Vilnius University

VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

Duomenų mokslo projektas - kursinis darbas

Respublikinio Priklausomybės Ligų Centro Vilniaus filialo pacientų pasveikimo veiksnių modeliavimas

Modeling of recovery factors for patients of the Vilnius branch of the Republican Center for Addictive Disorders

Rytis Baltaduonis
Justinas Pipiras
Ugnius Vilimas

VILNIUS 2024

DUOMENŲ MOKSLO STUDIJŲ PROGRAMA
III kursas

Darbo vadovas:	Povilas Treigys
Konsultantas:	Emilis Subata
Autorių kontaktai:	el. paštas Rytis.Baltaduonis@mif.stud.vu.lt
	el. paštas Justinas.Pipiras@mif.stud.vu.lt
	el. paštas Ugnius.Vilimas@mif.stud.vu.lt

Turinys

Respublikinio Priklausomybės Ligų Centro Vilniaus filialo pacientų pasveikimo veiksnių modeliavimas

Santrauka

Raktiniai žodžiai:

- Gradientas - vektorius, apibūdinantis skaliarinį lauką. Skaitine reikšme ir kryptimi apibūdina didžiausią skaliarinio dydžio u = u(x, y, z) kitimo greitį.
- Binarinio atsako modelis - toks modelis, kuris yra skirtas klasifikuoti 2 atsako reikšmes, pvz.: „Taip“ arba „Ne“.
- Modelio sąveika - tai yra dviejų modelio regresorių sandauga.
- Modelio tikslumas - tai tikimybė arba procentas, kuris nusako, kiek modelis tiksliai atspėja reikšmių.

1 Anotacija

Šis kursinis darbas nagrinėja Respublikinio Priklausomybės Ligų Centro Vilniaus filialo pacientų pasveikimo veiksnius. Tyrimo tikslas yra įvertinti gydymo poveikį pacientų būklei ir nustatyti sėkmingiausius gydymo metodus. Darbe naudojami įvairūs statistiniai ir mašininio mokymosi modeliai, įskaitant logistinės regresijos logit, probit ir XGBoost modelius. Darbe bus bandoma gerinti modelių tikslumą pasitelkiant visokiausiais būdais. Analizė atskleidžia svarbiausius veiksnius, darančius įtaką pacientų galutinei būklei, tokius kaip amžius, užimtumas ir sutrikimo tipas. Tyrimo rezultatai gali būti naudingi tobulinant gydymo metodus ir gerinant pacientų gerovę. Šiame darbe pateikiamos rekomendacijos sveikatos priežiūros sistemai, siekiant padidinti gydymo efektyvumą ir pacientų pasveikimo rodiklius.

2 Įvadas

2.1 Įžanga

Šiuo metu priklausomybės yra viena iš aktualiausių temų ne tik Lietuvoje, bet ir visame pasaulyje. Ši problema neapsiriboja tik tradicinėmis priklausomybėmis nuo psichotropinių medžiagų ar alkoholio; ji taip pat apima naujas priklausomybes, tokias kaip socialiniai tinklai ar netgi kasdienės veiklos, kurios, nors ir neatrodo tokios pavojingos kaip narkotikai ar alkoholis, gali turėti žalingų pasekmių žmogaus psichinei ir fizinės sveikatai bei socialiniam gyvenimui.

Respublikinis Priklausomybių Ligų Centras Vilniuje teikia pagalbą tiems, kuriems reikalinga pagalba susidūrus su priklausomybės problema. Suprantant, kad piktnaudžiavimas medžiagomis arba elgesio sutrikimais gali ne tik paveikti asmenį, bet ir turėti neigiamų padarinių jo artimiesiems ir visuomenei apskritai, yra būtina imtis veiksmų šioje srityje.

Šiame projektiniame darbe bus vertinamas gydymo efektyvumas, siekiant nustatyti geriausius gydymo metodus ir identifikuoti svarbiausius veiksnius, turinčius įtakos pacientų galutinei būklei. Mūsų tyrimo rezultatai gali turėti didelę reikšmę sveikatos priežiūros srityje, padedant gerinti pacientų gerovę, gydymo kokybę ir didinant sveikatos priežiūros sistemos efektyvumą. Pasauliui besikeičiant, yra esminis poreikis tobulinti mūsų sveikatos priežiūros sistemas, ir mes tikimės, kad šis projektinis darbas prisidės prie šio tikslo įgyvendinimo.

Kursiniame darbe planuojame atlikti pirminę duomenų analizę, patikrinti hipotezes, sudaryti regresinius ir mašininio mokymosi modelius ir bus bandoma sukurti keletas spėjimo modelių, kuriuos lyginsime ir tikrinsime kuris veiks efektyviausiai.Tai gali būti naudinga priklausomybių gydymui.

2.2 Darbo tikslas

Įvertinti atlikto gydymo poveikį Respublikinio Priklausomybės Ligų Centro Vilniaus filialo stacionaro skyriuose gydytų pacientų būklei ir gydymo rezultatams, nustatant sėkmingiausius gydymo metodus ir identifikuojant svarbiausius veiksnius, darančius įtaką pacientų galutinei būklei.

2.3 Uždaviniai

- Susipažinti su gautais duomenimis
- Atlikti literatūros analizę
- Sutvarkyti duomenis, atlikti pradinę analizę ir juos pavaizduoti
- Patikrinti statistines hipotezes
- Sudaryti statistinius modelius/taikyti mašininio mokymosi algoritmus
- Palyginti modelių gerumą ir gautus rezultatus
- Pateikti išvadas ir rekomendacijas

3 Literatūros analizė

Šioje dalyje mes atliekame aktualios Literatūros analizę, kurioje pateiksime keletą šaltinių, kuriuose ieškojome idėjų ir įžvalgų. Radome 3 straipsnius susijusius su mūsų tema ir turimais duomenimis. Visų straipsnių pagrindinė idėja buvo sukurti spėjimo modelius, kurie galėtų spėti priklausomo kintamojo binarinį atsaką bei juos palyginti.

3.1 COVID-19 pasveikimų modeliavimas Brazilijoje

Fernanda Sumika Hojo De Souza, Natália Satchiko Hojo-Souza, Edimilson Batista Dos Santos, Cristiano Maciel Da Silva ir Daniel Ludovico Guidoni straipsnyje:„Ligos rezultatų prognozavimas pacientams, sergantiems COVID-19, naudojant mašininį mokymąsi: retrospektyvus kohortinis tyrimas su Brazilijos duomenimis“ (angl. „Predicting the Disease Outcome in COVID-19 Positive Patients Through Machine Learning: A Retrospective Cohort Study With Brazilian Data") aiškinama apie Brazilijoje vykusį tyrimą, kurio metu buvo bandoma sukurti spėjimo modelį, COVID-19 užsikrėtimo baigčiai spėti.

Tyrimas buvo atliktas 2020-2021 metais, Brazilijoje. Straipsnyje minima, kad COVID-19 patvirtinti daugiau nei 672000 atvejai ir mažiausiai 36000 mirtys. Tiksli pacientų diagnozė, sergančių COVID-19 yra labai svarbi, tam kad būtų pasiūlytas tinkamas gydymas ir nebūtų perkrauta sveikatos priežiūros sistema. Pacientų charakteristikos, tokios kaip amžius, gretutinės ligos ir įvairūs klinikiniai simptomai, gali padėti klasifikuoti infekcijos sunkumo lygį, numatyti ligos baigtį ir hospitalizacijos poreikį. Straipsnio autorius pristato tyrimą, skirtą numatyti prastą teigiamų COVID-19 pacientų prognozę ir galimus rezultatus naudojant mašininį mokymąsi. Tyrimo duomenų rinkinį sudaro 8443 pacientų informacija apie baigtus atvejus dėl išgydymo ar mirties. eksperimentiniai rezultatai rodo, kad ligos baigtį galima numatyti, kai ROC kreivė AUC yra 0,92, jautrumas 0,88 ir specifiškumas 0,82, taikant geriausią prognozavimo modelį. Tai preliminarus retrospektyvus tyrimas, kurį galima patobulinti įtraukus daugiau duomenų. Autoriaus išvada: Mašininio mokymosi metodai, pagrįsti demografiniais ir klinikiniais duomenimis, kartu su pacientų gretutinėmis ligomis, gali padėti prognozuoti ir priimti gydytojo sprendimus, o tai leidžia greičiau reaguoti ir prisidėti prie sveikatos priežiūros sistemų neperkrovimo.

Peržvelgus visą straipsnį, gavome geresnį supratimą, ką galime padaryti su savo duomenimis ir kaip turėtų atrodyti tokio tipo tyrimas. Pastebėjome, kad mūsų gauti ir tyrimo naudoti duomenys buvo panašaus pobūdžio. Ir jie ir mes turėjome tik kategornius kintamuosius, taip pat sutapo dauguma charakteristikų apie pacientus, tokios kaip: žmogaus lytis, amžius, išsilavinimo lygis. Kadangi duomenys panašūs, supratome, kad bus galima naudoti panašius metodus ir mašininius modelius rezultatams spėti. Straipsnyje buvo panaudoti tokie metodai ir mašininiai modeliai: Logistinė regresija (angl. Logistic Regression), Tiesinė diskriminantinė analizė (angl. Linear Discriminant Analysis), Naive – Bajesas (angl. Naive Bayes), K-artimiausias kaimynas (angl. K-Nearest Neighbors), Sprendimų medis (angl. Decision Trees), XGBoost (angl. eXtreme Gradient Boosting), Atraminių vektorių klasifikatorius (angl. Support Vector Machine).

3.2 Alkoholio ir narkotikų priklausomybių modeliavimas Bangladešo populiacijai

Dr. Ariful Islam Arif, Saiful Islam Sany, Farah Sharmin, Dr. Sadekur Rahman, Dr. Tarek Habib straipsnyje: „Alkoholio ir narkotikų priklausomybių modeliavimas Bangladešo populiacijai" (angl. „Prediction of Addiction to Drugs and Alcohol Using Machine Learning: A Case Study on Bangladeshi Population") kalbama apie atliktą tyrimą Bangladeše, kurio metu buvo modeliuota rizika tapti priklausomu nuo alkoholio ir narkotikų pasitelkus mašininio mokymosi algoritmus.

Tyrimas buvo atliktas 2020-2021 metais, Bangladeše. Šiame darbe nagrinėjama mašininiu mokymusi pagrįstą būdą, kaip prognozuoti riziką tapti priklausomam nuo narkotikų, naudojant mašininio mokymosi algoritmus. Pirma, kai kuriuos reikšmingus priklausomybės veiksnius autoriai rado kalbėdami su gydytojais, nuo narkotikų priklausomais žmonėmis ir skaitydami atitinkamus straipsnius bei raštus. Tada buvo rinkti duomenys tiek iš priklausomų, tiek iš nepriklausomų žmonių. Logistinė regresija pranoksta visus kitus klasifikatorius pagal visus naudojamus rodiklius ir pasiekia 97,91% tikslumą.

Priešingai, CART rodo prastus rezultatus, kurių tikslumas artėja prie 59,37%, taikant pagrindinių komponentų analizę.

Išsianalizavus visą straipsnį dar labiau supratome tematikos svarbą ir reikšmingumą, kadangi šios problemos aktualumas yra ne tik Lietuvoje, bet ir visame pasaulyje. Iš visų rastų straipsnių, būtent šis yra artimiausias mūsų temai, kadangi čia tiriamos būtent mūsų įvardintos priklausomybės, tačiau duomenys buvo gauti per internetinį klausimyną, kurį galėjo užpildyti ir nepriklausomi ir priklausomi nuo psichotropinių medžiagų žmonės. Pagrinde tai buvo universiteto studentai ir klinikų pacientai. Tačiau mūsų duomenys buvo gauti iš oficialios įstaigos surinktų pacientų duomenų. Metodai, naudoti šiame straipsnyje: Logistinė regresija (angl. Logistic Regression), Naive – Bajesas (angl. Naive Bayes), K-artimiausias kaimynas (angl. K-Nearest Neighbors), Daugiasluoksnis perceptronas (angl. Multilayer perceptron), Atsitiktinumų miškas (angl. Random forest), XGBoost (angl. eXtreme Gradient Boosting), Atraminių vektorių klasifikatorius (angl. Support Vector Machine), CART (angl. Classification and regression tree), AdaBoost (angl. Adaptive boosting). Galime matyti, kad dauguma metodų sutampa su jau anksčiau analizuotų straipsnių.

3.3 Pasveikimų nuo psichologinių ligų modeliavimas

Katinka Franken, Peter ten Klooster, Ernst Bohlmeijer, Gerben Westerhof ir Jannis Kraiss straipsnyje: „Pasveikimų nuo psichologinių ligų modeliavimas" (angl. *„Predicting non-improvement of symptoms in daily mental healthcare practice using routinely collected patient-level data: a machine learning approach"*) buvo tiriami pacientai, turintys psicholiginių ligų, ir naudoti analizės metodai, norint išsiaiškinti ar pacientas turintis tam tikra sutrikimą galės pasveikti.

Tyrime nagrinėjamas mašininio mokymosi modelių veiksmingumas atisžvelgus į nepagerėjimą pacientams, gaunantiems įprastinę psichikos sveikatos priežiūrą dėl nerimo, nuotaikos, obsesinio-kompulsinio ar su traumomis susijusių sutrikimų. Buvo surinkti 755 pacientų duomenys, apie kuriuos buvo žinoma jų: diagnozė, sociodemografinę ir klinikinę informacija, taip pat pacientų pasveikimo buklę. Mašininio mokymosi algoritmai buvo išmokyti numatyti nereagavimą į simptominį distresą praėjus šešiems mėnesiams nuo tyrimo pradžios, atsižvelgus į ankstyvūs psichopatologijos ir gerovės pokyčių balais ir be jų, taip pat ir į paciento būklę. Rezultatai parodė, kad modeliai be ankstyvų pokyčių balų veikė prastai, tačiau įtraukus šiuos balus veikimas šiek tiek pagerėjo. Tačiau daug skaičiavimo reikalaujantys modeliai reikšmingai nepralenkė logistinės regresijos. Ankstyvųjų pokyčių balai nuolat buvo svarbūs skirtingų mašininio mokymosi algoritmų nepagerėjimo prognozėms. Nors mašininis mokymasis suteikia lankstumo, šiame tyrime jis nebuvo geresnis, palyginus su logistine regresija. Straipsnio gale minima, kad svarbu atsižvelgti į ankstyvųjų psichopatologijos ir gerovės pokyčių balus, kad būtų galima prognozuoti ilgalaikius simptomus.

Šis straipsnis taip pat buvo panašus į mūsų išsikeltus tikslus būtent savo kontekstu. Buvo tiriama žmonių grupė, turinti sutrikimų pasitelkus mašininio mokymosi algoritmus, taip pat logistinę regresiją. Šis straipsnis skyrėsi nuo mūsų, nes buvo atsižvelgta į laiko periodą (6 mėn), tačiau praėjus šitam periodui, buvo gauta, kad 70% reikšmių nepakito. Straipsnyje buvo naudoti keturi modeliai, į kuriuos galėsime atsižvelgti taip pat: Logistinė regresija (angl. Logistic Regression), Atsitiktinumų miškas (angl. Random forest) , XGBoost (angl. eXtreme Gradient Boosting), Atraminių vektorių klasifikatorius (angl. Support Vector Machine).

4 Duomenų analizė

4.1 Duomenys

Respublikinis Priklausomybių Ligų Centras pateikė 8196 skirtingų pacientų duomenis su 12 skirtingų kintamųjų. Apie pacientą yra žinoma tokia informacija: ataskaitiniai metai, lytis, amžius, gyvenamoji vieta, išsilavinimas, ar turi sveikatos draudimą, užimtumas, sutrikimas, vizito tipas, taikytas gydymas, į kurią vietą buvo išrašytas ir išrašyto paciento buklė.

