KAUNO TECHNOLOGIJOS UNIVERSITETAS

MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS

TAIKOMOSIOS MATEMATIKOS KATEDRA

Ingrida Grabauskytė

KAUNO MIESTO GYVENTOJŲ RIZIKOS SUSIRGTI ŠIRDIES IR KRAUJAGYSLIŲ LIGOMIS PROGNOZAVIMO MODELIAI

Magistro darbas

Vadovas
doc. dr. V. Janilionis

Konsultantas
prof. habil. dr. A. Tamošiūnas
(LSMU, Kardiologijos institutas)

KAUNAS, 2014

KAUNO TECHNOLOGIJOS UNIVERSITETAS

MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS

TAIKOMOSIOS MATEMATIKOS KATEDRA

TVIRTINU
Katedros vedėjas
doc. dr. N. Listopadskis
2014 06 03

KAUNO MIESTO GYVENTOJŲ RIZIKOS SUSIRGTI ŠIRDIES IR KRAUJAGYSLIŲ LIGOMIS PROGNOZAVIMO MODELIAI

Taikomosios matematikos magistro baigiamasis darbas

Vadovas
doc. dr. V. Janilionis
2014 06 02

Recenzentas
prof. dr. V. Šaferis
2014 06 03

Atliko
FMMM 2 gr. stud.
I. Grabauskytė
2014 05 30

KAUNAS, 2014

KVALIFIKACINĖ KOMISIJA

Pirmininkas:	Juozas Augutis, profesorius (VDU)
Sekretorius:	Eimutis Valakevičius, profesorius (KTU)
Nariai:	Arūnas Barauskas, dr., direktoriaus pavaduotojas (UAB „Danet Baltic“)
	Vytautas Janilionis, docentas (KTU)
	Zenonas Navickas, profesorius (KTU)
	Kristina Šutienė, docentė (KTU)
	Jonas Valantinas, profesorius (KTU)

Grabauskytė I. Kauno miesto gyventojų rizikos susirgti širdies ir kraujagyslių ligomis prognozavimo modeliai: Taikomosios matematikos magistro darbas/ vadovas doc. dr. V. Janilionis; Taikomosios matematikos katedra, Matematikos ir gamtos mokslų fakultetas, Kauno technologijos universitetas. – Kaunas, 2014. – 56 p.

SANTRAUKA

Širdies ir kraujagyslių ligos – tai dažniausiai pasitaikantys susirgimai ne tik Lietuvoje, bet ir Europoje [26]. Mirtingumas nuo šių ligų, taip pat yra vienas iš dažniausių. Todėl labai svarbu sukurti matematinius modelius, kurie leistų nustatyti įvairių veiksnių įtaką šių ligų išsivystymui ir įvertintų riziką susirgti. Šio darbo tikslas – panaudojus Lietuvos sveikatos mokslų universiteto sukauptus duomenis, sukurti statistinius modelius šių uždavinių sprendimui.

Darbe atlikta analizė metodų ir modelių, kurie dažniausiai taikomi Lietuvoje bei užsienyje širdies ir kraujagyslių ligų susirgimų bei mirčių nuo jų prognozavimui. Pagal Lietuvos sveikatos mokslų universiteto Kardiologijos instituto mokslininkų 1982-2013 metais atliktų Kauno miesto 45-72 m. amžiaus gyventojų tyrimų duomenis, sukurti regresiniai modeliai širdies ir kraujagyslių ligų susirgimų rizikos vertinimui ir mirčių dažnių prognozavimui. Nustatyti faktoriai, kurie labiausiai įtakoja susirgimus širdies ir kraujagyslių ligomis. Sukurti modeliai realizuoti su statistinės analizės sistema SAS. Sukurta internete veikianti skaičiuoklė, kuri leidžia kiekvienam asmeniui įsivertinti riziką susirgti širdies ir kraujagyslių ligomis.

Grabauskytė I. Risk Prediction Models for Cardiovascular Diseases Amongst Citizens of Kaunas City: Master‘s work in applied mathematics / supervisor dr. assoc. prof. V. Janilionis; Department of Applied mathematics, Faculty of Mathematics and Natural Sciences, Kaunas University of Technology. – Kaunas, 2014. – 56 p.

SUMMARY

Cardiovascular diseases are the most common diseases not only in Lithuania, but in all of Europe [26]. Mortality due to these diseases is one of the most common. So it‘s very important to create mathematical models, that could identify and evaluate the risk factors impacting the reason of these diseases. The main objective of this work is to use data collected from Lithuanian University of Health Sciences to create statistical models to solve these tasks.

We performed methods and models for analysis, which are often used in Lithuania and abroad identifying the risk to predict deaths from cardiovascular diseases. According to the Lithuanian University of Health Sciences, Institute of Cardiology, research data that were carried out from 1982 to 2013 years from Kaunas population, aged 45 to 64 years. There were developed regression models for cardiovascular diseases risk assessment and to predict death rates. Factors that influenced the cardiovascular diseases were identified. Created models realized with SAS statistical analysis system. An active online calculator that allows for each person to assess the risk of getting cardiovascular diseases was designed.

TURINYS

LENTELIŲ SĄRAŠAS

PAVEIKSLŲ SĄRAŠAS

ĮVADAS

Aktualus medicinos uždavinys – nustatyti veiksnius, kurie kelią riziką ligos išsivystymui, atsinaujinimui ar net mirčiai. Nors ir žinant šių rodiklių reikšmes, svarbu įvertinti ir ligos išsivystymo, atsinaujinimo, mirties tikimybę.

Širdies ir kraujagyslių ligos yra viena iš pagrindinių mirties priežasčių Lietuvoje ir visoje Europoje [26]. Europos regiono šalyse mirtingumas nuo šių ligų sudaro apie 40% visų mirčių, o Lietuvoje – apie 60% [32]. Todėl labai svarbu išsiaiškinti, kokie rizikos veiksniai daugiausiai daro įtaką šių ligų išsivystymui, kad būtų galima laiku užkirsti kelią širdies ir kraujagyslių ligų susirgimams. Šiam tikslui įgyvendinti atliekami tyrimai visame pasaulyje, taip pat ir Lietuvoje. Lietuvos sveikatos mokslų universiteto Kardiologijos instituto mokslininkai 2012 metais atliko tyrimą „Cardiovascular risk factors and cognitive function in middle aged and elderly Lithuanian urban population: results from the HAPIEE study“ [12].

Darbo tikslas – papildžius minėto tyrimo duomenis, sukurti modelius širdies ir kraujagyslių ligų rizikos vertinimui ir mirčių dažnio prognozavimui. Sukurtus modelius realizuoti statistinės analizės sistema SAS, bei sukurti rizikos susirgti širdies ir kraujagyslių ligomis skaičiuoklę virtualioje erdvėje (panaudojus įrankius: *bootstrap*, *jquery* ir *knockout*).

Panaudojus logistinės regresijos, logistinės regresijos retiems įvykiams metodus sudaryti širdies ir kraujagyslių ligų prognozavimo modeliai, o panaudojus diagialypės tiesinės regresijos ir kvantilinės regresijos metodus sudaryti regresijos modeliai Kauno miesto gyventojų mirtingumo prognozavimui.

Konferencijos ir publikacijos:

- XI Taikomosios matematikos konferencija 2013. Pranešimas: „Klinikinių ir genetinių veiksnių įtakos susirgti išemine širdies liga tyrimas“ [14];
- „Matematika ir matematikos dėstymas – 2014“. Pranešimas: „Kauno miesto gyventojų rizikos susirgti galvos smegenų insultu prognozavimo modelis“ [15];
- „Taikomoji matematika – 2014“. Pranešimas: „Kauno miesto 45-64 m. amžiaus gyventojų mirtingumo prognozavimas“ [16];
- XIII-oji Tarptautinė Lietuvos biochemikų konferencija 2014 m. birželio 18-20 d. Birštone. Pranešimas: „Transformuojančio faktoriaus TGF β1 koncentracijos kitimai esant kylančiosios aortos dilatacinei patalogijai“ [17];
- „Studentų mokslinės veiklos skatinimas“, 2014 m. birželio 26-26 dieną. Pranešimas: „Fibrilino geno polimorfizmu įtakojančių krūtinės aortos aneurizmos formavimąsi, tyrimas“ [18].

Pateikti straipsniai spaudai:

- “Does the association of *FBN1* SNPs with dilatative pathology of the ascending aorta is an extension of indications for surgical treatment?“ žurnalui *European Cardiovascular Research* (bendraautorė) [19];
- “Association of FBN1 SNPs rs2118181, rs1036477, rs10519177, rs755251 and rs4774517 with dilatative pathology of the ascending aorta (DPAA)” žurnalui *Journal of Cardiovascular surgery* (bendraautorė) [20];
- “The effect of clinical and genetic factor on early definite stent thrombosis” (bendraautorė) [21].

1. TEORINĖ DALIS

Šioje dalyje apžvelgti užsienio ir Lietuvos mokslininkų paskelbti straipsniai širdies ir kraujagyslių ligų prognozavimo tematika. Išanalizuoti šių uždavinių sprendimui dažniausiai naudojami statistiniai metodai ir programinė įranga.

1.1. ŠIRDIES IR KRAUJAGYSLIŲ LIGŲ PROGNOZAVIMO UŽDAVINYS

Vienas pagrindinių duomenų analizės uždavinių – parinkti tinkamiausią metodą turimiems duomenims. Medicinoje, norint prognozuoti susirgimo tikimybę tam tikra liga, mirčių skaičių nuo tos ligos ar atlikti kitas prognozes, dažniausiai naudojami įvairūs regresinės analizės modeliai.

Apibrėšime pagrindinius darbe naudojamus terminus:

Insultas – dar vadinamas *galvos smegenų insultu* yra smegenų kraujotakos pažeidimas, kurį sukelia staiga sutrikusi smegenų dalies kraujotaka arba kraujo išsiliejimas į smegenis. Smegenų insulto rūšys: *hemoraginis* (kraujo išsiliejimas į smegenis) ir *išeminis* (nutrūkęs arba ženkliai sumažėjęs kraujo patekimas į smegenis). Insultas gali sukelti kalbos problemas, protinius ir emocinius sutrikimus, paralyžių [31].

Išeminė širdies liga (IŠL) – tai širdies raumens funkcijos sutrikimas, atsiradęs dėl sumažėjusio aprūpinimo krauju. Išeminė širdies liga gali būti *ūminė* ir *lėtinė*. Dėl sumažėjusio aprūpinimo arteriniu krauju, atsirada širdies būklės: krūtinės angina, ūminis miokardo infarktas (dar vadinamas *širdies smūgiu* – širdies raumens ląstelių negrįžtamas pažeidimas, kurį sukelia nutrūkęs aprūpinimas krauju ir deguonimi), kartotinis miokardo infarktas, ūminio miokardo infarkto komplikacijos, lėtinė išeminė širdies liga. Išeminė širdies liga vystosi dėl vainikinių širdies arterijų aterosklerozės. Arterijų spindis gali siaurėti palaipsniui, tada, metams bėgant, organizmas prisitaiko ir vystosi lėtinė išeminė širdies liga. Jei aterosklerozinė plokštelė įtrūksta, virš jos gali formuotis trombas. Tokiu atveju gali įvykti ūminis miokardo infarktas ir net staigi mirtis [31].

Apžvelgsime Lietuvos bei užsienio straipsnius, kuriuose statistiniai metodai, taikomi širdies ir kraujagyslių ligų duomenų analizei.

Atlantoje (JAV) 1999-2006 m. buvo atliktas išeminės širdies ligos rizikos veiksnių ir aukšto cholesterolio kiekio kraujyje paplitimo tyrimas tarp jaunų suaugusiųjų (vyrai – 20-35 m. amžiaus, o moterys – 20-45m.). Tyrimo imties dydis – 2587. Atlikus tyrimą nustatyta, kad maždaug 65% jaunų suaugusiųjų serga išemine širdies liga, 26% sergančių turi 2 ar daugiau rizikos veiksnių, tokių kaip rūkymas, hipertenzija, nutukimas ar šeimos nariai sergantys šia liga; 12% – su vienu iš anksčiau

minėtų rizikos faktorių, ir 7% – be rizikos veiksnių turėjo didelį mažo tankio lipoproteinų cholesterolį. Duomenų analizei panaudota daugiamatė Kokso regresija [3].

Širdies ir kraujagyslių ligų prognozavimas 2010-2030 metams Jungtinėse Amerikos Valstijose atliktas analizuojant 1999-2006 metais surinktus duomenis. Atsižvelgiant į didėjančias sveikatos priežiūros išlaidas ir jų poveikį ekonomikai, apskaičiuota vidutinė kaina vienam asmeniui pagal amžiaus grupes (18-44 m., 45-64 m., 65-79 m. ir 80m.), lytį (vyrai, moterys) ir rasę. Sudarant modelius taikyta logistinė regresija [1].

Maskvoje (Rusija) 2011 m. atliktas išeminės širdies ligos diagnozės neuroniniais tinklais, įskaitant genetinius ir klinikinius parametrus modeliavimas. Į analizę įtraukti rizikos veiksniai: amžius, lytis, bendrasis cholesterolis, didelio/mažo tankio lipoproteinų cholesterolis, trigliceridų, cholesterolio santykis, gliukozė, arterinė hipertenzija, cukrinis diabetas, rūkymas, nutukimas, šeimos narių istorija apie išeminę širdies ligą [5].

Lietuvos sveikatos mokslų universiteto Kardiologijos instituto mokslininkų nustatytos 35-64 metų Kauno miesto gyventojų išeminės širdies ligos rizikos veiksnių vidurkių, didelės absoliučios rizikos susirgti ūminiu išeminiu sindromu bei sergamumo ūminiu miokardo infarktu pokyčių sasajos per 1983-2002 metų tyrimo laikotarpį. Analizuoti pagrindiniai išeminės širdies ligos rizikos veiksnių vidutiniai dydžiai: sistolinis/diastolinis kraujospūdžiai, surūkytų cigarečių skaičius per dieną, kūno masės indeksas, suvartoti alkoholinių gėrimų vienetai per mėnesį, cholesterolio koncentracija kraujo serume, gliukozės koncentracija nevalgius kapiliariniame kraujyje bei sergamumo ūminiu miokardo infarktu rodikliai. Duomenų analizė atlikti panaudojus netiesinė regresijos modelį [4].

Kauno miesto gyventojų mirtingumo nuo kraujotakos sistemos ligų pokyčius 2004-2008 m. atliko Lietuvos sveikatos mokslų universiteto Kardiologijos instituto mokslininkai. Tyrimo tikslas buvo nustatyti Kauno miesto 25-44 m. amžiaus ir 45-74m. amžiaus vyrų bei moterų mirtingumo nuo kraujotakos sistemos ligų rodiklių kitimo kryptis per 2004-2008 metus, atsižvelgiant į oficialiosios mirtingumo statistikos duomenis. Darbe panaudota netiesinė regresija [6].

Šveicarijoje 1990-2000 m. atliktas tyrimas parodė, jog mažesnis mirtingumas nuo išeminės širdies ligos ir insulto yra aukščiau jūros lygio esančioje Šveicarijos dalyje. Santykinė rizika buvo apskaičiuojama taikant kelių kintamųjų Puasono regresiją. Veiksniai, kurie buvo įtraukti į regresijos modelį: mokymosi trukmė, geros manieros, gyvenamos vietos aukštis virš jūros lygio ir skirtumas tarp gyvenamos vietos bei gimimo vietos [2].

Taip pat apžvelgsime Lietuvos bei užsienio skaičiuokles virtualioje erdvėje, kurios naudojamos įsivertinti riziką susirgti širdies ir kraujagyslių ligomis bei mirties, nuo šių ligų, tikimybę.

Australų sukurta skaičiuoklė skirta įsivertinti riziką susirgti širdies ir kraujagyslių ligomis. Reikalingi duomenys apie asmenį: lytis, amžius, sistolinis kraujospūdis, rūkymas, bendrasis cholesterolis, didelio tankio lipoproteinų cholesterolis, diabetas, kairiojo skilvelio hipertropija. Rizika susirgti pateikiama procentais [27].

Naudojant *Framingham Heart Study* informaciją, sukurta skaičiuoklė įvertinti riziką susirgti infarktu per 10 metų. Norint įsivertinti riziką, reikia pateikti asmens amžių, lytį, bendrąjį cholesterolį, didelio tankio lipoproteinų cholesterolį, rūkymą, sistolinį kraujo spaudimą ir informaciją, ar vartoja vaistus esant aukštam kraujo spaudimui [28].

Viskonsino medicinos koledže sukurta skaičiuoklė, nustatyti riziką susirgti koronarine širdies liga per ateinančius 10 metų bei palyginti su kitų, to paties amžiaus, asmenų rizika. Rizikos veiksniai: lytis, amžius, rūkymas, diabetas, kraujospūdis, bendrasis cholesterolis arba mažo tankio lipoproteinų cholesterolis, didelio tankio lipoproteinų cholesterolis. Pateikiami taškai, santykinė rizika bei rizika, išreikšta procentais [29].

Lietuvos širdies asociacijos internetiniame puslapyje pateikta skaičiuoklė, skirta įsivertinti mirties, nuo kardiovaskulinės ligos, riziką per 10 metų priklausomai nuo lyties, amžiaus, sistolinio kraujospūdžio, bendro cholesterolio ir rūkymo. Taip pat pateikiamas ir ne automatinis skaičiuoklės variantas [30].

Toliau aptariami širdies ir kraujagyslių ligų prognozavimui dažniausiai taikomų statistinės analizės metodų galimybės ir ribotumai.

