Kauno technologijos universitetas
Informatikos fakultetas

Intelektikos pagrindai (P176B101)

3 laboratorinio darbo ataskaita

Simonas Voronkevičius IFF-0/1

Studentas

lekt. Andrius Nečiūnas

Dėstytojas

Kaunas 2023

TURINYS

1. Įvadas

Tikslas yra sukurti sprendimų medžius ir jų miškus, skirtus pasirinktam duomenų rinkiniui, ir juos išbandyti. Darbo eiga apima šiuos žingsnius:

1. Reikia surasti tinkamą duomenų rinkinį, kuriame yra atributas, turintis kardinalumą intervale nuo 3 iki 10.
2. Pasirinktą duomenų rinkinį reikia apdoroti ir padalyti į dvi dalis.
3. Sudaryti sprendimų medį duomenų rinkiniui, jį pavaizduoti ir išbandyti.
4. Išbandyti kelis gylio apribojimus generuojant medį, pratestuoti sugeneruotus medžius (3-4 variacijos).
5. Sugeneruoti ir išbandyti miškus, kurie susideda iš 5 medžių su skirtingais gyliais (max gylis – gylis užfiksuotas eksperimento metu). Tada geriausią konfigūraciją reikia atrinkti.
6. Sugeneruoti ir pratestuoti miškus su medžiais nuo 3 iki 9 medžių, atrinkti geriausią konfigūraciją.
7. Palyginkite pirminio sprendimų medžio ir atsitiktinio miško gautus rezultatus.

2. Duomenys

Duomenų rinkinys paimtas iš https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles

Automobilio įvesties kintamieji:

1. Make – markė (pvz., Acura, Aston Martin, kt.)
2. Model – modelis (pvz., RLX, TSX, Q5, kt.)
3. Vehicle class – transporto priemonės klasė (pvz., sportinis (angl. *Suv-small*), kompaktiškas (angl. *Compact*), kt.)
4. Engine Size(L) – Variklio tūris (skaitinė reikšmė)
5. Cylinders – Cilindrų skaičius (skaitinė reikšmė)
6. Transmission – Pavarų dėžės tipas (pvz. AS8, A7, M6, kt.)
7. Fuel Consumption City (L/100 km) – Kuro sanaudos mieste (skaitinė reikšmė)
8. Fuel Consumption Hwy (L/100 km) – Kuro sąnaudos užmiestyje (skaitinė reikšmė)
9. Fuel Consumption Comb (L/100 km) – Vidutinės kuro sąnaudos (skaitinė reikšmė)
10. Fuel Consumption Comb (mpg) – Vidutinės kuro sąnaudos (skaitinė reikšmė)
11. CO2 Emissions(g/km) – Anglies dvideginio emisijos (skaitinė reikšmė)
12. Fuel type – Kuro tipas (pvz. Gamtinės dujos (angl. *Natural gas*), dyzelis (angl. *diesel*))

Sprendimų medžio išvesčiai pasirinktas automobilio kategorinis kintamasis yra **Kuro tipas** kurio kardinalumas yra 5.

Kuro tipo duomenys:

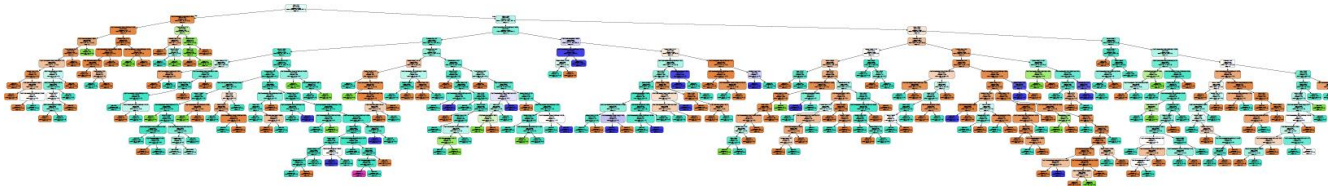
1. X - Įprastas benzinas (angl. *Regular gasoline*)
2. Z - Aukščiausios kokybės benzinas (angl. *Premium gasoline*)
3. D – Dyzelis (angl. *Diesel*)
4. E – Etanolis (angl. *Ethanol E85*)
5. N – Gamtinės dujos (angl. *Natural gas*)

