

Projektas

P175B101 Intelektikos pagrindai

Darbą atliko:

Aivaras Baranauskas IFF-4/3

Paulius Patalavičius IIF-4/3

Darbo vadovas: Germanas Budnikas

Kaunas

2017

Turinys

[1. Užduotis 3](#_Toc483516166)

[2. Metodai 4](#_Toc483516167)

[ Neural Network 4](#_Toc483516168)

[ Multiple Linear Regression 4](#_Toc483516169)

[ Regression Tree 4](#_Toc483516170)

[ Support Vector Regression 4](#_Toc483516171)

[ Feature Selection 4](#_Toc483516172)

[3. Darbo eiga 5](#_Toc483516173)

[Iš Lietuvos Statistikos Departamento pateikiamos viešai prieinamos duomenų rinkmenos gauname gyventojų metinio pajamų ir gyvenimo sąlygų statistinio tyrimo duomenis. 5](#_Toc483516174)

[SelectKBest metodu, naudojant chi2(chi-squared) funkciją atrenkami reikšmingiausi gyventojų atributai (pvz: amžus, lytis, santykinė padėtis, ekonominis aktyvumas ir t.t.) 5](#_Toc483516175)

[4. Rezultatai 6](#_Toc483516176)

[5. Išvados 8](#_Toc483516177)

# Užduotis

Lietuvos gyventojų pajamų prognozavimas 2015 metais pagal asmens amžių, lytį, santuokinę padėtį, ar šiuo metu mokosi, išsilavinimą, ekonominį aktyvumą, valandų praleistų darbe skaičių, kiek laiko asmuo dirba bei kitus kriterijus.

Duomenys imami iš Lietuvos Statistikos Departamento pateikiamos viešai prieinamos duomenų rinkmenos: <http://osp.stat.gov.lt/viesos-duomenu-rinkmenos/-/asset_publisher/i2LnhXkrXAbl/content/metinio-pajamu-ir-gyvenimo-salygu-statistinio-tyrimo-?redirect=http%3A%2F%2Fosp.stat.gov.lt%2Fviesos-duomenu-rinkmenos%3Fp_p_id%3D101_INSTANCE_i2LnhXkrXAbl%26p_p_lifecycle%3D0%26p_p_state%3Dnormal%26p_p_mode%3Dview%26p_p_col_id%3Dcolumn-1%26p_p_col_pos%3D3%26p_p_col_count%3D5>

Darbo metu prognozavimui bus naudojami keli skirtingi mokymosi su mokytoju algoritmai ir lyginami jų rezultatai. Naudojami algoritmai: „Regression tree“, „Support Vector Regression“, „Neural Network”, „Multiple Linear Regression”.

# Metodai

## Neural Network

Neuroniniam tinklui realizuoti buvo naudojamas daugiasluoksnis perceptronas – kelių neuroninių sluoksnių neuroninis tinklas. Pasirinktas modelis: n įėjimo neuronų, 1 paslėptas sluoksnis taip pat iš n neuronų ir 1 išėjimo neuronas. Neuroninio tinklo mokymui naudotas Broyden-Fletcher-Goldfarb-Shanno metodas. Modelis realizuotas naudojant python kalbą, numpy ir neurolab bibliotekas.

## Multiple Linear Regression

Kartotinė linijinė regresija – tai linijinės regresijos metodo taikymas kaip norima rasti ryšį tarp kelių kintamųjų ir atsakymo pritaikant tiesinę funkciją. Realizavimui naudota python programavimo kalba ir sklearn biblioteka.

## Regression Tree

Regresinis medis – tai sprendimų medžio modelio tipas, kai prognozuojama reikšmė – realus skaičius. Realizavimui naudota python programavimo kalba ir sklearn biblioteka.

## Support Vector Regression

Palaikymo vektoriaus regresija – tai palaikymo vektoriaus mašinos (SVM) tipas, skirtas regresijai prognozuoti. Realizavimui naudota python programavimo kalba ir sklearn biblioteka. Naudotas SVR metodas – NuSVR.