1.2. ŠIRDIES IR KRAUJAGYSLIŲ LIGŲ PROGNOZAVIMUI TAIKOMŲ STATISTIKOS METODŲ LYGINAMOJI ANALIZĖ

Daugialypė tiesinė regresinė analizė. Tarkime, kad priklausomas kintamasis yra *Y*, kurio *i*–ąją reikšmę Y_i norima prognozuoti, kai nepriklausomų kintamųjų reikšmės $x_{1i,}x_{2i},...,x_{ki}$ yra fiksuotos. Šio uždavinio sprendimui naudojamas diagialypės tiesinės regresijos modelis [9, 22]:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i , \qquad \textbf{(1.1)}$$

kur $\beta_0, \beta_1, \beta_2, ..., \beta_k$ – modelio koeficientai, ε_i – atsitiktinės modelio paklaidos.

Šiame modelyje daromos prielaidos, kad [9, 22]:

- ε_i – nepriklausomi atsitiktiniai dydžiai, turintys normalųjį pasiskirstymą su vidurkiu lygiu 0 ir dispersija σ^2;
- nepriklausomų kintamųjų reikšmės yra neatsitiktinės ir tiesiškai nepriklausomos (vieno nepriklausomo kintamojo reikšmių negalima išreikšti likusiųjų tiesine daugdara).

Nežinomų koeficientų $\beta_0, \beta_1, \beta_2, ..., \beta_k$ įverčiai $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$ randami mažiausių kvadratų metodu. Koeficientai $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$ parenkami taip, kad Y_i reikšmės būtų kuo arčiau modeliuotos reikšmės $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki}$ [22].

Vertinant, ar x_i reikšmingas regresiniame modelyje, tikrinama hipotezė:

$$\begin{cases} H_0 : \beta_i = 0 \\ H_a : \beta_i \neq 0 \end{cases} \quad \textbf{(1.2)}$$

Apskaičiuojama liekamųjų paklaidų kvadratų suma SSE_j. Tuomet kriterijaus statistika $t = \dfrac{\hat{\beta}_i}{\sqrt{MSE / SSE_j}}$. Nulinė hipotezė atmetama $(\beta_j \neq 0)$, jeigu $|t| > t_{\alpha/2}(n-k-1)$, kur $t_{\alpha/2}(n-k-1)$ - Stjudento skirstinio su $(n-k-1)$ laisvės laipsnių $\alpha/2$ lygmens kritinė reikšmė; α - reikšmingumo lygmuo. Nulinė hipotezė neatmetama, jeigu $|t| \leq t_{\alpha/2}(n-k-1)$ [9, 22].

Daugialypės regresijos modelio tinkamumui vertinti naudojamas koreguotas determinacijos koeficientas: $AdjR^2 = 1-(n-1)*SSE/(SST*(n-k-1))$, čia *SSE* – liekamųjų paklaidų kvadratų suma, *SST* – taip vadinama visa kvadratų suma. Šis koeficientas parodo, kuri *Y* dispersijos dalis paaiškinama *k* faktorių tiesine daugdara, atsižvelgiant į imties dydį *n* ir lygtyje esančių kintamųjų skaičių. Rekomenduojama naudoti tą regresinį modelį, kurio $AdjR^2$ yra didžiausias [9, 22]. Modelio tinkamumo įvertinimui taip pat naudojamas Malau statistika (angl. *Mallow C(p) statistic*). Modelis tinkamesnis, kuo statistikos reikšmė artimesnė $p-1$, kur *p* – nepriklausomų kintamųjų skaičius modelyje [11].

Sudarant regresijos modelius labai svarbu nustatyti išskirtis. Dažniausiai naudojamos išskirčių nustatymo statistikos (*p* – nepriklausomų kintamųjų skaičius, *n* – imties didumas) [8, 13]:

- *COOK'S D* – atsižvelgia į standartizuotą liekaną ir į stebėjimo įtakos indeksą $(>4/n)$;
- *LEVERAGE* – įtakos matas, nustatantis išskirtis nepriklausomų kintamųjų erdvėje $(>2p/n)$. Jeigu imtis pakankamai didelė, tai įtakos matas turėtų būti mažesnis $(>2(p+1)/n)$;
- *DFFITS* – standartinė stebėjimų įtaka prognozuojamoms reikšmėms $\left(>\left|2\cdot\sqrt{p/n}\right|\right)$;

- *RSTUDENT* – statistika, skaičiuojama kiekvienam stebėjimui (skirtumas tarp dviejų statistikų: viena apskaičiuota kiekvienam taškui, kita – be *i*-ojo taško, tada statistikos reikšmė lygi šiam skirtumui. $RSTUDENT = \frac{r_i}{s_{(i)}\sqrt{(1-h_i)}}$, čia $r_i = y_i - \widehat{y}_i$, dispersija $s_{(i)}$ įvertinama be *i*-ojo stebėjimo, h_i - *i*-ojo stebėjimo įtakos indeksas).

Daugialypės regresijos privalumai: pagal regresijos modelį galima nustatyti priežastinius ryšius; galima identifikuoti kintamuosius, kurie labiausiai įtakoja *Y*. Trūkumas: modelis turi tenkinti nemažai prielaidų, o tai dažniausiai būna didelė kliūtis analizuojant realius duomenis.

Logistinė regresija yra viena iš dažniausiai naudojamų širdies ir kraujagyslių ligų tyrime. Jos tikslas – kategorinio kintamojo reikšmių tikimybių prognozavimas pagal nepriklausomų kintamųjų reikšmes [9].

Esant kategoriniam priklausomam kintamajam (įgyjamos reikšmės yra 0 ir 1) bei fiksuotoms nepriklausomų kintamųjų reikšmėms, galime užrašyti daugialypės regresijos modelį:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_n x_{ni} + \varepsilon_i, \qquad \textbf{(1.3)}$$

čia Y_i – atsitiktinis dydis (dvireikšmis). Šio atsitiktinio dydžio tikimybė įgyti reikšmę 1 – $P(Y_i = 1) = \theta_i$, o įgyti reikšmę 0 – $P(Y_i = 0) = 1 - \theta_i$. $\beta_0, \beta_1, \beta_2, ..., \beta_n$ – regresijos parametrai, $x_{1i}, x_{2i} ..., x_{ni}$ – nepriklausomi kintamieji, ε_i – atsitiktinė paklaida [9].

Prognozėms negalima taikyti (1.3) regresijos modelio, nes netenkinamos tiesinės regresijos prielaidos. Kadangi Y_i įgyja dvi reikšmes, tai ir atsitiktinė paklaida ε_i įgis tik dvi reikšmes (nebus tenkinama normalumo prielaida). Kadangi $EY_i = \theta_i$, tai atliktume prognozę tikimybei, su kuria $Y_i = 1$, tačiau dešineje (1.3) lygybės pusėje galima gauti ir neigiamas reikšmes (taip pat reikšmes, kurios bus didesnės už 1), o tai prieštarauja tikimybės apibrėžimui [9].

Norint įvertinti tikimybės θ_i priklausomybę nuo kintamųjų $x_{1i}, x_{2i} ..., x_{ni}$ yra taikomas logistinės regresinės analizės modelis

$$\theta_i = \frac{\exp\left(\beta_0 + \sum_{n=1}^{k} \beta_n x_{ni}\right)}{1 + \exp\left(\beta_0 + \sum_{n=1}^{k} \beta_n x_{ni}\right)}, \qquad \textbf{(1.4)}$$

čia $\sum_{n=1}^{k} \beta_n x_{ni} = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_{ki} x_{ki}$ [9].

Iš (1.4) lygties:

$$\frac{\theta_i}{1-\theta_i} = \exp\left(\beta_0 + \sum_{n=1}^{k}\beta_n x_{ni}\right) = \exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_{ki} x_{ki}\right), \quad \textbf{(1.5)}$$

tada *logit funkcija* [9, 22]

$$\ln\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \sum_{n=1}^{k}\beta_n x_{ni} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_{ki} x_{ki}\,. \quad \textbf{(1.6)}$$

Funkcijai θ_i modeliuoti naudojama ne tik logistinė, bet ir kitos funkcijos [22]:

- *log-log* modelis: $\theta_i = 1 - \exp\left(-\exp\left(\beta_0 + \sum_{n=1}^{k}\beta_n x_{ni}\right)\right)$;
- *probit* modelis: $\theta_i = (2\pi)^{-1/2} \int_{-\infty}^{\beta_0 + \sum_{n=1}^{k}\beta_n x_{ni}} \exp\left(-y^2/2\right)dy$ ir kt.

Tarkime, yra stebėjimai $\left(Y_i, x_{1i}, x_{2i}, ..., x_{ni}\right)$, $i = \overline{1,k}$; čia Y_i – priklausomas kintamasis įgyjantis reikšmes 0 ir 1, $x_{1i}, x_{2i} ..., x_{ni}$ – nepriklausomų intervalinių arba pseudokintamųjų reiškmės. Parametrų $\beta_0, \beta_1, \beta_2, ..., \beta_n$ įverčius $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, ..., \widehat{\beta}_n$ parenkama taip, kad (1.4) modelis būtų kuo geriau suderintas su duomenimis (taikomas maksimalaus tikėtinumo metodas). Pagal (1.4) formulę, kiekvienam stebėjimui, apskaičiuojama tikimybė θ_i. Parametrų įverčius $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, ..., \widehat{\beta}_n$ reikia parinkti taip, kad tikėtinumo funkcija

$$L = \prod_{i=1}^{n} \widehat{\theta}_i^{Y_i} \prod_{i=1}^{n} \left(1 - \widehat{\theta}_i\right)^{1-Y_i} \quad \textbf{(1.7)}$$

būtų maksimali. Čia $\widehat{\theta}_i = \dfrac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_n x_{ni})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_n x_{ni})}$ [9, 22].

Dėl sudėtingo galimybės logaritmo pokyčių interpretavimo paprastai skaičiuojamas taip vadinamas *galimybių santykis* (angl. *odds ratio*). Jis parodo, kaip kinta priklausomojo kintamojo galimybė įgyti reikšmę 1 [9].

Skaičius $e^{\widehat{\beta}_1}$ yra galimybiu santykis:

$$\left(\frac{P(Y=1)}{P(Y=0)}\right)_{nauja} = e^{\widehat{\beta}_1}\left(\frac{P(Y=1)}{P(Y=0)}\right)_{sena} \quad \textbf{(1.8)}$$

Sakoma, kad nauja galimybė, tai tikimybių santykis, kai nepriklausomojo kintamojo reikšmė padidėja vienetu. Taip pat galima interpretuoti, kad didesnis už vienetą galimybių santykis parodo, kiek kartų (lyginant su ankstesne galimybe), labiau tikėtina, kad Y_i įgis reikšmę 1, o ne 0. Analogiškai, jeigu galimybių santykis mažesnis už vienetą, tai jo atvirkštinė reikšmė parodo, kiek kartų (lyginant su ankstesne galimybe), mažiau tikėtina, kad Y_i įgis reikšmę 1, o ne 0 [11].

Duomenys turi tenkinti šiuos reikalavimus [11]:

- Priklausomas kintamasis Y_i dvireikšmis. Geriausia, kad dauguma nepriklausomų kintamųjų būtų intervaliniai, o dvireikšmių nepriklausomų kintamųjų būtų nedaug.
- Negali dominuoti viena iš Y_i reikšmių. To reikalaujama, kad būtų galima nustatyti, kad būdinga kiekvienai kategorijai, t.y. reikia turėti informacijos apie abi kategorijas. Reikalaujama, kad Y_i reikšmių (0 ar 1) būtų bent penktadalis.
- Kategorinių nepriklausomų kintamųjų kategorijų deriniui turi būti ne mažiau 5 stebėjimai.

Logistinėje regresijoje nereikalaujama nepriklausomų kintamųjų normalumo, nereikalaujama normaliai pasiskirsčiusių paklaidų, taip pat nekalbama apie priklausomojo kintamojo homoskedastiškumą.

Rodikliai, parodantys modelio tikimą duomenims [11]:

- Klasifikacinė lentelė. Konkretiems stebėjimams prognozuojama Y_i reikšmė ir žiūrima, ar sutapo su tikrąja Y_i reikšme. Modelis laikomas geresnis, kuo daugiau yra sutapimų.
- Tikėtinumų santykio kriterijus ir jo *p*-reikšmė. Šis kriterijus parodo, ar modelyje yra bent vienas reikšmingas nepriklausomas kintamasis. Jei *p*-reikšmė didesnė už 0,05, tai modelio tinkamumas yra abejotinas.
- Determinacijos (pseudo) koeficientai (angl. *R square, pseudo* - R^2). Parodo bendrąjį modelio tikimą duomenims (įgyjamos reikšmės yra intervale [0;1]). Kuo šio koeficiento reikšmė didesnė, tuo modelis geriau tinkamas duomenims (paprastai rekomenduojama, kad $R^2 \geq 0{,}2$), tačiau, jei R^2 yra mažas, o visi kiti rodikliai logistinėje regresijoje tinka, vis vien tariama, kad modelis yra tinkamas.
- ROC kreivė (angl. *received operating charakterictic*). ROC kreivė yra jautrumo (angl. *sensitivity*) ir specifiškumo (angl. *specificity*) grafikas. Jautrumas (Y ašis) parodo modelio gebėjimą diagnozuoti susirgimą, jeigu asmuo iš tikrųjų serga. Specifiškumas (X ašis) parodo modelio gebėjimą nustatyti, jog susirgimo nėra, kai jo iš tikrųjų nėra. Modelis tuo geresnis, kuo didesni šie parametrai. Geresnis tas modelis, kurio plotas po ROC kreive yra didesnis.
- Akaike (angl. Akaike Information Criterion, AIC) ir Švarco (angl. *Schwarz Criterion*, SC) informaciniai kriterijai. Šie kriterijai yra skirti modelių palyginimui. Mažesnes AIC ir SC reikšmes turintis modelis bus laikomas geresniu.

Logistinėje regresijoje, stebėjimų rinkinys, kurio reikšmės nukrypsta nuo tikėtino intervalo ir to pasekoje atsiranda ypač didelės liekanos, vadinamas *išskirtimis*. Nustatyti išskirtis yra labai svarbu bet

kuriame modeliavimo uždavinyje. Nesugebėjimas aptikti išskirčių, gali iškreipti gautus rezultatus. Jeigu pašalinus stebėjimą iš esmės pasikeičia koeficientų įverčiai, tai jis laikomas išskirtimi [13].

Išskirčių nustatymui dažniausia naudojamos šios statistikos (p – nepriklausomų kintamųjų skaičius kintamųjų skaičius modelyje, n – imties didumas) [8, 13] :

- *LEVERAGE* – įtakos matas, nustatantis išskirtis nepriklausomų kintamųjų erdvėje $(> 2p/n)$. Jeigu imtis pakankamai didelė, tai įtakos matas turėtų būti mažesnis $(> 2(p+1)/n)$;
- *DFBETAS* – nustato išskirtis, kurios statistiškai reikšmingai įtakoja regresijos lygties koeficientus $\left(> 2/\sqrt{n}\right)$;
- *RESCHI* – Pearsono (Chi kvadrato) liekana $(|RESCHI| > 3)$;
- *RESDEV* – deviacijos liekana $(|RESDEV| > 3)$;
- *DIFCHI* – parodo Pearsono Chi kvadrato skirtumą išbraukus i-ąjį stebėjimą $\left(\left|\sqrt{DIFCHI}\right| > 2\right)$;
- *DIFDEV* – parodo deviacijos pokytį pašalinus *i*-ąjį stebėjimą iš imties $\left(\left|\sqrt{DIFDEV}\right| > 2\right)$

[8, 13].

Apskritai, pašalinus stebėjimus su didžiausiomis liekanomis ar daugiau ekstremaliomis reikšmėmis, beveik visada pagerina modelio tinkamumą [13].

Logistinės regresijos privalumai: nereikalaujama liekanų normalumo; aiškiai suprantami ir lengvai interpretuojami gauti rezultatai. Trūkumai: turint didelę duomenų imtį ir stipriai dominuojančią vieną priklausomojo kintamojo kategoriją, gali būti problemų su klasifikavimu, t.y. visi stebėjimai gali būti priskirti dominuojančiai kategorijai.

Diskriminantinė analizė. Tai yra visuma metodų, skirtų sudaryti taisyklę naujiems stebėjimams klasifikuoti, remiantis pradine stebėjimų klasifikacija [22]. Iš anksto žinomas grupių skaičius, yra „mokymo“ imtis, kurioje yra informacija apie dalies tiriamų objektų požymius ir priklausomybę vienai ar kitai grupei [9].

Diskriminantinė analizė yra plačiai taikoma įvairiose srityse, tame tarpe ir medicinoje. Pavyzdžiui, gydytojas naudodamasis ligonių tyrimų duomenimis diagnozuoja susirgimą.

Nustatome fiksuotą grupių skaičių k. Grupėse atsitiktinai atrenkame po $n_1, n_2, ..., n_k$ stebėjimų ir kiekvienam jų nustatome p kintamųjų $X^{(1)}, X^{(2)}, ..., X^{(p)}$ reikšmių. *i*-osios grupės *j*-ojo stebėjimo *p*-mačio kintamojo $X = \left(X^{(1)}, X^{(2)}, ..., X^{(p)}\right)$ reikšmės pažymimos $x_{ij} = \left(x_{ij}^{(1)}, x_{ij}^{(2)}, ..., x_{ij}^{(p)}\right)$. Naudojamų duomenų struktūra pateikta 1.1 lentelėje [22].

Diskriminantinės analizės tikslas – remiantis pradine stebėjimų klasifikacija (reikšmėmis x_{ij}, $j=1,2,...,n$, $i=1,2,...,k$) sudaryti taisyklę, kuri leistų klasifikuoti naujus stebėjimus. Klasifikavimo taisyklė sudaroma pagal duomenų statistinį modelį [9, 22].

