KAUNO TECHNOLOGIJOS UNIVERSITETAS
Informatikos fakultetas

P176B101 Intelektikos pagrindai

Laboratorinio darbo Nr. 1 ataskaita

Dėstytojai:
jaun. asist. Nakrošis Arnas
asist. BUDNIKAS Germanas

Studentai:
Rokas Gudžiūnas IFF-2/1

KAUNAS, 2025

1. Turinys

2. Lentelių sąrašas

3. Paveikslėlių sąrašas

4. Įvadas

Šios užduoties tikslas yra išanalizuoti pacientų medicininius duomenis, atlikti duomenų analizę, bei apdorojimą. Duomenų rinkinys yra susijęs su insulto rizikos veiksniais, jame pateikiami pacientų demografiniai ir sveikatos duomenys. Rinkinį sudaro 10 atributų, tačiau pašalinus nereikalingus, jų lieka 9:

- Gender
- Age
- Hypertension
- Ever_married
- Work_type
- Residence_type
- Avg_glucose_level
- Bmi
- Smoking_status

5. Duomenų rinkinio kokybės analizė

Prieš atliekant šią analizę yra išmetamas „id“ stulpelis su unikaliomis reikšmėmis, kuris neturi įtakos modeliui.

5.1. Tolydinio tipo atributai

Lentelė 1 Tolydinio tipo atributų kokybės analizės lentelė

Atributo pavadinimas	Kiekis (Eilučių sk.)	Trūkstamos reikšmės, %	Kardinalumas	Minimali reikšmė	Maksimali reikšmė	1-asis kvartilis	3-iasis kvartilis	Vidurkis	Mediana	Standartinis nuokrypis
age	5110	0 %	104	0.08	82	25	61	43.2266	45	22.612647
avg_glucose_level	5110	0 %	3979	55.12	271.74	77.245	114.09	106.1476	91.8849	45.283560
bmi	5110	3.933%	418	10.3	97.6	23.5	33.1	28.893237	28.1	7.854067

5.2. Kategorinio tipo atributai

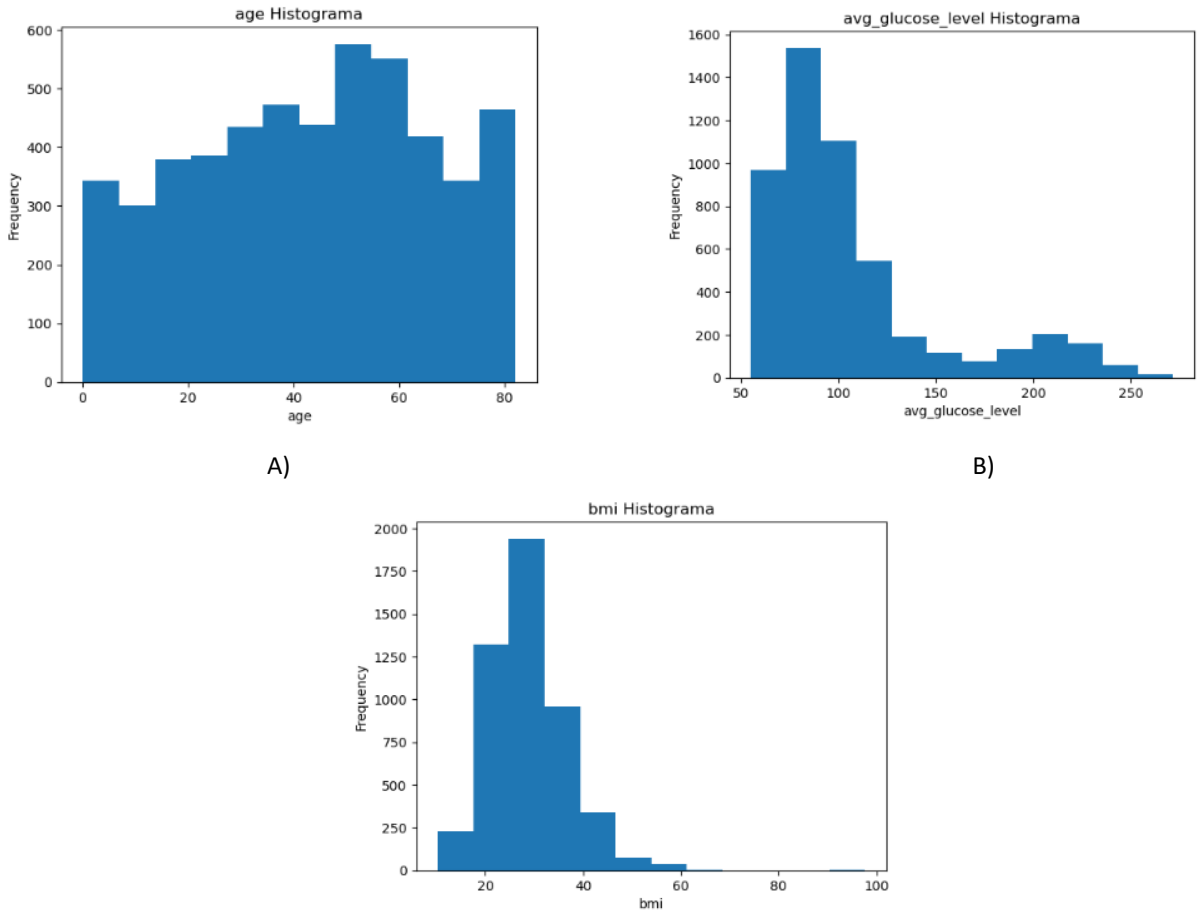
Lentelė 2 Kategorinio tipo atributų kokybės analizės lentelė