Visi šie duomenys yra pateikti kategoriniu pavidalu. Ataiskaitiniai metai yra nuo 2016 iki 2022 metų. Amžius yra suskirstytas į intervalus kas penkerius metus, pvz: 15-19, 20-24 ir t.t iki 70+ metų,

vyriausias pacientas nėra įvardintas. Gyvenamoji vieta yra sužymėta kiekviena apskritimis, išskyrus Vilniaus miestą ir užsienį. Sutrikimai yra žymimi mediciniais terminais prasidedančiais F raide. Iš viso yra 26 skirtingi sutrikimai, tačiau mes juos sugrupuosime, pvz: F10,0, F10,1, F10,2 yra visi sutrikimai susiję su alkoholiu, todėl mes juos apjungsime, taip ir su visais kitais sutrikimais, kurie yra mūsų duomenyse.

Lytis	**Ataskaitiniai metai**	**Amžius**	**Gyvenamoji vieta**	**Išsilavinimas**
Vyras	2019	55-59	Vilniaus miestas	Aukštesnysis
Vyras	2019	40-44	Kauno apskritis	Vidurinis (12 kl.)
Vyras	2018	35-39	Vilniaus miestas	Aukštasis
Vyras	2017	45-49	Vilniaus miestas	Pagrindinis (9-10kl.)

1 lentelė: Pacientų duomenys - dalis 1

Sveikatos draudimas	**Užimtumas**	**Sutrikimas**	**Vizitas**
True	Neįgalus	F10.2	Pirminis
True	Registruotas darbo biržoje	F10.3	Pirminis
True	Registruotas darbo biržoje	F10.3	Pirminis
True	Registruotas darbo biržoje	F11.3	Pirminis

2 lentelė: Pacientų duomenys - dalis 2

Taikytas gydymas	**Išrašytas**	**Išrašyto paciento būklė**
Nemedikamentis gydymas	NA	Sėkmingai baigė gydymą
Medikamentinis abstinencijos gydymas	NA	Sėkmingai baigė gydymą
Medikamentinis abstinencijos gydymas	Perkeltas	Sėkmingai baigė gydymą
Medikamentinis abstinencijos gydymas	Perkeltas	Sėkmingai baigė gydymą

3 lentelė: Pacientų duomenys - dalis 3

4.2 Duomenų vizualizavimas ir tvarkymas

Pasitelkę R ir Weka programas bus atlikta pradinė duomenų analizė, ten kur yra verta duomenys bus vizualizuoti. Pradinėje duomenų analizėje bus žiūrima į kintamųjų pasiskirstymus, išskirtis. Ten kur reikia kategoriniai kintamieji bus sugrupuoti, o kai kur net bus skirti rangai. Tai aptarsime kalbant apie kiekvieną kintamąjį.

Išmetę pacientus, kurie buvo nepateikę tam tikrų duomenų ir savo eilutėje turėjo praleistų reikšmių, mes liekame su 7562 pacientais. Pirmiausia galime pasižiūrėti kaip pasiskirstęs mūsų priklausomas kintamasis.

Pasveikę pacientai	**Nepasveikę pacientai**
6782	780

4 lentelė: Pasveikusių ir nepasveikusių pacientų pasiskirstymas

4 lentelėje Galime aiškiai matyti, kad mūsų priklausomas kintamasis yra labai blogai subalansuotas ir sukūrus modelius galime turėti problemų dėl perteklinių vientetų spėjimų. Pasveikę pacientai buvo pažymėti skaičiumi 1, o nepasveikę - 0. pasveikusių ir nepasveikusių pacientų santykis yra 90/10.

Toliau pažiūrėsime nepriklausomus kintamuosius ir jų pasiskirstymą bei dažnį mūsų duomenų pakete. Pirmiausia patikrinkime kategorinius parametrus, kurie turi tik 2 galimus variantus, tai yra 0 ir 1. Tokių nepriklausomų kintamųjų buvo 3 - Lytis (vyras - 1, moteris - 0), Sveikatos draudimas (turi - 1, neturi - 0) ir vizitas (pakartotinis - 1, pirminis - 0).

Vyras	**Moteris**
5471	2091

5 lentelė: Pacientų lyčių pasiskirstymas

Turi sveikatos draudimą	**Neturi sveikatos draudimo**
6714	848

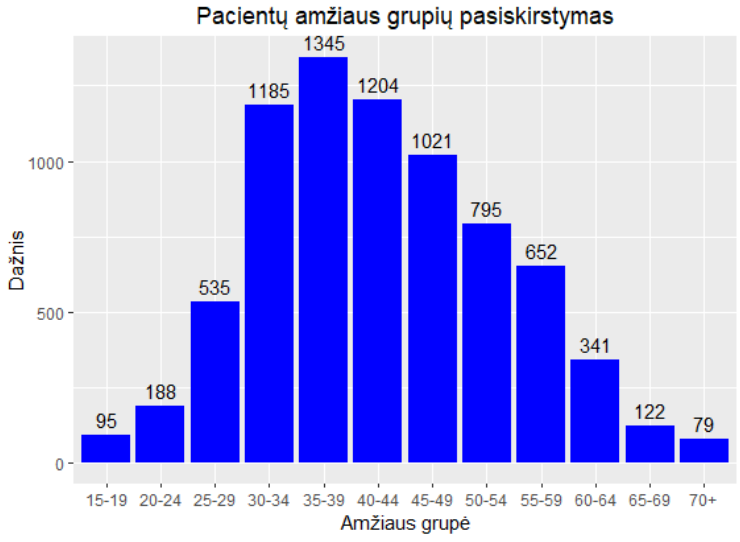
6 lentelė: Pacientų sveikatos draudimo pasiskirstymas

Pirminis vizitas	**Pakartotinis vizitas**
7182	380

7 lentelė: Pirminių ir pakartotinių pacientų vizitų dažnis

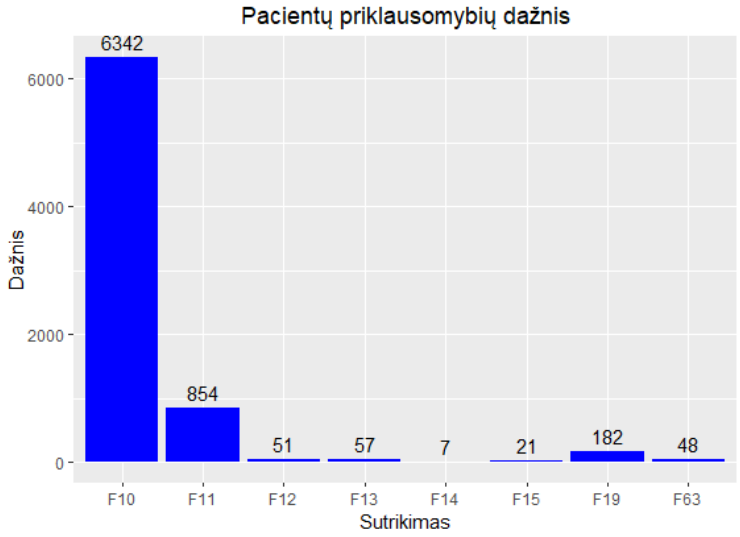
Iš visų lentelių matome, kad nesubalansuotumo problema išlieka, tačiau kai ji yra nepriklausomiems kintamiesiems, tai nebūtinai yra blogai.

Toliau galime apžvelgti kitus kategorinius kintamuosius, kuriuose jau yra 3 arba daugiau kategorijų. Tokiems kintamiesiems sukūrėme paprasčiausias histogramas, kad būtų galima lengviau įsivaizduoti gautus duomenis ir paprasčiau susidaryti įspūdį. Histogramas kūrėme tokiems nepriklausomiems kintamiesiems: paciento sutrikimas, amžiaus grupė, ataskaitiniai metai, gyvenamoji vieta, išsilavinimas, taikytas gydymas, užimtumas ir išrašymo vieta.



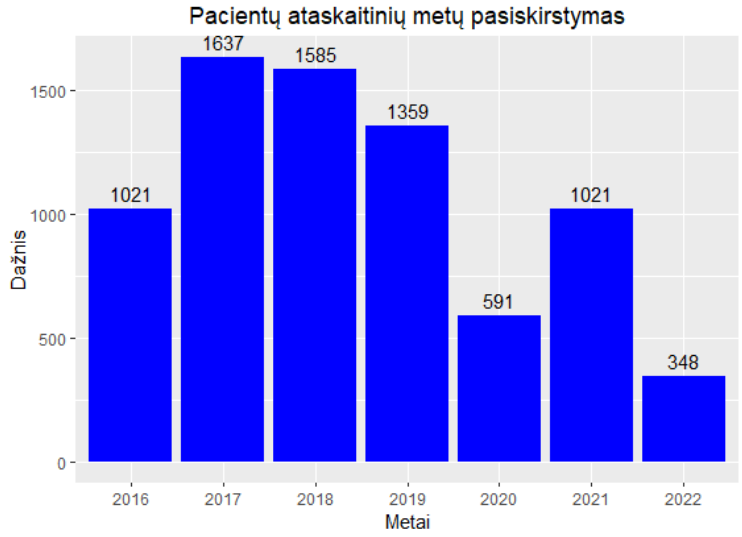
1 pav.: Visų pacientų pasiskirstymas pagal amžiaus grupes

Čia matome, kad mūsų pacientų amžiaus grupės yra pasiskirsčiusios panašiai lyg pagal normalųjį skirstinį, nes mūsų histograma panaši į varpo formą, kur mažiaužiai pacientų yra 15-19 metų grupėje bei 70 ir daugiau metų grupėje, o daugiausiai per vidurį, kur matome pacientus nuo 35 iki 39 metų. Atlikinėjant pradine analizę, taip pat nusprendėme, kad ranguosime pacientų amžiaus intervalus, kur patys jauniausi bus žymimi skaičiumi 1, o vyriausi - skaičiumi 11.



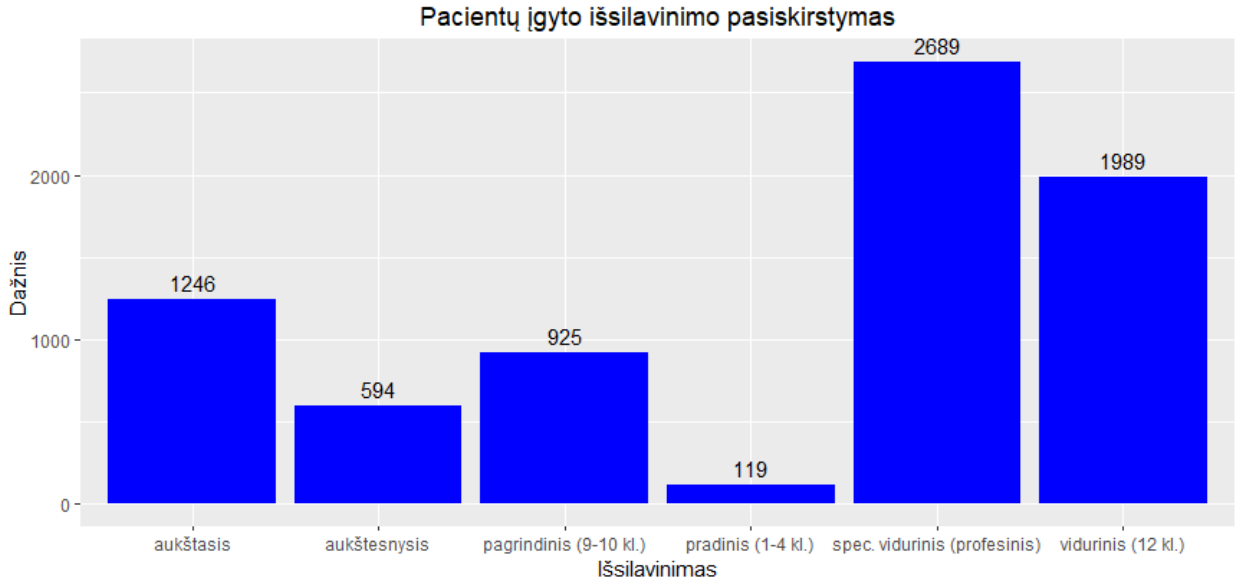
2 pav.: Visų pacientų pasiskirstymas pagal sutrikimus

Kadangi sutrikimai turi tą pačią identifikacinią kodo pradžią, mes juos apjungėme, kad būtų viskas konkrečiau. Sutrikimų reikšmės: F10 - Sutrikimas dėl alkoholio vartojimo, F11 - Sutrikimas dėl opioidų vartojimo, F12 - Sutrikimas dėl kanapių vartojimo, F13 - Sutrikimas dėl raminamųjų ir raminančiųjų medžiagų vartojimo, F14 - Sutrikimas dėl kokaino vartojimo, F15 - Sutrikimas dėl stimuliatorių ir kofeino vartojimo, F19 - Sutrikimas dėl kelių medžiagų vartojimo, F63 - Įpročių ir potraukių sutrikimai. Iš mūsų gautų duomenų matome, kad pagrindinis sutrikimas yra susijęs su alkoholių, yra net 5 kartus daugiau pacientų su šia problema, negu su betkuria kita. Antroje vietoje yra pacientai patekę, dėl opioidų vartojimo sutrikimo. Mažiausiai yra pacientų su kokaino vartojimo sutrikimu.



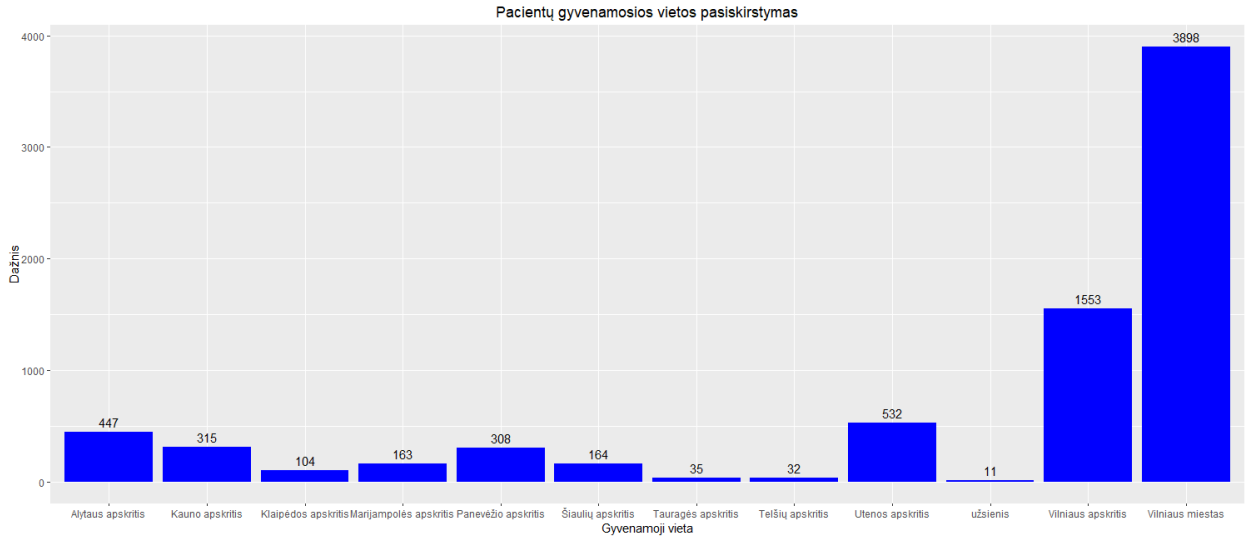
3 pav.: Visų pacientų pasiskirstymas pagal ataskaitinius metus

Peržiūrėjome metus, kada pacientai pateko į centrą. Matome, kad daugiausiai žmonių pateko 2017 metais, o nuo tada tas skaičius mažėjo. Tarp 2019 ir 2020 metų buvo didelis kritimas, manoma dėl to, kad prasidėjo karantinas ir žmonės negalėjo taip laisvai eiti į centrus gydytis. 2021 šis skaičius išaugo, bet 2022 metais jis vėl nukrito.