Pirmiausia duomenų masyvo žodinės įvestis buvo pakeistos skaitinėmis. Po to, duomenų rinkinys buvo padalintas į treniravimo ir testavimo rinkinius. Treniravimo rinkinys buvo sudarytas iš 70% duomenų, o testavimo rinkinys - iš likusių 30% duomenų. Užduoties tikslas yra nuspėti kuro tipą pasitelkiant sprendimų medžius, todėl jį atskyrėme nuo visų duomenų rinkinių.

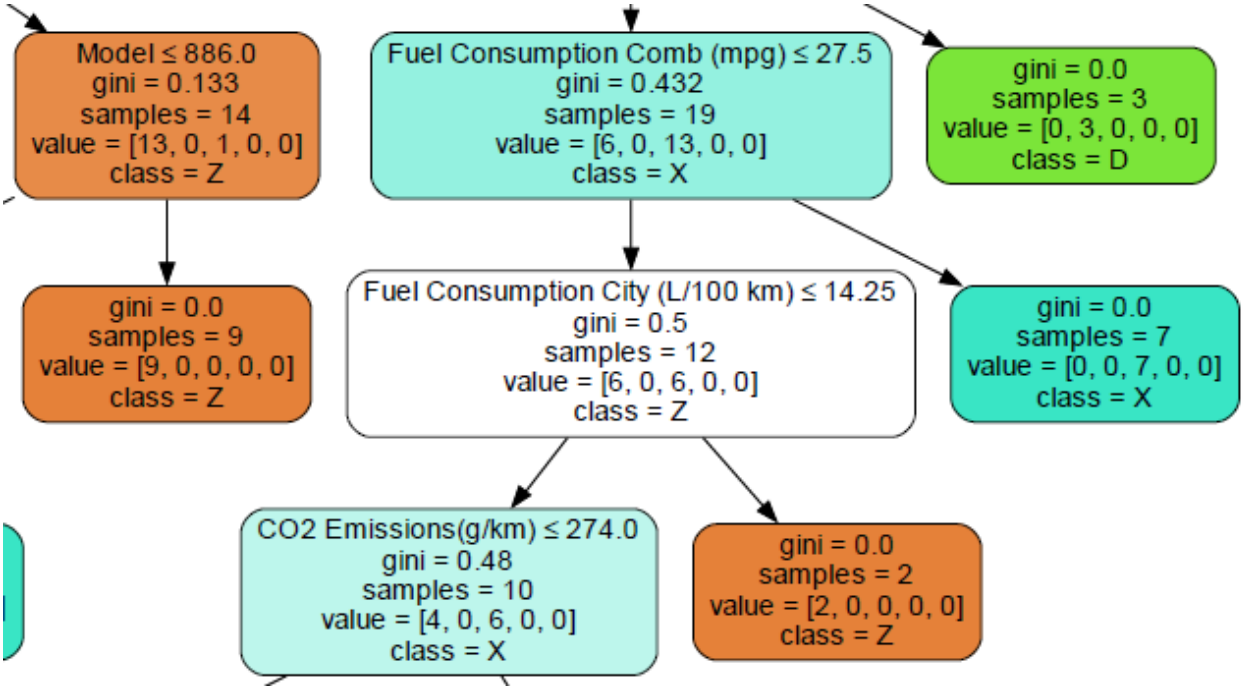
3. Sprendimų medžių realizacija

3.1. Neapriboto sprendimų medžio sudarymas

Sprendimo medis yra sukurtas neapribojant jo gylio, t.y. šakų medžio sudarymui gali būti labai daug. Gautas medžio gylis yra labai didelis (žr. psl. 5, Nuotrauka 1), todėl sprendimo medžiui atvaizduojame tik dalį medžio lapų (žr. psl. 5, Nuotrauka 2).