1.1 lentelė

Diskriminantinės analizės duomenų matrica

Grupė	**Kintamieji**			
	$x^{(1)}$	$x^{(2)}$	...	$x^{(p)}$
1	$x_{11}^{(1)}, x_{12}^{(1)}, ..., x_{1n1}^{(1)}$	$x_{11}^{(2)}, x_{12}^{(2)}, ..., x_{1n1}^{(2)}$	...	$x_{11}^{(p)}, x_{12}^{(p)}, ..., x_{1n1}^{(p)}$
2	$x_{21}^{(1)}, x_{22}^{(1)}, ..., x_{2n2}^{(1)}$	$x_{21}^{(2)}, x_{22}^{(2)}, ..., x_{2n2}^{(2)}$	...	$x_{21}^{(p)}, x_{22}^{(p)}, ..., x_{2n2}^{(p)}$
...	...	...	...	...
k	$x_{k1}^{(1)}, x_{k2}^{(1)}, ..., x_{knk}^{(1)}$	$x_{k1}^{(2)}, x_{k2}^{(2)}, ..., x_{knk}^{(2)}$	...	$x_{k1}^{(p)}, x_{k2}^{(p)}, ..., x_{knk}^{(p)}$

Duomenys turi tenkinti diskriminantinės analizės prielaidas [9]:

- grupių skaičius turi būti baigtinis;
- grupės nepriklausomos ir neturi bendrų objektų (objektas negali priklausyti vienu metu kelioms grupėms);
- kiekvienas diskriminantinis kintamasis, kiekvienoje klasėje yra normaliai pasiskirstęs arba jo skirstinys yra artimas normaliajam;
- nė vienas diskriminavimo kintamasis negali būti kitų kintamųjų tiesinė daugdara;
- diskriminavimo kintamųjų kovariacijų matricos grupėse lygios;
- pageidautina, kad grupių didumai labai nesiskirtų.

Esant itin dideliems nuokrypiams nuo normalaus skirstinio, geriausia alternatyva diskriminantinei analizei yra logistinė regresija, dispersinė analizė arba kai kurie neparametriniai metodai.

Diskriminantinės analizės privalumai: iš anksto yra žinomos grupės. Trūkumai: ši analizė reikalauja diskriminavimo kintamųjų normalumo, o tai susiaurina šios regresijos naudojimą (dažnai realūs duomenys šios sąlygos netenkina). Taip pat ši analizė nenaudojama, kai grupių didumai ženkliai skiriasi (širdies ir kraujagyslių ligų prognozavimo atveju būtent taip ir yra, t.y. ligonių grupė yra ženkliai mažesnė nei sveikų asmenų).

Kai netenkinamos tiesinės regresijos prielaidos, galima taikyti **kvantilinę regresiją.** Medianos regresijos idėja (medianos regresija yra atskiras atvejis kvantilių regresijos) – modeliuoti ne priklausomojo kintamojo vidutinę reikšmę parinktų regresorių atžvilgiu, o medianos reikšmę [11].

Tiesinėje regresijoje prognozuojama vidutinė priklausomojo kintamojo reikšmė. Tarkime, kad tiesinės regresijos modelį naudojame gydymo trukmės prognozei. Įstačius konkretaus paciento tyrimų rezultatus į modelį prognozuojame kiek vidutiniškai laiko gyja ligoniai. Medianos regresijos interpretacija: pusė visų ligonių gydėsi iki trijų savaičių [11].

Ši regresija taikoma, kai duomenys turi išskirčių (medianos regresija nejautri išskirtims, kai jos nedidelės); netenkinama normalumo prielaida; heteroskedastiškumas didelės įtakos neturi [11], pseudo R^2 - vienas iš statistinių rodiklių medianos regresijoje [11].

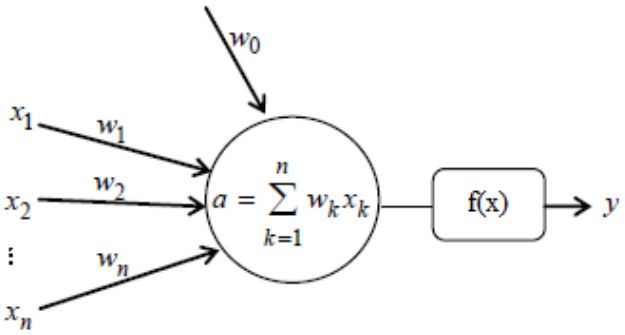
Privalumas: medianos regresiją galima taikyti, kai netenkinamos tiesinės regresijos prielaidos. Trūkumai: mažėja prognozavimo tikslumas ir patikimumas.

Dėl savo lankstumo, galimybės greitai apdoroti didelius kiekius informacijos plačiai naudojami **neuroniniai tinklai**. Neuroniniai tiklai gali nustatyti paslėptas priklausomybes bei pagerinti informacijos apdorojimo savybes remiantis mokymu. Jie toleruoja klaidas ir gali interpretuoti informacijos netikslums bei prisitaikyti prie naujų situacijų [24].

Neuroniniai tinklai sukurti remiantis panašaus veikimo principu kaip žmogaus smegenys. Abiejų veikimas grindžiamas neuronų tarpusavio sąveika [24]. Neuronas yra sudarytas iš sumatoriaus ir aktyvavimo funkcijos. Aktyvacija tiesiogiai proporcinga įėjimo signalams (a – neurono aktyvacijos lygis):

$$a = a_0 + \sum_{k=1}^{n} w_k x_k \ , \qquad \textbf{(1.9)}$$

čia n – neuronų įėjimų skaičius, x_k – i-asis signalas, w_k – i-osios jungties svorinis koeficientas, a_0 – neurono aktyvacija ramybės būsenoje [24].



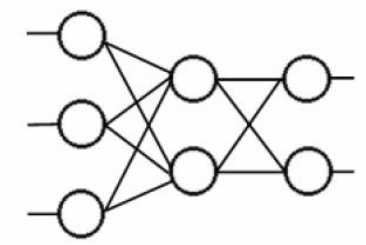
1.1 pav. Dirbtinis neuronas

Galimos neurono aktyvacijos funkcijos [23, 24]:

- sigmoidinė funkcija $f(x) = \dfrac{1}{1+e^{-x}}$;

- hiperbolinis tangentas $f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ ir kt.

Neuronai yra jungiami į neuroninius tinklus. Toks neuroninis tinklas gali būti vaizduojamas kaip grafas, kurio viršūnės yra neuronai, o lankai (su svoriais) jungia neuronų išėjimus su kitų neuronų įėjimais. Iš pradžių neuronai grupuojami į sluoksnius, tada kiekvieno sluoksnio neuronas sujungiamas su gretimų sluoksnių neuronais. Pirmas sluoksnis – įvesties sluoksnis, paskutinis – išvesties sluoksnis, o likę – paslėptieji sluoksniai [23, 24].



1.2 pav. Neuroninis tinklas

Neuroninis tinklas mokomas keičiant jungčių tarp neuronų svorius. Reikia turėti apmokymo aibę su įvedamo ir rezultatų duomenimis[24].

Kvadratinė paklaida naudojama neuroninio tinklo kokybei vertinti [23, 24]:

$$D = \frac{1}{2}\sum_{k}^{N}\sum_{p}^{m}\left(y_{kp} - f_{kp}\right)^2 \qquad \textbf{(1.10)}$$

čia f_{kp} - faktinis išėjimo signalas *p*-ajame elemente, y_{kp} - stebėtas signalas tame elemente. Minimizuoti paklaidai dažniausiai taikomi pirmos ir antros eilės gradientiniai metodai.

Bendru atveju mokymo algoritmas:

- neuronų svoriams priskiriamos atsitiktinės reikšmės;
- įvedamas mokymo imties p-asis įeities ir išvesties vektorius;
- apskaičiuojamas neuroninių tinklu išėjimo vektorius;
- apskaičiuojama išėjimo signalo paklaida;
- koreguojami ivesties sluoksnio svoriai:

$$w'_{ip} = w_{ip} - \mu q_p \varphi_{ip} - \alpha(w_{ip} - w^*_{ip}), \qquad \textbf{(1.11)}$$

čia $\mu > 0$, $0 < \alpha < 1$, q_p - paklaida *p*-ojo neurono išėjime, φ_{ip} - netiesinės neurono išėjimo funkcijos išvestinė pagal i-ąjį svorį, w_{ip} - nekoreguoti svoriai, w'_{ip} - koreguoti svoriai, w^*_{ip} - svoriai prieš tai buvusioje iteracijoje.

- Paklaidos sklidimo procesas iki įėjimo sluoksnio, kai kuoreguojami šių sluoksnių svoriai pagal analogišką (1.11) formulę.

- Algoritmas stabdomas, kai mokymui suvedama visa imtis, o paklaida tampa mažesnė už norimą reikšmę [24].

Neuroninių tinklų privalumai yra tie, kad juos galima naudoti, kai pažeistos klasikinių metodų prielaidos; galimybė pasiekti norimą tikslumą. Trūkumai: gauti rezultatai neišreiškiami analizine išraiška; gana didelis kompiuterinio skaičiavimo ir atminties poreikis; blogas pradinių reikšmių parinkimas gali nulemti ilgus skaičiavimus.

1.3. PROGRAMINĖS ĮRANGOS PASIRINKIMAS

Daugelis statistikos uždavinių yra pakankamai sudėtingi ir turint didelius duomenų kiekius sprendimas gali užtrukti gana ilgai. Pateikiama keletas programinės įrangos paketų, kurie pagreitina skaičiavimus, leidžia interpretuoti gautus rezultatus grafiškai.

Duomenų analizės programa SAS. Sistema SAS (angl. *Statistical Analysis System*) – vystoma nuo 1976 metų. Apima visus reikalingus duomenų analizės etapus (duomenų įvedimą, tvarkymą, saugojimą, analizę). SAS gali veikti įvairiose operacinėse sistemose. Pagrindinis pranašumas – SAS realizuota daug statistinių algoritmų. Taip pat galima nesunkiai susikurti vartotojui reikiamus algoritmus [10].

IBM SPSS Statistics programinė įranga. IBM SPSS Statistics - specializuota statistinė programinė įranga, leidžianti vartotojams: įkelti duomenis iš įvairių šaltinių, pertvarkyti juos, išanalizuoti duomenis statistiniais metodais, lengvai ir supranamai pateikti gautus rezultatus grafiškai bei analitinėmis lentelėmis, eksportuoti rezultatus [25].

Statistica. Programos Statistica kūrėja yra amerikiečių kompanija StatSoft – viena didžiausių statistikos ir analizės programinės įrangos gamintojų pasaulyje. Ši programa, kaip ir SAS ar SPSS apima visus reikalingus duomenų analizės etapus.

Atlikti statistinei analizei pasirinkta duomenų analizės sistema SAS, nes yra visi reikalingi metodai (procedūros); padeda taupyti darbo sąnaudas (automatizuotos procedūros, kurias kitose statistinės analizės programose reikėtų programuoti) ir užtikrina greitus skaičiavimus. Sukurtų modeliu realizavimui pasirinktos šios pagrindinės procedūros: LOGISTIC, GLM, GLMSELECT, UNIVARIATE, QUANTREG ir GPLOT.

PROC LOGISTIC procedūros MODEL dalyje aprašome logistinės regresijos modelis.

Jei norime taikyti logistinę regresiją retiems įvykiams, reikia naudoti *Firth's Penalized Likelihood* metodą.

PROC GLM analizuoja duomenis, taikydama apibendrintą tiesinės regresijos modelį. PROC GLM naudoja modelius, susijusius su vieno ar keleto tolydžiųjų kintamųjų, su vienu ar keletu nepriklausomų kintamųjų. Nepriklausomi kintamieji gali būti arba kokybiniai arba kiekybiniai, kurių

stebėjimus suskirstome į atskiras grupes, arba tolydūs kintamieji. PROC GLMSELECT procedūra nuo GLM skririasi tik tuo, kad nepriklausomi kintamieji į lygtį įtraukiami panaudojant pažingsninius algoritmus, paliekami tik statistiškai reikšmingi kintamieji.

Taigi, GLM procedūra gali būti naudojama taikant:

- daugialypę regresiją;
- dispersinę analizę (ANOVA);
- kovariacinę analizę;
- daugiamatę dispersinę analizę (MANOVA) ir kt [8].

Procedūros sintaksė analogiška logistinei regresinei analizei.

Procedūra PROC UNIVARIATE naudojama kai norime atlikti detalią kitamojo analizę. Galima atlikti grafinę analizę.

PROC QUANTREG procedūra analizę galima atlikti nurodant skirtingus kvantilius. Procedūra GPLOT naudojama grafiniam rezultatų realizavimui. Šia procedūra, nubraižomas jautrumo ir specifiškumo grafikas (ROC kreivė) [8].

Sukurti skaičiuoklei virtualioje erdvėje buvo pasirinktas teksto redaktorius Notepad++ (dėl skaičiuoklės techninio paprastumo) ir naršyklė Mozilla Firefox (dėl lengvo naudojimosi); vartotojo sąsajos formavimo palengvinimui buvo naudojami šie įrankiai: *bootstrap*, *jquery* ir *knockout.*

1.4. DARBE SPRENDŽIAMI UŽDAVINIAI

Darbe analizuojami Lietuvos sveikatos mokslų universiteto Kardiologijos instituto mokslininkų surinkti duomenys.

Pagrindinis šio darbo tikslas:

- atlikti metodų ir modelių, taikomų širdies ir kraujagyslių ligų susirgimų ir mirčių nuo jų prognozavimui Lietuvoje bei užsienyje, analizę;
- pagal Lietuvos sveikatos mokslų universiteto Kardiologijos instituto mokslininkų 1982-2013 metais atliktų tyrimų metu surinktus Kauno miesto gyventojų duomenis, sukurti modelius širdies ir kraujagyslių ligų susirgimų rizikos vertinimui ir mirčių dažnių prognozavimui;
- sukurtus modelius realizuoti su statistinės analizės sistema SAS bei sukurti rizikos susirgti širdies ir kraujagyslių ligomis skaičiuoklę internete.

2. TIRIAMOJI DALIS IR REZULTATAI

Šioje dalyje pateikiami sudaryti modeliai širdies ir kraujagyslių ligų rizikai vertinti, taip pat Kauno miesto 45-64m. amžiaus gyventojų mirtingumui prognozuoti.

2.1. KAUNO MIESTO GYVENTOJŲ ŠIRDIES IR KRAUJAGYSLIŲ LIGŲ LOGISTINĖS REGRESIJOS PROGNOZAVIMO MODELIAI

Šiame darbe analizuojami Lietuvos sveikatos mokslų universiteto Kardiologijos instituto mokslininkų 2006-20013 m. sukaupti tyrimo duomenys apie Kauno miesto gyventojų 45-72 m. amžiaus širdies ir kraujagyslių ligas.

Lietuvos sveikatos mokslų universiteto Kardiologijos instituto mokslininkai 2012 metais atliko tyrimą „Cardiovascular risk actors and cognitive function in middle aged and elderly Lithuanian urban population: results from the HAPIEE study“ [12]. Duomenys paimti iš tyrimo, atlikto tarptautinės HAPIEE (angl. *Health, Alcohol and Psychosocial Factors in Eastern Europe*) studijos. Atrinkti 7 087 respondentai, kurie dalyvavo šiame sveikatos tyrime 2006-2008 metais. Respondentų duomenys buvo renkami naudojant standartinio klausimyno (amžius, subjektyvus sveikatos vertinimas, išsilavinimas, rūkymas, alkoholio vartojimas, fizinis aktyvumas, svoris, ūgis, šeimyninė padėtis, šeimos narių susirgimai tirtomis ligomis ir t.t.) atsakymus [12].

Analizuojant rizikos veiksnius naudota kelių kintamųjų logistinė regresinė analizė ir SPSS 13.4 programinė įranga. Gauti statistiškai reikšmingi kintamieji: rūkymas (vyrai/moterys), išeminė širdies liga (vyrai), gyvenimo kokybės vertinimas (vyrai/moterys), subjektyvus sveikatos vertinimas (vyrai/moterys), depresiniai simptomai (vyrai/moterys) [12].

Tyrimo rezultatai parodė, kad ryšys tarp širdies ir kraujagyslių ligų rizikos veiksnių bei mažesnės kognityvinės funkcijos (*Kognityvinė funkcija* – asmens gebėjimas gauti, suvokti, išanalizuoti, išlaikyti, atgaminti tam tikrą informaciją; orientacija laiko, vietos atžvilgiu; kalba, tarimas; trumpalaikė atmintis) tarp sveikų vidutinio ir vyresnio amžiaus suaugusiųjų labai priklauso nuo lyties [12].

Papildžius 2006-2008m. atlikto tyrimo duomenis, buvo tirti 7 115 žmonių, iš kurių 158 buvo diagnozuotas galvos smegenų insultas (76 moterims ir 82 vyrams), o išeminė širdies liga ir miokardo infarktas – 343 asmenims (152 moterims ir 191 vyrui).

Tyrimui buvo naudojami LSMU medicinos ekspertų nurodyti rizikos veiksniai (žr. 2.1 lentelė).