## Feature Selection

Atributų parinkimas – tai reikšmingiausių atributų parinkimas taikant matematinį modelį. Atliekant atributų vertinimą buvo naudojamas SelectKBest metodas iš python sklearn bibliotekos. Jis pagal norimą funkciją įvertina atributus ir išrenka juos pagal k didžiausią vertinimą. Vertinimui naudota chi2 (chi-squared) funkcija.

# Darbo eiga

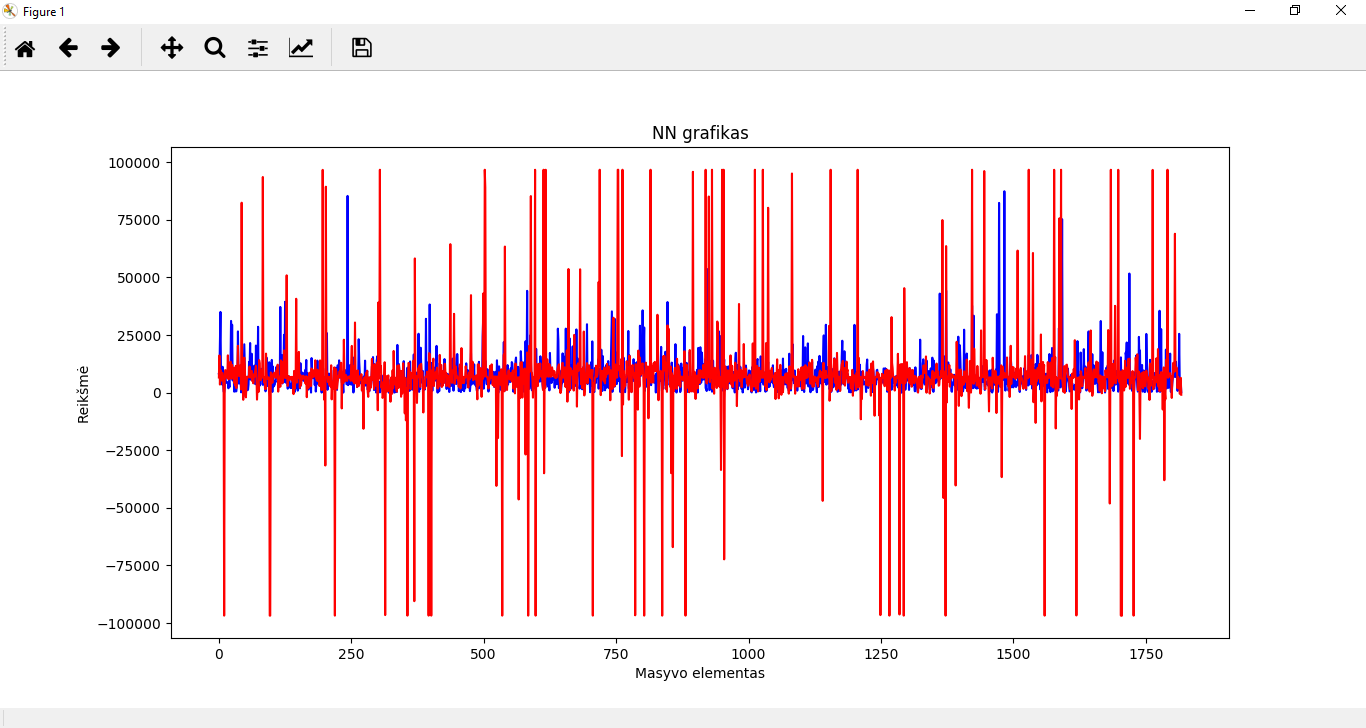
Iš Lietuvos Statistikos Departamento pateikiamos viešai prieinamos duomenų rinkmenos gauname gyventojų metinio pajamų ir gyvenimo sąlygų statistinio tyrimo duomenis.