Atributo pavadinimas	Kiekis (Eilučių sk.)	Trūkstamos reikšmės, %	Kardinalumas	Moda	Modos dažnumas	Moda, %	2-oji Moda	2-osios modos dažnumas	2-oji moda, %
gender	5110	0%	3	Female	2994	58.59099%	Male	2115	41.3894%
hypertension	5110	0%	2	0	4612	90.2544%	1	498	9.74559%
heart_disease	5110	0%	2	0	4834	94.5988%	1	276	5.4011%
ever_married	5110	0%	2	Yes	3353	65.6164%	No	1757	34.3835%
work_type	5110	0%	5	Private	2925	57.2407%	Self-Employed	819	16.027%

Residence_type	5110	0%	2	Urban	2596	50.8023%	Rural	2514	49.19765%
smoking_status	5110	0%	4	Never smoked	1892	37.02544%	Unknown	1544	30.21526%

6. Atributų histogramos

6.1. Tolydinio tipo histogramos



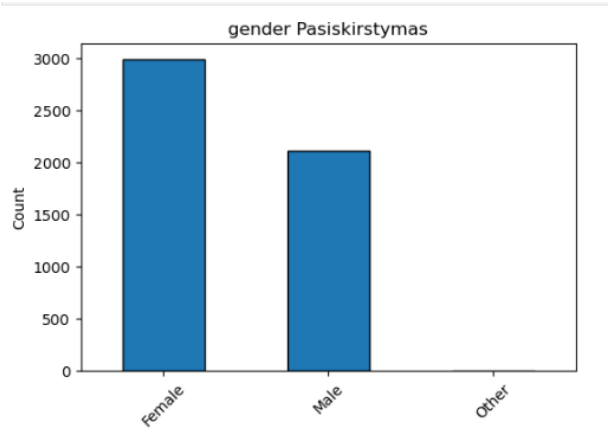
A) B)

C)

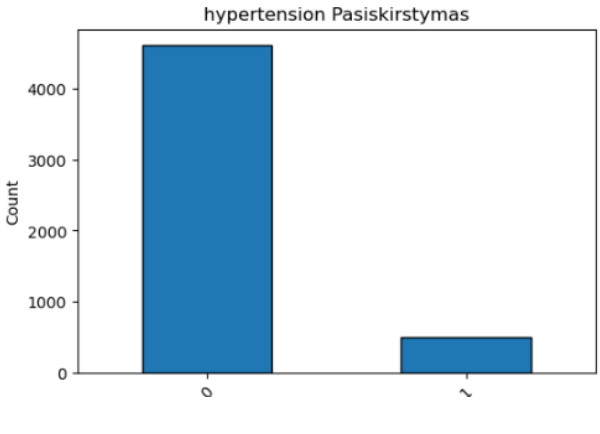
1 pav. Tolydinių atributų histogramos

Iš histogramų matome, jog „age“ atributo pasiskirstymas yra vienmodalis ir nėra pastebimų išskirčių. Tačiau „avg_glucose_level“ ir „bmi“ yra eksponentiškai pasiskirstę duomenys ir yra pastebimos išskirtys. Tikėtina, kad buvo nutukusių ir diabetu sergančių žmonių.

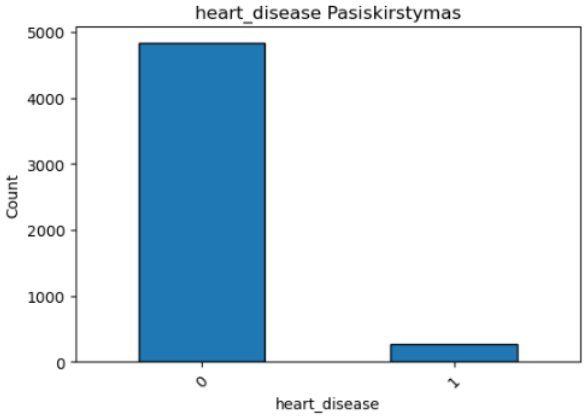
6.2. Kategorinio tipo stulpelinės diagramos



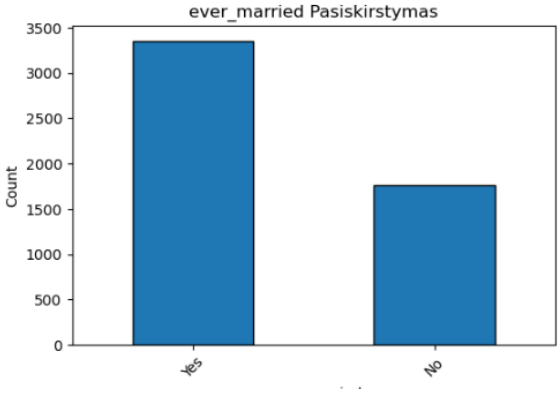
A)



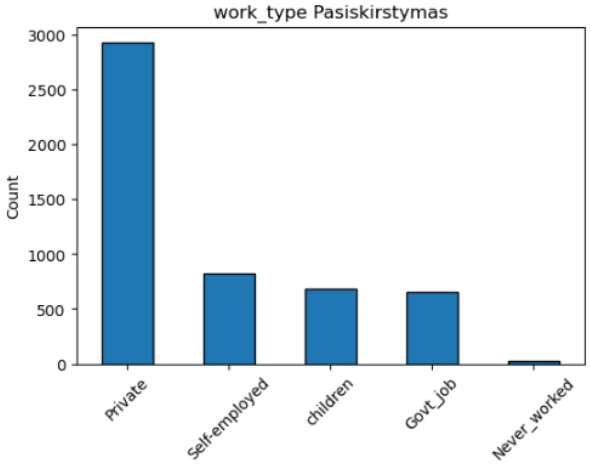
B)