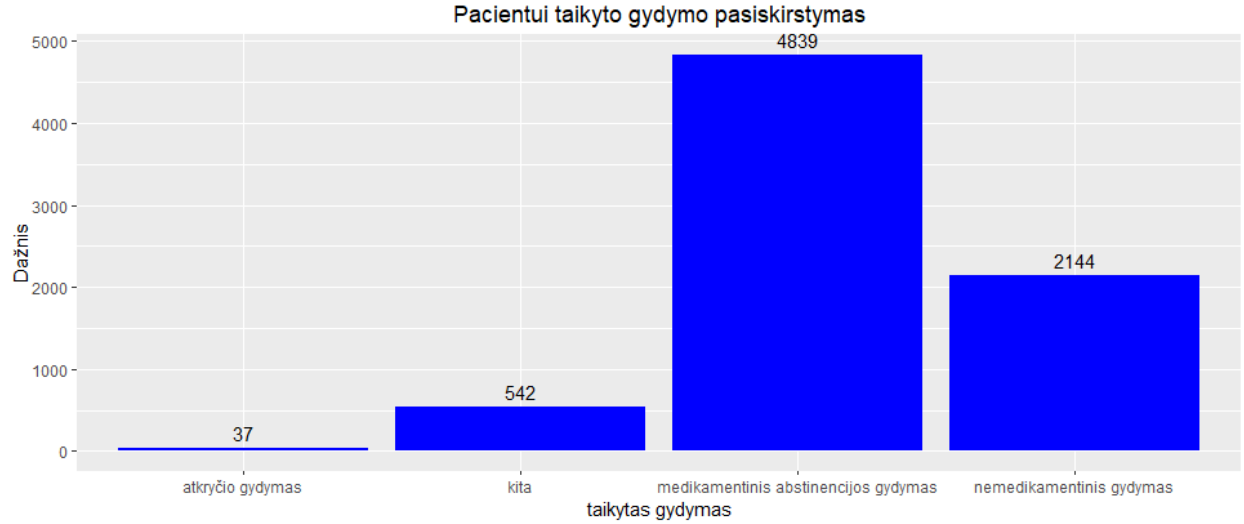
4 pav.: Visų pacientų pasiskirstymas pagal išsilavinimą

Toliau tikrinsime kokį išsilavinimą turintys žmonės kreipiasi į centrą pagalbos. Iš diagramos matome, kad daugiausiai pacientų turi profesinį išsilavinimą. Antroje vietoje yra pacientai turinys vidurinį išsilavinimą. Mažiausiai yra pacientų, kurie turi tik pradinį išsilavinimą. Atliekant pradinę analizę nusprendėme, kad, kaip ir su amžiaus intervalais, mes galime ranguoti išsilavinimą. Mūsų nuožiura suranguosime išsilavinimą nuo pačio lengviausio iki pačio sudėtingiausio, kur skaičiumi 1 bus žymima pacientai su pradiniu išsilavinimu, o skaičiumi 6 bus žymimi pacientai su aukštuoju išsilavinimu.



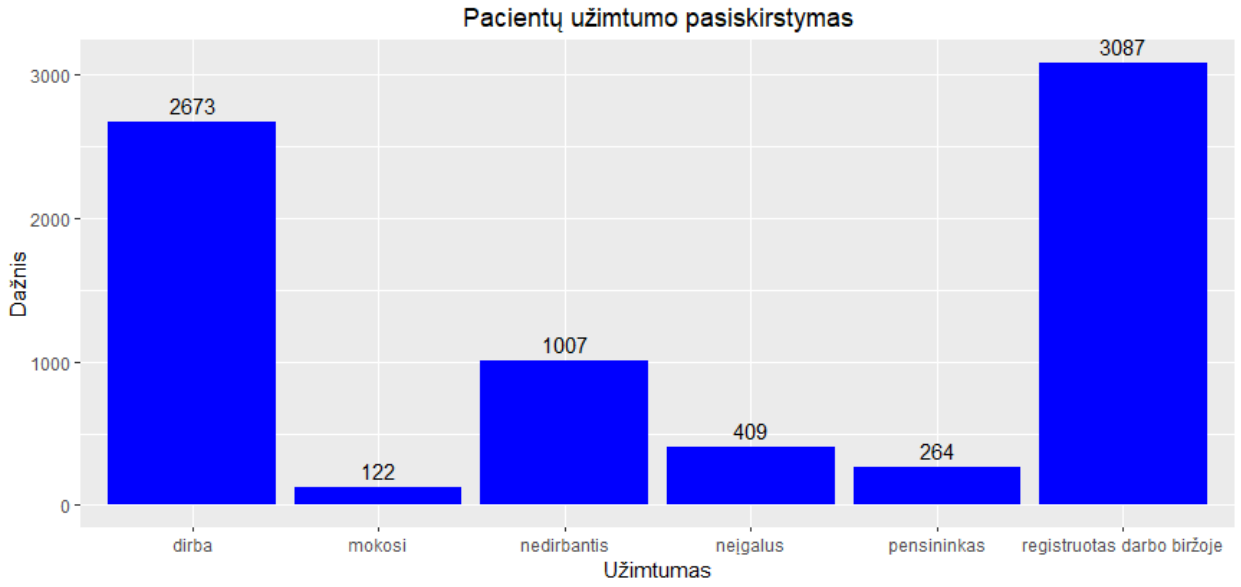
5 pav.: Visų pacientų pasiskirstymas pagal gyvenamąją vietą

Pagal gyvenamąją vietą daugiausiai pacientų yra iš Vilniaus miesto ir Vilniaus apskrities. Pagrinde šitos dvi vietos sudaro didžiausią dalį visuose duomenyse. Tai visiškai logiška, nes gauti duomenys buvo iš Vilniaus filialo priklausomybių centro. Mažiausiai žmonių kreipėsi iš užsienio. Visos kitos apskritys pasiskirsčiusios daugmaž panašiai tarp 32 ir 447 pacientų.



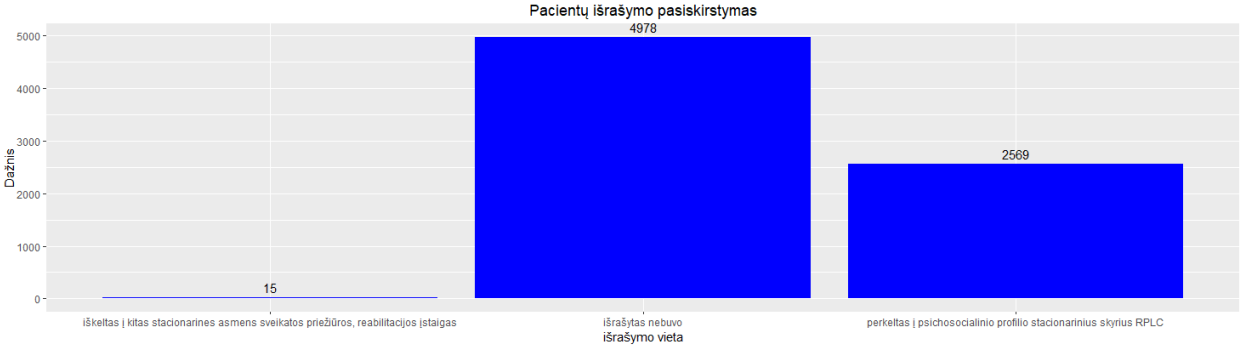
6 pav.: Visų pacientų pasiskirstymas pagal jiems taikytą gydymą

Medikamentinis abstinencijos gydymas yra populiauriasias gydymo metodas tarp visų duomenyse esančių pacientų. Atkryčio gydymas buvo rečiausias taikytas gydymo metodas. Pagrinde visi pacientai buvo gydyti medikamentiniu abstinencijos gydymo metodu.



7 pav.: Visų pacientų pasiskirstymas pagal užimtumą

Daugiausiai pacientų yra registruoti darbo biržoje (3087). Antroje vietoje yra dirbantys žmonės, o mažiausiai pacientų mokosi. Pagal užimtumą daugiausiai pacientai yra arba registruoti darbo biržoje arba dirba. Greičiausiai, kad šiuos duomenis reikės modifikuoti ir išsiskirstyti žmones tik į dvi grupes, dirbantys ir nedirbantys, kad turėtumėme dar vieną reikšme iš dviejų kintamųjų. jų reikšmės būtų: dirbantys - 2673, nedirbantys - 4889.



8 pav.: Visų pacientų pasiskirstymas pagal išrašymo vietą

Toliau buvo žiūrima ar pacientui reikėjo papildomo perrašymo į kitą įstaiga, t.y: ar jam reikėjo papildomos priežiūros. Daugumai pacientų nereikėjo, tačiau 2569 buvo perkelti į psischosocialinio profilio stacionarius skyrius RPLC, o 15 pacientų prireikė perkėlimo į reabilitacijos įstaigas.

5 Modelių kūrimas

5.1 „Logit“ ir „Probit“ binarinio atsako modeliai

Logit ir probit modeliai yra du pagrindiniai statistiniai metodai, naudojami analizuoti binarinius priklausomuosius kintamuosius. Tai reiškia, kad jie skirti modeliuoti situacijas, kuriose rezultatas gali turėti tik dvi galimas reikšmes, pavyzdžiui, „taip“ arba „ne“, „sėkmė“ arba „nesėkmė“, o mūsų atveju - „pasveiko“ arba „nepasveiko“.

Šioje dalyje mes sukursime „Logit“ ir „Probit“ modelius, patikrinsime abiejų modelių tinkamumo prielaidas, pažiūrėsime kaip gerai modelis spėlioja abi reikšmes ir palyginsime rezultatus

5.1.1 „Logit“ Modelis

Pirmiausia turime išsiaiškinti kas yra „Logit“ modelis ir kaip jis veikia bei kuo skiriasi nuo „Probit“ modelio.

- Priklausomas kintamasis Y įgyja vieną iš dviejų reikšmių:

$$Y = \begin{cases} 1 & \text{su tikimybe } p, \\ 0 & \text{su tikimybe } 1-p. \end{cases}$$

- Regresijos modelis gaunamas užrašant tikimybę p tam tikra regresorių funkcija. Pažymėkime $\mathbf{x}$ – regresorių vektorius, $\mathbf{x}'$ - vektoriaus eilutė ir $\boldsymbol{\beta}$ – parametrų vektorius (vektorius stulpelis: $K \times 1$).

$$p = \mathbb{P}(Y = 1 \mid \mathbf{x}) = F(\mathbf{x}'\boldsymbol{\beta}) = \mathbb{E}(Y \mid \mathbf{x}).$$

- Logit modelis:

$$F(\mathbf{x}'\boldsymbol{\beta}) = \frac{\exp\{\mathbf{x}'\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}'\boldsymbol{\beta}\}} = \Lambda(\mathbf{x}'\boldsymbol{\beta}) \quad \text{(logistinio skirstinio pasiskirstymo funkcija).}$$

$$\ln\left(\frac{p}{1-p}\right) = \mathbf{x}'\boldsymbol{\beta} \quad \text{(logit modelis)}$$

Pradžioje padalinsime duomenis į dvi skirtingas grupes - duomenys, kurie bus apmokami ir, kurie duomenys bus testuojami. Atitinkamai domenys bus padalinti atsitiktinai į 80% ir 20%. Toliau galime susidaryti „Logit“ modelį su visais kintamaisiais ir atlikus pažingsninę regresiją pasilikti tik statistiškai reikšmingus parametrus. Taigi susidarome lentelę, kurioje matysime visus modelyje esančias kovariantes ir jų p reikšmes, kurios indikuos, ar parametras statistiškai reikšmingas ar ne. Naudosime 0,05 reikšmingumo lygmenį.

Regresorius	**P reikšmė**
(Intercept)	0,08689
ataskaitiniai_metai	0,09627
lytis	0,91882
amzius_gr	0,03083 *
gyvenamoji_vietaKauno apskritis	0,76174
gyvenamoji_vietaKlaipėdos apskritis	0,21568
gyvenamoji_vietaMarijampolės apskritis	0,06470 .
gyvenamoji_vietaPanevėžio apskritis	0,38371
gyvenamoji_vietaŠiaulių apskritis	0,59791
gyvenamoji_vietaTauragės apskritis	0,81343
gyvenamoji_vietaTelšių apskritis	0,85304
gyvenamoji_vietaUtenos apskritis	0,34312
gyvenamoji_vietaUžsienis	0,93295
gyvenamoji_vietaVilniaus apskritis	0,52221
gyvenamoji_vietaVilniaus miestas	0,88322
issilavinimas	0,71546
sveikatos_draudimas	0,56307
uzimtumas	0,00184 **
sutrikimasF11	< 2e-16 ***
sutrikimasF12	5,39e-09 ***
sutrikimasF13	0,10068
sutrikimasF14	0,95649
sutrikimasF19	1,72e-14 ***
sutrikimasF63	0,61791
vizitas	0,35595
taikytas_gydymaskita	0,92088
taikytas_gydymasmedikamentinis abstinencijos gydymas	0,06467 .
taikytas_gydymasnemedikamentinis gydymas	0,09460 .

8 lentelė: „Logit“ modelis su visomis kovariantėmis ir p reikšmėmis

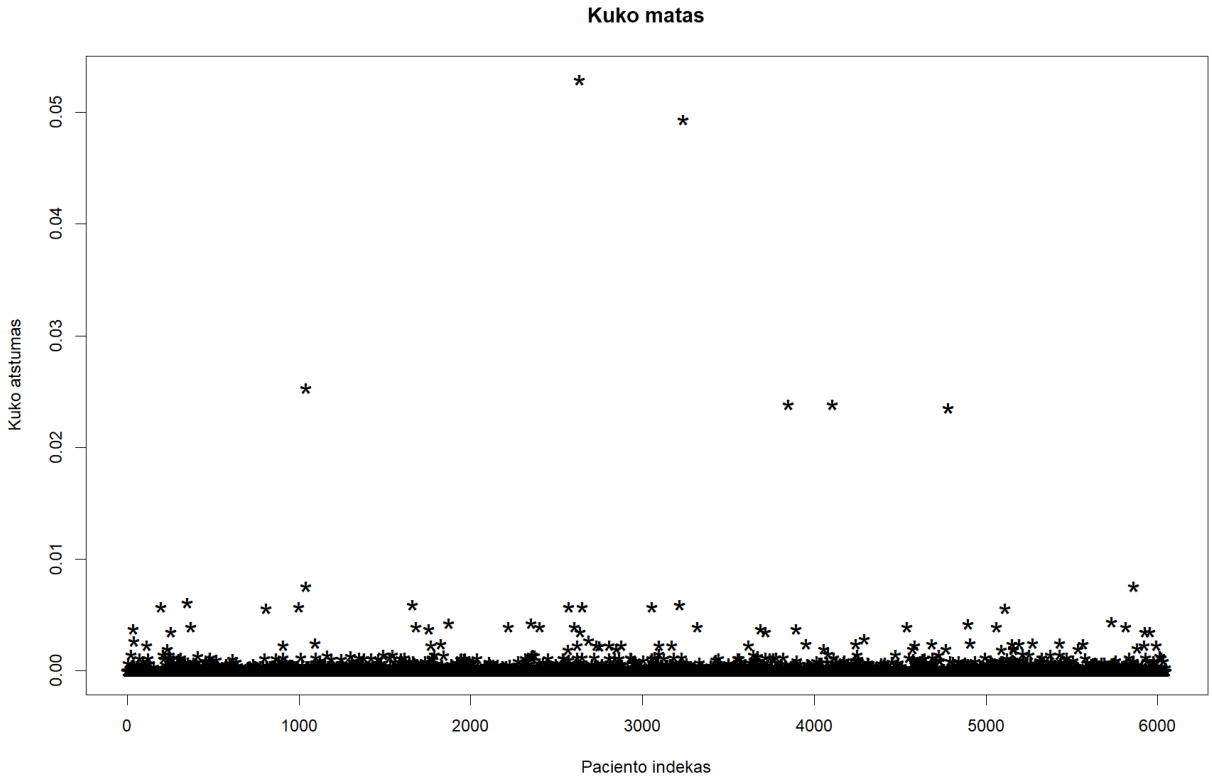
Simboliai prie p reikšmių indikuoja kokiame intervale yra p reikšmė, kuri leidžia lengviau pamatyti kurie regresoriai yra statistiškai reikšmingi. Žymėjimas: „0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1“. Iš lentelės matome, kad turime tik 3 reikšmingus regresorius, kurių p reikšmės yra mažesnės negu mūsų nustastytas reikšmingumo lygmuo - 0,05. Tai būtų „uzimtumas“, kuris indikuoja ar žmogus dirba ar ne. „amzius_gr“, kuris nurodo paciento amžiaus grupę ir „sutrikimas“, kuris nurodo, kokią priklausomybę turėjo pacientas.