Nuotrauka 1 Neapriboto sprendimų medžio atvaizdavimas



Nuotrauka 2 Neapriboto sprendimų medžio dalies atvaizdavimas

Gautam medžiui apskaičiuojame vidutinį absoliutų paklaidos kvadratą (žr. lentelė 1), aboliučią reikšmę ir nuokrypį (angl. *MSE, MAE, RMSE*). Tai leidžia įvertinti prognozuotų reikšmių nuokrypį nuo tikrųjų reikšmių bei įvertinti sprendimo medžio modelio tikslumą prognozuojant reikšmei, šiuo atveju kuro tipui.

Algoritmas	**Gauta reikšmė**
Mean Squared Error (**MSE**)	0.52
Mean Absolute Error (**MAE**)	0.27
Root Mean Squared Error (**RMSE**)	0.73

lentelė 1 Neriboto sprendimų medžio apskaičiuoti statistiniai rodikliai

Šie trys statistiniai rodikliai - vidutinis kvadratinis paklaidos nuokrypis (MSE), vidutinis absoliutus paklaidos nuokrypis (MAE) ir šaknis iš vidutinio kvadratinio paklaidos nuokrypio (RMSE), rodo, jog gauname didelę paklaidą nuo tikrųjų numanomų reikšmių.

Sudarykime maišos lentelę, kuri parodys kiek neteisingai suskirstė testinės aibės elementų duotam sprendimo medžiui (žr. lentelė 2). Tada galime įvertinti kiekvieno atributo tikslumą pasitelkę šia formule $tikslumas = \frac{geri\ nustatymai}{(geri\ nustatymai + blogi\ nustatymai)}$ (žr. lentelė 3).

Kuro tipas	**Z**	**D**	**X**	**E**	**N**
Z	**968**	2	50	13	0
D	2	**44**	3	0	0
X	189	26	**830**	26	0
E	4	0	0	**58**	0
N	0	0	0	0	**0**

lentelė 2 Maišos lentelės rezultatai neapriboto gylio sprendimo medžiui

Kuro tipas	Z	D	X	E	N
Tikslumas, h = ∞	93.7%	89,8%	77,5%	92%	0%

lentelė 3 Maišos lentelės rezultatų išvedimas į tikslumą kiekvienai kategorijai

Iš lentelės matyti jog sprendimų medžio modelis Z kuro tipą gali numatyti 93,7 % tikslumu, D kuro tipą – 89,8 % tikslumu ir pan. Paskutinis N stulpelis rodo, kad modelis negali numatyti N kuro rūšies.

3.2. Riboto gylio medis

Kadangi praeitas sprendimų medis gali turėti neribotą kiekį šakų, jis tampa per daug treniruotas treniruojamų duomenų aibei, todėl norint įsitinkinti, ar galima išgauti geresnius rezultatus, mes jį galime „apkarpyti“, t.y. nustatyti maksimalų viso medžio gylį. Kad ištestuotume apribojimų efektyvumą, mes sukursime 4 medžius, kurie vienas iš jų turės atitinkamus gylio apribojimus: 4, 7, 10 ir 12. Visa darbo seka atitiks neapriboto sprendimų medžio darbo eigai.

3.2.1 4 šakų gylio medis

Pirmam medžiui taikome 4 šakų gylio ribojimą. Gautos MSE, MAE, MSER reikšmės yra prastesnės nei neriboto gylio modelio medžiui (žr. lentelė 1, 4).

Algoritmas	**Gauta reikšmė**
Mean Squared Error (**MSE**)	1.09
Mean Absolute Error (**MAE**)	0.54
Root Mean Squared Error (**RMSE**)	1.05

lentelė 4 Riboto sprendimų medžio (h=4) apskaičiuoti statistiniai rodikliai

Kuro tipas	**Z**	**D**	**X**	**E**	**N**
Z	**669**	0	363	1	0
D	12	**0**	37	0	0
X	174	0	**897**	0	0
E	23	0	5	**35**	0
N	0	0	0	0	**0**

lentelė 5 Maišos lentelės rezultatai apriboto gylio (h=4) sprendimo medžiui

Kuro tipas	Z	D	X	E	N
Tikslumas, h = 4	64.8%	0%	83.75%	55.55%	0%

lentelė 6 Maišos lentelės rezultatų išvedimas į tikslumą kiekvienai kategorijai

Iš gautų rezultatų matome (žr. lentelė 5, 6), jog apribojus medžio gyli iki 4 tik X kuro tikslumas testiniams duomenimis buvo gaunami geresni rezultatai nei neapriboto medžio modelyje. Visu kitu atvėju gauti prastesti rezultatai.