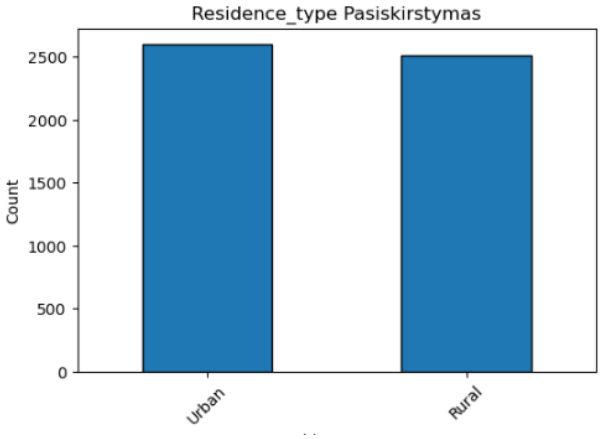
C)



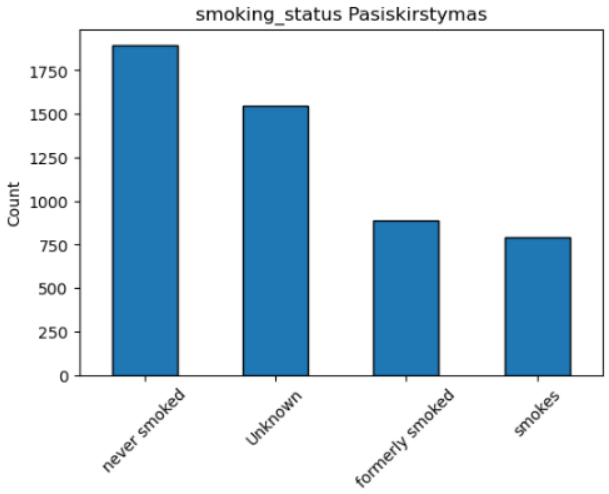
D)



E)



F)



G)

2 pav. Kategorinių tipų stulpelinės pasiskirstymo diagramos

Iš stulpelinių diagramų galime pastebėti tai, jog yra tik vienas įrašas, kuriame paciento lytis yra „Other“. Tai yra nereikalingas triukšmas duomenų rinkinyje, todėl šią reikšmę pašalinsime.

7. Duomenų rinkinio kokybės problemos ir sprendimo būdai

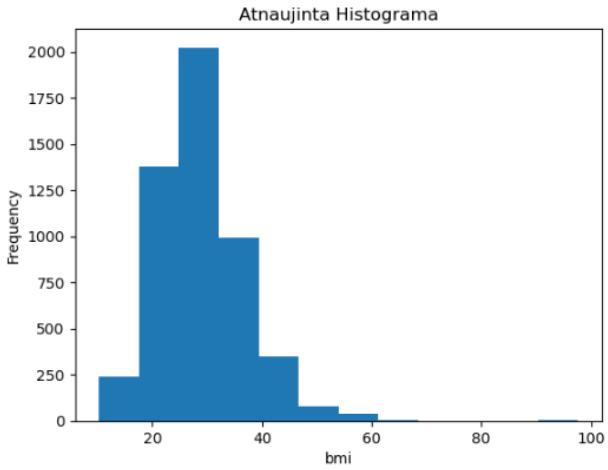
Viena iš problemų yra ta, jog 3,933% „bmi“ atributo reikšmių (201 įrašas) yra traktuojami, kaip N/A. Vienas iš galimų sprendimų yra šiuos įrašus ištrinti arba skaičiuoti populiacijos vidurkį, tačiau taip galima labai stipriai iškreipti duomenis. Geriausias būdas yra ištreniruoti maišinio mokymosi metodais grįstą algoritmą pagal kitus atributus ir taip prognozuoti „bmi“ reikšmes. Iš pradžių buvo svarstoma pritaikyti tiesinės regresijos modelį, tačiau „bmi“ atributas turi labai silpną koreliaciją su kitais atributais:

```
bmi                              1.000000
ever_married                     0.341695
age                              0.333398
work_type_Private                0.208029
avg_glucose_level                0.175502
hypertension                     0.167811
smoking_status_never smoked      0.107964
smoking_status_formerly smoked   0.107031
smoking_status_smokes            0.088324
work_type_Self-employed          0.072701
stroke                           0.042374
heart_disease                    0.041357
id                               0.003084
Residence_type                  -0.000122
gender                          -0.026678
work_type_Never_worked          -0.028602
work_type_children              -0.448674
Name: bmi, dtype: float64
```

3 pav. Koreliacijos koeficientai

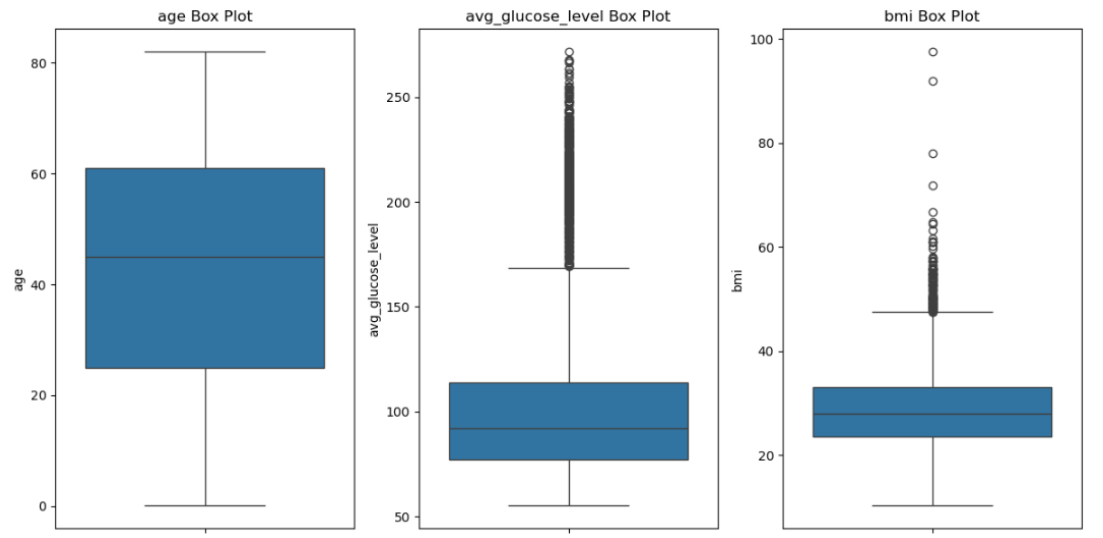
Tad buvo nuspręsta išbandyti „RandomForestRegressor“, bei XGBRegressor modelius, kuris gali geriau aptikti netiesines priklausomybes tarp atributų. Tačiau rezultatai nebuvo teigiami – modelių R^2 reikšmė pasiekdavo tik 0,2-0,3, kas reiškia, kad tik 20-30% atvejų priklausomo kintamojo kitimo regresijos modelyje paaiškinama nepriklausomais kintamaisiais. Kūno masės indeksas priklauso nuo