2.1 lentelė

Medicinos ekspertų nurodyti širdies ir kraujagyslių ligų rizikos veiksniai

Nr.	Kintamasis	Kintamojo reikšmės	Matavimo vienetai
1.	amžius	Tolydus kintamasis	metai
2.	ūgis	Tolydus kintamasis	centimetras
3.	svoris	Tolydus kintamasis	kilogramas
4.	lytis	0 – moteris, 1 – vyras	-
5.	diabetas	0 – neserga, 1 – serga	-
6.	kūno masės indeksas	1 – iki 25, 2 – 25,01-29,9, 3 – daugiau už 30; tolydus kintamasis	kg/m^2
7.	alkoholio vartojimas	1 – negeria arba metė, 2 – ne daugiau kaip kartą per savaitę, 3 – daugiau kaip kartą per savaitę	-
8.	trigliceridų padidėjimas	0 – 0,299-2,3, 1 – 2,301-18; tolydus	mmol/l
9.	didelio tankio lipoproteinų cholesterolis	0 – 0,901-5, 1 – 0,24-0,9; tolydus	mmol/l
10.	mažo tankio lipoproteinų cholesterolis	0 – 0,1-3,3, 1 – 3,301-10; tolydus	mmol/l
11.	rūkymas	0 – nerūko arba metė, 1 – rūko	-
12.	fizinis aktyvumas	0 – fiziškai aktyvus, 1 – fiziškai neaktyvus	-
13.	arterinis kraujospūdis	0 – normalus, 1 – padidėjęs	-
14.	gliukozė	Tolydus kintamasis	mmol/l
15.	cholesterolis	0 – 0,99-5,2, 1 – 5,201-12,001; tolydus	mmol/l
16.	subjektyvus sveikatos vertinimas	1 – labai bloga, 2 – vidutinė, 3 – labai gera	-
17.	išeminė širdies liga prieš insultą arba miokardo infarktą	0 – nesirgo, 1 – sirgo	-
18.	šeimyninė padėtis	0 – nevedęs(-usi), našlys(-ė), išsiskyręs(-usi), 1 – vedęs(-usi)	-
19.	išsilavinimas	0 – neuniversitetinis, 1 – universitetinis	-
20.	sistolinis kraujospūdis	Tolydus kintamasis	mm Hg
21.	diastolinis kraujospūdis	Tolydus kintamasis	mm Hg
22.	juosmens apimtis	Tolydus kintamasis	centimetras
23.	šeimos narių susirgimai širdies ligomis	0 – nesirgo, 1 – sirgo	-
24.	šeimos narių susirgimai insultu	0 – nesirgo, 1 – sirgo	-
25.	depresiniai simptomai	0 – nėra, 1 – yra	-
26.	pažintinių gebėjimų sutrikimai	0 – nėra, 1 – yra	-

Kintamieji, kurie turi daugiau nei dvi kategorijas, buvo perkoduoti į dvireikšmius pseudokintamuosius.

Norint išsiaiškinti, kokie rizikos veiksniai daro įtaką konkrečiai ligai, buvo sudaryti logistiniai modeliai trims ligų grupėms:

1) galvos smegenų insultui;
2) išeminei širdies ligai ir miokardo infarktui;
3) *širdies ir kraujagyslių ligoms* (bendrai apjungus 1 ir 2 grupę).

Kiekvienu minėtu atveju, logistinės regresijos modeliai buvo sudaryti naudojant visus turimus duomenis, pašalinus iš duomenų išskirtis ir retiems įvykiams (kai buvo pašalintos išskirtys). Iš viso buvo sudaryta po 42 modelius kiekvienai ligai. Gauti modeliai buvo lyginami pagal šiuos kriterijus: Akaike informacinį kriterijų ir Švarco informacinį kriterijų, Pseudo R^2 bei plotą po ROC kreive.

Galvos smegenų insultas. Buvo sudaryti logistinės regresijos modeliai prognozuoti galvos smegenų insulto tikimybę pagal klinikinių, gyvensenos ir sociodemografinių nepriklausomų kintamųjų reikšmes. 2 priede pateikti visi gauti modeliai. Iš jų atrinkti du (paryškinti): geriausias gautas modelis (4) ir blogiausias gautas modelis (9). Detali informacija apie modelius pateikta 2.2 lentelėje. Taip pat pateikti ir modelio tikimumą apibūdinantys kriterijai (AIC, SC, Pseudo R^2, plotas po ROC kreive).

Iš 2.2 lentelės matome, jog susirgti galvos smegenų insultu reikšmingi šie rizikos veiksniai: *amžius, rūkymas, išeminė širdies liga prieš insultą (IŠL_p), didelio tankio lipoproteinų cholesterolis (D_DTL), gliukozė, sistolinis/diastolinis kraujospūdžiai (DK/SK).*

Atrinktas geriausias gautas modelis retiems įvykiams, kurio sudarymui panaudoti duomenys su pašalintomis išskirtimis. Šio modelio pseudo R^2 ir plotas po ROC kreive reikšmės skiriasi labai mažai, tačiau informacinių kriterijų reikšmės ženkliai mažesnės nei kitų modelių.

Pateiksime geriausią gautą logistinės regresijos modelį prognozuoti galvos smegenų insultui:

$$\ln\frac{\hat{P}(insultas=1)}{\hat{P}(insultas=0)} = -12{,}94 + 0{,}07 * Amžius + 0{,}78 * Rukymas + 0{,}63 * IŠL\_p + 0{,}22 * Gliukozė + \\ + 0{,}03 * DK.$$

Didesnės nepriklausomų kintamųjų *Amžius*, *Rukymas*, *IŠL_p*, *Gliukozė* ir *Dk* reikšmės didina tikimybę susirgti galvos smegenų insultu.

Kintamojo *Rukymas* galimybių santykis yra 2,2, t.y. rūkančiam asmeniui, lyginant su nerūkančiu, susirgti galvos smegenų insultu galimybė padidėja net 2,2 karto, o 95 % pasikliautinasis intervalas - $PI_{0,95}(\beta_1) = (1{,}564; 3{,}057)$; kintamojo *IŠL_p* galimybių santykis yra 1,9, t.y. asmeniui, sirgusiam išemine širdies liga prieš insultą, lyginant su nesirgusiu, susirgti galvos smegenų insultu galimybė padidėja 1,9 karto ($PI_{0,95}(\beta_2) = (1{,}321; 2{,}682)$).

2.2 lentelė

Insulto prognozavimo logistinės regresijos modeliai

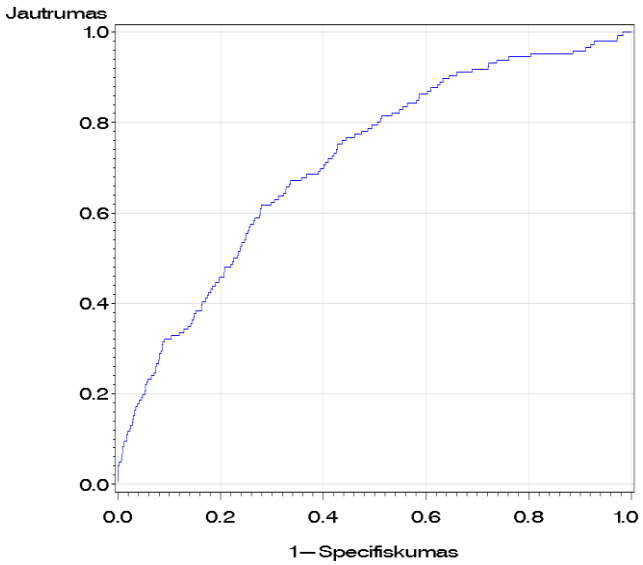
(Kiekvienam regresijos modeliui pateikiami regresijos koeficientų įverčiai, standartinės koeficientų įverčių paklaidos ir hipotezių apie koeficientų lygybę nuliui tikrinimo rezultatai)

Kintamasis	Logistinės regresijos modeliai					
	Geriausias gautas modelis			Blogiausias gautas modelis		
	Visi duomenys	Duomenys, pašalinus išskirtis		Visi duomenys	Duomenys, pašalinus išskirtis	
		Logistinis modelis	Retiems įvykiams		Logistinis modelis	Retiems įvykiams
Konstanta	-12,0254*** (1,1162)	-13,0218*** (1,1511)	**-12,9381*** (1,1317)**	-11,0351*** (0,9605)	-11,7774*** (0,9901)	-11,6996*** (0,9710)
Amžius	0,0729*** (0,0134)	0,0753*** (0,0135)	**0,0741*** (0,0133)**	0,0680*** (0,0133)	0,0686*** (0,0134)	0,0675*** (0,0131)
Rukymas	0,7422*** (0,1728)	0,7900*** (0,1736)	**0,7822*** (0,1710)**	0,7099*** (0,1722)	0,7317*** (0,1726)	0,7066*** (0,1702)
IŠL_p	0,5558*** (0,1837)	0,6428*** (0,1829)	**0,6324*** (0,1806)**	-	-	-
D_DTL	-2,1184** (1,0096)	-	-	-	-	-
Gliukozė	0,1349*** (0,0479)	0,2206*** (0,0558)	**0,2199*** (0,0550)**	0,1305*** (0,0483)	0,2280*** (0,0576)	0,2243*** (0,0564)
SK	-	-	-	0,0130*** (0,00364)	0,0140*** (0,00373)	0,0141*** (0,00368)
DK	0,0264*** (0,00656)	0,0296*** (0,00681)	**0,0298*** (0,00671)**	-	-	-
Modelio parametrai						
AIC	1312,42	1293,52	**1275,26**	1331,35	1317,33	1308,83
SC	1360,02	1334,20	**1315,99**	1365,35	1351,29	1342,96
Pseudo R^2	0,0146	0,0165	**0,0164**	0,0112	0,0130	0,0126
Plotas po ROC kreive	0,718	0,720	**0,718**	0,696	0,697	0,698

* – $p < 0{,}1$; ** – $p < 0{,}05$; *** – $p < 0{,}01$;

Modelio jautrumo ir specifiškumo grafikas (ROC kreivė) pateiktas 2.1 paveiksle.

Iš 2 priede pateiktos lentelės taip pat matome, jog dažniausiai geresni rodikliai buvo modelių, naudojant logistinę regresiją, kai duomenims pašalintos išskirtys.



2.1 pav. Galvos smegenų insulto logistinės regresijos modelio ROC kreivė (plotas po ROC kreive 0,718)

Išeminė širdies liga ir miokardo infarktas (IŠL_MI). Buvo sudaryti logistinės regresijos modeliai, skirti prognozuoti išeminės širdies ligos (IŠL) ir miokardo infarkto (MI) tikimybę pagal klinikinių, gyvensenos ir sociodemografinių nepriklausomų kintamųjų reikšmes (modeliai sudaryti tokiu pat principu, kaip ir galvos smegenų insulto prognozei). 3 priede pateikti visi gauti modeliai. Iš jų atrinkti du (paryškinti): geriausias gautas modelis (8) ir blogiausias gautas modelis (10), kurie pateikti 2.3 lentelėje. Taip pat pateikti ir modelio tikimą apibūdinatys kriterijai.

Gauta , kad susirgti išemine širdies liga ar miokardo infarktu reikšmingiausi šie rizikos veiksniai (2.3 lentelė): *amžius, lytis, išeminė širdies liga prieš miokardo infarktą ar pasikartojusį IŠL (IŠL_p), kūno masės indeksas (KMI), subjektyvus sveikatos vertinimas (SV), didelio tankio lipoproteinų cholesterolis (DTL), arterinė hipertenzija (AH), cholesterolis (Chol), mažo tankio lipoproteinų cholesterolis (MTL, dvireikšmis D_MTL),diabetas, sistolinis kraujospūdis (SK)*.

Atrinktas geriausias gautas modelis, kuriam sudaryti buvo pašalintos išskirtys. Šio modelio pseudo R^2 ir plotas po ROC kreive reikšmės skiriasi labai nedaug, tačiau informacinių kriterijų reikšmės ženkliai mažesnės. Pateiksime geriausią gautą logistinės regresijos modelį prognozuoti IŠL ir miokardo infarktui:

$$\ln\frac{\hat{P}(I\check{S}L\_MI=1)}{\hat{P}(I\check{S}L\_MI=0)} = -6{,}60 + 0{,}03 * Amžius + 0{,}65 * Lytis + 0{,}03 * KMI + 0{,}42 * I\check{S}L\_p + 0{,}39 * SV +$$
$$+ 0{,}2 * MTL + 0{,}66 * Diabetas.$$

2.3 lentelė

Išeminės širdies ligos ir miokardo infarkto prognozavimo logistinės regresijos modeliai (Kiekvienam regresijos modeliai pateikiami regresijos koeficientų įverčiai, standartinės koeficientų įverčių paklaidos ir hipotezių apie koeficientų lygybę nuliui tikrinimo rezultatai)

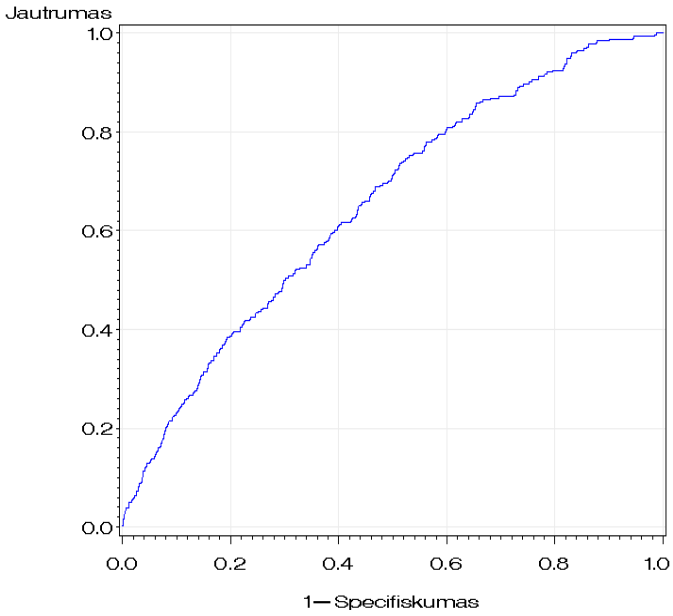
Kintamasis	Logistinės regresijos modeliai					
	Geriausias gautas modelis			Blogiausias gautas modelis		
	Visi duomenys	Duomenys, pašalinus išskirtis		Visi duomenys	Duomenys, pašalinus išskirtis	
		Logistinis modelis	Retiems įvykiams		Logistinis modelis	Retiems įvykiams
Konstanta	-5,1959*** (0,6707)	**-6,6028*** (0,6638)**	-6,7858*** (0,5561)	-5,8521*** (0,5951)	-6,1318*** (0,6042)	-6,6797*** (0,5066)
Amžius	0,0234*** (0,00849)	**0,0258*** (0,00847)**	0,0363*** (0,00708)	0,0262*** (0,00846)	0,0263*** (0,00849)	0,0355*** (0,00711)
Lytis	0,4562*** (0,1238)	**0,6511*** (0,1228)**	0,6048*** (0,1015)	0,4736*** (0,1211)	0,5232*** (0,1223)	0,4494*** (0,1009)
KMI	-	**0,0282** (0,0113)**	0,0302*** (0,00941)	-	-	0,3507*** (0,0874)
IŠL_p	0,3134** (0,1350)	**0,4245*** (0,1343)**	0,5377*** (0,1093)	-	-	-
SV	0,3121*** (0,3121)	**0,3920*** (0,1027)**	0.2744*** (0,0885)	0,3671*** (0,1013)	0,4496*** (0,1013)	0,3507*** (0,0874)
DTL	-0,4112*** (0,1686)	-	-	-	-	-
AH	0,3152** (0,1468)	-	-	-	-	-
Chol	0,1344*** (0,0509)	-	-	-	-	-
MTL	-	**0,1952*** (0,0578)**	0,1583*** (0,0481)	-	-	-
D_MTL	-	-	-	0,3190** (0,1308)	0,3475*** (0,1315)	0,3234*** (0,1085)
SK	-	-	-	0,00525** (0,00265)	0,00699** (0,00273)	0,00989*** (0,00223)
Diabetas	0,3672** (0,1840)	**0,6606*** (0,1887)**	0,5712*** (0,1601)	0,4919*** (0,1809)	0,7671*** (0,1813)	0,6959*** (0,1540)
Modelio tinkamumo parametrai						
AIC	2461,77	**2433,46**	3210,35	2472,97	2448.27	3229,73
SC	2529,78	**2494,53**	3271,64	2527,38	2502.59	3284,25
Pseudo R^2	0,0122	**0,0149**	0,0226	0,0099	0.0126	0,0196
Plotas po ROC kreive	0,651	**0,656**	0,665	0,634	0.643	0,652

* – $p < 0{,}1$; ** – $p < 0{,}05$; *** – $p < 0{,}01$

Didesnės nepriklausomų kintamųjų *Amžius, Lytis, KMI, IŠL_p, SV, MTL* ir *Diabetas* reikšmės didina tikimybę susirgti išemine širdies liga ar miokardo infarktu.

Apskaičiuoti šio logistinio modelio galimybių santykio įverčiai ir 95% pasikliautinieji intervalai: kintamojo *Lytis* galimybių santykis yra 1,9, t.y. vyrams, lyginant su moterimis, susirgti IŠL ir MI galimybė padidėja 1,9 karto ($PI_{0,95}(\beta_1) = (1{,}507; 2{,}440)$); kintamojo *IŠL_p* galimybių santykis yra 1,5, t.y. asmeniui, sirgusiam išemine širdies liga prieš šias ligas, lyginant su nesirgusiu, susirgti galimybė padidėja 1,5 karto ($PI_{0,95}(\beta_2) = (1{,}175; 1{,}989)$); o kintamojo *Diabetas* galimybių santykis yra 1,9, t.y. sergantiems diabetu, susirgti IŠL ar MI, galimybė padidėja 1,9 karto nei nesergantiems diabetu ($PI_{0,95}(\beta_3) = (1{,}337; 2{,}802)$).

Modelio jautrumo ir specifiškumo grafikas (ROC kreivė) pateiktas 2.2 paveiksle.



2.2 pav. Išeminės širdies ligos ir miokardo infarkto logistinės regresijos modelio ROC kreivė (plotas po ROC kreive 0,656)

Iš 3 priede pateiktos lentelės taip pat matome, jog dažniausiai geresni rodikliai buvo modelių, naudojant logistinę regresiją, kai duomenims pašalintos išskirtys.