3.2.1 7 šakų gylio medis

Antram medžiui taikome 7 šakų gylio ribojimą. Gautos MSE, MAE, MSER reikšmės tokios pačios kaip ir neriboto gylio medžio modelyje (žr. lentelė 1, 7).

Algoritmas	**Gauta reikšmė**
Mean Squared Error (**MSE**)	0.52
Mean Absolute Error (**MAE**)	0.27
Root Mean Squared Error (**RMSE**)	0.73

lentelė 7 Riboto sprendimų medžio (h=7) apskaičiuoti statistiniai rodikliai

Kuro tipas	**Z**	**D**	**X**	**E**	**N**
Z	**900**	0	126	7	0
D	10	**10**	29	0	0
X	129	0	**940**	2	0
E	4	0	9	**50**	0
N	0	0	0	0	**0**

lentelė 8 Maišos lentelės rezultatai apriboto gylio (h=7) sprendimo medžiui

Kuro tipas	Z	D	X	E	N
Tikslumas, h = 7	87.1%	20.4%	87.76%	79.36%	0%

lentelė 9 Maišos lentelės rezultatų išvedimas į tikslumą kiekvienai kategorijai

Iš gautų rezultatų matome (žr. lentelė 8, 9), jog apribojus medžio gyli iki 7 tik X kuro tikslumas testiniams duomenimis buvo gaunami geresni rezultatai nei neapriboto medžio modelyje. Visu kitu atvėju gauti prastesti rezultatai. Tačiau palyginus su medžiu, kurio ribotas gylis yra 4, gaunami kur kas geresni rezultatai. Z, X kategorijos spėjimai yra kur kas tikslesni. Taip pat Šis modelis sugeba atpažinti ir D kategorijos kurą su ~20% tikslumu. Tačiau visvien yra kur tobulėti.

3.2.1 10 šakų gylio medis

Trečiam medžiui taikome 10 šakų gylio ribojimą. Gautos MSE, MAE, MSER reikšmės tokios pačios kaip ir neriboto gylio medžio modelyje (žr. lentelė 1, 7).

Algoritmas	**Gauta reikšmė**
Mean Squared Error (**MSE**)	0.54
Mean Absolute Error (**MAE**)	0.28
Root Mean Squared Error (**RMSE**)	0.73

lentelė 10 Riboto sprendimų medžio (h=10) apskaičiuoti statistiniai rodikliai

Kuro tipas	Z	D	X	E	N
Z	**953**	13	60	7	0
D	1	**23**	25	0	0
X	196	24	**849**	2	0
E	4	0	1	**58**	0
N	0	0	0	0	**0**

lentelė 11 Maišos lentelės rezultatai apriboto gylio (h=10) sprendimo medžiui

Kuro tipas	Z	D	X	E	N
Tikslumas, h = 10	92.25%	46.94%	79.27%	92.06%	0%

lentelė 12 Maišos lentelės rezultatų išvedimas į tikslumą kiekvienai kategorijai

Iš gautų rezultatų matome (žr. lentelė 11, 12), jog apribojus medžio gyli iki 10 gauname geresnius rezultatus nei praeitoje lentelėje. Išskyrus X atributą, kurio tikslumas yra ~8% sumažėjęs. Tačiau kur kas geriau įvertina E atributą, tikslumas padidėjęs nuo ~79% iki ~92%, bei D atributą, tikslumas padidėjes nuo ~20% iki ~47%.

3.2.1 12 šakų gylio medis

Paskutiniam medžiui taikome 12 šakų gylio ribojimą. Gautos MSE, MAE, MSER reikšmės yra nežymiai geresnės nei neriboto gylio medžio modelyje (žr. lentelė 1, 7).