kitų parametrų: ūgio, svorio, fizinio pasiruošimo ir panašiai, o tokių atributų duomenų rinkinyje nėra. Tad buvo nuspręsta trūkstamas reikšmes užpildyti pagal esamų „bmi“ atributo reikšmių distribuciją. Gautas rezultatas:



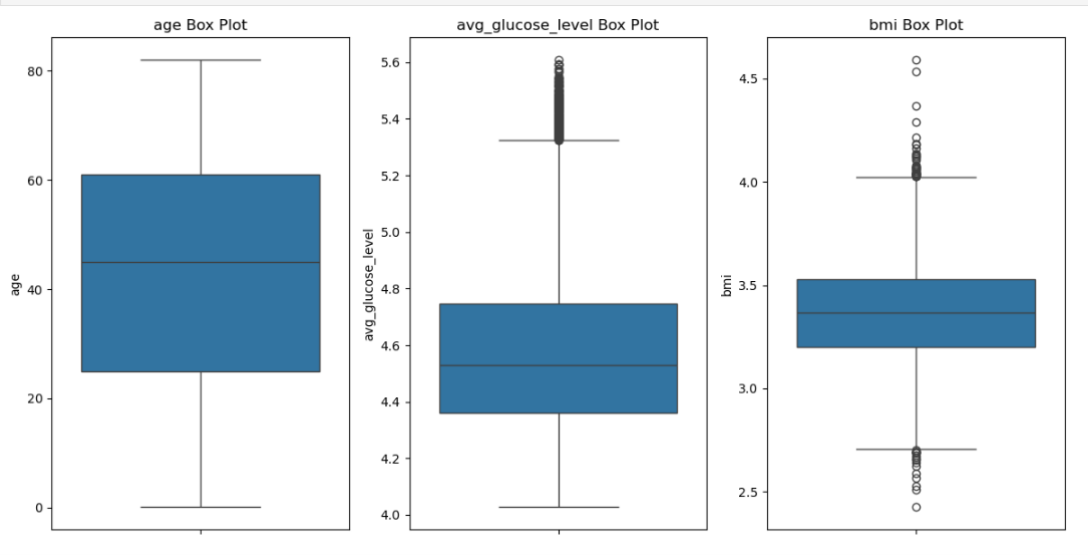
4 pav. Histograma su užpildytomis reikšmėmis

Kita problema yra ta, jog yra daug atributų „avg_glucose_level“ ir „bmi“ išskirčių:



5 pav. „Box plot “ diagramos

Šiuo atveju išskirtys yra medicinine prasme tikros ir ištrinti jų paprastai negalime. Kad sumažinti išskirtį šių duomenų pritaikysime log transformaciją. Gauti tokie rezultatai:



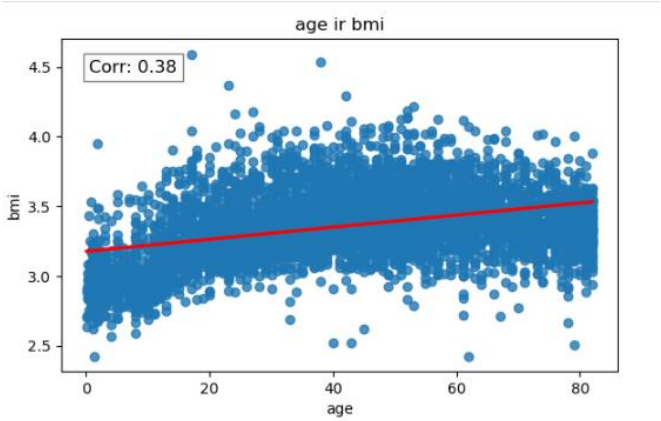
6 pav. „Box plot“ diagramos po log transformacijos

Matome, kad šis metodas pasitvirtino ir pačių išskirčių skaičius, bei atstumas tarp kvartilių tapo ženkliai mažesnis, kas bus tikrai naudinga treniruojant modelius.

8. Sąryšiai tarp atributų naudojant vizualizacijos būdus

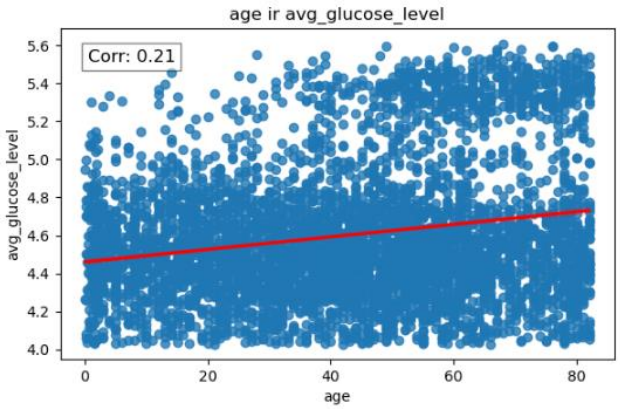
8.1. Tolydinio tipo atributų sąryšiai

Kadangi turime 3 tolydinius atributus, nubrėšime tris „scatter plot“ diagramas ir bandysime nustatyti ar yra koks ryšys tarp tų atributų.