Toliau atliksime pažingsninę regresiją, kurios metu bus pašalinami statistiškai nereikšmingi regresoriai ir paliekami tik reikšmingi, kurie geriau leis modeliui spėti nepriklausomojo kintamojo reikšmes. Gauname tokį modelį:

$$\begin{aligned}\text{israsyto\_paciento\_bukle} = {} & \underset{(0,12)}{2,09} + \underset{(0,02)}{0,045}\,\text{amzius\_gr} + \underset{(0,10)}{0,48}\,\text{uzimtumas} - \underset{(0,11)}{1,13}\,\text{sutrkimasF11} \\ & - \underset{(0,33)}{2,34}\,\text{sutrkimasF12} + \underset{(0,52)}{0,008}\,\text{sutrkimasF13} + \underset{(217,59)}{11,12}\,\text{sutrkimasF14} \\ & - \underset{(0,76)}{0,30}\,\text{sutrkimasF15} - \underset{(0,18)}{1,75}\,\text{sutrkimasF19} - \underset{(0,35)}{1,42}\,\text{sutrkimasF63}\end{aligned}$$

Toliau, susidarę modelį, reiktų žiūrėti ar neturime multikolinearumo problemos, tačiau žinome, kad mūsų visi regresoriai yra kategoriniai kintamieji, kas leidžia daryti išvadą, kad tokios problemos neturėsime.

Dabar turime patikrinti modelio išskirtis. Tai darysime Kuko mato grafiko pagalba.



9 pav.: Kuko mato grafikas su „Logit“ modeliu

Iš 9 paveikslėlio matome, kad išskirčių neturime, nes nei vienas taškas neviršija 1. Toliau pradėsime tikrinti modelio tinkamumą ir pradėsime nuo tikėtinumo santykių kriterijaus hipotezės.

Hipotezių formulavimas tikėtinumo santykio kriterijui

$$H_0 : \beta_1 = 0, \beta_2 = 0, \ldots, \beta_k = 0$$

(visi modelio regresoriai nereikalingi)

$$H_1 : \text{Bent vienas iš } \beta_i \neq 0, \text{ kur } i \in \{1, 2, \ldots, k\}$$

(bent vienas modelio regresorius reikalingas)

Tikėtinumo santykis yra gaunamas 283,23, o p reikšmė yra labai arti nulio, t.y. $2,2e^{-16}$, tai reiškia, kad atmetame nulinę hipotezę ir priimame alternatyvą. Tai mums leidžia daryti išvadas, kad modelis yra tinkamas ir bent vienas modelio regresorius yra reikalingas.

Hipotezių formulavimas Voldo kriterijui

H_0 : Visi regresoriai yra reišmingi

H_1 : Bent vienas regresorius yra nereikšmingas

Voldo kriterijaus statistika p gaunama taip pat labai arti nulio, t.y. $2.2e^{-16}$, tai reiškia, kad atmetame nulinę hipotezę ir priimame alternatyvą. Iš to galime daryti išvadas, kad ši prielaida yra tenkinama ir mūsų galutinio modelio visi regresoriai yra reikšmingi.

Norint pažiūreti modelio tinkamumą duomenims, reikia suskaičiuoti jo determinacijos koeficientą, kuris būna intervale [0,1]. Nėra gerai, kai determinacijos koeficientas < 0.2, tai parodo, kad modelis nebūtinai tinka duomenims. Tačiau, nors ir determinacijos koeficientas yra viena iš prielaidų, jis vaidina tik pagalbinį vaidmenį logistinėje regresijoje su „Logit“. Mūsų gautas determinacijos koeficientas $R^2 = 0{,}07$.

Dabar liko paskutinis žingsnis modelio tinkamumui patikrinti, tai yra klasifikavimo lentelė, kuri parodys kaip modelis spėja 0 ir 1 bei parodys kokia tikimybe pataiko teisingai.

	Atsakas	
	0	1
FALSE	5	10
TRUE	630	5405

9 lentelė: Klasifikavimo lentelė

	Atsakas	
	0	1
FALSE	0,008	0,002
TRUE	0,992	0,998

10 lentelė: Klasifikavimo tikimybės

Iš 9 ir 10 lentelės matome, kad mūsų sukurtas modelis labai gerai spėja pasveikimus, net 99,8% tikslumu, o nepasveikimus spėja labai blogai, tik 0,02% tikslumu. Tai yra visiška tragedija ir galime daryti išvadą, kad mūsų modelis spėja tik, kad pacientas pasveiks. Nors ir bendras modelio tikslumas gaunamas 89,4%, atsižvelgus į kaip modelis spėja nepasveikimus, modelio naudojimas bereikšmis.

5.1.2 „Probit“ Modelis

Pirmiausia turime išsiaiškinti kas yra „Probit“ modelis ir kaip jis veikia bei kuo skiriasi nuo „Logit“ modelio.

Probit modelyje modeliuojant binarinį atsaką, kintamasis Y priklauso nuo X, Z ir W. Kintamasis Y yra priklausomas kintamasis, o X, Z ir W yra nepriklausomi kintamieji arba regresoriai. Probit modelis yra sudaromas ne priklausomam kintajamam Y, o jo tikimybei P(Y=0):

$$\mathrm{P}(Y = 0) = \Phi(C + b_1 X + b_2 Z + b_3 W) \tag{5.1}$$

Reiktų paminėti, kad jeigu koeficientas prie kažkurio regresoriaus yra teigiamas ir tas regresorius didėja, tai tikimybė didėja ir tikimybė įgyti 0. Jeigu koeficientas prie kažkurio regresoriaus yra neigiamas ir tas regresorius didėjas, tai mažėja tikimybė įgyti 0 arba kitaip - didėja tikimybė įgauti 1.

Jeigu $b_1 > 0$, X didėjant, didėja ir tikimybė $\mathrm{P}(Y = 0)$.

Jeigu $b_1 < 0$, X didėjant, didėja ir tikimybė $\mathrm{P}(Y = 1)$.

Pagrindiniai du skirtumai, kuo Probit modelis skiriasi nuo Logit modelio, tai kad skirtingai negu logistinėje regresijoj nėra skaičiuojami galimybių santykiai ir skiriasi pasiskirstymo funkcijos. Logit modelyje tikymbė yra skaičiuojama pagal logistinio skirstinio pasiskirstymo funkcija, o Probit pagal standartinio normaliojo skirstinio pasiskirstymo funkcija.

Probit modelis

$$F(\mathbf{x}'\boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}} \varphi(z)dz = \Phi(\mathbf{x}'\boldsymbol{\beta})$$

(standartinio normaliojo skirstinio pasiskirstymo funkcija).

Toliau galime susidaryti „Probit“ modelį su visais kintamaisiais ir vėl atlikus pažingsninę regresiją pasilikti tik statistiškai reikšmingus parametrus. Taigi susidarome lentelę, kurioje matysime visus modelyje esančias kovariantes ir jų p reikšmes, kurios indikuos, ar parametras statistiškai reikšmingas ar ne. Kaip ir su „Logit“ modeliu, naudosime 0,05 reikšmingumo lygmenį.

Regresorius	**P reikšmė**
(Intercept)	0,10597
ataskaitiniai_metai	0,11784
lytis	0,92701
amzius_gr	0,02539 *
gyvenamoji_vietaKauno apskritis	0,80950
gyvenamoji_vietaKlaipėdos apskritis	0,21174
gyvenamoji_vietaMarijampolės apskritis	0,06228 .
gyvenamoji_vietaPanevėžio apskritis	0,35078
gyvenamoji_vietaŠiaulių apskritis	0,60294
gyvenamoji_vietaTauragės apskritis	0,76818
gyvenamoji_vietaTelšių apskritis	0,86512
gyvenamoji_vietaUtenos apskritis	0,38766
gyvenamoji_vietaUžsienis	0,95658
gyvenamoji_vietaVilniaus apskritis	0,57701
gyvenamoji_vietaVilniaus miestas	0,89526
issilavinimas	0,08840
sveikatos_draudimas	0,51395
uzimtumas	0,00156 **
sutrikimasF11	< 2e-16 ***
sutrikimasF12	2,23e-08 ***
sutrikimasF13	0,09953
sutrikimasF14	0,93995
sutrikimasF19	8,38e-14 ***
sutrikimasF63	0,09328
vizitas	0,45939
taikytas_gydymaskita	0,32952
taikytas_gydymasmedikamentinis abstinencijos gydymas	0,90874 .
taikytas_gydymasnemedikamentinis gydymas	0,05170 .

11 lentelė: „Probit“ modelis su visomis kovariantėmis ir p reikšmėmis

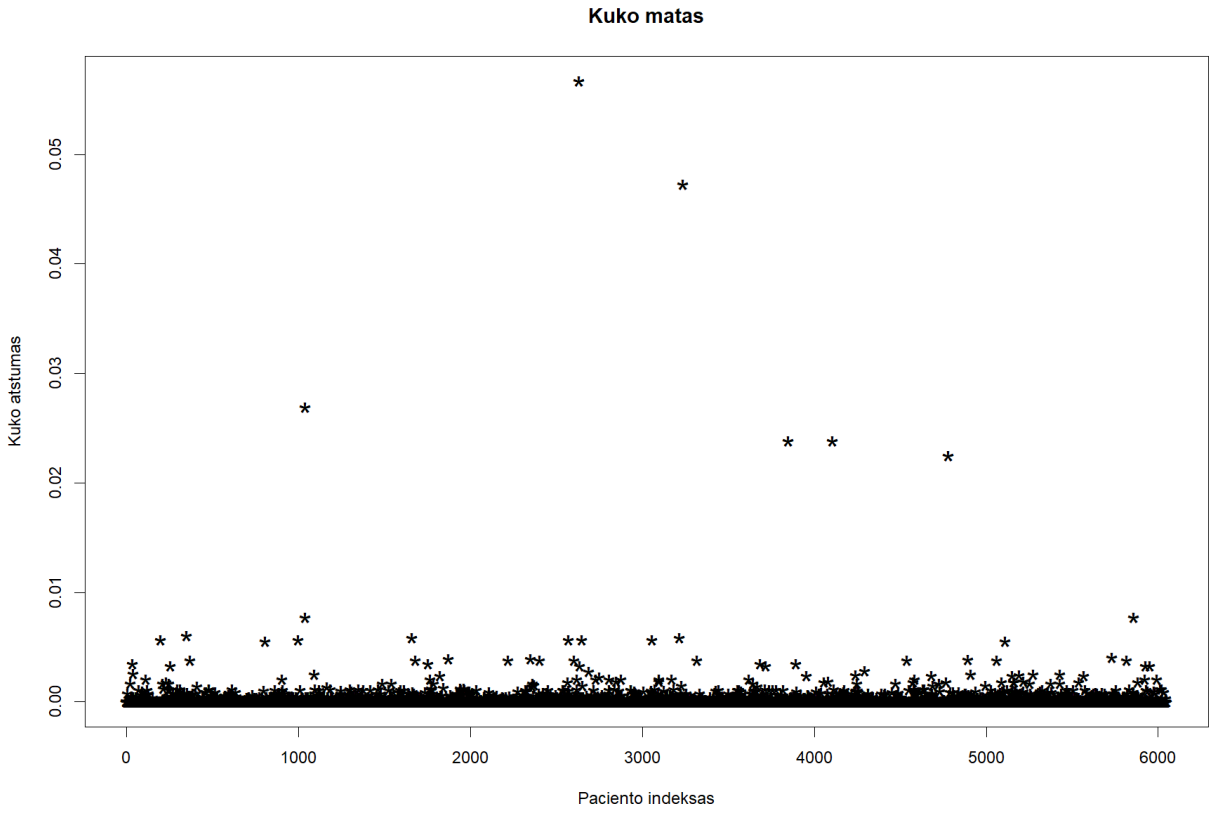
Simboliai prie p reikšmių indikuoja kokiame intervale yra p reikšmė, kuri leidžia lengviau pamatyti kurie regresoriai yra statistiškai reikšmingi. Žymėjimas: „0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1“. Iš lentelės matome, kad turime tik 3 reikšmingus regresorius, kurių p reikšmės yra mažesnės negu mūsų nustastytas reikšmingumo lygmuo - 0,05. Tai būtų „uzimtumas“, kuris indikuoja ar žmogus dirba ar ne. „amzius_gr“, kuris nurodo paciento amžiaus grupę ir „sutrikimas“, kuris nurodo, kokią priklausomybę turėjo pacientas.

Toliau atliksime pažingsninę regresiją, kurios metu bus pašalinami statistiškai nereikšmingi regresoriai ir paliekami tik reikšmingi, kurie geriau leis modeliui spėti nepriklausomojo kintamojo reikšmes. Gauname tokį modelį:

$$\begin{aligned}\text{israsyto\_paciento\_bukle} = {} & \underset{(0,06)}{1,22} + \underset{(0,01)}{0,02}\,\text{amzius\_gr} + \underset{(0,05)}{0,24}\,\text{uzimtumas} - \underset{(0,05)}{0,61}\,\text{sutrkimasF11} \\ & - \underset{(0,21)}{1,34}\,\text{sutrkimasF12} + \underset{(0,27)}{0,01}\,\text{sutrkimasF13} + \underset{(59,63)}{3,44}\,\text{sutrkimasF14} \\ & - \underset{(0,4)}{0,16}\,\text{sutrkimasF15} - \underset{(0,11)}{0,98}\,\text{sutrkimasF19} - \underset{(0,21)}{0,77}\,\text{sutrkimasF63}\end{aligned}$$

Toliau, susidarę modelį, reiktų žiūrėti ar neturime multikolinearumo problemos, tačiau žinome, kad mūsų visi regresoriai yra kategoriniai kintamieji, kas leidžia daryti išvadą, kad tokios problemos neturėsime.

Dabar turime patikrinti modelio išskirtis. Tai darysime Kuko mato grafiko pagalba.



10 pav.: Kuko mato grafikas su Probit modeliu

Iš 10 paveikslėlio matome, kad išskirčių neturime, nes nei vienas taškas neviršija 1. Toliau pradėsime tikrinti modelio tinkamumą ir pradėsime nuo tikėtinumo santykių kriterijaus hipotezės, taip pat kaip ir su „Logit“ modeliu.

Hipotezių formulavimas tikėtinumo santykio kriterijui

$$H_0 : \beta_1 = 0, \beta_2 = 0, \ldots, \beta_k = 0$$

(visi modelio regresoriai nereikalingi)

$$H_1 : \text{Bent vienas iš } \beta_i \neq 0, \text{ kur } i \in \{1, 2, \ldots, k\}$$

(bent vienas modelio regresorius reikalingas)

Tikėtinumo santykis yra gaunamas 283,94, o p reikšmė yra labai arti nulio, t.y. $2,2e^{-16}$, tai reiškia, kad ir vėl atmetame nulinę hipotezę ir priimame alternatyvą. Tai mums leidžia daryti išvadas, kad modelis yra tinkamas ir bent vienas modelio regresorius yra reikalingas.