Širdies ir kraujagyslių ligos (ŠKL). Sudarysime logistinės regresijos modelius, skirtus prognozuoti širdies ir kraujagyslių ligų tikimybę (modeliai sudaryti tokiu pat principu, kaip ir pirmuose dviejuose atvejuose). 4 priede pateikti visi gauti modeliai. Iš jų atrinkti du (paryškinti): geriausias gautas modelis (11) ir blogiausias gautas modelis (12), kurie pateikti 2.4 lentelėje. Taip pat pateikti ir modelio tikimą apibūdinatys kriterijai.

Gauta , kad susirgti širdies ir kraujagyslių ligomis reikšmingiausi šie rizikos veiksniai (2.4 lentelė): *amžius, lytis, išeminė širdies liga prieš ŠKL (IŠL_p), rūkymas, kūno masės indeksas (KMI), subjektyvus sveikatos vertinimas (SV), mažo tankio lipoproteinų cholesterolis (D_MTL), diabetas,*

*sistolinis/diastolinis kraujospūdžiai (SK/DK), kūno masės indekso ir sistolinio kraujospūdžio sąveika (KMI*SK), amžiaus ir lyties sąveika, juosmens apimtis (JA).*

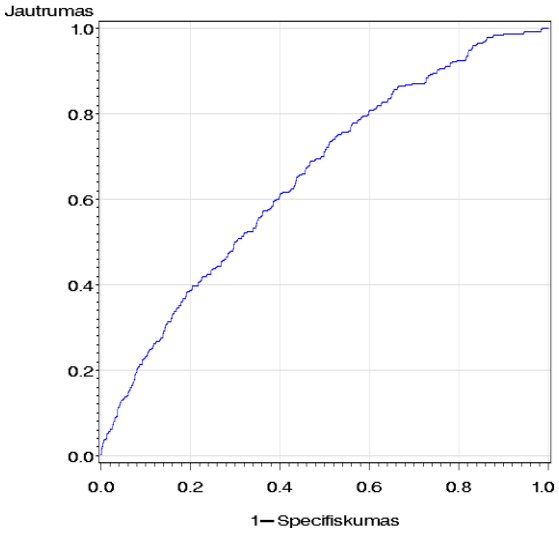
Atrinktas geriausias gautas modelis, kuriam sudaryti buvo pašalintos iš duomenų išskirtys. Pateiksime geriausią gautą logistinės regresijos modelį prognozuoti širdies ir kraujagyslių ligas:

$$\ln\frac{\hat{P}(\check{S}KL=1)}{\hat{P}(\check{S}KL=0)} = -8{,}92 + 0{,}06 * Amžius + 0{,}45 * Rukymas + 2{,}39 * Lytis + 0{,}03 * KMI + 0{,}49 * IŠL\_p +$$
$$+ 0{,}27 * SV + 0{,}29 * D\_MTL + 0{,}01 * SK + 0{,}64 * Diabetas - 0{,}03 * Amžius * Lytis.$$

Didesnės nepriklausomų kintamųjų *Amžius, Rukymas, Lytis, KMI, IŠL_p, SV, D_MTL, SK* ir *Diabetas* reikšmės didina tikimybę susirgti širdies ir kraujagyslių ligomis, o kintamųjų *Amžius* ir *Lytis* sąveika šią tikimybę mažina.

Kintamojo *Lytis* galimybių santykis yra 10,9, t.y. vyrams, lyginant su moterimis, susirgti ŠKL galimybė padidėja 10,9 karto ($PI_{0,95}(\beta_1) = (1{,}811; 65{,}906)$); kintamojo *IŠL_p* galimybių santykis yra 1,6, t.y. asmeniui, sirgusiam išemine širdies liga prieš ŠKL, lyginant su nesirgusiu, susirgti galimybė padidėja 1,6 karto ($PI_{0,95}(\beta_2) = (1{,}303; 2{,}028)$); o kintamojo *Diabetas* galimybių santykis yra 1,9, t.y. sergantiems diabetu, susirgti ŠKL, galimybė padidėja 1,9 karto nei nesergantiems diabetu ($PI_{0,95}(\beta_3) = (1{,}386; 2{,}614)$).

Modelio jautrumo ir specifiškumo grafikas pateiktas 2.3 paveiksle.



2.3 pav. Širdies ir kraujagyslių ligų logistinės regresijos modelio ROC kreivė (plotas po ROC kreive 0,680)

2.4 lentelė

Širdies ir kraujagylių ligų prognozavimo logistinės regresijos modeliai

(Kiekvienam regresijos modeliai pateikiami regresijos koeficientų įverčiai, standartinės koeficientų įverčių paklaidos ir hipotezių apie koeficientų lygybę nuliui tikrinimo rezultatai)

Kintamasis	Logistinės regresijos modeliai					
	Geriausias gautas modelis			Blogiausias gautas modelis		
	Visi duomenys	Duomenys, pašalinus išskirtis		Visi duomenys	Duomenys, pašalinus išskirtis	
		Logistinis modelis	Retiems įvykiams		Logistinis modelis	Retiems įvykiams
Konstanta	-12,2325*** (1,9329)	**-8,9160*** (0,7876)**	-8,9045*** (0,7726)	-8,3128*** (0,6423)	-8,2033*** (0,6525)	-8,1114*** (0,6228)
Amžius	0,0553*** (0,0112)	**0,0564*** (0,0113)**	0,0565*** (0,0111)	0,6423*** (0,00743)	0,0446*** (0,00747)	0,0456*** (0,00713)
Rukymas	0,3999*** (0,1188)	**0,4474*** (0,1206)**	0,4319*** (0,1183)	0,4883*** (0,1037)	0,4773*** (0,1045)	0,4415*** (0,1002)
Lytis	2,4162*** (0,9088)	**2,3911*** (0,9169)**	2,5181*** (0,8962)	-	-	-
IŠL_p	0,4271*** (0,1117)	**0,4858*** (0,1128)**	2,4814*** (0,8953)	-	-	-
KMI	0,1524** (0,0592)	**0,0309*** (0,0103)**	0,0298** (0,0101)	-	-	-
SV	0,2022** (0,0899)	**0,2723*** (0,0904)**	0,2567*** (0,0886)	0,2502*** (0,0902)	0,2713*** (0,0903)	0,2428*** (0,0862)
D_MTL	0,2597** (0,1097)	**0,2654** (0,1108)**	0,2949*** (0,1092)	0,2258** (0,1115)	0,2272** (0,1120)	0,2429** (0,1072)
SK	0,0338*** (0,0126)	**0,00873*** (0,00236)**	0,00881** (0,00232)	-	-	-
Diabetas	0,3258** (0,1585)	**0,6436*** (0,1618)**	0,6408*** (0,1594)	-	-	-
KMI*SK	-0,00089** (0,000406)	-	-	-	-	-
Amžius*Lytis	-0,0346** (0,0144)	**-0,0340** (0,0145)**	-0,0352** (0,0142)	-	-	-
DK	-	-	-	0,0151*** (0,00413)	0,0140*** (0,00424)	0,0131*** (0,00406)
JA	-	-	-	0,0119*** (0,00393)	0,0130*** (0,00402)	0,0124*** (0,00386)
Modelio parametrai						
AIC	3201,59	**3126,26**	3139,30	3067,83	3039,96	3235,93
SC	3290,01	**3207,58**	3220,92	3121,80	3093,91	3290,51
Pseudo R^2	0,0235	**0,0284**	0,0286	0,0193	0,0182	0,0171
Plotas po ROC kreive	0,669	**0,680**	0,680	0,653	0,651	0,646

* – $p < 0{,}1$; ** – $p < 0{,}05$; *** – $p < 0{,}01$

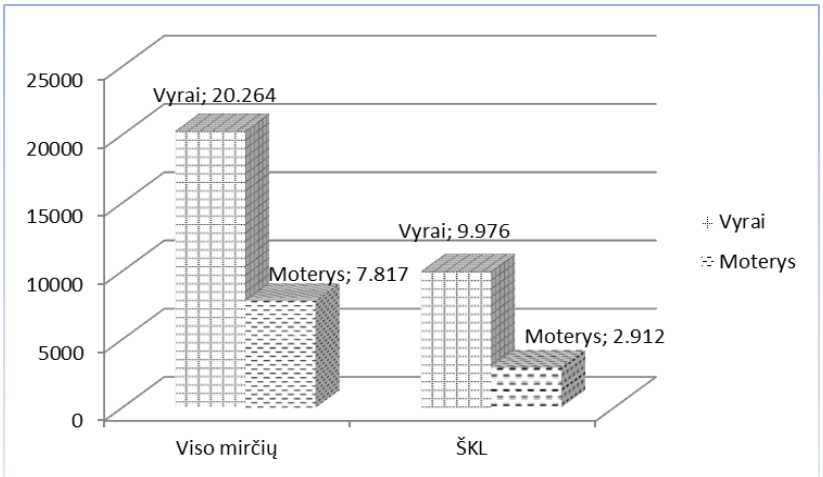
Iš 4 priede pateiktos lentelės taip pat matome, jog dažniausiai geresni rodikliai buvo modelių, naudojant logistinę regresiją, kai duomenims pašalintos išskirtys.

Taip pat buvo bandoma analizuoti duomenis neuroninių tinklų pagalba. Dėl mažo susirgimų skaičiaus ir didelio sveikų (širdies ir kraujagyslių ligų atžvilgiu) asmenų skaičiaus, buvo 100 % klasifikuojami stebėjimai, jog tai bus sveikas žmogus. Todėl neuroninių tinklų buvo atsisakyta, kaip ir atsisakyta diskriminantinės analizės (šioje analizėje reikalaujama, kad grupių didumai ženkliai nesiskirtų, o nagrinėjamu atveju, grupių didumai skiriasi ženkliai).

2.2. KAUNO MIESTO GYVENTOJŲ MIRTINGUMO DAŽNIO PROGNOZAVIMO MODELIAI

Sudaryti regresijos modelius Kauno miesto gyventojų mirtingumo prognozavimui buvo analizuojami Lietuvos sveikatos mokslų universiteto Kardiologijos instituto mokslininkų 1982-2013 m. surinkti Kauno miesto 45-64 m. amžiaus gyventojų mirtingumo nuo įvairių ligų duomenys.

Per analizuojamus 32 metus Kauno mieste buvo užregistruota 30 240 (2.4 pav.) mirčių 45-64 m. amžiaus gyventojų grupėje (iš jų – 20 264 vyrai ir 9 976 moterys). Iš jų dėl širdies ir kraujagyslių ligų buvo 10 729 mirtys (7 817 vyrų ir 2 912 moterų).



2.4 pav. Kauno miesto 45-64 m. amžiaus gyventojų mirčių skaičius per 32 metus

Tyrime buvo naudojami detalūs duomenys apie mirčių dažnius kiekvienais metais pagal mirties priežastį ir lytį. Mirtingumo nuo širdies ligų per metus (100 tūkst. gyventojų) prognozavimo modelio sudarymui, taikyti daugialypės tiesinės regresinės analizės modeliai bei SAS procedūra GLM ir

kvantilinė regresinė analizė bei SAS prosedūra QUANTREG. Kauno miesto gyventojų skaičius (100 tūkst. gyventojų) nuo 1984 m. iki 2013 m. pateiktas 2.5 lentelėje [7].

2.5 lentelė

Kauno miesto gyventojų skaičius 1984-2013m.

Metai	Gyventojų skaičius, 100 tūkst.	Metai	Gyventojų skaičius, 100 tūkst.	Metai	Gyventojų skaičius, 100 tūkst.
1984	3,998	**1994**	4,23978	**2004**	3,63952
1985	4,050	**1995**	4,16605	**2005**	3,56954
1986	4,104	**1996**	4,08706	**2006**	3,48506
1987	4,171	**1997**	4,02061	**2007**	3,43932
1988	4,229	**1998**	3,95555	**2008**	3,39535
1989	4,28495	**1999**	3,90623	**2009**	3,35393
1990	4,33938	**2000**	3,85620	**2010**	3,29542
1991	4,35393	**2001**	3,79706	**2011**	3,17319
1992	4,35233	**2002**	3,74176	**2012**	3,10773
1993	4,30388	**2003**	3,70242	**2013**	3,06888

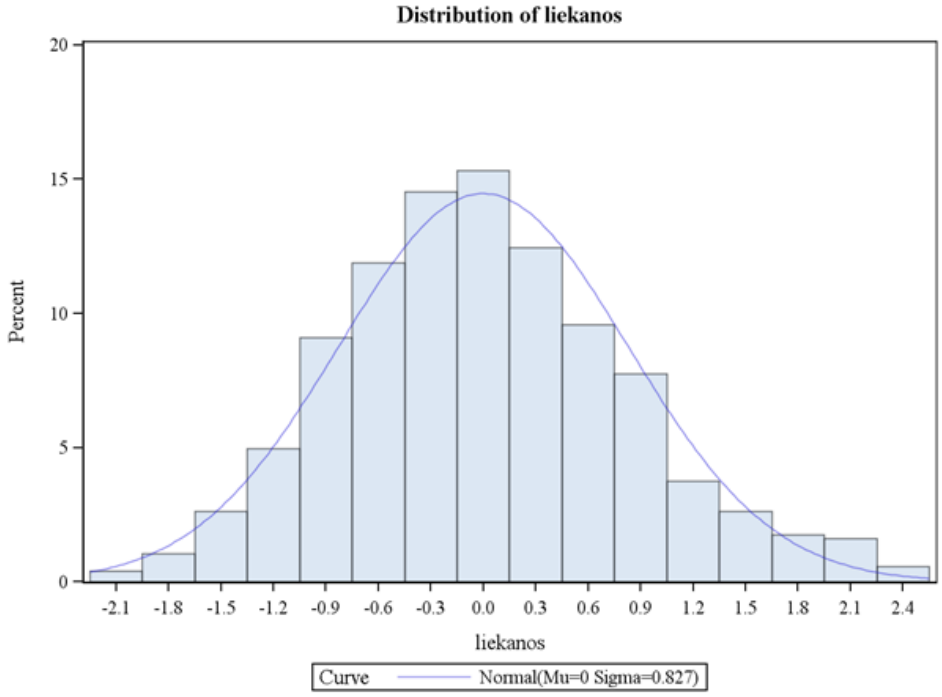
Atlikus daugialypę tiesinę regresinę analizę, gautas modelis, kurio lygtis

$$Mirtingumo\_dažnis = -6{,}78 + 1{,}93 * lytis + 0{,}15 * amžius.$$

Iš šios lygties matome, jog kintamieji *lytis* ir *amžius* didina mirtingumo tikimybę nuo širdies ir kraujagyslių ligų. Pavyzdžiui, vyrams (fiksavus amžių), vidutiniškai tenka 1,93 mirtys 100 tūkst. gyventojų. Modelio Malau statistika $C(p) = 2{,}39$, $Adj\,R^2 = 0{,}69$. Tai gana neblogi modelio tikimą duomenims parodantys rodikliai.

2.5 paveiksle pateikiamas modelio liekanų skirstinio grafikas.

Nors ir suderinamumo hipotezės testas atmetamas, bet vizualinė modelio liekanų santykinių dažnių histograma artima normaliojo skirstinio tankio funkcijai. Didelių imčių atveju, vietoj suderinamumo hipotezės, rekomenduojama patikrinti prielaidas nubraižant liekanų histogramą.



2.5 pav. Modelio liekanų santykinių dažnių histograma ir normaliojo skirstinio tankio funkcija

Tikriname nulinę hipotezę, kad duomenys yra homoskedastiški, kadangi White‘s kriterijus (158,5) ir Breusch-Pagan kriterijus (150,0) atmeta nulinę hipotezę ($p<0,0001$), t.y. duomenys yra heteroskedastiški. Tai reiškia, kad regresijos lygties parametrų įverčių standartinės paklaidos yra netikslios.

Kadangi netenkinamos daugialypės tiesinės regresijos prielaidos, buvo pasirinkta kvantilinė regresija. Kvantilinės regresijos modelis sudarytas taikant 0,5 kvantilį, t.y. medianą. Gautas modelis:

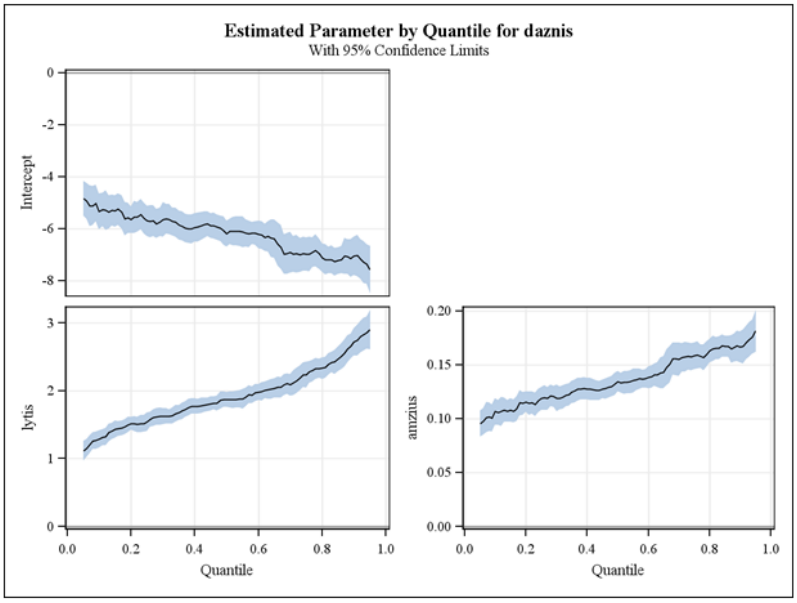
$$Mirtingumo\_dažnis = -6{,}19 + 1{,}87 * lytis + 0{,}13 * amžius$$

Iš šios lygties matome, jog kintamieji *lytis* ir *amžius* didina mirtingumo dažnį nuo širdies ir kraujagyslių ligų. Koeficiento prie kintamojo *lytis* 95 % pasikliautinasis intervalas $PI_{0,95}(\beta_1) = (1{,}761; 1{,}984)$, o koeficiento prie kintamojo *amžius* – $PI_{0,95}(\beta_2) = (0{,}125; 0{,}144)$. Modelio *ADJ* $R^2 = 0{,}64$. Mirties metai nebuvo statistiškai reikšmingi.