SelectKBest metodu, naudojant chi2(chi-squared) funkciją atrenkami reikšmingiausi gyventojų atributai (pvz: amžus, lytis, santykinė padėtis, ekonominis aktyvumas ir t.t.)

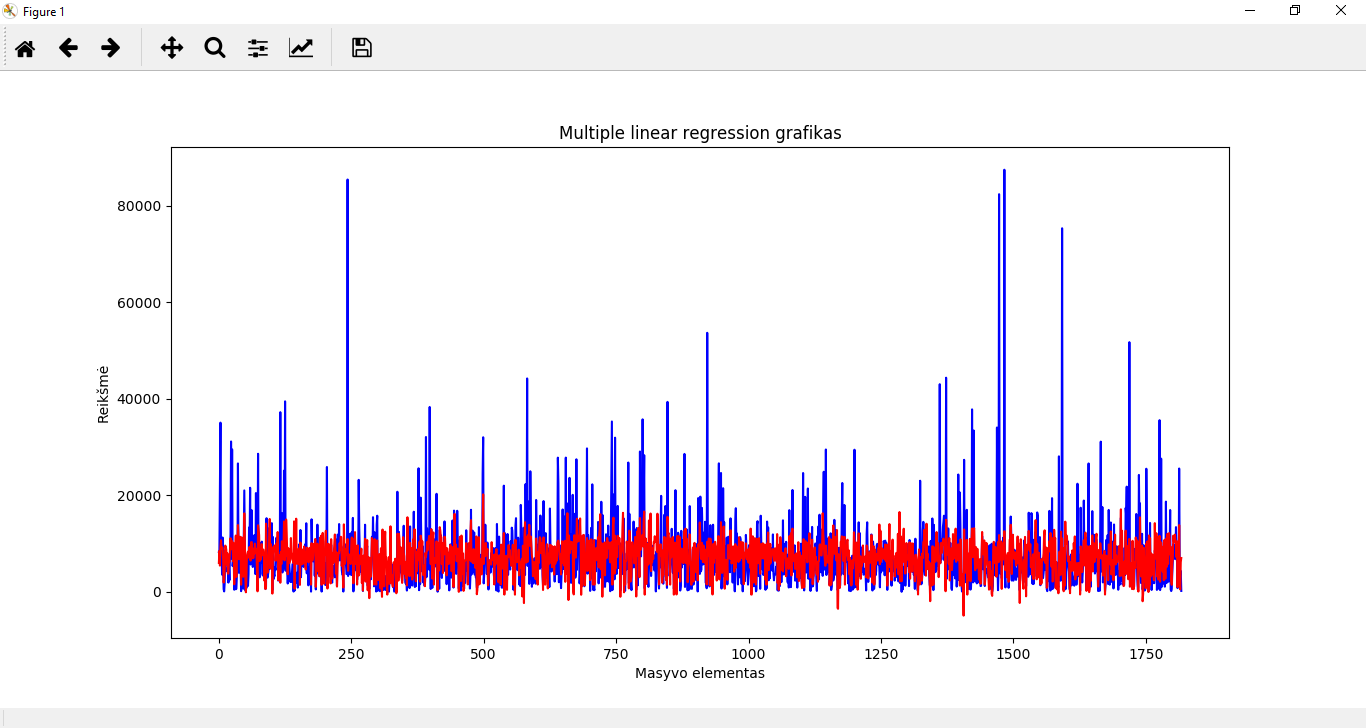
Duomenis išdaliname į 3 dalis: apsimokymo, validavimo ir testavimo. Naudodami 4 skirtingus mokymosi su mokytoju algoritmus atskirai atliekame apsimokymą, validavimą ir testavimą bei prognozojame rezultatus ir skaičiuojame vidutinį kvadratinį nuokrypį (MSE). Rezultatus apipavidaliname ir pateikiame grafikais.