Hipotezių formulavimas Voldo kriterijui

H_0 : Visi regresoriai yra reišmingi

H_1 : Bent vienas regresorius yra nereikšmingas

Voldo kriterijaus statistika p gaunama taip pat labai arti nulio, t.y. $2.2e^{-16}$, tai reiškia, kad atmetame nulinę hipotezę ir priimame alternatyvą. Iš to galime daryti išvadas, kad ši prielaida yra tenkinama ir mūsų galutinio modelio visi regresoriai yra reikšmingi.

Norint pažiūreti modelio tinkamumą duomenims, reikia suskaičiuoti jo determinacijos koeficientą, kuris būna intervale [0,1]. Nėra gerai, kai determinacijos koeficientas < 0.2, tai parodo, kad modelis nebūtinai tinka duomenims. Tačiau, nors ir determinacijos koeficientas yra viena iš prielaidų, jis vaidina tik pagalbinį vaidmenį logistinėje regresijoj su „Probit“ modeliu. Mūsų gautas determinacijos koeficientas $R^2 = 0{,}07$.

Dabar liko paskutinis žingsnis „Probit“ modelio tinkamumui patikrinti, tai yra klasifikavimo lentelė, kuri parodys kaip modelis spėja 0 ir 1 bei parodys kokia tikimybe pataiko teisingai.

	Atsakas	
	0	1
FALSE	5	10
TRUE	630	5405

12 lentelė: Klasifikavimo lentelė

	Atsakas	
	0	1
FALSE	0,008	0,002
TRUE	0,992	0,998

13 lentelė: Klasifikavimo tikimybės

Iš 12 ir 13 lentelės matome, kad mūsų sukurtas modelis ir vėl labai gerai spėja pasveikimus, net 99,8% tikslumu, o nepasveikimus spėja labai blogai, tik 0,02% tikslumu. Tai ir vėl yra mums nepriimtina ir galime daryti išvadą, kad mūsų modelis spėja tik, kad pacientas pasveiks. Nors ir bendras modelio tikslumas gaunamas ganėtinai geras - 89,4%, atsižvelgus į kaip modelis spėja nepasveikimus, modelio naudoti negalime.

5.1.3 Modelių palyginimas

Patikrinus abiejų modelių tinkamumo hipotezes, priėjome išvadą, kad abu modeliai veikia labai blogai ir negalime tęsti su jais darbo, tačiau vistiek galime pasižiūrėti į modelių galimybių santykių lenteles ir ROC kreives. Iš klasifikavimo lentelių, galime pastebėti, kad gavome identiškus rezultatus, todėl nusprendėme pasirinkti interpretuoti „Logit“ modelį su testavimo duomenų aibe. Taip pat AIC indeksai abiejų modelių sutapo, tai irgi neparodo, kuris tinkamesnis naudojimui.

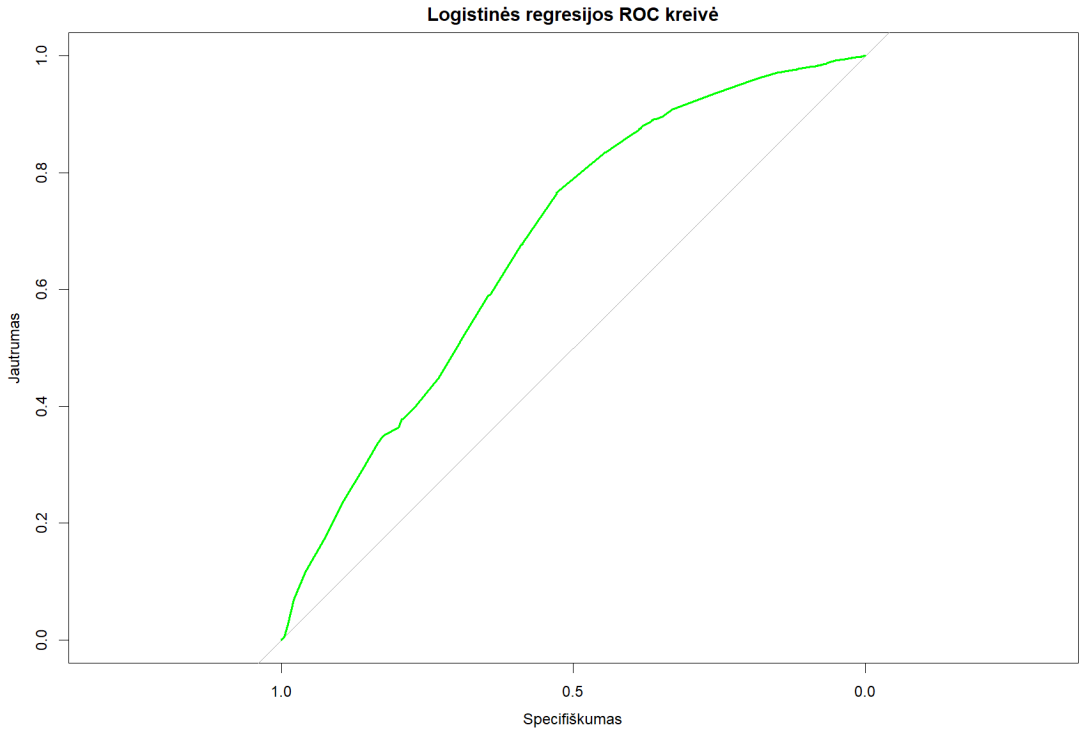
Pirmiausia pavaizduosime kiekvieno reikšmingo „Logit“ modelio regresorius, jų galimybių santykius ir pasikliovimo intervalus.

Regresorius	**Galimybių santykis**	**5%**	**95%**
(Intercept)	8,083	6,602	9,897
amzius_gr	1,047	1,012	1,082
uzimtumas	1,611	1,358	1,910
sutrikimasF11	0,324	0,269	0,391
sutrikimasF12	0,096	0,057	0,166
sutrikimasF13	1,008	0,423	2,404
sutrikimasF14	67589	$2{,}493\times10^{-151}$	$1{,}832\times10^{160}$
sutrikimasF15	0,743	0,214	2,581
sutrikimasF19	0,174	0,129	0,233
sutrikimasF63	0,242	0,137	0,430

14 lentelė: „Logit“ modelio galimybių santykiai ir pasikliovimo intervalai

Iš 14 lentelės galima matyti, kad sutrikimo žymimo F14 galimybių santykis yra labai didelis, dėl to, kad šitas sutrikimas yra rečiausias ir kol kas visi buvę pacientai nuo jo pasveiko. Galimybių santykį galime interpretuoti taip: jeigu užimtumo statusas pakinta iš dirbančio į nedirbantį, toks pacientas turi apie 61% šansą nepasveikti nuo tam tikros priklausomybės.

Toliau pasižiūrėsime į ROC kreivės grafika, sudarytą pagal „Logit“ modelį.



11 pav.: Logistinės regresijos ROC kreivė

Iš ROC kreivės galime matyti klasifikatoriaus specifiškumo ir jautrumo sąryšį.

Atlikome modelio testavimą ant testinių duomenų ir gavome tokius rezultatus

	spėta, kad nepasveiko	spėta, kad pasveiko
nepasveiko	1	144
pasveiko	3	1364

15 lentelė: „Logit“ modelio spėjimai su testiniais duomenimis

Modelis testiniams duomenims spėjo ne itin gerai. iš 145 nepasveikusių pacientų atspėjo tik 1, kas yra <1% ir yra labai blogai, tačiau pasveikusių žmonių atspėjo net 1364 iš 1367, kas yra 99,8%. Bendras modelio tikslumas gaunasi 90,3%, tačiau modelio naudoti negalime, nes matome, kad jis spėlioja tik, kad žmogus pasveiks, toks modelis yra visiškai nenaudingas, bent jau spėja pozityviai!

5.2 „Logit“ ir „Probit“ binarinio atsako modeliai su svoriais

Kaip ir žinome iš pradinės duomenų analizės, kad mūsų nepriklausomas kintamasis yra labai blogai subalansuotas, santykiu 90/10, kur dauguma yra pasveikusių pacientų, kuo tikrai džiaugiamės, tačiau tai nedaro mūsų darbo lengvesniu. Jau praeitame analizės skyriuje pastebėjome, kad ši problema sutrukdė gauti norimus rezultatus, todėl turime sugalvoti, kaip tai išspręsti. Vienas to būdas yra pritaikyti mūsų nepriklausomam kintamajam svorius. Pritaikę svorius, mes galime arba sumažinti didesnės nepriklausomų kintamųjų grupės įtaką, arba padidinti mažesnės nepriklausomų kintamųjų grupės įtaką, šiuo atveju mes renkamės pirmą variantą. Kad būtų lengviau įsivaizduoti kaip tai padarome, parašėme formulę.

$$\text{Svorių formulė} = \begin{cases} \frac{\sum(\text{X=0})}{\sum(\text{X=1})} & \text{, kai X} = 1 \\ 1 & \text{, kai X} = 0 \end{cases}$$

Čia X = 0 yra nepriklausomas kintamasis kai pacientas nepasveiko, o X = 1, kai pacientas pasveiko.

Naujus suskaičiuotus svorius įtrauksime į modelį ir stebėsime kaip keisis rezultatai, bet pirmiausia turime praeiti visus modelio tinkamumo etapus.

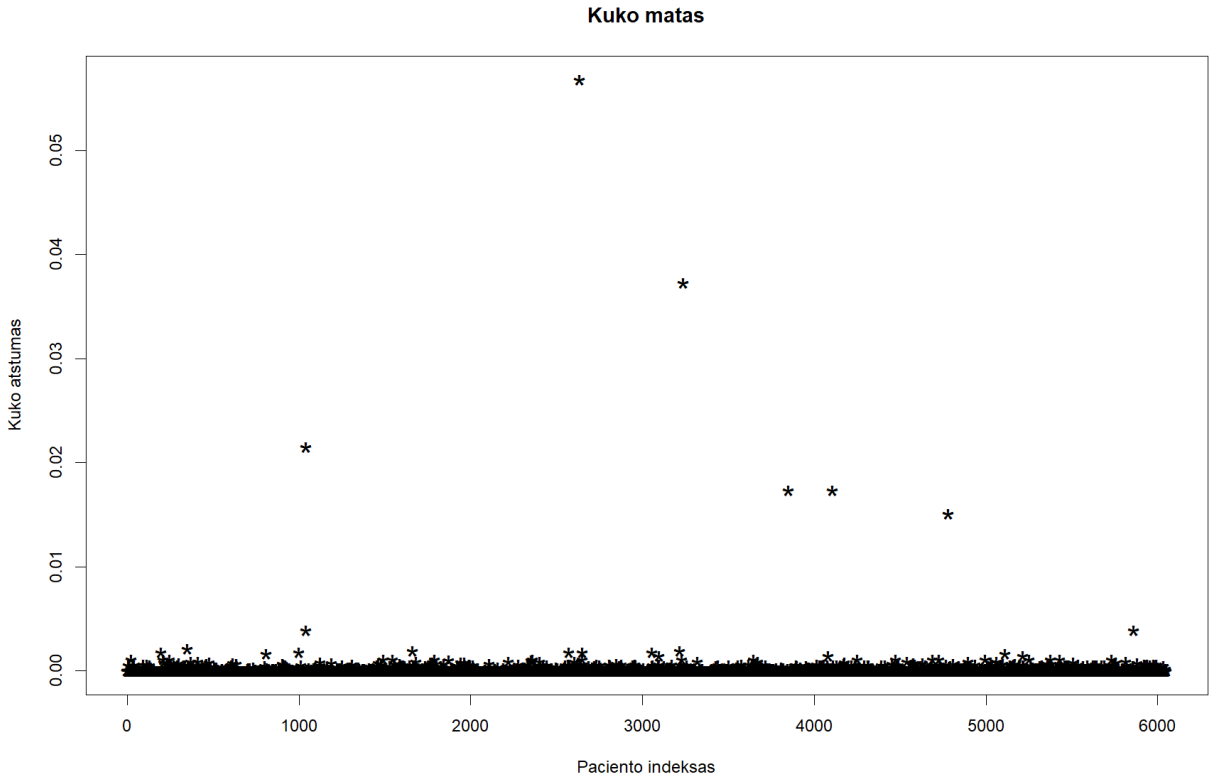
5.2.1 „Logit“ Modelis su svoriais

Atlikę pažingsinę regresiją su svoriais, likome prie tokių pačių reikšmingų kintamųjų, dabar žiūrėsime ar naujai sukurtas modelis tenkins prielaidas. Taip vaizduojamas mūsų naujai sukurtas modelis.

$$\begin{aligned} \text{israsyto\_paciento\_bukle} = & \underset{(0,16)}{-0,1} + \underset{(0,13)}{0,49}\,\text{amzius\_gr} + \underset{(0,027)}{0,05}\,\text{uzimtumas} - \underset{(0,17)}{1,1}\,\text{sutrkimasF11} \\ & - \underset{(0,7)}{2,42}\,\text{sutrkimasF12} + \underset{(0,68)}{0,005}\,\text{sutrkimasF13} + \underset{(323.25)}{11,93}\,\text{sutrkimasF14} \\ & - \underset{(1,04)}{0,33}\,\text{sutrkimasF15} - \underset{(0,33)}{1,79}\,\text{sutrkimasF19} - \underset{(0,6)}{1,32}\,\text{sutrkimasF63} \end{aligned}$$

Toliau, susidarę modelį, reiktų žiūrėti ar neturime multikolinearumo problemos, tačiau žinome, kad mūsų visi regresoriai yra kategoriniai kintamieji, kas leidžia daryti išvadą, kad tokios problemos neturėsime.

Dabar turime patikrinti modelio išskirtis. Tai ir vėl darysime Kuko mato grafiko pagalba.



12 pav.: Kuko mato grafikas su Logit modeliu ir svoriais

Tikėtinumo santykis yra gaunamas 2445,2, o p reikšmė yra labai arti nulio, t.y. $2,2e^{-16}$, tai reiškia, kad ir vėl atmetame nulinę hipotezę ir priimame alternatyvą. Tai mums leidžia daryti išvadas, kad modelis su svoriais yra tinkamas ir bent vienas modelio regresorius yra reikalingas.

Voldo kriterijaus statistika p gaunama taip pat labai arti nulio, t.y. $2.2e^{-16}$, tai reiškia, kad atmetame nulinę hipotezę ir priimame alternatyvą. Iš to galime daryti išvadas, kad ši prielaida yra tenkinama ir mūsų galutinio modelio su svoriais visi regresoriai yra reikšmingi.

Ir vėl norint pažiūreti modelio tinkamumą duomenims, reikia suskaičiuoti jo determinacijos koeficientą, kuris būna intervale [0,1]. Nėra gerai, kai determinacijos koeficientas < 0.2, tai parodo, kad modelis nebūtinai tinka duomenims. Tačiau, nors ir determinacijos koeficientas yra viena iš prielaidų, jis vaidina tik pagalbinį vaidmenį logistinėje regresijoj su „Logit“ modeliu su svoriais. Mūsų gautas determinacijos koeficientas $R^2 = 0{,}08$.

Dabar liko paskutinis žingsnis „Logit“ modelio su svoriais tinkamumui patikrinti, tai yra klasifikavimo lentelė, kuri parodys kaip modelis spėja 0 ir 1 bei parodys kokia tikimybe pataiko teisingai.

	Atsakas	
	0	1
FALSE	256	740
TRUE	379	4675

16 lentelė: Klasifikavimo lentelė

	Atsakas	
	0	1
FALSE	0,403	0,137
TRUE	0,597	0,863

17 lentelė: Klasifikavimo tikimybės

Iš 16 ir 17 lentelės matome, kad mūsų sukurtas modelis jau blogiau spėja pasveikimus, 86,3% tikslumu, o nepasveikimus spėja jau daug geriau, net 40,3% tikslumu. Tai jau daug geriau negu turėjome prieš tai bandytuose modeliuose be svorių, nors ir bendras modelio tikslumas mažesnis - 82%, vis dar atsižvelgus į kaip modelis spėja nepasveikimus, modelio naudoti negalime, nes yra žinoma taisyklė, kad modelis turi spėti abu įvykius bent 50% tikslumu.