Algoritmas	Gauta reikšmė
Mean Squared Error (**MSE**)	0.51
Mean Absolute Error (**MAE**)	0.25
Root Mean Squared Error (**RMSE**)	0.71

lentelė 13 Riboto sprendimų medžio (h=12) apskaičiuoti statistiniai rodikliai

Kuro tipas	Z	D	X	E	N
Z	**962**	0	61	10	0
D	0	**41**	8	0	0
X	182	30	**859**	0	0
E	4	0	1	**58**	0
N	0	0	0	0	**0**

lentelė 14 Maišos lentelės rezultatai apriboto gylio (h=12) sprendimo medžiui

Kuro tipas	Z	D	X	E	N
Tikslumas, h = 12	93.12%	83.67%	80.20%	92.06%	0%

lentelė 15 Maišos lentelės rezultatų išvedimas į tikslumą kiekvienai kategorijai

Iš gautų rezultatų matome (žr. lentelė 14, 15), jog apribojus medžio gyli iki 12 gauname geriausius rezultatus iš visų ribotų medžio modelių (žr. lentelė 16). Pasiektas geresnis arba toks pats tikslumas X ir E atributams, tačiau Z ir D reikšmės nežymiai atsiliko nuo neriboto sprendimų medžio modelio rezultatų. Galime teigti, jog patikimiausiai kategorizuoja atributus medis su gyliu 12.

Kuro tipas	Z	D	X	E	N
Tikslumas, h = ∞	93.7%	89,8%	77,5%	92%	0%
Tikslumas, h = 4	64.8%	0%	83.75%	55.55%	0%
Tikslumas, h = 7	87.1%	20.4%	87.76%	79.36%	0%
Tikslumas, h = 10	92.25%	46.94%	79.27%	92.06%	0%
Tikslumas, h = 12	93.12%	83.67%	80.20%	92.06%	0%

lentelė 16 Nustatytas tikslumas maišos lentelių su nustatytais medžių gyliais

3.3. Sprendimų medžių miškai

3.3.1 Geriausio gylio paieška

Tikslas yra sugeneruoti ir išbandyti miškus, kurie susideda iš 5 medžių su skirtingais gyliais. Maksimalus gylis yra riboto gylio geriausių rezultatų šakų gylis. Šią konfiguracija automatiškai paleidžiame ciklu, kuris beskaičiuojant tikrina, ar MAE rezultatas yra geresnis už paskutinį geriausią rezultatą.

Po ciklo gauname, jog Gylis 11 yra geriausias šiai konfiguracijai (žr. lentelę 17). Palyginus su gyliu 12 matome, jog MSE pagerėjo 0.10, MAE – 0.05, RMSE – 0.07.

Algoritmas	**Gauta reikšmė h = 11**	**Gauta reikšmė h = 12**
Mean Squared Error (**MSE**)	0.39	0.49
Mean Absolute Error (**MAE**)	0.20	0.25
Root Mean Squared Error (**RMSE**)	0.63	0.70

lentelė 17 Ribotų sprendimų medžių (h=11;12) apskaičiuoti statistiniai rodikliai

Palygindami Maišos lentelės rezultatus (žr. lentelė 18, 19) matome, jog sprendimų medžio modelis gyliu 11 didesniu tikslumu priskiria automobilio kuro tipą nei sprendimų medžio modelis gyliu 12. Nors ir duomenų parinkimo tikslumas D tipo kategorijai yra ~13% mažesnis, tačiau šių duomenų aibė yra maža ir nežymi visai duomenų imčiai.

Kuro tipas	**Z**	**D**	**X**	**E**	**N**
Atitikimas, h = 11	971	32	913	58	0
Atitikimas, h = 12	967	41	860	58	0

lentelė 18 Palyginti maišos lentelės rezultatai apriboto gylio (h=12;11) sprendimo medžiams

Kuro tipas	Z	D	X	E	N
Tikslumas, h = 11	93.99%	65.30%	85.25%	92.06%	0%
Tikslumas, h = 12	93.12%	83.67%	80.20%	92.06%	0%

lentelė 19 Maišos lentelės rezultatų išvedimas į tikslumą kiekvienai kategorijai

Galime teigti, jog medis gyliu 11 sudaro geriausią sprendimų medžio modelį. Tačiau tokius skirtingus MSE, MAE ir RMSE rezultatus galėjo lemti keli veiksniai. Vienas iš jų – miško skaičiaus mažinimas. Pagal nutylėjimą algoritmo miško dydis yra 1, kuris ir buvo naudojamas neribotam bei ribotam 4, 7, 10, 12 gylio medžiams, o šiuo atveju buvo partinktas atitinkamas 5 medžių miškas, kas galėjo sudaryti palankesnių rezultatų išvestį.