# Rezultatai

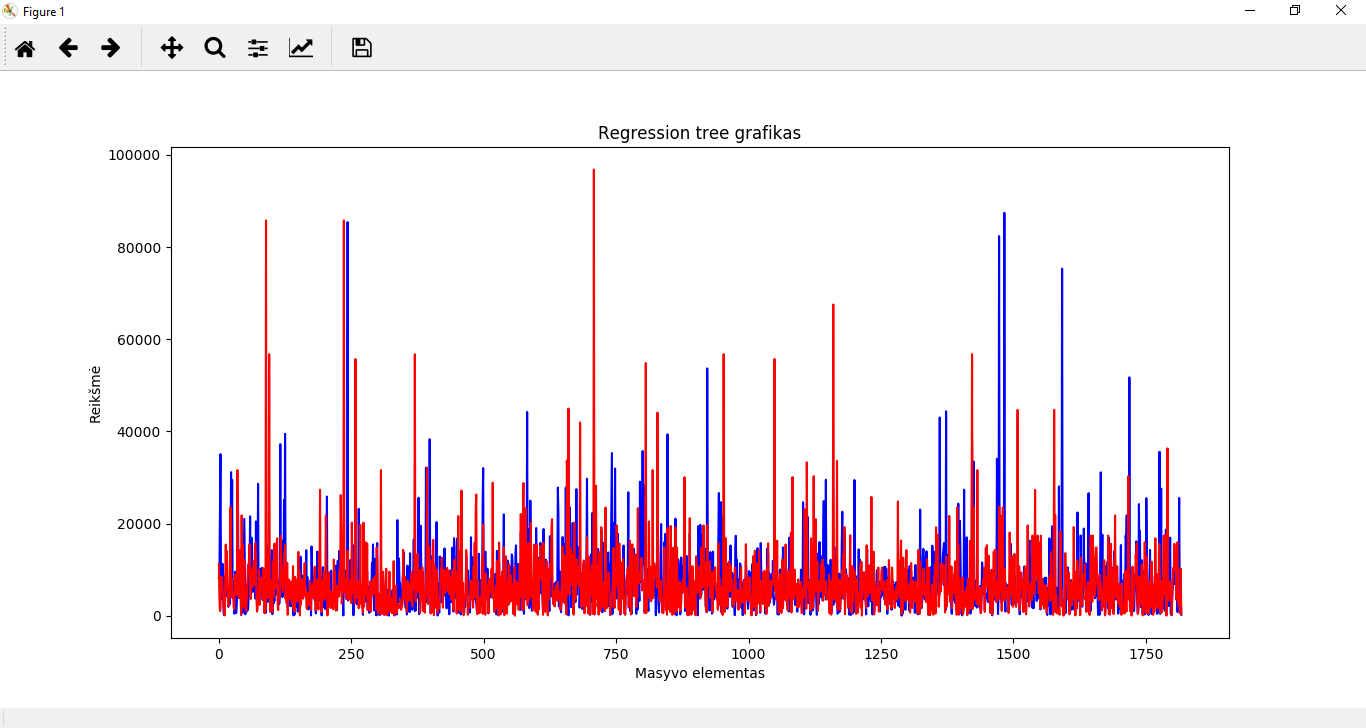
Atlikę skaičiavimus kiekvienam metodui nubraižėme po grafiką: raudona spalva vaizduoja prognozuojamas reikšmes, o mėlyna reikšmes, kurias tikimąsi gauti:



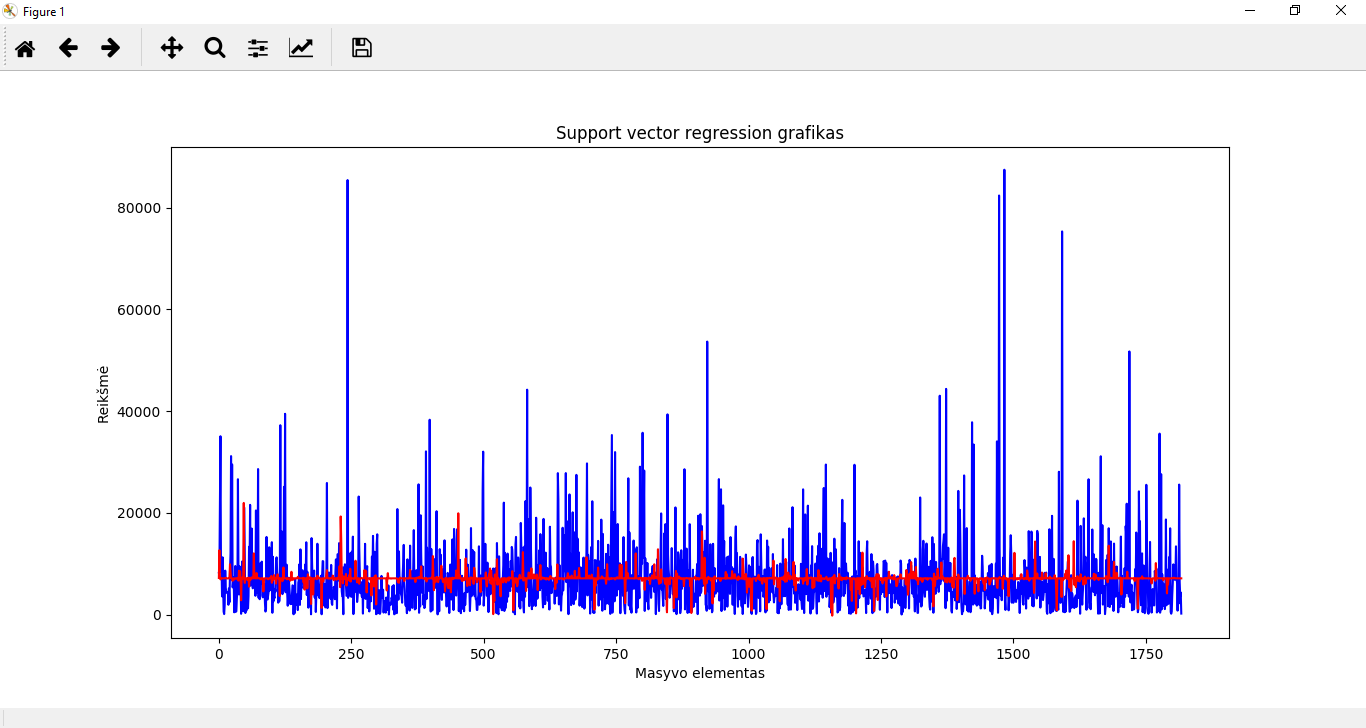
pav. 1 Neural network metodo grafikas



pav. 2 Multiple linear regression metodo grafikas



pav. 3 Regression tree metodo grafikas



pav. 4 Support vector regression metodo grafikas

Taip pat apskaičiavome kiekvieno metodo vidutinį kvadratinį nuokrypį (MSE).

MSE(Neral Network) = 20256.179078;

MSE(Multiple Linear Regression) = 6442.043060;

MSE(Regression Tree) = 9289.764933;

MSE(Support Vector Regression) = 7180.157860;

# Išvados

Projektinį darbą pavyko įgyvendinti sklandžiai: susipažinome su keturiais skirtingais mokymosi su mokytoju algoritmais: „Regression tree“, „Support Vector Regression“, „Neural Network”, „Multiple Linear Regression”. Sėkmingai pasinaudojome šiais metodais prognozuodami Lietuvos gyventojų pajamas pagal įvairius kriterijus (amžių, lytį, santuokinę padėtį, išsilavinimą, sveikatos būklę ir t.t.). Išsiaiškinome, jog efektyviausiai prognozuojamos reikšmės naudojant „Multiple Linear Regression” metodą, o Neuroninio tinkle realizacija pasirodė mažiausiai tinkama uždaviniui spręsti.