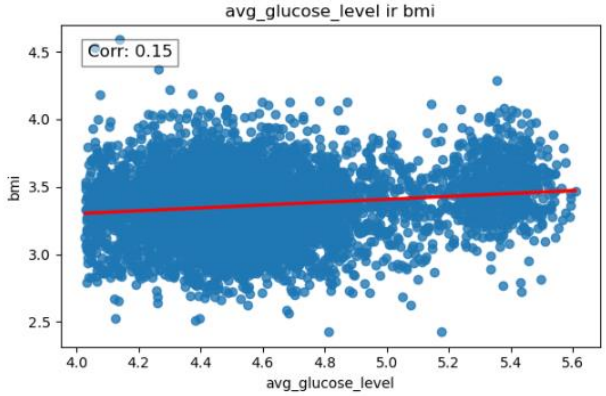
7 pav. „age“ ir „bmi“ atributų priklausomybė

Matome, kad tarp šių atributų koreliacija yra silpna, bet nedaug trūksta iki vidutinės koreliacijos diapazono. Tai logiška, nes su amžiumi dažnai mažėja fizinis aktyvumas, o metabolizmas lėtėja, todėl gali didėti kūno masės indeksas. Visgi, pastebimas didelis duomenų išsibarstymas, kas rodo, jog yra daug kitų veiksnių, darančių įtaką BMI, pavyzdžiui, mitybos įpročiai, genetika, lytis ar sveikatos būklė.



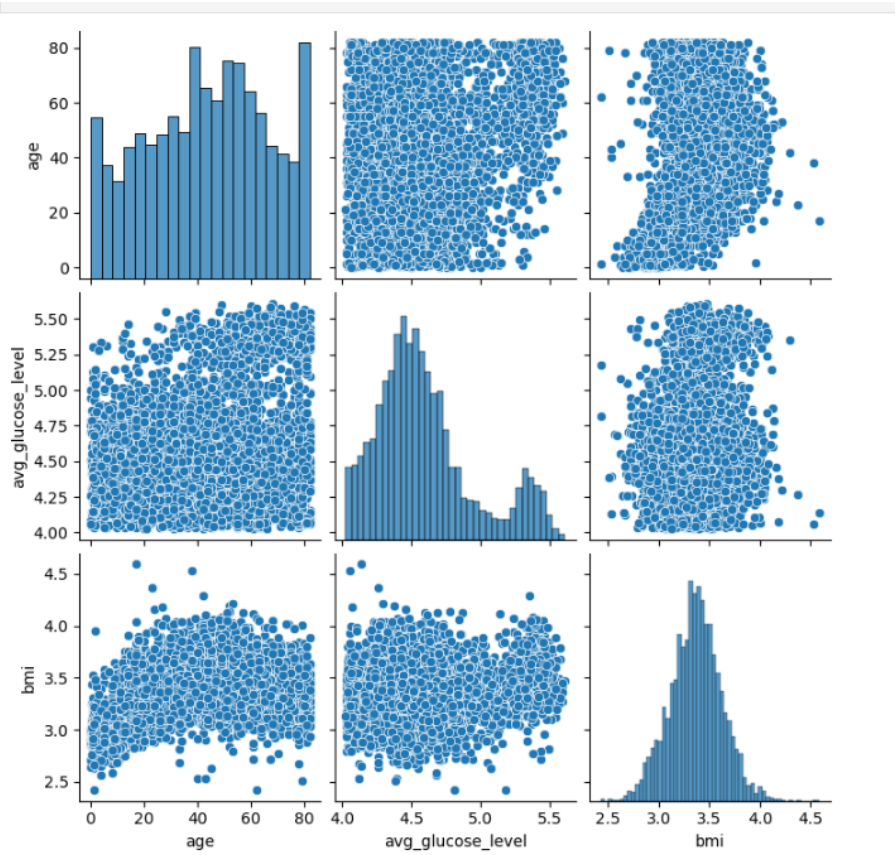
8 pav. „age“ ir „avg glucose level“ atributų priklausomybė

Šioje diagramoje matome, kad tarp amžiaus (age) ir vidutinio gliukozės lygio (avg_glucose_level) egzistuoja labai silpna teigiama koreliacija (0.21). Tai reiškia, kad su amžiumi vidutinis gliukozės lygis šiek tiek didėja, tačiau šis ryšys nėra stiprus ir duomenys yra gana išsibarstę.



9 pav. „avg glucose level“ ir „bmi“ atributų priklausomybė

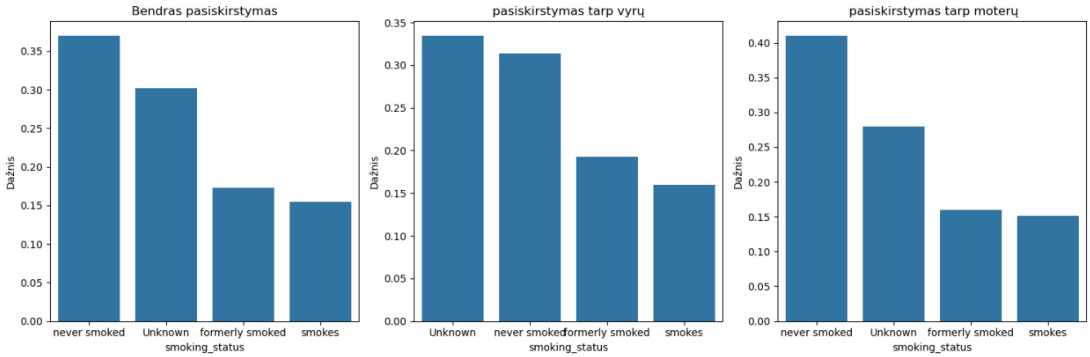
Žiūrint į šią diagramą gali pasirodyti keista, nes ryšys tarp BMI ir gliukozės lygio turėtų būti stipresnis. Tačiau įtaką BMI gali turėti kiti veiksniai, tokie kaip fizinis aktyvumas, mityba. Galimas ir netiesinis ryšys – BMI gali smarkiai kisti tik tada, kai gliukozės lygis kraujyje pasiekia tam tikrą ribą, todėl linijinė koreliacija neatskleidžia tikrojo ryšio tarp šių atributų.