Iš 2.6 paveikslo matome, jog vyrų mirtingumas didesnis negu moterų; asmens amžius taip pat didina mirtingumą.

Pritaikius sudarytus daugialypės regresijos ir kvantilinės regresijos modelius gavome, jog lytis (vyrai) ir didesnis amžius statistiškai reikšmingai didina tikimybę mirti nuo širdies ir kraujagyslių ligų.

Gautas kvantilinės regresijos modelis, kuris gali būti panaudoti prognozuojant 45-64 m. amžiaus Kauno miesto gyventojų mirtingumą nuo širdies ligų pagal lytį ir amžių.



2.6 pav. Kvantilinės regresijos parametrų įverčių ir pasikliautinųjų intervalų priklausomybė nuo kvantilių

2.3. PROGRAMINĖ REALIZACIJA

Panaudojus SAS programavimo kalbą sukurtos duomenų analizės programinės priemonės skirtos:

- logistinės, daugialypės, kvantilinės regresijų statistinių modelių taikymui;
- modelių tikimo analizuojamiems duomenims vertinimui;
- gautų rezultatų grafiniam vaizdavimui;
- atliktos analizės ataskaitos rengimo automatizavimui ir rašymui *rtf* formato failuose.

Programinės priemonės, pateiktos 5 priedo 1 lentelės pirmoje dalyje, darbo eiga:

- PROC IMPORT procedūra importuojami duomenys;
- atliekamas kintamųjų perkodavimas ir sujungimas;
- PROC LOGISTIC procedūra atliekama logistinė regresinė analizė, apskaičiuojamos liekanos, modelio informaciniai kriterijai ir kt.;
- standartizuojamos Pearsono ir deviacijos liekanos;
- nustatomos ir pašalinamos išskirtys;

- PROC LOGISTIC procedūra atliekama logistinė regresinė analizė, apskaičiuojami modelio informaciniai kriterijai ir kt.;
- atliktos analizės rezultatai įrašomi *rtf* formato faile.

Šia programine priemone buvo analizuojami galvos smegenų insulto, išeminės širdies ligos ir miokardo infarkto, bei širdies ir kraujagyslių ligų (bendrai apjungus pirmus du susirgimus) duomenys, ir gauti modeliai, įvertinantys tikimybę susirgti šiomis ligomis.

Programinės priemonės, pateiktos 5 priedo 1 lentelės antroje dalyje, darbo eiga:

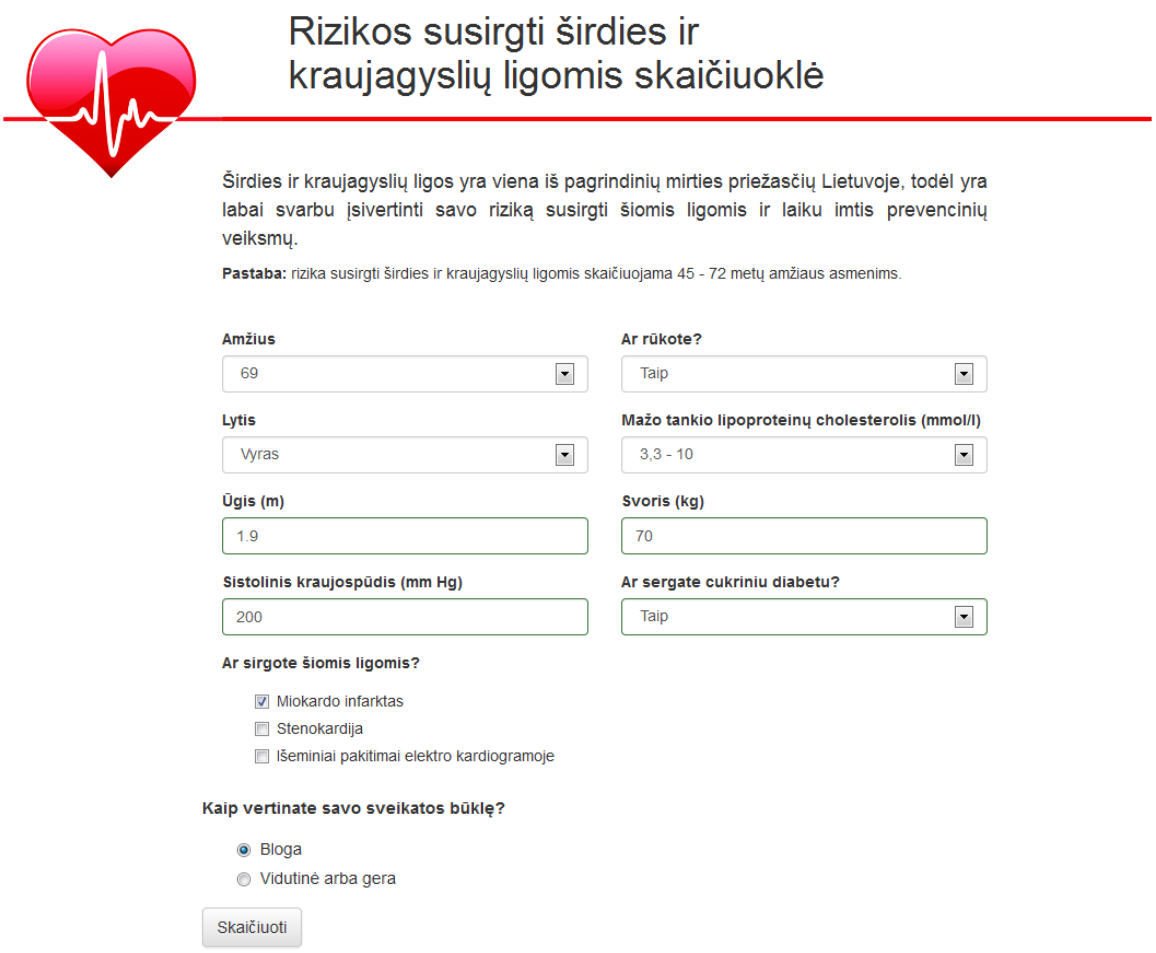
- PROC IMPORT procedūra importuojami faile esantys duomenys;
- PROC GLM procedūra atliekama daugialypė tiesinė regresinė analizė, apskaičiuojamos liekanos, modelio informaciniai kriterijai ir kt.;
- nustatomos ir pašalinamos išskirtys;
- PROC GLMSELECT procedūra atliekama pažingsninė daugialypė tiesinė regresinė analizė, apskaičiuojami modelio informaciniai kriterijai ir pan.;
- PROC UNIVARIATE tiriamas regresijos modelio liekanų normalumas;
- tiriamas duomenų heteroskedastiškumas;
- atliktos analizės rezultatai įrašomi *rtf* formato faile.

Šia programine priemone buvo analizuojami Kauno miesto 45-64 m. amžiaus gyventojų mirtingumo duomenys ir gautas modelis, kurio pagalba galima prognozuoti gyventojų mirtingumą (100 tūkst. gyventojų).

Šiame darbe atlikta susirgimų nuo širdies ir kraujagyslių ligų duomenų regresinė analizė. Taip pat atlikta ir mirtingumo nuo šių ligų daugialypė tiesinė regresinė analizė bei kvantilinė regresinė analizė. Gauti modeliai bei sukurtos programinės priemonės gali būti nesunkiai pritaikytos kitų ligų analizei.

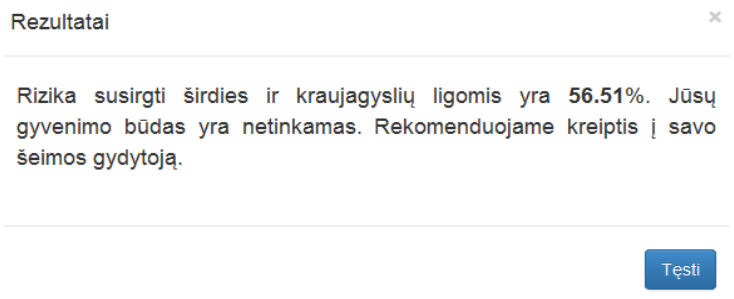
2.7 paveiksle pateikiamas rizikos susirgti širdies ir kraujagyslių ligomis skaičiuoklės pagrindinis langas virtualioje erdvėje (http://donluc.stud.if.ktu.lt/sirdies-ligos/). Skaičiuoklėje realizuotas šiame darbe gautas ŠKL prognozavimo modelis. Suvedus asmeninius rizikos veiksnius: amžių (45-72m.), lytį (vyras/moteris), rūkymą (rūko/nerūko), mažo tankio lipoproteinų cholesterolį (0,1-3,3 mmol/l arba 3,31-10 mmol/l), ūgį (metrais; apribojimai: 1,3-2,2 m), svorį (kilogramais; apribojimai: 30-200 kg), sistolinį kraujospūdį (mm Hg; apribojimai: 80-250 mm Hg), sergamumą diabetu (taip/ne), sirgimą miokardo infarktu, stenokardija ar esant išeminiams pakitimams elektro kardiogramoje, bei subjektyvų

sveikatos vertinimą (bloga/vidutinė arba gera), galima įsivertinti riziką susirgti širdies ir kraujagyslių ligomis.



2.7 pav. Rizikos susirgti širdies ir kraujagyslių ligomis skaičiuoklė

Rezultatas pateikiamas procentais. Iki 50% prognozuojama, jog asmuo nesirgs ŠKL, o nuo 50 % prognozuojama, jog asmuo gali susirgti ŠKL. Taip pat pateikiamos rekomendacijos, jog vertėtų kreiptis pas šeimos gydytoją, kai rizika būna labai didelė.



2.8 pav. Skaičiuoklės rezultatų langas

Pavyzdžio (žr. 2.7 lentelę) rezultatai pateikti 2.8 paveiksle. Matome, kad rizika susirgti širdies ir kraujagyslių ligomis viršija 50%, todėl vertėtų atkreipti dėmesį į savo gyvenimo būdą.

PADĖKA

Dėkoju Lietuvos sveikatos mokslų universiteto Kardiologijos instituto mokslininkams už suteiktus duomenis. Profesoriui Abdonui Tamošiūnui už konsultacijas.

IŠVADOS

1. Atlikta užsienio ir Lietuvos literatūros apžvalga parodė, kad analizuojant širdies ir kraujagyslių susirgimus dažniausiai taikomi logistinės, Kokso regresijos modeliai ir neuroniniai tinklai, o mirtingumo nuo šių ligų analizei – Puasono regresijos ir netiesinės regresijos modeliai.

2. Lietuvos sveikatos mokslų universiteto Kardiologijos instituto mokslininkų surinktiems 2006-2013 m. duomenims apie Kauno miesto 45-72 m. amžiaus gyventojų širdies ir kraujagyslių ligas sudaryti logistinės regresijos prognozavimo modeliai, kurie leidžia įvertinti susirgimo tikimybes.

3. Atlikus tyrimus nustatyta, kad rūkymas, diabetas, išeminė širdies liga prieš susirgimą ir lytis daugiausiai turi įtakos širdies ir kraujagyslių ligoms.

4. Panaudojus Kauno miesto gyventojų 45-64 m. amžiaus duomenis ir daugialypę tiesinę regresinę analizę bei kvantilinę regresiją, sudaryti Kauno miesto gyventojų mirtingumo dažnio prognozavimo modeliai.

5. Nustatyta, kad mirtingumas nuo širdies ir kraujagyslių ligų 45-64 m. vyrams ženkliai didesnis nei moterims (vyrų dažnis 11,37 mirčių 100 tūkst. gyventojų).

6. Sukurta skaičiuoklė virtualioje erdvėje, leidžianti kiekvienam asmeniui įsivertinti tikimybę susirgti širdies ir kraujagyslių ligomis.

LITERATŪRA

1. Heidenreich P., Trogdon J. G. ir kt., “Forecasting the Future of Cardiovascular Disease in the United States: A Policy Statement From the American Heart Association”, 2011; 933-944; originally published online January 24, 2011. Print ISSN: 009-7322.
2. Faeh D., Gutzwiller F., Bopp M. “Lower Mortality From Coronary Heart Disease and Stroke at Higher Altitudes in Switzerland”, 2009; Print ISSN: 0009-7322. Online ISSN:1524-4539.
3. Kuklina E. V., Yoon P. W. ir kt., “Prevalence of Coronary Heart Disease Risk Factos and Screening for High Cholesterol Levels Among Young Adults, United States, 1999-2006”; Annanls Of Family Medicine, vol. 8.
4. Buivydaitė K., A. Tamošiūnas ir kt., „Sergamumas ūminiu miokardo infarktu, rizikos veiksniai ir išemijos rizika“, *Medicina (Kaunas)*, 2005; 41 (2).
5. Yu. Oleg, Gorokhova S. G. ir kt., “Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters”, Juornal of Cardiology (2012) 59, 190-194.
6. Bernotienė G., Radišauskas R., A. Tamošiūnas ir kt., „Kauno miesto gyventojų mirtingumo nuo kraujotakos sistemos ligų pokyčiai 2004 – 2008 m.“, *Lietuvos bendrosios praktikos gydytojas*, 2011 m. lapkritis (T.15, Nr. 9).
7. Lietuvos statistikos departamentas. Prieiga per internetą: http://www.stat.gov.lt/.
8. SAS Analytical Products 9.22 documentation. Prieiga per internetą: http://support.sas.com/documentation/.
9. Čekanavičius V., Murauskas G. „Statistika ir jos taikymai II“, Vilnius: TEV, 2009.
10. Levulienė R., „Statistikos taikymai naudojant SAS“, VU l-kla 2009.
11. Čekanavičius V., Murauskas G. „Taikomoji regresinė analizė socialiniuose tyrimuose“, Vilnius, ISBN 978-609-459-300-0; 2014.
12. Tamosiunas A., Baceviciene M. ir kt., “Cardiovascular Risk Factors and Cognitive Function in Middle Aged and Elderly Lithuanian Urban Population: Results from the HAPPIEE Study”, Tamosiunas et al. BMC Neurology 2012, 12:149.
13. Sarkar S. K., Midi H. and Rana Sohel “Detection of Outliers and Influential Observations in Binary Logistic Regression: An Empirical Study”, Journal of Applied Sciences 11 (1); 26-35, 2011.
14. Grabauskytė I. „Klinikinių ir genetinių veiksnių įtakos susirgti išemine širdies liga tyrimas“ XI Taikomosios matematikos konferencija 2013.
15. Grabauskytė I. „Kauno miesto gyventojų rizikos susirgti galvos smegenų insultu prognozavimo modelis“, konferencija *Matematika ir matematikos dėstymas – 2014*.

16. Grabauskytė I. „Kauno miesto 45-64 m. amžiaus gyventojų mirtingumo prognozavimas“, konferencija *Taikomoji matematika – 2014*.

17. Lesauskaitė V., Tamošiūnas A., Grabauskytė I. ir kt., „Transformuojančio faktoriaus TGF β1 koncentracijos kitimai esant kylančiosios aortos dilatacinei patalogijai“ XIII-oji Tarptautinė Lietuvos biochemikų konferencija 2014 m. birželio 18-20 d. Birštone.

18. Stanevičiūtė Ž., Šepetienė R., Grabauskytė I. „Fibrilino geno polimorfizmu įtakojančių krūtinės aortos aneurizmos formavimąsi, tyrimas“ *Studentų mokslinės veiklos skatinimas*, 2014 m. birželio 26-26 d.

19. Lesauskaitė V., Tamošiūnas A., Grabauskytė I. ir kt., “Does the association of *FBN1* SNPs with dilatative pathology of the ascending aorta is an extension of indications for surgical treatment?“, 2014.

20. Lesauskaitė V., Tamošiūnas A., Grabauskytė I. ir kt., “Association of FBN1 SNPs rs2118181, rs1036477, rs10519177, rs755251 and rs4774517 with dilatative pathology of the ascending aorta (DPAA)”, 2014.

21. Kupstytė N., Lesauskaitė V., Tatarūnas V., Grabauskytė I. ir kt.,“The effect of clinical and genetic factor on early definite stent thrombosis”, 2014.

22. Venclovienė J. „Statistiniai metodai medicinoje“, VDU, Kaunas, 2010 m.

23. Dzemyda G., Kurasova O., J. Žilinskas „Daugiamačių duomenų vizualizavimo metodai“, Mokslo Aidai, Vilnius, 2008 m.

24. Sakalauskas L., „Duomenų gavyba“, Vilnius, 2009.

25. http://www.insol.lt/software/statistics/ .

26. Rinkūnienė E., Petrulionienė Ž., Zdanevičiūtė I., Dženkevičiūtė V. „Mirtingumo nuo širdies ir kraujagyslių ligų tendencijos Lietuvoje ir Europos Sąjungos šalyse“, *Medicinos teorija ir praktika* 2013-T. 19 (Nr.2), 130-136p.