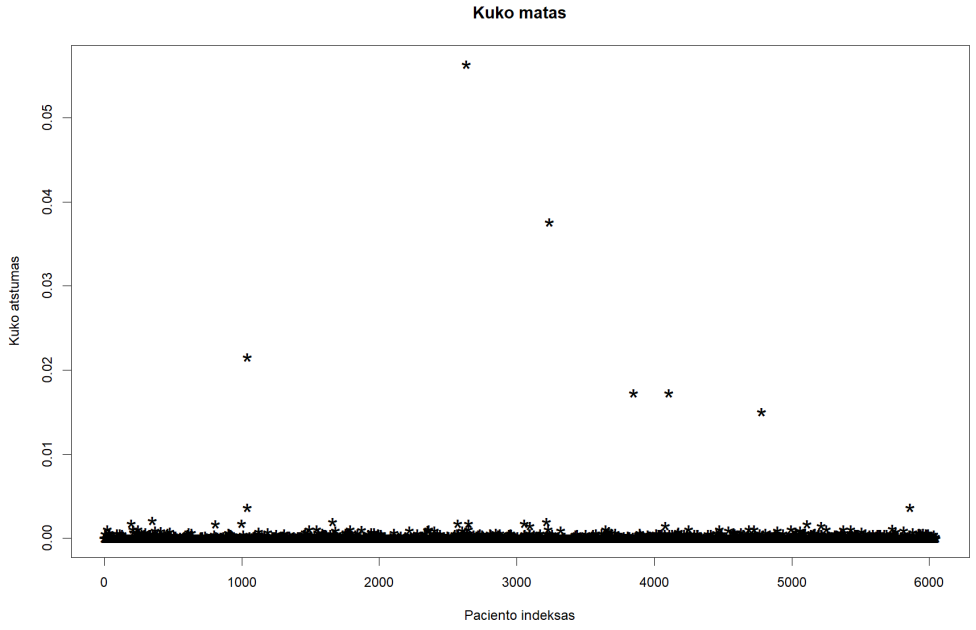
5.2.2 „Probit“ Modelis su svoriais

Atlikę pažingsinę regresiją su svoriais, likome prie tokių pačių reikšmingų kintamųjų, dabar ir vėl žiūrėsime ar naujai sukurtas modelis tenkins prielaidas. Taip vaizduojamas mūsų naujai sukurtas modelis.

$$\begin{aligned}\text{israsyto\_paciento\_bukle} = &\underset{(0,1)}{-0,05} + \underset{(0,03)}{0,02}\,\text{amzius\_gr} + \underset{(0,3)}{0,08}\,\text{uzimtumas} - \underset{(0,1)}{0,68}\,\text{sutrkimasF11} \\ &- \underset{(0,37)}{1,42}\,\text{sutrkimasF12} + \underset{(0,43)}{0,002}\,\text{sutrkimasF13} + \underset{(94,4}{4,4}\,\text{sutrkimasF14} \\ &- \underset{(0,65)}{0,21}\,\text{sutrkimasF15} - \underset{(0,19)}{1,08}\,\text{sutrkimasF19} - \underset{(0,35)}{0,82}\,\text{sutrkimasF63}\end{aligned}$$

Toliau, susidarę „Probit“ modelį su svoriais, reiktų žiūrėti ar neturime multikolinearumo problemos, tačiau žinome, kad mūsų visi regresoriai yra kategoriniai kintamieji, kas leidžia daryti išvadą, kad tokios problemos neturėsime.

Dabar turime patikrinti modelio išskirtis. Tai ir vėl darysime Kuko mato grafiko pagalba.



13 pav.: Kuko mato grafikas su Probit modeliu ir svoriais

Tikėtinumo santykis yra gaunamas 2444,9, o p reikšmė yra labai arti nulio, t.y. $2,2e^{-16}$, tai reiškia, kad ir vėl atmetame nulinę hipotezę ir priimame alternatyvą. Tai mums leidžia daryti išvadas, kad modelis su svoriais yra tinkamas ir bent vienas modelio regresorius yra reikalingas.

Voldo kriterijaus statistika p gaunama taip pat labai arti nulio, t.y. $2.2e^{-16}$, tai reiškia, kad atmetame nulinę hipotezę ir priimame alternatyvą. Iš to galime daryti išvadas, kad ši prielaida yra tenkinama ir mūsų galutinio modelio su svoriais visi regresoriai yra reikšmingi.

Ir vėl norint pažiūreti modelio tinkamumą duomenims, reikia suskaičiuoti jo determinacijos koeficientą, kuris būna intervale [0,1]. Nėra gerai, kai determinacijos koeficientas < 0.2, tai parodo, kad modelis nebūtinai tinka duomenims. Tačiau, nors ir determinacijos koeficientas yra viena iš prielaidų, jis vaidina tik pagalbinį vaidmenį logistinėje regresijoj su „Probit“ modeliu su svoriais. Mūsų gautas determinacijos koeficientas $R^2 = 0{,}08$.

Dabar liko paskutinis žingsnis „Probit“ modelio su svoriais tinkamumui patikrinti, tai yra klasifikavimo lentelė, kuri parodys kaip modelis spėja 0 ir 1 bei parodys kokia tikimybe pataiko teisingai.

	Atsakas	
	0	1
FALSE	256	740
TRUE	379	4675

18 lentelė: Klasifikavimo lentelė

	Atsakas	
	0	1
FALSE	0,403	0,137
TRUE	0,597	0,863

19 lentelė: Klasifikavimo tikimybės

Iš 16 ir 17 lentelės matome, kad mūsų sukurtas modelis jau blogiau spėja pasveikimus, 86,3% tikslumu, o nepasveikimus spėja jau daug geriau, net 40,3% tikslumu. Tai jau daug geriau negu turėjome prieš tai bandytuose modeliuose be svorių, nors ir bendras modelio tikslumas mažesnis - 82%, vis dar atsižvelgus į kaip modelis spėja nepasveikimus, modelio naudoti negalime, nes yra žinoma taisyklė, kad modelis turi spėti abu įvykius bent 50% tikslumu.

5.2.3 Modelių su svoriais palyginimas

Patikrinus abiejų modelių tinkamumo hipotezes, priėjome išvadą, kad abu modeliai su svoriais veikia geriau negu paprasti „Logit“ ir „Probit“, t.y. bendras modelių tikslumas šiek tiek prastesnis, tačiau beveik išsisprendė per daug vienetų spėjimo problema, tačiau vistiek negalime tęsti darbo su šiais modeliais su svoriais, nes šiek tiek neatitiko modelių tinkamumo reikalavimai. Vistiek galime pasižiūrėti į modelių galimybių santykių lenteles ir ROC kreives. Iš klasifikavimo lentelių, galime pastebėti, kad gavome identiškus rezultatus, todėl nusprendėme pasirinkti interpretuoti „Logit“ modelį su testavimo duomenų aibe. Taip pat AIC indeksai abiejų modelių sutapo, tai irgi neparodo, kuris tinkamesnis naudojimui.

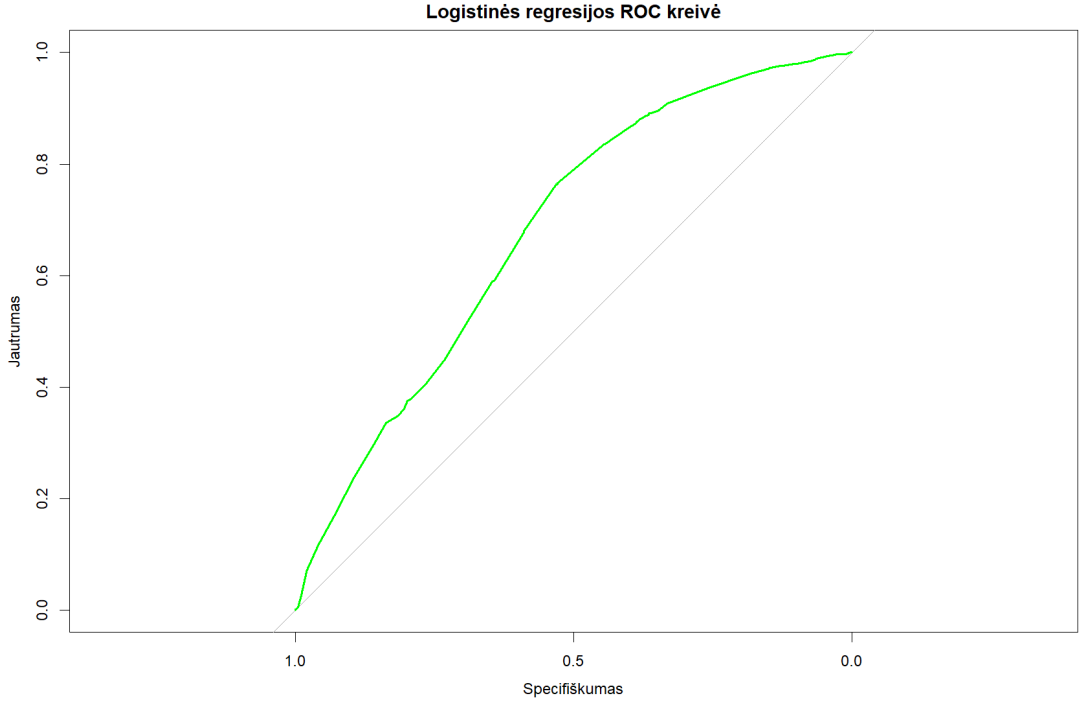
Pirmiausia pavaizduosime kiekvieno reikšmingo „Logit“ modelio su svoriais regresorius, jų galimybių santykius ir pasikliovimo intervalus.

Regresorius	Galimybių santykis	5%	95%
(Intercept)	0,903	0,690	1,181
uzimtumas	1,626	1,306	2,023
sutrikimasF11	0,332	0,249	0,442
sutrikimasF12	0,089	0,028	0,282
sutrikimasF13	1,004	0,326	3,095
sutrikimasF14	152000	$1{,}843\times10^{-236}$	$1{,}253\times10^{236}$
sutrikimasF15	0,716	0,129	3,976
sutrikimasF19	0,168	0,098	0,288
sutrikimasF63	0,267	0,099	0,72
amzius_gr	1,056	1,011	1,103

20 lentelė: „Logit“ modelio su svoriais galimybių santykiai ir pasikliovimo intervalai

Iš 20 lentelės galima matyti, kad sutrikimo žymimo F14 galimybių santykis yra labai didelis, dėl to, kad šitas sutrikimas yra rečiausias ir kol kas visi buvę pacientai nuo jo pasveiko. Galimybių santykį galime interpretuoti taip: jeigu užimtumo statusas pakinta iš dirbančio į nedirbantį, toks pacientas turi apie 63% šansą nepasveikti nuo tam tikros priklausomybės.

Toliau pasižiūrėsime į ROC kreivės grafika, sudarytą pagal „Logit“ modelį su svoriais.



14 pav.: Logistinės regresijos ROC kreivė su svoriais

Iš ROC kreivės galime matyti klasifikatoriaus specifiškumo ir jautrumo sąryšį.

Atlikome modelio testavimą ant testinių duomenų ir gavome tokius rezultatus

	spėta, kad nepasveiko	spėta, kad pasveiko
nepasveiko	58	87
pasveiko	176	1191

21 lentelė: „Logit“ modelio su svoriais spėjimai su testiniais duomenimis

Modelis testiniams duomenims spėjo geriau, bet ne taip kaip norime, tačiau matome, kad modelis nebespėja vien pasveikimų. Iš 145 nepasveikusių pacientų, modelis teisingai atspėjo 58, tai yra 40%, o iš 1367 pasveikusių, pataikė 1191, tai yra 87,1%. bendras modelio tikslumas yra 82,6%. Tai yra jau daug geriau negu prieš tai gautame modelyje, tačiau vis tiek iki galo neišsprendėme mažo nepasveikusių žmonių spėjimo tikslumo.

5.3 „Logit“ ir „Probit“ binarinio atsako modeliai su svoriais ir sąveikomis

Pastebėjus, kad „Logit“ ir „Probit“ modelius naudojant su svoriais, gaunamas šiek tiek geresnis rezultatas, gali dar ieškoti kaip galime patobulinti modelį, o tai mums gali padėti sąveikų įtraukimas. Sąveika yra tiesiog paprasta sandauga tarp kintamųjų, kurią galima labai paprastai įdėtį į modelį. Patikrinus daug sąveikų porų, radome tik vieną, kuri modelyje tampa statistiškai reikšminga, todėl dabar patikrinsime abiejų modelių su svoriais bei rasta sąveika veikimą.

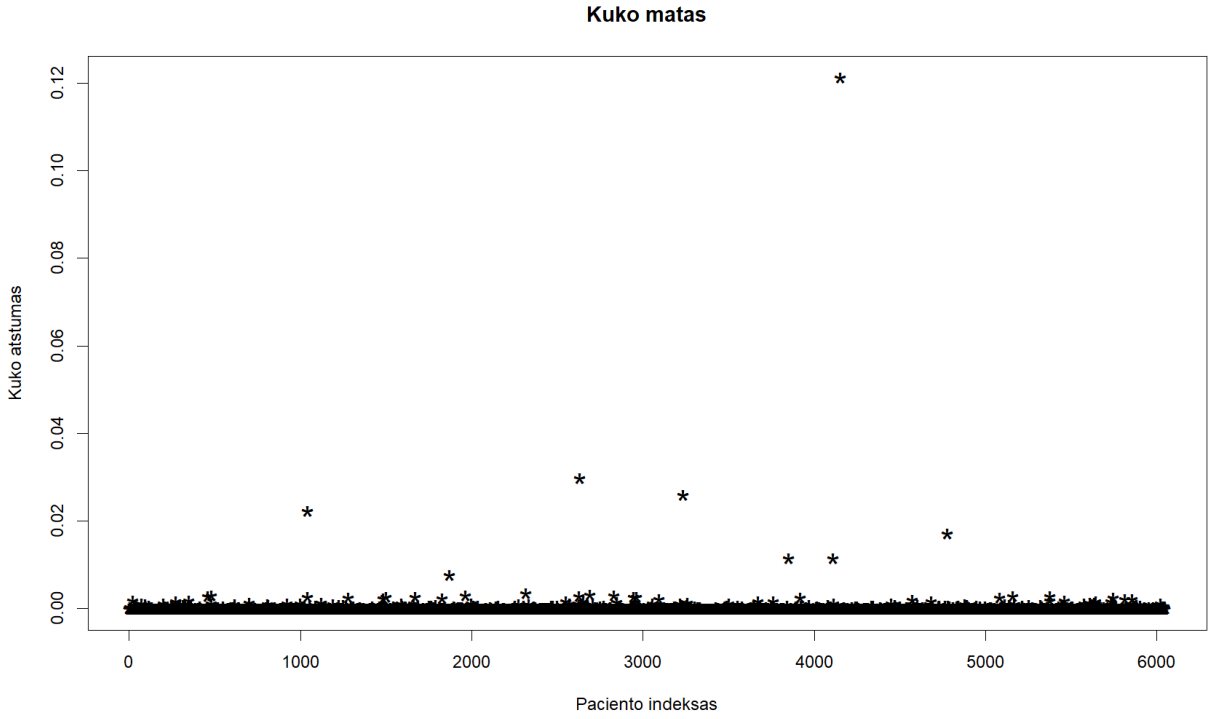
5.3.1 „Logit“ Modelis su svoriais ir sąveikomis

Atlikę pažingsinę regresiją su svoriais, likome prie tokių pačių reikšmingų kintamųjų ir prisideda mūsų įtraukta sąveika tarp taikyto gydymo tipo ir išsilavinimo. dabar žiūrėsime ar naujai sukurtas modelis tenkins tas pačias prielaidas. Taip vaizduojamas mūsų naujai sukurtas modelis.

$$\begin{aligned}\text{israsyto\_paciento\_bukle} = &\underset{(0,25)}{-0,32} - \underset{(0,612)}{0,57}\,\text{amzius\_gr} + \underset{(0,14)}{0,36}\,\text{uzimtumas} - \underset{(0,2)}{0,36}\,\text{sutrkimasF11}\\ &- \underset{(0,73)}{2,23}\,\text{sutrkimasF12} - \underset{(0,71)}{1,17}\,\text{sutrkimasF13} + \underset{(300,21)}{12,16}\,\text{sutrkimasF14}\\ &- \underset{(1,1)}{0,001}\,\text{sutrkimasF15} - \underset{(0,35)}{1,7}\,\text{sutrkimasF19} - \underset{(0,61)}{0,58}\,\text{sutrkimasF63}\\ &+ \underset{(0,32)}{2,4}\,\text{taikytas\_gydymasatkryčio gydymas:issilavinimas}\\ &+ \underset{(0,07)}{0,05}\,\text{taikytas\_gydymaskita:issilavinimas}\\ &+ \underset{(0,06)}{0,3}\,\text{taikytas\_gydymasmedikamentinis abstinencijos gydymas:issilavinimas}\\ &- \underset{(0,56)}{0,15}\,\text{taikytas\_gydymasnemedikamentinis gydymas:issilavinimas}\end{aligned}$$

Toliau, susidarę modelį, reiktų žiūrėti ar neturime multikolinearumo problemos, tačiau žinome, kad mūsų visi regresoriai yra kategoriniai kintamieji, kas leidžia daryti išvadą, kad tokios problemos neturėsime.