10 pav. SPLOM diagrama

8.2. Kategorinio tipo atributų sąryšiai

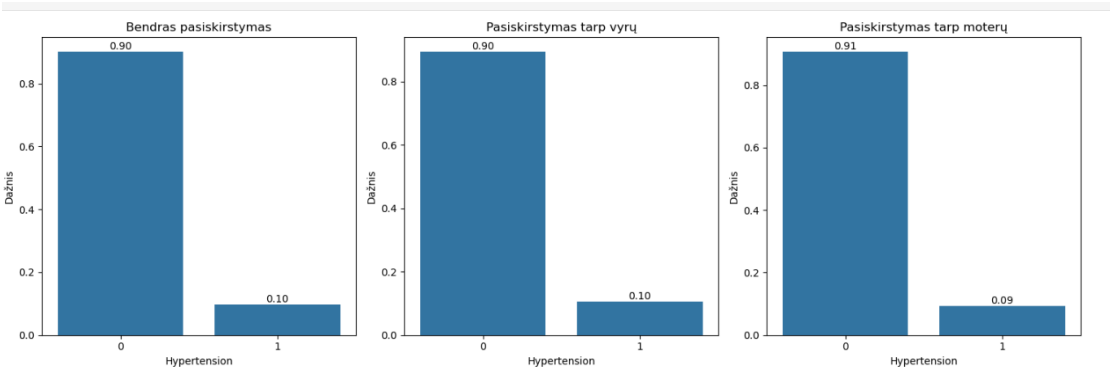
Iš pradžių pasižiūrėkime, koks yra rūkančių žmonių tipų pasiskirstymas tarp lyčių.



11 pav. Rūkymo statuso pasiskirstymas tarp lyčių

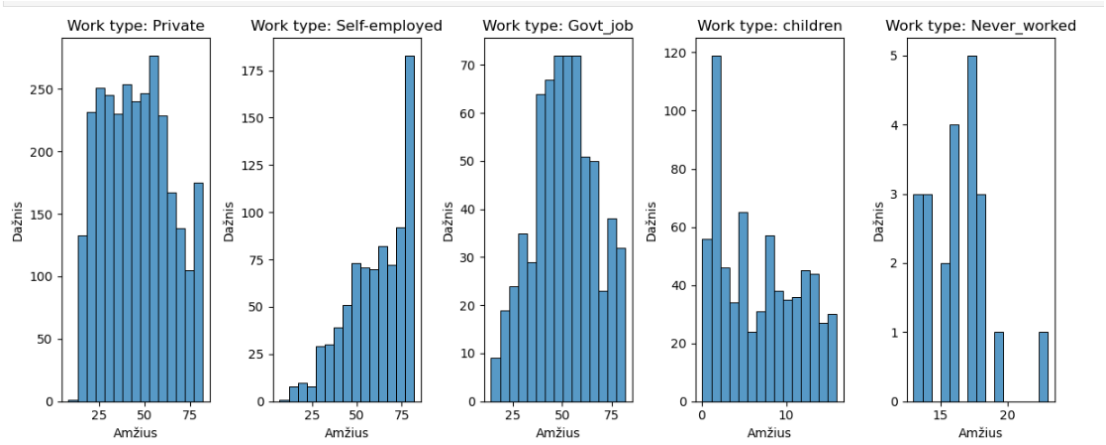
Šie grafikai aiškiai parodo, jog rūkymo įpročiai skiriasi tarp vyrų ir moterų, o vyrų grupėje dažniau pasitaiko anksčiau rūkiusių ar šiuo metu rūkančių žmonių.

Tada reikėtų ištirti koks yra hipertenzijos ligos pasiskirstymas tarp lyčių:



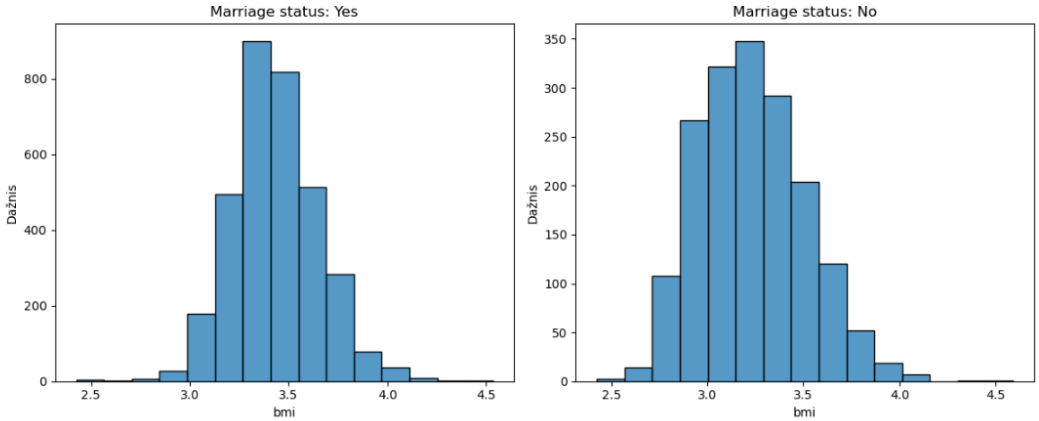
12 pav. Hipertenzijos pasiskirstymas tarp lyčių

Iš grafikų galime daryti prielaidą, jog hipertenzija (ilgalaikis kraujospūdžio padidėjimas) šiame duomenų rinkinyje pasireiškia maždaug 10% žmonių ir nėra stipraus pagrindo manyti, jog ši liga priklauso nuo lyties.