27. http://www.cvdcheck.org.au/

28. http://cvdrisk.nhlbi.nih.gov/

29. http://www.mcw.edu/calculators/Coronary-Heart-Disease-Risk.htm

30. http://www.heart.lt/pagrindinis_meniu/suzinok_savo_rizika/3983/

31. Medicina žmogui. Prieiga per internetą: http://www.sos03.lt/

32. European Cardiovascular Disease Statistics - European Heart Network and European Society of Cardiology, September 2012. Prieiga per internetą: http://www.escardio.org/about/documents/eu-cardiovascular-disease-statistics-2012.pdf

PRIEDAI

1 PRIEDAS. DUOMENŲ MATRICOS STRUKTŪRA

1 lentelė

Kintamųjų žymėjimai modelyje ir SAS duomenų faile

Nr.	Kintamojo žymėjimas	Kintamojo vardas SAS duomenų faile	Pastaba
1.	**-**	ID_Nr	Paciento numeris
2.	amžius	amzius	-
3.	ūgis	ugis	-
4.	svoris	svoris	-
5.	lytis	lytis	-
6.	IŠL_p	isl_prad	Išeminė širdies liga prieš susirgimą
7.	kmi	kmi	Kūno masės indeksas
8.	Rukymas	ruk	Rūkymas
9.	SV	subj_sve	Subjektyvus sveikatos vertinimas
10.	Chol	chol	Cholesterolis
11.	trigl	trigl	Trigliceridai
12.	DTL	hdl	Didelio tankio lipoproteinų cholesterolis
13.	MTL	ldl	Mažo tankio lipoproteinų cholesterolis
14.	FA	fiz_akt	Fizinis aktyvumas
15.	Diabetas	diabetas	-
16.	SK	sks_omr_vid3	Sistolinis kraujospūdis
17.	DK	dks_omr_vid3	Diastolinis kraujospūdis
18.	AH	AH_omron	Arterinė hipertenzija
19.	D_Chol	chol_susk	Suskirstytas kintamasis į dvi kategorijas
20.	trigl_susk	trigl_susk	Suskirstytas kintamasis į dvi kategorijas
21.	hdl_susk	hdl_susk	Suskirstytas kintamasis į dvi kategorijas
22.	ldl_susk	ldl_susk	Suskirstytas kintamasis į dvi kategorijas
23.	KMI3	KMI3gr	Suskirstytas kintamasis į tris kategorijas
24.	Gliukozė	gluk	-
25.	Insultas	insult	-
26.	ISL_MI	ISL_MI	Išeminė širdies liga ir miokardo infarktas

27.	alkoh_d	alkoh_d	-
28.	Išsilav	issilav	Išsilavinimas
29	Šeim_pad	seim_pad	Šeimyninė padėtis
30.	ŠKL	visos_ligos	Širdies ir kraujagyslių ligos
31.	Juosm_ap	juosm_ap	Juosmens apimtis
32.	Šeima_ŠKL	seima_skl	Šeimos narių susirgimai širdies ir kraujagyslių ligomis
33.	Šeima_insultas	seima_insult	Šeimos narių susirgimas insultu
34.	Šeima_diabetas	seima_diab	Šeimos narių sirgimas diabetu
35.	Depresija	depresija	-
36.	Paž_geb	paz_geb	Pažintinių gebėjimų sutrikimai

2 lentelė

Logistinių modelių kintamieji

Nr.	Kintamieji	Nr.	Kintamieji
1.	alkoh_d, amzius, ruk, lytis isl_prad, kmi, subj_sve, chol_susk, trigl_susk, ldl_susk, hdl_susk, fiz_akt, diabetas, seim_pad, issilav, gyv_kok, sks_omr_vid3, gluk	2.	amzius, ruk, lytis, isl_prad, kmi, subj_sve, chol, trigl, ldl,hdl, alkoh_d, fiz_akt, diabetas, seim_pad, issilav, sks_omr_vid3, gluk
3.	amzius, ruk, lytis, isl_prad, KMI3gr, subj_sve, chol_susk, trigl_susk, ldl_susk, hdl_susk, fiz_akt, diabetas, seim_pad, issilav, sks_omr_vid3, gluk	4.	amzius, ruk, lytis, isl_prad, kmi, subj_sve, chol_susk, trigl_susk, ldl_susk, hdl_susk, alkoh_d, fiz_akt, diabetas, seim_pad, issilav, dks_omr_vid3, gluk
5.	amzius, ruk, lytis, isl_prad, kmi, subj_sve, chol, trigl, ldl, hdl, alkoh_d, fiz_akt, diabetas, seim_pad, issilav, dks_omr_vid3, gluk	6.	amzius, ruk, lytis, isl_prad, KMI3gr, subj_sve, chol_susk, trigl_susk, ldl_susk, hdl_susk, alkoh_d, fiz_akt, diabetas, seim_pad, issilav, dks_omr_vid3, gluk
7.	amzius, ruk, lytis, isl_prad, kmi, subj_sve, trigl_susk, ldl_susk, hdl_susk, alkoh_d, fiz_akt, diabetas, seim_pad, issilav, AH_omron, gluk	8.	amzius, ruk, lytis, isl_prad, kmi, subj_sve, chol, trigl, ldl, hdl, alkoh_d, fiz_akt, diabetas, seim_pad, issilav, AH_omron, gluk
9.	alkoh_d, amzius, ruk, lytis, kmi, subj_sve, chol, trigl, ldl, hdl, fiz_akt, diabetas, seim_pad, issilav, sks_omr_vid3, gluk	10.	alkoh_d, amzius, ruk, lytis, KMI3gr, subj_sve, chol_susk, trigl_susk, ldl_susk, hdl_susk, fiz_akt, diabetas, seim_pad, issilav, sks_omr_vid3, gluk
11.	amzius, ruk, lytis, isl_prad, kmi, subj_sve, chol_susk, trigl_susk, ldl_susk,	12.	subj_sve, chol_susk, trigl_susk, ldl_susk hdl_susk, alkoh_d, fiz_akt,

	hdl_susk, fiz_akt, diabetas, seim_pad,issilav, sks_omr_vid3, gluk,sks_omr_vid3|kmi, lytis|amzius, kmi|fiz_akt, alkoh_d|ruk|lytis @2		seim_pad, issilav, dks_omr_vid3, gluk, juosm_ap, seima_skl, seima_insult, seima_diab, depresija, paz_geb
13.	diabetas, isl_prad, amzius, ruk, lytis, subj_sve, chol, trigl, ldl,hdl, alkoh_d, fiz_akt, seim_pad, issilav, sks_omr_vid3, gluk, juosm_ap, seima_skl, depresija, paz_geb	14.	diabetas, isl_prad, amzius, kmi, ruk, lytis, subj_sve, chol, trigl, ldl, hdl, alkoh_d, fiz_akt, seim_pad, issilav, AH_omron, gluk, juosm_ap, seima_skl

2 PRIEDAS. GALVOS SMEGENŲ INSULTO PROGNOZAVIMO MODELIŲ ANALIZĖS REZULTATAI

1 lentelė

Galvos smegenų insulto prognozavimui reikšmingi kintamieji ir modelių tinkamumo rodikliai

Nr.		Reikšmingi kintamieji galvos smegenų insulto prognozavimui	AIC	SC	Pseudo R^2	Plotas po ROC kreive
1.	Pradinis	sks_omr_vid3, amzius, ruk, gluk, isl_prad, hdl_susk	1317.05	1364.65	0.0139	0.714
	Pašalinus išskirtis	sks_omr_vid3, amzius, ruk, gluk, isl_prad	1299.75	1340.43	0.0156	0.714
	Su retais	sks_omr_vid3, amzius, ruk, gluk, isl_prad	1280.35	1321.10	0.0156	0.714
2.	Pradinis	sks_omr_vid3, amzius, ruk, gluk, isl_prad	1324.72	1365.52	0.0124	0.706
	Pašalinus išskirtis	sks_omr_vid3, amzius, ruk, gluk, isl_prad	1307.62	1348.37	0.0147	0.711
	Su retais	sks_omr_vid3, amzius, ruk, gluk, isl_prad	1296.13	1337.08	0.0142	0.709
3.	Pradinis	sks_omr_vid3, amzius, ruk, gluk, isl_prad, hdl_susk	1317.22	1364.83	0.0139	0.714
	Pašalinus išskirtis	sks_omr_vid3, amzius, ruk, gluk, isl_prad	1298.06	1338.74	0.0158	0.715
	Su retais	sks_omr_vid3, amzius, ruk, gluk, isl_prad	1278.16	1318.89	0.0158	0.715
4.	**Pradinis**	**amzius, dks_omr_vid3, gluk, ruk,**	**1312.42**	**1360.02**	**0.0146**	**0.718**

		isl_prad, hdl_susk				
	Pašalinus išskirtis	**amzius, dks_omr_vid3, gluk, ruk, isl_prad**	**1334.20**	**1293.52**	**0.0165**	**0.720**
	Su retais	**amzius, dks_omr_vid3, gluk, ruk, isl_prad**	**1275.26**	**1315.99**	**0.0164**	**0.718**
5.	Pradinis	amzius, dks_omr_vid3, gluk, ruk, isl_prad	1319.94	1360.74	0.0132	0.711
	Pašalinus išskirtis	amzius, dks_omr_vid3, gluk, ruk, isl_prad	1304.32	1345.06	0.0152	0.713
	Su retais	amzius, dks_omr_vid3, gluk, ruk, isl_prad	1294.67	1335.62	0.0146	0.712
6.	Pradinis	amzius, dks_omr_vid3, gluk, ruk, isl_prad, hdl_susk	1312.42	1360.02	0.0146	0.718
	Pašalinus išskirtis	amzius, dks_omr_vid3, gluk, ruk, isl_prad	1293.52	1334.20	0.0165	0.720
	Su retais	amzius, dks_omr_vid3, gluk, ruk, isl_prad	1275.26	1315.99	0.0164	0.718
7.	Pradinis	amzius, ruk, gluk, isl_prad, hdl_prad, issilav	1323.50	1371.10	0.0129	0.706
	Pašalinus išskirtis	amzius, ruk, gluk, isl_prad, issilav	1309.26	1349.95	0.0142	0.706
	Su retais	amzius, ruk, gluk, isl_prad, issilav	1299.23	1340.00	0.0140	0.704
8.	Pradinis	amzius, ruk, isl_prad, gluk, issilav	1330.66	1371.46	0.0116	0.698
	Pašalinus išskirtis	amzius, ruk, isl_prad, gluk, issilav	1318.81	1359.57	0.0131	0.699
	Su retais	amzius, ruk, isl_prad, gluk, issilav	1316.26	1357.24	0.0125	0.698
9.	**Pradinis**	**sks_omr_vid3, amzius, ruk, gluk**	**1331.35**	**1365.35**	**0.0112**	**0.696**
	Pašalinus išskirtis	**sks_omr_vid3, amzius, ruk, gluk**	**1317.33**	**1351.29**	**0.0130**	**0.697**
	Su retais	**sks_omr_vid3, amzius, ruk, gluk**	**1308.83**	**1342.96**	**0.0126**	**0.698**
10.	Pradinis	sks_omr_vid3, amzius, ruk, gluk, hdl_susk	1324.51	1365.32	0.0125	0.704
	Pašalinus išskirtis	sks_omr_vid3, amzius, ruk, gluk	1310.22	1344.13	0.0137	0.704
	Su retais	sks_omr_vid3, amzius, ruk, gluk	1293.50	1327.46	0.0138	0.704
11.	Pradinis	amzius, ruk, sks_omr_vid3, isl_prad, gluk, hdl_susk	1317.22	1364.83	0.0139	0.714
	Pašalinus išskirtis	amzius, ruk, sks_omr_vid3, isl_prad, gluk	1305.22	1345.93	0.0149	0.712
	Su retais	amzius, ruk, sks_omr_vid3, isl_prad, gluk	1285.56	1326.33	0.0149	0.712
12.	Pradinis	amzius, dks_omr_vid3, ruk, seima_diab, paz_geb	1263.16	1303.64	0.0120	0.705
	Pašalinus išskirtis	amzius, dks_omr_vid3, ruk, seima_diab, paz_geb	1253.63	1294.08	0.0133	0.707
	Su retais	amzius, dks_omr_vid3, ruk, seima_diab, paz_geb	1315.43	1356.35	0.0140	0.713
13.	Pradinis	sks_omr_vid3, amzius, ruk, isl_prad,	1260.496	1307.741	0.0128	0.703

		paz_geb, seima_sl				
	Pašalinus išskirtis	sks_omr_vid3, amzius, ruk, isl_prad, paz_geb, seima_sl	1246.308	1293.487	0.0148	0.706
	Su retais	sks_omr_vid3, amzius, ruk, isl_prad, paz_geb, seima_sl	1302.049	1349.760	0.0154	0.709
14.	Pradinis	amzius, ruk, isl_prad, gluk, seima_sl	1321.601	1362.358	0.0115	0.701
	Pašalinus išskirtis	amzius, ruk, isl_prad, gluk, seima_sl	1310.894	1351.608	0.0129	0.701
	Su retais	amzius, ruk, isl_prad, gluk, seima_sl	1308.565	1349.503	0.0123	0.698

3 PRIEDAS. IŠEMINĖS ŠIRDIES LIGOS PROGNOZAVIMO MODELIŲ ANALIZĖS REZULTATAI

1 lentelė

Išeminės širdies ligos prognozavimui reikšmingi kintamieji ir modelių tinkamumo rodikliai

Nr.		**Reikšmingi kintamieji išeminės širdies ligos prognozavimui**	**AIC**	**SC**	**Pseudo R^2**	**Plotas po ROC kreive**
1.	Pradinis	amzius, lytis, kmi, isl_prad, subj_sve, ldl_susk	2466.91	2521.31	0.0108	0.643
	Pašalinus išskirtis	amzius, lytis, kmi, isl_prad, subj_sve, ldl_susk, diabetas	2458.25	2519.42	0.0121	0.646
	Su retais	amzius, lytis, kmi, isl_prad, subj_sve, ldl_susk, diabetas	2474.40	2529.01	0.0115	0.644
2.	Pradinis	amzius, hdl, isl_prad, lytis, subj_sve, chol, diabetas	2464.60	2525.81	0.0115	0.647
	Pašalinus išskirtis	amzius, ldl, isl_prad, lytis, subj_sve, kmi, diabetas	2436.27	2497.36	0.0146	0.655
	Su retais	amzius, ldl, isl_prad, lytis, subj_sve, kmi, diabetas	2450.32	2504.84	0.0139	0.652
3.	Pradinis	amzius, ldl_susk, isl_prad, lytis, subj_sve, diabetas	2469.64	2524.04	0.0104	0.639
	Pašalinus išskirtis	amzius, ldl_susk, isl_prad, lytis, subj_sve, diabetas	2443.92	2498.24	0.0133	0.647
	Su retais	amzius, ldl_susk, isl_prad, lytis, subj_sve, diabetas	2468.85	2516.58	0.0128	0.645
4.	Pradinis	amzius, ldl_susk, isl_prad, lytis, subj_sve,	2467.44	2521.85	0.0108	0.643

		kmi				
	Pašalinus išskirtis	amzius, ldl_susk, isl_prad, lytis, subj_sve, kmi, diabetas	2458.83	2520.01	0.0121	0.646
	Su retais	amzius, ldl_susk, isl_prad, lytis, subj_sve, kmi, diabetas	2474.40	2529.01	0.0115	0.644
5.	Pradinis	amzius, hdl, isl_prad, lytis, subj_sve, chol, diabetas	2464.60	2525.81	0.0115	0.647
	Pašalinus išskirtis	amzius, ldl, isl_prad, lytis, subj_sve, kmi, diabetas	2436.27	2497.36	0.0146	0.655
	Su retais	amzius, ldl, isl_prad, lytis, subj_sve, kmi, diabetas	2450.32	2504.84	0.0139	0.652
6.	Pradinis	amzius, lytis, subj_sve, isl_prad, diabetas, ldl_susk	2469.64	2524.04	0.0104	0.639
	Pašalinus išskirtis	amzius, lytis, subj_sve, isl_prad, diabetas, ldl_susk	2449.06	2503.40	0.0128	0.645
	Su retais	amzius, lytis, subj_sve, isl_prad, diabetas, ldl_susk	2473.63	2521.38	0.0122	0.643
7.	Pradinis	amzius, lytis, kmi, subj_sve, isl_prad, ldl_susk, AH_omron	2465.41	2526.62	0.0114	0.644
	Pašalinus išskirtis	amzius, lytis, kmi, subj_sve, isl_prad, ldl_susk	2458.27	2512.64	0.0117	0.645
	Su retais	amzius, lytis, kmi, subj_sve, isl_prad, ldl_susk	2469.78	2517.54	0.0111	0.642
8.	**Pradinis**	**amzius, hdl, AH_omron, isl_prad, lytis, subj_sve, chol, diabetas**	**2461.77**	**2529.78**	**0.0122**	**0.651**
	Pašalinus išskirtis	**diabetas, isl_prad, lytis, subj_sve, ldl, amzius, kmi**	**2433.46**	**2494.53**	**0.0149**	**0.656**
	Su retais	**diabetas, isl_prad, lytis, subj_sve, ldl, amzius, kmi**	**3210.35**	**3271.64**	**0.0226**	**0.665**
9.	Pradinis	amzius, hdl, subj_sve, lytis, chol, diabetas	2468.45	2522.85	0.0106	0.638
	Pašalinus išskirtis	amzius, hdl, subj_sve, lytis, chol, diabetas	2445.67	2499.98	0.0131	0.644
	Su retais	amzius, hdl, subj_sve, lytis, chol, diabetas	2517.65	2565.46	0.0140	0.646
10.	**Pradinis**	**amzius, lytis, subj_sve, diabetas, ldl_susk, sks_omr_vid3**	**2472.97**	**2527.38**	**0.0099**	**0.634**
	Pašalinus išskirtis	**amzius, lytis, subj_sve, diabetas, ldl_susk, sks_omr_vid3**	**2448.27**	**2502.59**	**0.0126**	**0.643**
	Su retais	**amzius, lytis, subj_sve, diabetas, ldl_susk, sks_omr_vid3**	**3229.73**	**3284.25**	**0.0196**	**0.652**
11.	Pradinis	amzius, lytis, kmi, subj_sve, isl_prad, amzius*lytis, ldl_susk	2463.70	2524.91	0.0116	0.642