Dabar turime patikrinti modelio išskirtis. Tai ir vėl darysime Kuko mato grafiko pagalba.



15 pav.: Kuko mato grafikas su Logit modeliu, svoriais ir sąveika

Tikėtinumo santykis yra gaunamas 2610,4, o p reikšmė yra labai arti nulio, t.y. $2,2e^{-16}$, tai reiškia, kad ir vėl atmetame nulinę hipotezę ir priimame alternatyvą. Tai mums leidžia daryti išvadas, kad modelis su svoriais yra tinkamas ir bent vienas modelio regresorius yra reikalingas.

Voldo kriterijaus statistika p gaunama taip pat labai arti nulio, t.y. $2.2e^{-16}$, tai reiškia, kad atmetame nulinę hipotezę ir priimame alternatyvą. Iš to galime daryti išvadas, kad ši prielaida yra tenkinama ir mūsų galutinio modelio su svoriais visi regresoriai yra reikšmingi.

Ir vėl norint pažiūreti modelio su svoriais ir sąveika tinkamumą duomenims, reikia suskaičiuoti jo determinacijos koeficientą, kuris būna intervale [0,1]. Nėra gerai, kai determinacijos koeficientas $<$ 0.2, tai parodo, kad modelis nebūtinai tinka duomenims. Tačiau, nors ir determinacijos koeficientas yra viena iš prielaidų, jis vaidina tik pagalbinį vaidmenį logistinėje regresijoj su „Logit“ modeliu su svoriais. Mūsų gautas determinacijos koeficientas $R^2 = 0{,}17$.

Dabar liko paskutinis žingsnis „Logit“ modelio su svoriais ir sąveika tinkamumui patikrinti, tai yra klasifikavimo lentelė, kuri parodys kaip modelis spėja 0 ir 1 bei parodys kokia tikimybe pataiko teisingai.

	Atsakas	
	0	1
FALSE	509	1925
TRUE	126	3490

22 lentelė: Klasifikavimo lentelė

	Atsakas	
	0	1
FALSE	0,802	0,356
TRUE	0,198	0,644

23 lentelė: Klasifikavimo tikimybės

Iš 22 ir 23 lentelės matome, kad mūsų sukurtas modelis jau dar blogiau spėja pasveikimus, tik 64,4% tikslumu, o nepasveikimus spėja jau net geriau negu pasveikimus, net 80,1% tikslumu. Tai jau daug geriau negu turėjome prieš tai bandytuose modeliuose be sąveikų, nors ir bendras modelio tikslumas dar labiau sumažėjo - 66%, jau galime daryti išvadą, kad modelis spėja gana gerai ir galėtumėme jį naudoti, nes yra žinoma taisyklė, kad modelis turi spėti abu įvykius bent 50% tikslumu teisingai, kas ir yra daroma.

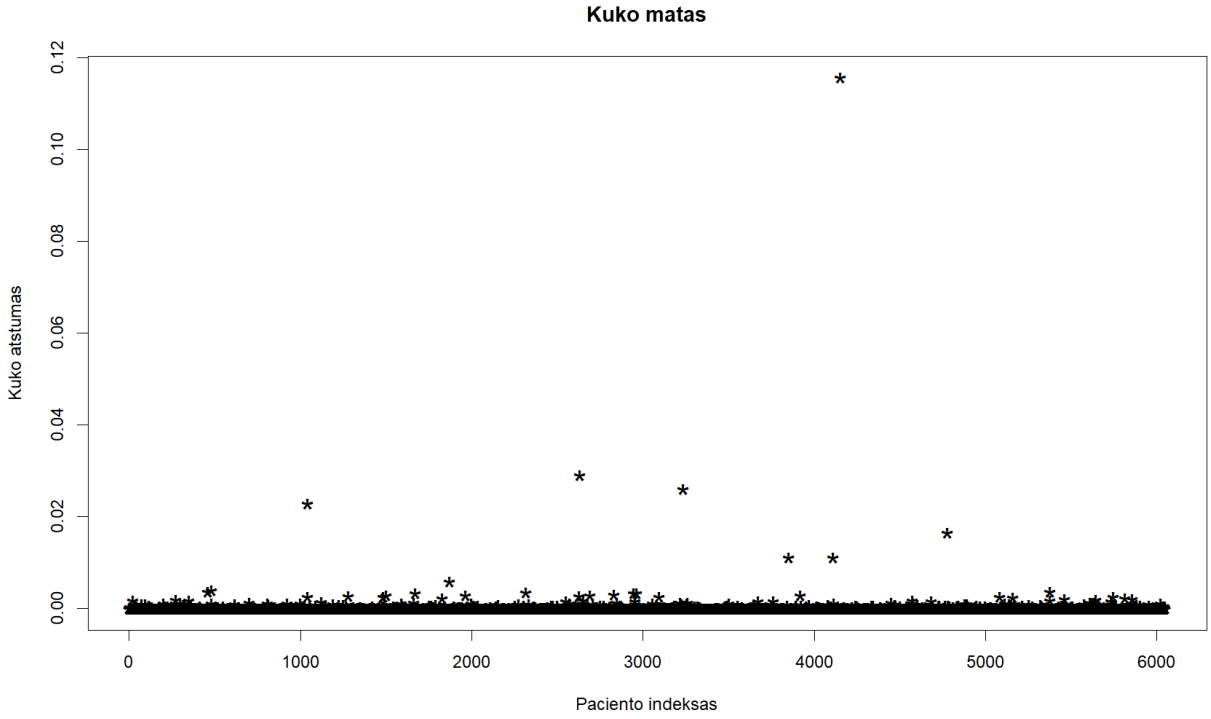
5.3.2 „Probit“ Modelis su svoriais ir sąveikomis

Atlikę pažingsinę regresiją su svoriais, likome prie tokių pačių reikšmingų kintamųjų ir prisideda mūsų įtraukta sąveika tarp taikyto gydymo tipo ir išsilavinimo. dabar žiūrėsime ar naujai sukurtas modelis tenkins tas pačias prielaidas. Taip vaizduojamas mūsų naujai sukurtas modelis.

$$
\begin{aligned}
\text{israsyto\_paciento\_bukle} = & \underset{(0,15)}{-0,17} - \underset{(0,035)}{0,17}\,\text{amzius\_gr} + \underset{(0,087)}{0,22}\,\text{uzimtumas} - \underset{(0,12)}{1,1}\,\text{sutrkimasF11} \\
& - \underset{(0,39)}{1,28}\,\text{sutrkimasF12} - \underset{(0,7)}{0,44}\,\text{sutrkimasF13} + \underset{(87,7)}{4,73}\,\text{sutrkimasF14} \\
& - \underset{(0,67)}{0,008}\,\text{sutrkimasF15} - \underset{(0,19)}{0,99}\,\text{sutrkimasF19} - \underset{(0,36)}{0,37}\,\text{sutrkimasF63} \\
& + \underset{(0,19)}{0,15}\,\text{taikytas\_gydymasatkryčio gydymas:issilavinimas} \\
& + \underset{(0,03)}{0,05}\,\text{taikytas\_gydymaskita:issilavinimas} \\
& + \underset{(0,035)}{0,2}\,\text{taikytas\_gydymasmedikamentinis abstinencijos gydymas:issilavinimas} \\
& - \underset{(0,03)}{0,09}\,\text{taikytas\_gydymasnemedikamentinis gydymas:issilavinimas}
\end{aligned}
$$

Toliau, susidarę modelį, reiktų žiūrėti ar neturime multikolinearumo problemos, tačiau žinome, kad mūsų visi regresoriai yra kategoriniai kintamieji, kas leidžia daryti išvadą, kad tokios problemos neturėsime.

Dabar turime patikrinti modelio išskirtis. Tai ir vėl darysime Kuko mato grafiko pagalba.



16 pav.: Kuko mato grafikas su Logit modeliu, svoriais ir sąveika

Tikėtinumo santykis yra gaunamas 2608,6, o p reikšmė yra labai arti nulio, t.y. $2,2e^{-16}$, tai reiškia, kad ir vėl atmetame nulinę hipotezę ir priimame alternatyvą. Tai mums leidžia daryti išvadas, kad modelis su svoriais yra tinkamas ir bent vienas modelio regresorius yra reikalingas.

Voldo kriterijaus statistika p gaunama taip pat labai arti nulio, t.y. $2.2e^{-16}$, tai reiškia, kad atmetame nulinę hipotezę ir priimame alternatyvą. Iš to galime daryti išvadas, kad ši prielaida yra tenkinama ir mūsų galutinio modelio su svoriais visi regresoriai yra reikšmingi.

Ir vėl norint pažiūreti modelio su svoriais ir sąveika tinkamumą duomenims, reikia suskaičiuoti jo determinacijos koeficientą, kuris būna intervale [0,1]. Nėra gerai, kai determinacijos koeficientas $<$ 0.2, tai parodo, kad modelis nebūtinai tinka duomenims. Tačiau, nors ir determinacijos koeficientas yra viena iš prielaidų, jis vaidina tik pagalbinį vaidmenį logistinėje regresijoj su „Probit“ modeliu su svoriais. Mūsų gautas determinacijos koeficientas $R^2 = 0{,}17$.

Dabar liko paskutinis žingsnis „Probit“ modelio su svoriais ir sąveika tinkamumui patikrinti, tai yra klasifikavimo lentelė, kuri parodys kaip modelis spėja 0 ir 1 bei parodys kokia tikimybe pataiko teisingai.

	Atsakas	
	0	1
FALSE	506	1921
TRUE	129	3494

24 lentelė: Klasifikavimo lentelė

	Atsakas	
	0	1
FALSE	0,797	0,355
TRUE	0,203	0,645

25 lentelė: Klasifikavimo tikimybės

Iš 24 ir 25 lentelės matome, kad mūsų sukurtas modelis jau dar blogiau spėja pasveikimus, tik 64,5% tikslumu, o nepasveikimus spėja jau net geriau negu pasveikimus, net 79,7% tikslumu. Tai jau daug geriau negu turėjome prieš tai bandytuose modeliuose be sąveikų, nors ir bendras modelio tikslumas dar labiau sumažėjo - 66%, jau galime daryti išvadą, kad modelis spėja gana gerai ir galėtumėme jį naudoti, nes yra žinoma taisyklė, kad modelis turi spėti abu įvykius bent 50% tikslumu teisingai, kas ir yra daroma.

5.3.3 Modelių su svoriais ir sąveikomis palyginimas

Patikrinus abiejų modelių tinkamumo hipotezes, priėjome išvadą, kad abu modeliai su svoriais ir sąveikomis veikia jau daug geriau negu paprasti „Logit“ ir „Probit“ bei modeliai tik su svoriais, t.y. bendras modelių tikslumas šiek tiek prastesnis, tačiau visiškai išsisprendė per daug vienetų spėjimo problema, tačiau vistiek negalime tęsti darbo su šiais modeliais su svoriais, nes šiek tiek neatitiko modelių tinkamumo reikalavimai. Vistiek galime pasižiūrėti į modelių galimybių santykių lenteles ir ROC kreives. Iš klasifikavimo lentelių, galime pastebėti, kad gavome identiškus rezultatus, todėl nusprendėme pasirinkti interpretuoti „Logit“ modelį su testavimo duomenų aibe. Taip pat AIC indeksai abiejų modelių sutapo, tai irgi neparodo, kuris tinkamesnis naudojimui.

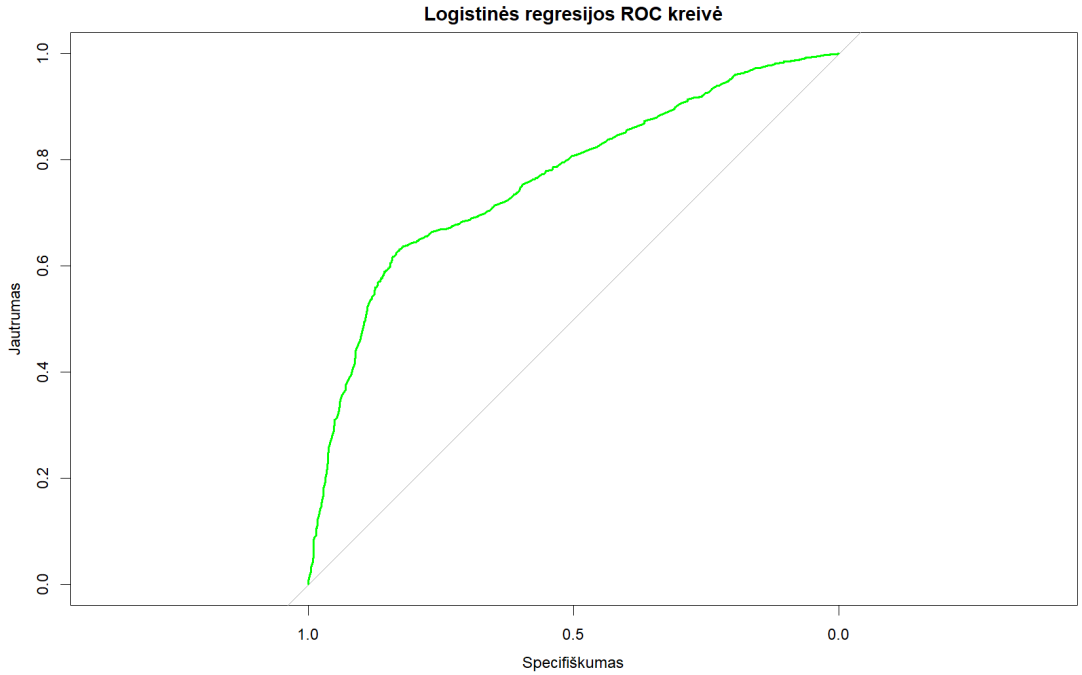
Pirmiausia pavaizduosime kiekvieno reikšmingo „Logit“ modelio su svoriais ir sąveika regresorius, jų galimybių santykius ir pasikliovimo intervalus.

