 Vilniaus Universitetas

Regresinė analizė

Projektinis darbas

Darbą atliko:

Vainius Gataveckas, Matas Gaulia, Dovydas Martinkus

Duomenų Mokslas

3 kursas 2 gr.

Vilnius, 2022

**Turinys**

[Naudoti metodai 3](#_Toc102929215)

[Duomenys ir jų šaltiniai 4](#_Toc102929216)

[Tikslas ir uždaviniai 5](#_Toc102929217)

[Atliktos analizės aprašymas 6](#_Toc102929218)

[1. Naudojant R 6](#_Toc102929219)

# Naudoti metodai

Šiame darbe naudota tiesinė ir kvantilių regresijos. Taip pat naudoti apibendrintieji tiesiniai modeliai su glodniaisiais splainais, multinominis logistinės regresijos modelis. Darbas atliktas naudojant R.

Naudoti R paketai:

*tidyverse*

*rsample*

*corrplot*

*car*

*effects*

*lm.beta*

*quantreg*

*mgcv*

*gratia*

*effect*

*yardstick*

*mnet*

*themis*

*recipes*

# Duomenys ir jų šaltiniai

Gyventojų skaičiaus prieaugio prognozavimas: atskirai tiriamas natūralus pokytis ir pokytis dėl migracijos.

Duomenų aibės sudarymui panaudoti duomenys iš skirtingų šaltinių. Duomenų šaltiniai:

Our World in Data. Natūralaus ir bendro gyventojų prieaugio šalims duomenys. Prieiga per internetą: <https://ourworldindata.org/world-population-growth>

UNData. Įvairūs ekonominiai, socialiniai, su aplinkosauga ir infrastruktūra susiję šalių indikatoriai. Prieiga per internetą[: https://www.kaggle.com/datasets/sudalairajkumar/undata-country-profiles/code](:%20https:/www.kaggle.com/datasets/sudalairajkumar/undata-country-profiles/code)

World Happiness Report. Apklausomis paremti gyvenimo kokybės šalyse įvertinimų pagal skirtingus kriterijus duomenys. Prieiga per internetą: <https://www.kaggle.com/datasets/mathurinache/world-happiness-report-20152021?select=2017.csv>

Atlikus pirminį apdorojimą duomenų aibę sudaro šie požymiai:

“employment\_industry\_percent\_of\_employed" – dalis dirbančių industriniame sektoriuje.

"unemployment\_percent\_of\_labour\_force" - bedarbių dalis visoje darbo rinkoje.

"agricultural\_production\_index\_2004\_2006\_100" – šalyje pagamintų agrikultūros produktų inkesas pasvertas pagal kainas.

"urban\_population\_percent\_of\_total\_population" - miestuose gyvenanti gyventojų dalis.

"health\_total\_expenditure\_percent\_of\_gdp" – dalis BVP išleidžiama sveikatos apsaugai.

"education\_primary\_gross\_enrol\_ratio\_f\_m\_per\_100\_pop" - gyventojų, lankančių pradinį mokymąsį skaičius 100 gyventojų.

"education\_tertiary\_gross\_enrol\_ratio\_f\_m\_per\_100\_pop" – - gyventojų, lankančių aukštajį (ar kitą trečio lygio) mokymąsį skaičius 100 gyventojų.

"pop\_using\_improved\_drinking\_water\_urban" - miesto gyventojų, naudojančių geros kokybės geriamą vandenį skaičius 100 gyventojų.

"freedom" – asmeninės laisvės įvertinimas.

"generosity" – dosnumo įvertinimas (labdara, savanoriavimas ir t.t.).

"trust\_government\_corruption" - pasitikėjimo vyriausybe ir korupcijos lygio įvertinimas.

"migration\_growth" – procentinis gyventojų skaičiaus prieaugis dėl migracijos.

"natural\_growth” – procentinis natūralus gyventojų skaičiaus prieaugis (vien tik dėl mirčių ir gimimų šalyje).

"category" – vėliau sudarytas kategorinis kintamasis priskiriantis šalims klases pagal migracijos ir natūralaus gyventojų skaičiaus prieaugio reikšmes.

# Tikslas ir uždaviniai

Tikslas: Atlikti regresinę analizę natūraliam ir migracijos gyventojų prieaugiui pagal ekonominius ir socialinius šalių indikatorius.

Uždaviniai:

Požymių aibės sudarymas.