	Pašalinus išskirtis	amzius, lytis, kmi, subj_sve, isl_prad, amzius*lytis, ldl_susk	2457.52	2518.70	0.0123	0.645
	Su retais	amzius, lytis, kmi, subj_sve, isl_prad, amzius*lytis, ldl_susk	2474.04	2528.65	0.0117	0.641
12.	Pradinis	juosm_ap, amzius, lytis, subj_sve, seima_sl, chol_susk	2351.56	2405.54	0.0111	0.640
	Pašalinus išskirtis	juosm_ap, amzius, lytis, subj_sve, seima_sl, chol_susk	2349.11	2403.08	0.0113	0.640
	Su retais	juosm_ap, amzius, lytis, subj_sve, seima_sl, chol_susk	2560.89	2608.82	0.0102	0.632
13.	Pradinis	juosm_ap, amzius, lytis, subj_sve, seima_sl, ldl, isl_prad	2355.02	2415.76	0.0119	0.647
	Pašalinus išskirtis	juosm_ap, amzius, lytis, subj_sve, seima_sl, ldl, isl_prad	2347.23	2407.94	0.0128	0.650
	Su retais	juosm_ap, amzius, lytis, subj_sve, seima_sl, ldl, isl_prad	2445.22	2499.75	0.0117	0.645
14.	Pradinis	juosm_ap, amzius, isl_prad, lytis, subj_sve, ldl, seima_sl	2443.22	2504.36	0.0113	0.644
	Pašalinus išskirtis	juosm_ap, amzius, isl_prad, lytis, subj_sve, ldl, seima_sl	2435.49	2496.59	0.0121	0.647
	Su retais	juosm_ap, amzius, isl_prad, lytis, subj_sve, ldl, seima_sl	2445.15	2499.69	0.0117	0.645

4 PRIEDAS. ŠIRDIES IR KRAUJAGYSLIŲ LIGŲ PROGNOZAVIMO MODELIŲ ANALIZĖS REZULTATAI

1 lentelė

Širdies ir kraujagyslių ligų prognozavimui reikšmingi kintamieji ir modelių tinkamumo rodikliai

Nr.		**Reikšmingi kintamieji širdies ir kraujagyslių ligų prognozavimui**	**AIC**	**SC**	**Pseudo R^2**	**Plotas po ROC kreive**
1.	Pradinis	amzius, lytis, kmi, isl_prad, subj_sve, ldl_susk	2466.91	2521.31	0.0108	0.643
	Pašalinus išskirtis	amzius, lytis, kmi, isl_prad, subj_sve, ldl_susk, diabetas	2458.25	2519.42	0.0121	0.646

	Su retais	amzius, lytis, kmi, isl_prad, subj_sve, ldl_susk, diabetas	2474.40	2529.01	0.0115	0.644
2.	Pradinis	amzius, hdl, isl_prad, lytis, subj_sve, chol, diabetas	2464.60	2525.81	0.0115	0.647
	Pašalinus išskirtis	amzius, ldl, isl_prad, lytis, subj_sve, kmi, diabetas	2436.27	2497.36	0.0146	0.655
	Su retais	amzius, ldl, isl_prad, lytis, subj_sve, kmi, diabetas	2450.32	2504.84	0.0139	0.652
3.	Pradinis	amzius, ldl_susk, isl_prad, lytis, subj_sve, diabetas	2469.64	2524.04	0.0104	0.639
	Pašalinus išskirtis	amzius, ldl_susk, isl_prad, lytis, subj_sve, diabetas	2443.92	2498.24	0.0133	0.647
	Su retais	amzius, ldl_susk, isl_prad, lytis, subj_sve, diabetas	2468.85	2516.58	0.0128	0.645
4.	Pradinis	amzius, ldl_susk, isl_prad, lytis, subj_sve, kmi	2467.44	2521.85	0.0108	0.643
	Pašalinus išskirtis	amzius, ldl_susk, isl_prad, lytis, subj_sve, kmi, diabetas	2458.83	2520.01	0.0121	0.646
	Su retais	amzius, ldl_susk, isl_prad, lytis, subj_sve, kmi, diabetas	2474.40	2529.01	0.0115	0.644
5.	Pradinis	amzius, hdl, isl_prad, lytis, subj_sve, chol, diabetas	2464.60	2525.81	0.0115	0.647
	Pašalinus išskirtis	amzius, ldl, isl_prad, lytis, subj_sve, kmi, diabetas	2436.27	2497.36	0.0146	0.655
	Su retais	amzius, ldl, isl_prad, lytis, subj_sve, kmi, diabetas	2450.32	2504.84	0.0139	0.652
6.	Pradinis	amzius, lytis, subj_sve, isl_prad, diabetas, ldl_susk	2469.64	2524.04	0.0104	0.639
	Pašalinus išskirtis	amzius, lytis, subj_sve, isl_prad, diabetas, ldl_susk	2449.06	2503.40	0.0128	0.645
	Su retais	amzius, lytis, subj_sve, isl_prad, diabetas, ldl_susk	2473.63	2521.38	0.0122	0.643
7.	Pradinis	amzius, lytis, kmi, subj_sve, isl_prad, ldl_susk, AH_omron	2465.41	2526.62	0.0114	0.644
	Pašalinus išskirtis	amzius, lytis, kmi, subj_sve, isl_prad, ldl_susk	2458.27	2512.64	0.0117	0.645
	Su retais	amzius, lytis, kmi, subj_sve, isl_prad, ldl_susk	2469.78	2517.54	0.0111	0.642
8.	**Pradinis**	**amzius, hdl, AH_omron, isl_prad, lytis, subj_sve, chol, diabetas**	**2461.77**	**2529.78**	**0.0122**	**0.651**
	Pašalinus	**diabetas, isl_prad, lytis, subj_sve, ldl,**	**2433.46**	**2494.53**	**0.0149**	**0.656**

	išskirtis	**amzius, kmi**				
	Su retais	**diabetas, isl_prad, lytis, subj_sve, ldl, amzius, kmi**	**3210.35**	**3271.64**	**0.0226**	**0.665**
9.	Pradinis	amzius, hdl, subj_sve, lytis, chol, diabetas	2468.45	2522.85	0.0106	0.638
	Pašalinus išskirtis	amzius, hdl, subj_sve, lytis, chol, diabetas	2445.67	2499.98	0.0131	0.644
	Su retais	amzius, hdl, subj_sve, lytis, chol, diabetas	2517.65	2565.46	0.0140	0.646
10.	**Pradinis**	**amzius, lytis, subj_sve, diabetas, ldl_susk, sks_omr_vid3**	**2472.97**	**2527.38**	**0.0099**	**0.634**
	Pašalinus išskirtis	**amzius, lytis, subj_sve, diabetas, ldl_susk, sks_omr_vid3**	**2448.27**	**2502.59**	**0.0126**	**0.643**
	Su retais	**amzius, lytis, subj_sve, diabetas, ldl_susk, sks_omr_vid3**	**3229.73**	**3284.25**	**0.0196**	**0.652**
11.	Pradinis	amzius, lytis, kmi, subj_sve, isl_prad, amzius*lytis, ldl_susk	2463.70	2524.91	0.0116	0.642
	Pašalinus išskirtis	amzius, lytis, kmi, subj_sve, isl_prad, amzius*lytis, ldl_susk	2457.52	2518.70	0.0123	0.645
	Su retais	amzius, lytis, kmi, subj_sve, isl_prad, amzius*lytis, ldl_susk	2474.04	2528.65	0.0117	0.641
12.	Pradinis	juosm_ap, amzius, lytis, subj_sve, seima_sl, chol_susk	2351.56	2405.54	0.0111	0.640
	Pašalinus išskirtis	juosm_ap, amzius, lytis, subj_sve, seima_sl, chol_susk	2349.11	2403.08	0.0113	0.640
	Su retais	juosm_ap, amzius, lytis, subj_sve, seima_sl, chol_susk	2560.89	2608.82	0.0102	0.632
13.	Pradinis	juosm_ap, amzius, lytis, subj_sve, seima_sl, ldl, isl_prad	2355.02	2415.76	0.0119	0.647
	Pašalinus išskirtis	juosm_ap, amzius, lytis, subj_sve, seima_sl, ldl, isl_prad	2347.23	2407.94	0.0128	0.650
	Su retais	juosm_ap, amzius, lytis, subj_sve, seima_sl, ldl, isl_prad	2445.22	2499.75	0.0117	0.645
14.	Pradinis	juosm_ap, amzius, isl_prad, lytis, subj_sve, ldl, seima_sl	2443.22	2504.36	0.0113	0.644
	Pašalinus išskirtis	juosm_ap, amzius, isl_prad, lytis, subj_sve, ldl, seima_sl	2435.49	2496.59	0.0121	0.647
	Su retais	juosm_ap, amzius, isl_prad, lytis, subj_sve, ldl, seima_sl	2445.15	2499.69	0.0117	0.645

5 PRIEDAS. SUKURTŲ PROGNOZAVIMO MODELIŲ SAS PROGRAMOS

1 lentelė

SAS programos kodas

```
                        /* 1 dalis. Logistinė regresija*/
ods listing close;
ods rtf file='C:\Users\Ingrida\Desktop\magistrinis\SKL.rtf';
ods graphics on;
/*nuskaitomas duomenų failas*/
proc import datafile='C:\Users\Ingrida\Desktop\magistrinis\duomenys.xls'
out=magistrinis
DBMS=Excel2000 replace;
getnames=yes;
run;
data magistrinis; set magistrinis;
if  ISL_MI= '1' then visos_ligos=1;
if  ISL_MI= '0' then visos_ligos=0;
if  insult= '1' then visos_ligos=2;
run;
data magistrinis; set magistrinis;
if  visos_ligos='1' or visos_ligos='2' then bendrai=1;
else bendrai=0;
run;
/*logistinė regresija*/
proc logistic data=magistrinis outest=betas covout plots(only label)=
(dfbetas influence leverage phat dpc);
class alkoh_d subj_sve;
model bendrai (event='1')= amzius ruk lytis  isl_prad kmi subj_sve chol_susk
trigl_susk ldl_susk hdl_susk  fiz_akt diabetas seim_pad issilav sks_omr_vid3
gluk       sks_omr_vid3 |kmi lytis|amzius kmi|fiz_akt alkoh_d|ruk|lytis @2
/ link=logit selection=stepwise RSQ CLODDS=WALD outroc=roc1 CTABLE details
lackfit;
output out=paklaidos RESCHI=chires RESDEV=devres lower=lcl upper=ucl
difchisq=difchisq difdev=difdev h=lev  dfbetas=dfb c=c;
title "Dvinare logistine regresija";
run;

data paklaidos2;
set paklaidos;
zchires=chires;
zdevres=devres;
run;
/*standartizuojamos paklaidos*/
proc standard data=paklaidos2 mean=0 std=1 out=zpaklaidos2;
var  zchires zdevres;
run;

proc means data=zpaklaidos2;
run;
/* išskirčių nufiltravimas*/
data isskirtys;
set zpaklaidos2;
levv=(2*20+2)/7115;
difchisqq=sqrt(difchisq);
difdevv=sqrt(difdev);
dfba=abs(dfb);
dfbb=(2/sqrt(7115));
if lev>levv then lev_out=1;
else lev_out=0;
```

```
if zchires<-3 then zchires_out=1;
else zchires_out=0;
if zchires>3 then zchires_out=1;
else zchires_out=0;
if zdevres<-3 then zdevres_out=1;
else zdevres_out=0;
if zdevres>3 then zdevres_out=1;
else zdevres_out=0;
if difchisqq<-2 then difchisq_out=1;
else difchisq_out=0;
if difchisqq>2 then difchisq_out=1;
if difdevv<-2 then difdev_out=1;
else difdev_out=0;
if difdevv>2 then difdev_out=1;
else difdev_out=0;
if dfba>dfbb then dfb_out=1;
else dfb_out=0;

if sum(of lev_out dfb_out zchires_out zdevres_out difchisq_out difdev_out
)=1 then outlier=1;
else outlier=0;
run;
/*išskirčių pašalinimas*/
data atrinkta;
set isskirtys;
if outlier=1 then delete;
run;
/* logistinė regresija pašalinus išskirtis*/
proc logistic data=atrinkta outest=betas covout plots(only label)=
(dfbetas influence leverage phat dpc) ;
class alkoh_d subj_sve;
model bendrai (event='1')= amzius ruk lytis  isl_prad kmi subj_sve chol_susk
trigl_susk ldl_susk hdl_susk  fiz_akt diabetas seim_pad issilav sks_omr_vid3
gluk sks_omr_vid3 |kmi lytis|amzius kmi|fiz_akt alkoh_d|ruk|lytis @2
/ link=logit  RSQ CLODDS=WALD outroc=roc1 CTABLE details lackfit;
output out=paklaidos3 RESCHI=chires RESDEV=devres lower=lcl upper=ucl
difchisq=difchisq
difdev=difdev h=lev  dfbetas=dfb c=c ;
title "Dvinare logistine regresija";
run;
/* logistinė regresija retiems įvykiams*/
proc logistic data=paklaidos3;
class  subj_sve;
model bendrai(event='1') = amzius lytis  isl_prad kmi  ldl diabetas
subj_sve
 / firth  clodds=pl outroc=roc1 RSQ ;
ods output cloddspl=orf;
run;

goptions reset=all  border cback=white
         colors=(black) vpos=40
         ftitle=swissb ftext=swiss htitle=1 targetdevice=pscolor;
/*braižoma ROC kreivė*/
proc gplot data=roc1;
   symbol1 i=join v=none c=blue;
   title font ='Times New Roman Baltic';
      title 'ROC kreive';
      plot _sensit_*_1mspec_=1 / vaxis=0 to 1 by .2
            haxis=0 to 1 by .2 grid
cframe=white;
label _1mspec_='1-Specifiskumas'
 _sensit_= 'Jautrumas';
run;
```

```
quit;

ods rtf close;
ods listing;

                        /*2 dalis. Daugialypė regresija*/

ods rtf file='C:\Users\Ingrida\Desktop\magistrinis\mirtingumas.rtf';
ods graphics on;
/*nuskaitomas duomenų failas*/
proc import datafile='C:\Users\Ingrida\Desktop\magistrinis\duomenys_D.xls'
DBMS=Excel2000 replace;
getnames=yes;
run;
/*atliekama procedūra GLM*/
proc glm data = veiksniai;
 model suskaiciuota=lytis amzius metai /solution ;
 output out=paklaidos predicted=p rstudent=r cookd=cd dffits=df H=lev ;
run; quit;
/*nustatomos išskirtys*/
data isskirtys;
set paklaidos;
if abs(r)>2.5 then r_out=1;
else r_out=0;
if lev>0.00625 then lev_out=1;   /* 2*4/1280= 0.00625  2*3/1200 = 0.005      ,
2*2/1200=0.0033 2*4/1200=0.00667
/*p=14, n= 7297, lev=0.003837, kai p=13, lev=0.003563 su formule 2p/n;
           30/
o su formule (2p+2)/n kai p=14 ir n=7297, lev=0.004111; kai p=13,
lev=0.003837, kai n=3, 0.00082, kai n=4 0.001096*/
else lev_out=0;
if cd>0.003125 then cd_out=1; /*4/7297,   4/1200=0.0033 4/120= 0.003125  */
else cd_out=0;
if abs(df)>0.1155  then df_out=1; /* kai p=3, n=1200 0.01   , p=2,0.0816
kai p=4, 0.1155    *p=14,n=7297, df=0.087604; o kai p=13, df=0.084417, kai
n=3, 0.04055, kai n=4, 0.04683*/
else df_out=0;
if sum(of r_out lev_out cd_out df_out)>=3 then outlier=1;
else outlier=0;
run;
/*ištrinamos išskirtys*/
data atrinkta;
set isskirtys;
if outlier=1 then delete;
run;
/* pašalinus išskirtis atliekama procedūra GLMSELECT*/
proc glmselect data= atrinkta /*plots(stepAxis=number)=all*/;
 model suskaiciuota=  lytis  amzius metai
 /          details=all stats=all selection=stepwise (stop=ADJRSQ);
 output out=gauta residual= liekanos ;  /*stdr=r residual=r*/
run;

goption reset=all;
/*tiriamas liekanų skirstinio normalumas, braižoma histograma*/
proc univariate data=gauta normal;
var liekanos;
histogram liekanos/normal(mu=est sigma=est);
run;
/* Homoskedastiškumo tyrimas*/
proc model data=gauta;
parms a1 b1 b2;
suskaiciuota=a1+b1*lytis+b2*amzius;
fit  suskaiciuota/ white breusch= (1 lytis amzius );
```

```
run;quit;

ods rtf close;
ods listing;


                        /*3 dalis. Kvantilinė regresija*/
ods rtf file='C:\Users\Ingrida\Desktop\magistrinis\mirtingumas_K.rtf';
ods graphics on;
/*nuskaitomas duomenų failas*/
proc import datafile='C:\Users\Ingrida\Desktop\magistrinis\duomenys.xls'
DBMS=Excel2000 replace;
getnames=yes;
run;

ods graphics on;
/*atliekama kvantilinė regresija*/
proc quantreg data=veiksniai ci=resamling alpha=0.05 plots=(rdplot ddplot
reshistogram ) algorithm=SIMPLEX ;
model daznis = lytis amzius /
             quantile= 0.05 to 0.5 by 0.01 CovB CorrB seed=12345 diagnostics
leverage(cutoff=8)  plot=quantplot;
test   lytis amzius;
output out=rezultatai residual=resid predicted=predict;
run;
proc gplot data=rezultatai;
  plot resid*predict;
run;
quit;
proc gplot data=rezultatai;
  plot resid*daznis;
run;
quit;

ods graphics off;
/*rezultatai išvedami į .xls failą*/
proc export outfile= "C:\Users\InGridute\Desktop\data_05.xls"
data=rezultatai
dbms=excel ;
run ;

ods rtf close;
ods listing;
```