Regresorius	**Galimybių santykis**	**5% CI**	**95% CI**
(Intercept)	0,726	0,483	1,092
uzimtumas	1,436	1,130	1,823
sutrikimasF11	0,161	0,017	0,227
sutrikimasF12	0,108	0,032	0,356
sutrikimasF13	0,309	0,096	0,995
sutrikimasF14	190741	$6{,}682\times10^{-210}$	$5{,}445\times10^{219}$
sutrikimasF15	0,999	0,163	6,094
sutrikimasF19	0,182	0,103	0,322
sutrikimasF63	0,562	0,206	8,612
amzius_gr	1,061	1,012	1,113
taikytas_gydymasatkryčio_gyd:issilavinimas	1,275	0,755	2,152
taikytas_gydymaskita_issilavinimas	1,015	0,930	1,188
taikytas_gydymasmed_abs_gydymas:issilavinimas	1,356	1,229	1,460
taikytas_gydymasnemed_gydymas:issilavinimas	0,861	0,786	0,944

26 lentelė: „Logit“ modelio su svoriais ir sąveikomis galimybių santykiai ir pasikliovimo intervalai

Iš 26 lentelės galima matyti, kad sutrikimo žymimo F14 galimybių santykis yra labai didelis, dėl to, kad šitas sutrikimas yra rečiausias ir kol kas visi buvę pacientai nuo jo pasveiko. Galimybių santykį galime interpretuoti taip: jeigu užimtumo statusas pakinta iš dirbančio į nedirbantį, toks pacientas turi apie 44% šansą nepasveikti nuo tam tikros priklausomybės.

Toliau pasižiūrėsime į ROC kreivės grafika, sudarytą pagal „Logit“ modelį su svoriais ir sąveika.



17 pav.: Logistinės regresijos ROC kreivė su svoriais ir sąveika

Iš ROC kreivės galime matyti klasifikatoriaus specifiškumo ir jautrumo sąryšį. Atlikome modelio testavimą ant testinių duomenų ir gavome tokius rezultatus

	spėta, kad nepasveiko	spėta, kad pasveiko
nepasveiko	125	20
pasveiko	470	897

27 lentelė: „Logit“ modelio su svoriais ir sąveikomis spėjimai su testiniais duomenimis

Modelis testiniams duomenims spėjo gerai ir jau taip kaip norime. Teisingai atspėjo net 125 nepasveikusius pacientus iš 145, tai yra 86,2%, o pasveikusių žmonių pataikė 897 iš 1367, tai yra 64,3%. Bendras modelio tikslumas gaunamas 67,6%. Tai jau yra daug geriau negu turėjo anksčiau, nes modelis spėja ir nulius ir vienetus bent 50% tikslumu, kas mums jau leidžia daryti išvadą, kad modelis nespėlioja ir yra galimas naudoti interpretacijai.

5.4 XGBoost modelis

XGBoost (Angl. Extreme Gradient Boosting) yra labai galinga mašininio mokymosi technika, kuri turi didelį našumą ir efektyvumą, ir kuri gali daug geriau prognozuoti mūsų nepriklausomojo kintamojo reikšmes, negu anksčiau aptarti modeliai ir metodai. Jis pagrįstas gradiento didinimo (Angl. gradient boosting) algoritmu, kuris naudojamas regresijos ir klasifikavimo užduotims spręsti, būtent tai, ką mes ir darome.

Vienas iš naudojamų metodų XGBoost modeliuose yra gradiento didinimas. Gradientas tai yra vektorius, apibūdinantis skaliarinį lauką. Skaitine reikšme ir kryptimi apibūdina didžiausią skaliarinio dydžio u = u(x, y, z) kitimo greitį. Jo esmė yra sukurti stiprų modelį, kombinuojant daug silpnų modelių (dažniausiai medžių). Kiekviename žingsnyje modelis bando sumažinti klaidas, padarytas ankstesniuose žingsniuose ir taip mokosi daryti tikslesnius spėjimus.

XGBoost optimizuoja tikslinę funkciją, kuri apima klaidų funkciją (loss function) ir reguliavimo terminą:

$$\mathcal{L}(t) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) \tag{5.2}$$

kur:

- l yra klaidų funkcija,
- y_i yra tikroji reikšmė,
- $\hat{y}_i^{(t)}$ yra modelio prognozė t-uoju žingsniu,
- Ω yra reguliavimo terminas, kuris padeda išvengti persimokymo.

Reguliavimo terminas yra apibrėžiamas kaip:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{5.3}$$

kur:

- γ yra parametras, kontroliuojantis lapų skaičių T,
- λ yra parametras, kontroliuojantis svorių w_j dydį.

Kiekviename žingsnyje, naujas modelis f_t pridedamas siekiant minimizuoti likusią klaidą:

$$f_t = -\eta \nabla_{\hat{y}^{(t-1)}} \mathcal{L} \tag{5.4}$$

kur η yra mokymosi greitis (learning rate).

XGBoost yra galingas ir efektyvus algoritmas, tinkamas įvairioms mašininio mokymosi užduotims ir kurį bandėme pritaikyti ir mes. Jo gebėjimas optimizuoti tikslinę funkciją ir išvengti persimokymo daro jį patraukliu pasirinkimu tiek akademinėje, tiek pramoninėje aplinkoje. Ir būtent tai šiame skyriuje mes bandysime realizuoti su mūsų duomenų rinkiniu.

5.4.1 XGBoost modelio kūrimas

Priešingai negu anksčiau bandytuose metoduose ir modeliuose, mums nebereikia tikrinti jokių modelio tinkamumo hipotezių bei išskirčių, XGBoost modelis dažniausiai nesusiduria su jokiomis problemomis, kurios gali daryti įtakos rezultatams. Vienintelis dalykas kurį supratome iš savo ankstesnių stebėjimu ir tarpinių rezultatų, kad nepriklausomojo kintamojo svorių keitimas gali būti naudingas, todėl šiuo metodu bandysime pasinaudoti ir šioje dalyje.

Pirmiausia mes susidarėme matricą, kurioje visi priklausomi kategoriniai parametrai yra perkoduojami į dvejatainį fiktyvųjį kintamąjį (Angl. one-hot encoding).

Tada mes perkoduojam mūsų nepriklausomąjį kintamąjį, kuris rodo ar pacientas pasveiko ar ne, į skaitinį parametrą, nes toks modelis reikalauja tokio formato, kuris padės mūsų modelio mokymui.

Tada mes sukuriame bendrą XGBoost modeliui pritaikytą matricą su strukturizuotais mūsų duomenimis, į kuriuos įeina buvę 2 etapai. Tai padės modeliui greičiau ir efektyviau dirbti ir mokytis.

Toliau atliekame svorių keitimą. Kaip ir esame mynėje ankstesnėse dalyse, turime labai nesubalansuotą nepriklausomąjį kintamąjį, kas labai trukdo atlikti tinkamus spėjimus ir modeliai spėja per daug vienos grupės reikšmių. Taigi šioje dalyje darysime tapatį, pagal tokią pat svorio formulę.

Sekantis žingsnis būtų susikurti parametrų aibę, kuri ir padės modeliui klasifikuoti. Į parametrų aibę įtraukiame svorių perskaičiavimo funkciją, kuri padės gerinti rezultatus. Taip pat turime nurodyti kaip apmokysime mūsų modelį. Mes pasirinkome naudoti jau išbandytą logistinės regresijos binarinio atsako modelį, kuris bus testuojamas daug kartų iki kol bus gaunamas reikiamas rezultatas. Šioje vietoje atsiranda persimokymo problema.

Taigi atlikę šiuos veiksmus, jau galime susidaryti modelį, kuris galės daug kartu vykdyti logistinę regresiją ir sudaryti galimybių medį, pagal kurį klasifikuos nepriklausomąjį kintamąjį.

Dabar galime pasižiūrėti kaip gerai modelis spėja pacientų pasveikimą su mūsų apmokymo duomenimis. Tai padarysime susikurę pačią paprasčiausią spėjimų lentelę, iš kurios galėsime išsiskaičiuoti procentus kiek modelis atspėjo pasveikusių ir nepasveikusių žmonių.

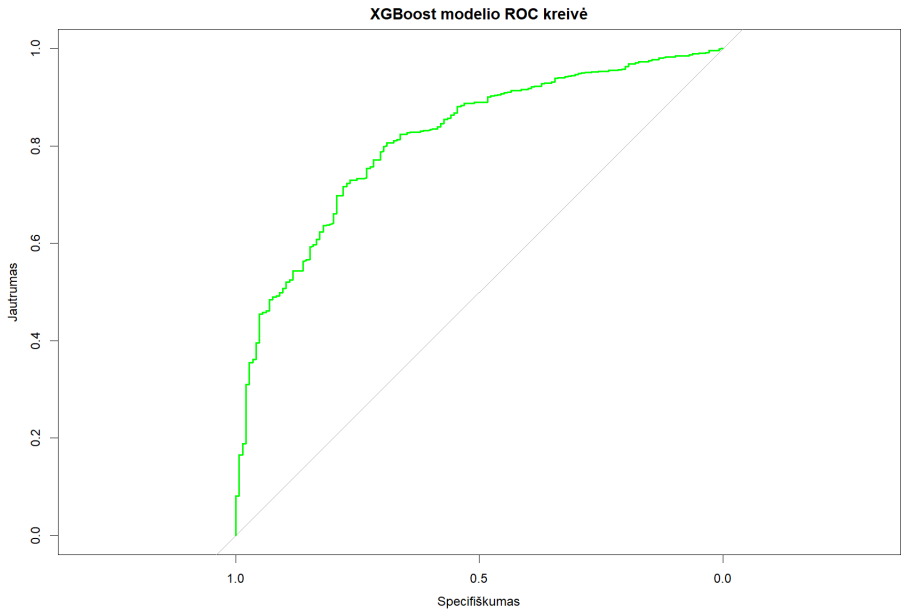
	spėta, kad nepasveiko	spėta, kad pasveiko
nepasveiko	576	59
pasveiko	979	4436

28 lentelė: XGBoost modelio spėjimai su apmokymo duomenimis

28 lentelėje matome, kad modelis jau daug geriau spėja abi nepriklausomo kintamojo grupes, iš 635 pacientų, kurie nepasveiko, mūsų modelis atspėjo net 576 nepasveikusių pacientų, kas išreiškus procentais yra net 90,7%, tačiau jau pažiūrėjus kaip modelis spėja pasveikusius žmones, matome šiek tiek prastesnį rezultatą, kad iš 5415 pacientų teisingai pataikė 4436, kas išreiškus procentais yra 82%. Tai mums duoda 82,8% bendrą modelio tikslumą, nes mažesnis procentas, atspėtų pasveikusių pacientų skaičiaus, daro didesnę įtaką bendram modelio tikslumui. Bet tai yra daug geresnis rezultatas negu turėjome prieš tai atliktuose bandymuose.

5.4.2 XGBoost modelio testavimas

Susidarius modelį su apmokymo duomenimis, dabar galime jį patikrinti su jau padalintais testavimo duomenimis. Primename, kad dalinome santykiu 80/20, taigi šioje dalyje testuosime savo modelį su testavimo duomenimis. Pirmiausia pasižiūrėsime kaip atrodo ROC kreivė ir kaip ji skiriasi nuo anksčiau bandytų modelių ROC kreivių.



18 pav.: XGBoost modelio ROC kreivė

XGBoost modelio ROC kreivė yra aiškiai geresnė negu mūsų praeitų modelių, nes matome, kad plotas po kreive yra jau daug didesnis. Sekantis dalykas ka galime pasižiūrėti tai yra klasifikavimo lentelė su naujais duomenimis.

	spėta, kad nepasveiko	spėta, kad pasveiko
nepasveiko	100	45
pasveiko	265	1102

29 lentelė: XGBoost modelio spėjimai su apmokymo duomenimis

29 lentelėje matome šiek tiek blogesnius rezultatus negu tikėjomės. Mūsų sukurtas modelis nelabai gerai prisitaikė prie testinių duomenų. Iš 145 nepasveikusių pacientų, modelis atspėjo tik 100, kas yra tik apie 69%, kas yra žymiai blogiau negu su apmokymo duomenimis. Tapatį galime pastebėti ir su pasveikimų spėjimu, pasveikusių žmoniu modelis atspėjo 1102 iš 1367, tai yra 80,6%. Tačiau kaip ir žinome, kad pasveikusių pacientų geresni spėjimai, daro daugiau įtakos bendram modelio tikslumui, taigi apskaičiavus jį gauname 79,5% tikslumą.

6 Rezultatai

Atlikome duomenų ir regresinę analizę su gautais duomenimis. Gauti rezultatai mums parodė, kad paprasti „Probit“ ir „Logit“ modeliai pagrinde spėja daugiau pasveikųsių pacientų, kas yra neblogai sveikatos atžvilgiu, tačiau blogai modelio interpretacijai. Kursinio darbo metu su R programa buvo bandoma patobulinti modelio kokybę, kad kuo tiksliau būtų spėjama ar pacientas sėkmingai ar nesėkmingai baigs gydymą. Lentelėje pateikiamas kiekvienas sukurtas modelis su jo bendru tikslumu, kiek teisingai atspėjo pacientų, kurie nepasveiko ir, kurie pasveiko.

Modelio pavadinimas	**Nepasveikimo tikslumas**	**Pasveikimo tikslumas**	**Bendras tikslumas**
„Logit“ modelis	<1%	99.8%	90,3%
„Logit“ modelis su svoriais	40%	87,1%	82,6%
„Logit“ modelis su svoriais ir sąveika	86,2%	64,3%	67,6%
XGBoost	69%	80,6%	79,5%

30 lentelė: Modelio tikslumai

Taigi matome, kad modelių bendras tikslumas dar neparodo, kad modelis yra geras. Galime aiškiai matyti, kad į modelį įtraukus sąveikas ir svorius, gauname žymiai geresnius rezultatus, kuriuos jau galima interpretuoti ir naudoti. Pats geriausias modelis gavosi XGBoost, nes jo bendras tikslumas buvo geriausias ir nepasveikimus ir pasveikimus, spėjo bent 50%, koks ir žinomas reikalavimas.

7 Išvados bei rekomendacijos

Taigi, galutinės darbo išvados yra, kad paprasti binarinio atsako modeliai mūsų duomenims visiškai netiko ir negalime jų naudoti, nes kaip matome iš rezultatų, jie nors ir tiksliai pataiko pasveikusius pacientus, visiškai arba beveik nepataiko nepasveikųsių. Mūsų atveju, mes stengiamės šituos modelius patobulinti, todėl įtraukus svorius ir sąveikas šitie modeliai pradėjo spėti geriau ir net tiko naudojimui. Tačiau ties tuo mūsų darbas nesibaigė, nes pasinaudojus iš literatūros šaltinių XGBoost modeliu gavome dar geresnius rezultatus, kuriuos taip pat galima naudoti. Norint gauti dar geresnius rezultatus ir tikslesnius spėjimus, yra keli variantai, vienas iš jų, daugiau duomenų apie pacientus ir apie jų sveikatos būklę. Taip XGBoost modelį galima tobulinti keičiant jo hiperparametrus ir ieškant geriausių. Dar vienas patobulinimo būdas būtu naudoti kryžminę patikrą (Angl. Cross validation). To mūsų darbe nebuvo, tačiau vistiek gauti rezultatai yra gana geri ir tiknami naudojimui.

8 Literatūros sąrašas

- Vydas Čekanavičius, Gediminas Murauskas. *Taikomoji regresinė analizė socialiniuose tyrimuose.* – Vilnius: Vilniaus universitetas, 2014. – 561 p.
- Dirbtinio intelekto programinė įranga.
- De Souza FSH, Hojo-Souza NS, Dos Santos EB, Da Silva CM and Guidoni DL (2021) Predicting the Disease Outcome in COVID-19 Positive Patients Through Machine Learning: A Retrospective Cohort Study With Brazilian Data.
- Franken K, ten Klooster P, Bohlmeijer E, Westerhof G and Kraiss (2023) Predicting non-improvement of symptoms in daily mental healthcare practice using routinely collected patient-level data: a machine learning approach.
- Rūta Levulienė *Binarinio atsako modeliai.* – Vilnius: Vilniaus universitetas, 2024.

9 Priedai

- https://github.com/RytisBalt/Kursinis_Darbas/blob/main/KursinisR(1).R