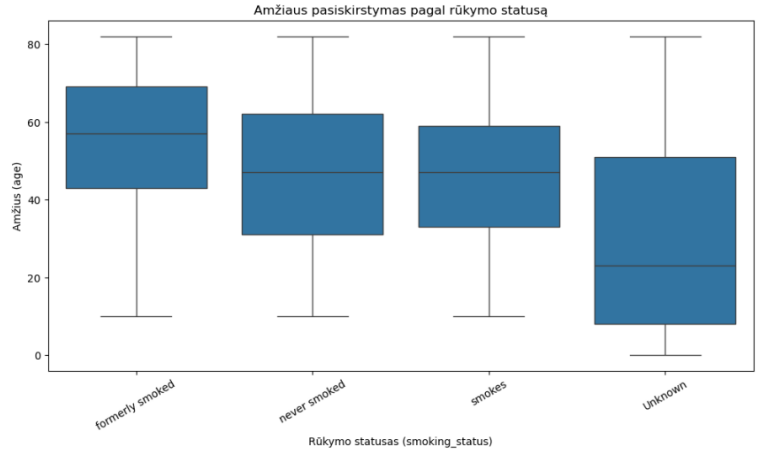
13 pav. Sąryšis tarp amžiaus ir darbo pobūdžio

Iš histogramos gan aiškiai matosi, kad darbo pobūdis stipriai priklauso nuo amžiaus, su aiškiu atskyrimu tarp dirbančių suaugusiųjų ir jaunų asmenų, kurie dar nedirba.



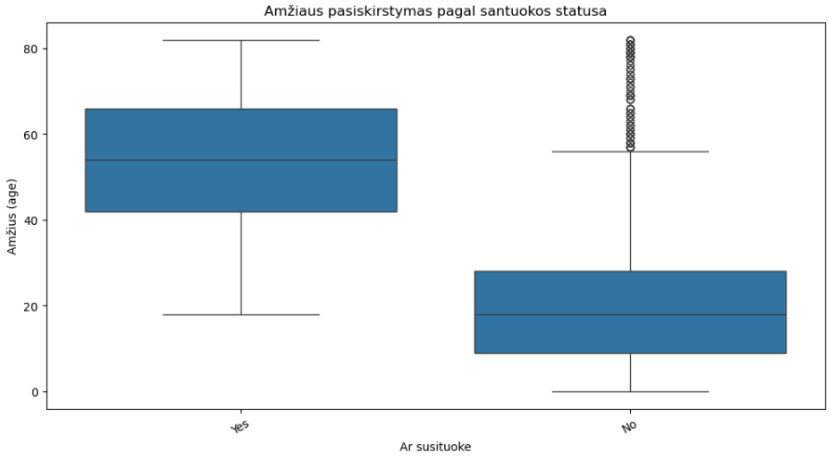
14 pav. Sąryšis tarp BMI ir santuokos statuso

Iš histogramos galime įžvelgti, kad nesusituokę žmonės yra linkę turėti šiek tiek mažesnį BMI, nei susituokę.



15 pav. Sąryšis tarp kategorinio „smoking status“ ir tolydinio atributo „age“

Rūkymo statusas ir amžius turi gan aiškų sąryšį – vyresni žmonės dažniau yra buvę rūkaliai, jaunesni žmonės dažniau priklauso „Unknown“ ar „smokes“ kategorijai.



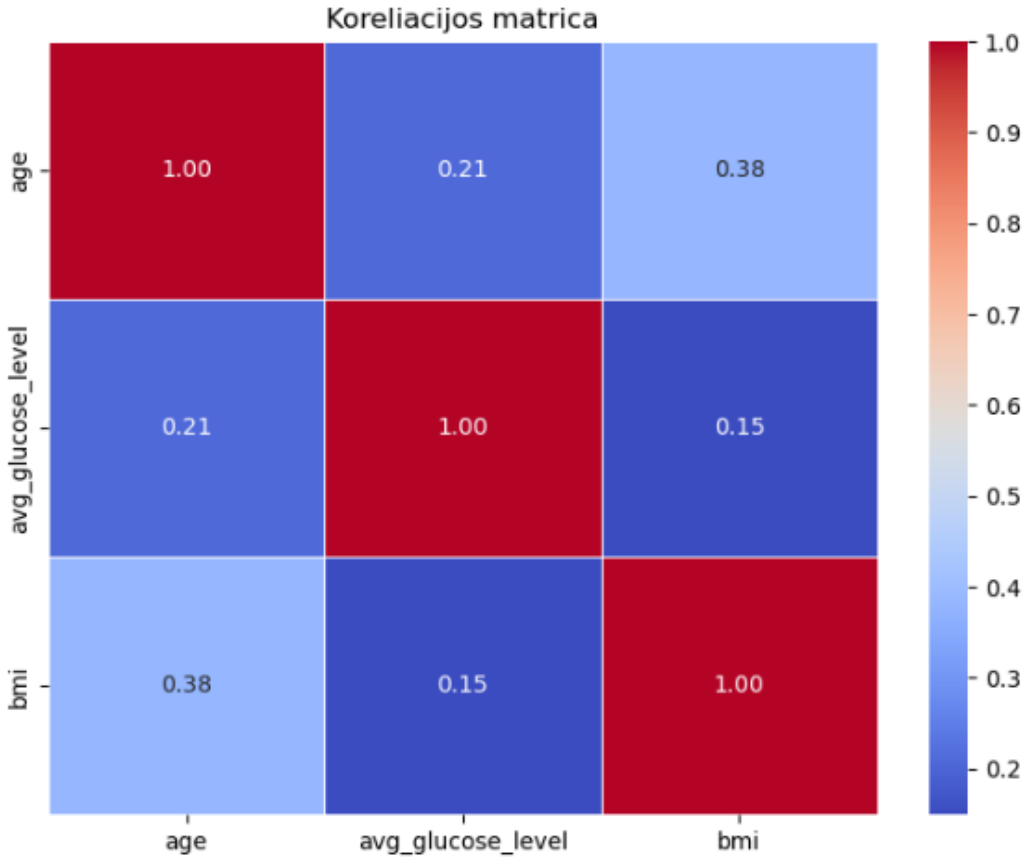
16 pav. Sąryšis tarp kategorinio „ever married“ ir tolydinio atributo „age“