3.3.1 Geriausio miško dydžio paieška

Tikslas yra sugeneruoti iki 9 miškų ir rasti kuris miškas geriausiai apdoroja duomenis su jau duotu 11 šakų gyliu. Šią konfiguracija automatiškai paleidžiame ciklu, kuris beskaičiuojant tikrina, ar MAE rezultatas yra geresnis už paskutinį geriausią rezultatą.

Po ciklo gauname, jog geriausias medžių skaičius miškui yra 7 (žr. lentelę 18).

Algoritmas	**Gauta reikšmė h = 11 m.sk = 5**	**Gauta reikšmė h = 11, m.sk = 7**
Mean Squared Error (**MSE**)	0.39	0.26
Mean Absolute Error (**MAE**)	0.20	0.13
Root Mean Squared Error (**RMSE**)	0.63	0.51

lentelė 20 Ribotų sprendimų medžių (aukšis = 11, medžių sk. = 5) bei (aukšis = 11, medžių sk. = 7) apskaičiuoti statistiniai rodikliai

Palyginus su praeitais rezultatais, radome optimaliausią sprendimų medžių konfiguraciją, kuri išskirsto duomenis tiksliausiai. Visų skirtingų atributų procentalus išskirstymas pagerėjo su kiekvienu atributu.

Kuro tipas	**Z**	**D**	**X**	**E**	**N**
Atitikimas, h = 11, m.sk = 5	971	32	913	58	0
Atitikimas, h = 12, m.sk = 7	995	37	966	59	0

lentelė 21 Palyginti maišos lentelės rezultatai apriboto gylio (h=12;11) sprendimo medžiams

Kuro tipas	Z	D	X	E	N
Tikslumas, m.sk = 5	93.99%	65.30%	85.25%	92.06%	0%
Tikslumas, m.sk = 7	96.32%	75.51%	90.20%	93.65%	0%

lentelė 22 Maišos lentelės rezultatų išvedimas į tikslumą kiekvienai kategorijai

Efektyviausio miško maksimalus gylis yra 11 šakų, o miško dydis 7, šis miškas daug efektyviau bendrai geba nuspėti atributus palyginus su bet kuriuo prieš tai testuotu medžiu. Palyginus su pradiniu medžiu, kuris neturėjo nei gylio ribojimų, nei miško parametro (žr. lentelė 1, 2, 3) surinko kur kas prastesnius rezultatus. Z randa 93.7% tikslumu, kuomet nurodyto gylio ir dydžio miškas randa 96.32% tikslumu, taip pat geriau suskirsto X ir E atributus į atitinkamas kategorijas. Tik D kategorijoje geriausias rastas sprendimų medžio modelis atrėpia tikslumu 75.51%, kuomet be konfiguracijos sprendimų medžio modelis atlieka 89.8% tikslumu. Tačiau tai yra nežymus rodiklis, nes D atributo poaibis yra santykinai mažas palyginus su visa duomenu imtimi, todėl šis modelis yra prastas.

4. Šaltiniai

- https://pypi.org/project/distfit/
- https://mode.com/python-tutorial/python-histograms-boxplots-and-distributions/
- https://statisticsbyjim.com/hypothesis-testing/identify-distribution-data/
- https://plotly.com/python/splom/
- https://sciencetrends.com/what-does-no-correlation-mean-in-science/
- https://www.geeksforgeeks.org/how-to-convert-categorical-variable-to-numeric-in-pandas/
- https://pbpython.com/selecting-columns.html
- https://seaborn.pydata.org/tutorial/categorical.html
- https://www.shanelynn.ie/bar-plots-in-python-using-pandas-dataframes/
- https://www.geeksforgeeks.org/detect-and-remove-the-outliers-using-python/