Tiesinių regresijos modelių gyventojų prieaugiui sudarymas.

Kvantilių regresijos ir apibendrintųjų adityviųjų modelių su glodniaisiais splainais gyventojų prieaugiui sudarymas.

Multinominės logistinės regresijos modelio sudarymas šalių klasifikavimui į pagal gyventojų prieaugį sudarytas klases.

Modelių tinkamumo analizė.

Kovariančių įtakos gyventojų prieaugiui įvertinimas.

Modelių panaudojimas prognozėms gauti.

# Atliktos analizės aprašymas

Pradinis duomenų apdorojimas ir analizė

Tyrimui naudoti 2017 metų duomenys. Visi naudoti duomenų rinkiniai sujungti į vieną jungiant pagal šalies pavadinimą.

Iš duomenų aibės pašalinti su migracija, gimstamumu susiję požymiai, nes su jais tiesiogiai susijusios siekiamos prognozuoti reikšmės. Taip pat pašalinti su aplinkosauga, energijos sritimi susiję požymiai, nes jie laikyti ne tokie svarbūs išsikeltam tyrimo tikslui. Papildomai atsisakyta požymių su daug praleistų reikšmių, mažais požymio reikšmių skirtumais tarp šalių.

Sudaryta duomenų aibė padalinta į apmokymo ir testavimo aibes naudojant 85-15 santykį. Iš duomenų iš anksto pašalintos stipriai tarpusavyje koreliuojančios kovariantės naudojant 0.7 Pirsono koreliacijos ribą. Kovariantės, likusios po šio ir prieš tai aprašyto požymių filtravimo, aprašytos praėjusiame skyriuje.

Sudarytas kategorinis kintamasis šalims priskiriantis klases pagal jų natūralaus ir migracijos prieaugio reikšmes (klasė 0 – teigiamas migracijos prieaugis ir teigiamas natūralus, 1 – teigiamas migracijos ir neigiamas natūralus, 2 – neigiamas migracijos ir teigiamas natūralus, 3 – neigiamas migracijos ir neigiamas natūralus).

Nubraižytos koreliacijų diagramos. Pastebėta, kad didelė dalis požymių kurie teigiamai koreliuoja su migracijos gyventojų prieaugiu neigiamai koreliuoja su natūraliu prieaugiu (ir atvirkščiai). Taip pat atskirai natūraliam ir migracijos prieaugiams nubraižytos sklaidos diagramos su regresijos kreive pagal naudojamas kovariantes. Vėl pastebėta, kad kovariančių įtaką dažnai yra priešingos krypties migracijos prieaugiui, negu ji yra natūraliam prieaugiui. Ryšiai tarp atsako ir kovariančių dažnai tiesiniai, tačiau neretais atvejais paveikti šalių išskirčių. Tiesa, braižant prieš tai minėtas regresijos kreives nėra atsižvelgiama į kitų kovariančių reikšmes todėl šie grafikai gali tik sufleruoti apie šių kovariančių įtaką pilname modelyje.

Sudaryti atskiri tiesiniai modeliai natūraliam ir migracijos gyventojų prieaugiui. Migracijos modelyje pastebėtos išskirtys pagal visus diagnostinius grafikus. Šios išskirtis atitinka ... Kadangi ... šiuos šalys pašalintos iš modelio mokymo aibės. Atlikta pažingsninė regresija. Sumažintame modelyje gauta, kad migracijos prieaugis labiausiai priklauso nuo.

Tokia pati procedūra naudota ir sudarant natūralaus gyventojų prieaugio modelį. Šį kartą ryškių išskirčių nerasta

Tikėtina, kad kovariančių įtaka nėra pastovi lyginant šalis su dideliu teigiamu tam tikro tipo gyventojų prieaugiu ir neigiamu prieaugiu, todėl naudinga nagrinėti kvantilių regresiją.

Toks regresijos modelis leidžia nesudėtingai interpretuoti kovariančių įtaką gyventojų prieaugiui, tačiau modelyje neatsižvelgiama į galimus netiesinius ryšius tarp kovariančių ir atsako ir prognozuojant šiais reikšmes galimi gauti prasti rezultatai, todėl papildomai pasirinkta sudaryti netiesinės regresijos modelius, šiuo atveju naudojant apibendrintus adityviuosius modelius su glodniaisiais splainais.