Ši diagrama aiškiai parodo, kad jaunesni žmonės yra labiau linkę nesituokti, o susituokę žmonės yra vyresnio amžiaus, tačiau yra ir išskirčių.

9. Kovariacija ir koreliacija

	age	avg_glucose_level	bmi
age	511.331792	1.686503	2.214397
avg_glucose_level	1.686503	0.128678	0.013738
bmi	2.214397	0.013738	0.065584

17 pav. Kovariacijos matrica



18 pav. Koreliacijos matrica

Pagal šią koreliacijos diagramą matome, kad koreliacija tarp tolydinių atributų yra labai maža.

10.Duomenų normalizacija

Turint duomenų rinkinyje tolydines reikšmes, kurių galimų reikšmių diapazonas skiriasi, yra didelė tikimybė jog tai turės neigiamą poveikį modelio treniravimui. Šitame duomenų rinkinyje diapazonų skirtumas nėra itin didelis tarp tolydinių kintamųjų, tačiau vis tiek duomenis normalizuosime intervale [-1;1] ir naudosime formulę:

$$x_i^{norm} = \frac{x_i - \min(X)}{\max(X) - \min(X)}(b - a) + a$$

Čia min(X) – mažiausia reikšmė rinkinyje, max(X) – didžiausia reikšmė rinkinyje, a ir b – intervalo ribos.

Normalizavę duomenis gauname tokį rezultatą:

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	Male	0.633789	0	1	Yes	Private	Urban	0.782688	0.109928	formerly smoked	1
1	Female	0.487305	0	0	Yes	Self-employed	Rural	0.627737	-0.145903	never smoked	1
2	Male	0.951172	0	1	Yes	Private	Rural	-0.184595	0.003331	never smoked	1
3	Female	0.194336	0	0	Yes	Private	Urban	0.418494	0.054264	smokes	1
4	Female	0.926758	1	0	Yes	Self-employed	Rural	0.439545	-0.266875	never smoked	1
5	Male	0.975586	0	0	Yes	Private	Urban	0.523995	-0.098547	formerly smoked	1
6	Male	0.804688	1	1	Yes	Private	Rural	-0.700885	-0.149148	never smoked	1
7	Female	0.682617	0	0	No	Private	Urban	-0.328940	-0.312289	never smoked	1
8	Female	0.438477	0	0	Yes	Private	Rural	-0.597402	0.035824	Unknown	1

19 pav. Duomenų normalizavimas

11.Kategorinio tipo kintamųjų pavertimas į tolydinius

Turime 7 kategorinio tipo atributus, kuriuos reikia paversti į skaitines reikšmes: gender, ever_married, work_type, Residence_type, smoking_status.

Šiai užduočiai išspręsti bus naudojama „LabelEncoder“ biblioteka.

Gautas rezultatas išspausdinus pirmus 20 įrašų:

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	1	0.633789	0	1	1	2	1	0.782688	0.109928	1	1
1	0	0.487305	0	0	1	3	0	0.627737	-0.145903	2	1
2	1	0.951172	0	1	1	2	0	-0.184595	0.003331	2	1
3	0	0.194336	0	0	1	2	1	0.418494	0.054264	3	1
4	0	0.926758	1	0	1	3	0	0.439545	-0.266875	2	1
5	1	0.975586	0	0	1	2	1	0.523995	-0.098547	1	1
6	1	0.804688	1	1	1	2	0	-0.700885	-0.149148	2	1
7	0	0.682617	0	0	0	2	1	-0.328940	-0.312289	2	1
8	0	0.438477	0	0	1	2	0	-0.597402	0.035824	0	1
9	0	0.902344	0	0	1	2	1	-0.924530	-0.259518	0	1
10	0	0.975586	1	0	1	2	0	-0.529102	-0.077252	2	1
11	0	0.487305	0	1	1	0	0	-0.023302	0.114826	3	1
12	0	0.316406	0	0	1	2	1	-0.201388	-0.152405	3	1
13	1	0.902344	0	1	1	2	1	0.732983	-0.285527	0	1
14	0	0.926758	0	1	1	2	1	0.699610	-0.123501	2	1
15	0	0.218750	1	0	1	3	0	0.390121	-0.041852	2	1
16	1	0.560547	0	1	1	2	1	0.559968	0.131767	3	1
17	1	0.829102	1	0	1	2	1	0.741262	-0.202685	3	1
18	0	0.462891	0	0	0	2	1	-0.399430	0.138933	2	1
19	1	0.389648	0	1	0	0	1	0.717074	-0.234222	0	1

20 pav. Kintamųjų pavertimas į tolydinius

12. Išvados

1. Duomenų rinkinys turėjo trūkumų – buvo nemažai trūkstamų reikšmių ir išskirčių, tačiau pritaikius tinkamus metodus su šios problemos buvo išspręstos.
2. Koreliacija tarp kintamųjų buvo nedidelė – jie tikėtina priklauso nuo kitų veiksnių.
3. Pavyko pritaikyti duomenų normalizaciją ir kategorinių kintamųjų pavertimą tolydiniais.