
Kar98k	0.566532
2x Aimpoint Scope	0.563413

AKM	0.56336
Thumb Grip	0.562222
Large Supressor	0.561722
Angled Foregrip	0.561038
Large QuickDraw Mag	0.559911
SKS	0.556024
Sniper Rifle Bullet Loops	0.553744
Med kit	0.552102
Light Grip	0.550707
Sniper Rifle Compensator	0.548793
Sniper Rifle Flash Hider	0.546539
groggy	0.544439
is_in_blue_zone	0.542052
M24	0.541651
dmg_per_distance	0.539576
M16A4	0.538926
9mm Ammo	0.534843
Half Grip	0.534674
Flashbang	0.529766
dist_on_swim	0.524123
Medium Extended QuickDraw Mag	0.522816
Medium Extended Mag	0.520843
Beryl	0.520271
Medium Compensator	0.51867
.45 ACP Ammo	0.516703
squad	0.516592
Medium Quickdraw Mag	0.51406
Micro Uzi	0.508285
12 Gauge Ammo	0.507814
S686	0.505348
solo	0.504723
UMP9	0.504302
Tommy Gun	0.500752
is_in_red_zone	0.499988

G priedas

Atributų įtaka žaidėjo išlikimo laikui naudojant Pagreitinio įvykio laiko modelį