Pasirinkta sudaryti multinominės logistinės regresijos modelį, kuriuo siekiama gauti geresnius rezultatus negu prieš tai naudotais dviem atskirais regresijos modeliais, kai siekiama sužinoti tik kokiai klasei priklauso šalis (ar šalies natūralus/migrantų prieaugiai teigiami ar neigiamai). Akivaizdu, kad kitas šio modelio privalumas yra, kad gauti prognozei reikalingas tik vienas modelis vietoje dviejų.

Klasifikavimo modelio kokybė naudojant maišos matricas (angl. confusion matrices), bendrą tikslumą (angl. accuracy), F1-score. Kadangi turimas daugelio klasių (multiclass) uždavinys paskutinės minėtos modelio kokybės vertinimo metrikos bendra reikšmė gauta „macro“ vidurkinimą imant metrikos reikšmių visoms 4 klasėms vidurkį (taip kiekvienai klasei priskiriant tokį pat svorį). Taip pat nubraižytos ROC kreivės ir apskaičiuota AUC.

Kadangi duomenų rinkinyje turimos stipriai išbalansuotos klasės (klasės “1” - teigiamas migracijos prieaugis, neigiamas natūralus ir „3“ - neigiami abiejų tipų prieaugiai, sudaro mažiau nei po 5% stebėjimų), pasirinkta šią problemą spręsti sugeneruojant dirbtinų stebėjimų mažumos klasėms naudojant SMOTE algoritmą.

Naudojant testavimo duomenų aibę palyginti rezultatai gauti naudojant paprastą multinominį logistinės regresijos modelį, modelį gautą naudojant SMOTE algoritmu sugeneruotus dirbtinius stebėjimus ir klasifikavimui pritaikant du regresijos modelius.

Tiek pagal maišos matricą, tiek pagal modelio kokybės metrikas testavimo aibėje matomi geresni antrojo modelio rezultatai (modelio, kuriam apmokyti buvo naudojami SMOTE algoritmu sugeneruoti stebėjimai). Kaip ir iš anksto tikėtasi, klasifikavimui panaudojus du prieš tai sudarytus regresijos modelius gauti prasčiausi rezultatai.

Išvados:

Atlikus pirminę duomenų aibės analizę rasta, kad betono stipris gali netiesiškai priklausyti nuo jo mišinį sudarančių medžiagų, todėl laikyta, kad paprastas tiesinės regresijos modelis gali būti netinkamas prognozuoti betono stiprį.

Siekiant turėti palyginamąjį modelį sudarytas paprastas tiesinės regresijos modelis. Kaip vienas iš būdų pagerinti modelį buvo pasirinkta polinominė regresija. Į modelį buvo įtraukti antrieji laipsniai tų kintamųjų, kuriuose pagal tiesinės regresijos diagnostinius grafikus rasta neišnaudotos informacijos. Pasitelkiant grafines modelių diagnostikas ir statistinius testus laikyta, kad antru modeliu gaunami geresni rezultatai.

Siekiant dar labiau pagerinti rezultatus (ir parodyti, kad aukštesnių laipsnių įtraukimas į paprasta tiesinį modelį nėra pakankamas tikslioms prognozėms gauti) sudarytas apibendrintas adityvus regresijos modelis, naudojantis glodniuosius splainus. Pirmiausia naudotas numatytasis mazgų skaičius, tačiau naudojant modelio diagnostikas pasirinkta šį skaičių padidinti. Parinkus didesnį mazgų skaičių modelio diagnostikose beveik nerasta nukrypimų. Visos modelyje naudotos kovariantės modelyje statistiškai reikšmingos.

Testavimo aibė panaudota įvertinti visų šių modelių gebėjimą prognozuoti reikšmes. Kaip ir tikėtasi, geriausi rezultatai gauti naudojant glodniųjų splainų modelį (paklaidos pagal dvi skirtingas matavimo metrikas daugiau nei dvigubai mažesnės už tiesinį ir apie 50% mažesnės už polinominį modelį). Dėl šios priežasties daroma išvada, kad sudarytas glodniųjų splainų apibendrintas adityvus tinkamas prognozuoti betono stiprį, be to šiuo modeliu gaunami ryškiai geresni rezultatai negu naudojant paprastesnį metodą netiesiniams sąryšiams tarp atsako ir kovariančių modeliuoti, kuriuo paprastas tiesinės regresijos modelis papildomas kovariančių aukštesniais laipsniais.