KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS

Intelektikos pagrindai (P176B101)

Laboratorinis darbas Nr. 2

Atliko:

IFAi-0 gr. studentas

Martynas Tvaska

2022 m. lapkričio 4 d.

Priėmė:

Arnas Nakrošis

Agnė Paulauskaitė-Tarasevičienė

KAUNAS 2022

TURINYS

1. Duomenų rinkinys

Laboratoriniai darbui pasirinktas duomenų rinkinys duoda 200 eilučių apie pacientus ir kokie vaistai jiems tiko. Rinkinio atributai:

- Age - paciento amzius
- Sex - paciento lytis
- BP - spaudimas
- Cholesterol - cholesterolio lygis
- Na_to_K - kalis - natris
- Drug - Vaistas kuris padėjo

Duomenų pavyzdys:

	Age	Sex	BP	Cholesterol	Na_to_K	Drug
0	23	F	HIGH	HIGH	25.355	drugY
1	47	M	LOW	HIGH	13.093	drugC
2	47	M	LOW	HIGH	10.114	drugC
3	28	F	NORMAL	HIGH	7.798	drugX
4	61	F	LOW	HIGH	18.043	drugY

Nuoroda į duomenų rinkinį:

- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0101EN-SkillsNetwork/labs/Module%203/data/drug200.csv
- https://www.kaggle.com/datasets/pablomgomez21/drugs-a-b-c-x-y-for-decision-trees

2. Sprendimų medis

2.1. Pasiruošimas darbui

Darbui atlikti naudojama „sklearn“ biblioteka.

Sklearn bibliotekos funkcija „**DecisionTreeClassifier**“ medžio sudarymui naudoja CART algoritmą, medžio dalinimui numatytai GINI, bet galima rinktis tarp: GINI, ENTROPY, LOG_LOSS.

- **ID3** yra godus algoritmas. Jis kiekvienam mazgui renkasi kategorini atributą kuris suteiks daugiausia informacijos. Medžiai auginami iki maksimalaus jų dydžio. Po to medis genimas, kad pritaikyti nematytiems duomenims.

- **C4.5** yra ID3 patobulinimas, suteikia galimybę naudoti ne tik kategorinius atributus.

- **C5.0** yra tikslesnis, naudoja mažiau atminties.

- **CART** panašus į C4.5, išvestyje palaiko skaitines reikšmes.

Medžio sudarymui naudosiu **CART** algoritmą, medžio dalinimui **GINI**.

Duomenų rinkinį dalinu į dvi dalis, **70% mokymui 30% testavimui**.

Pasirenku **"Drug"** kaip prognozuojamą atributą. Jo **kardinalumas 5**

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

2.2. Medis (1)

Medžio gylis: 4

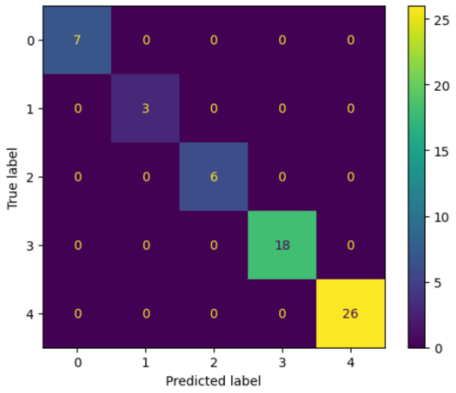
Lapų skaičius: 6

Tikslumas: 100%

Lyginami tikri duomenys su atspėtais:

	Actual	Predicted
95	drugX	drugX
15	drugY	drugY
30	drugX	drugX
158	drugC	drugC
128	drugY	drugY

Sumaišymo matrica:



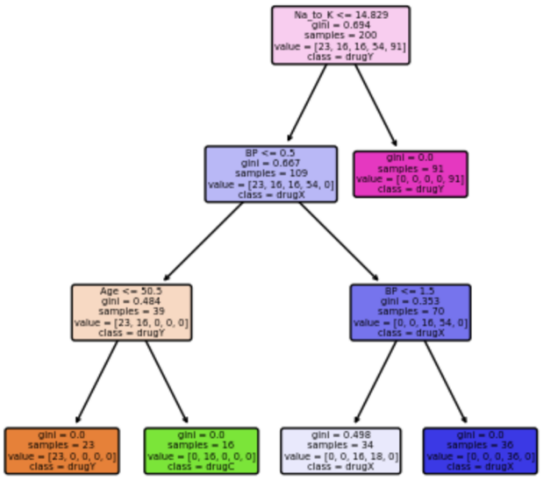
2.3. Medis (2)

Medžio gylis: 3

Lapų skaičius: 5

Tikslumas: 90%

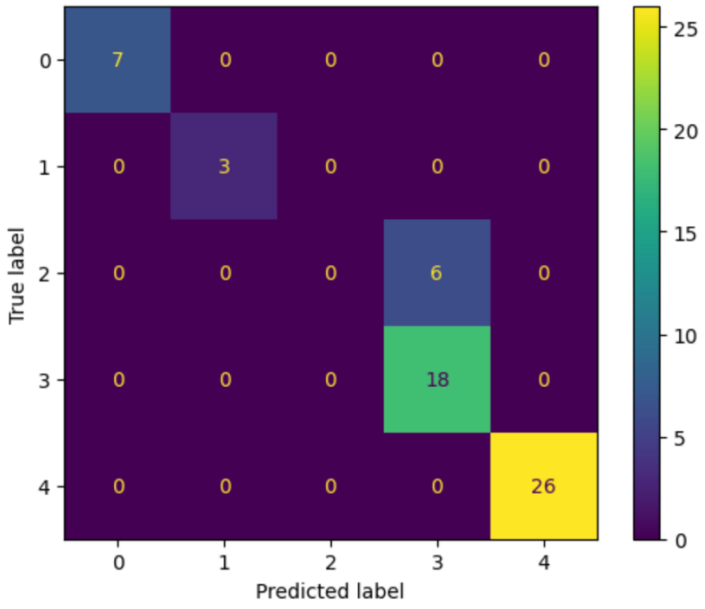
Trukmė: 0.2s



Lyginami tikri duomenys su atspėtais:

	Actual	Predicted
95	drugX	drugX
15	drugY	drugY
30	drugX	drugX
158	drugC	drugX
128	drugY	drugY

Sumaišymo matrica:



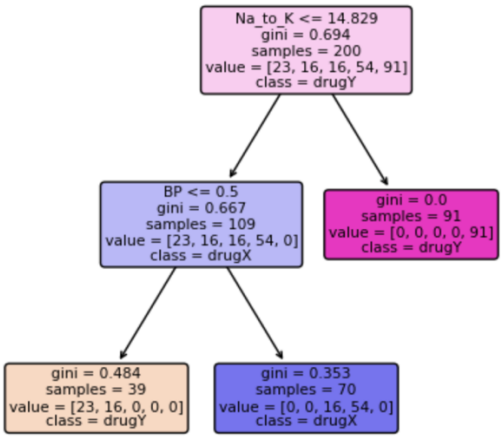
2.4. Medis (3)

Medžio gylis: 2

Lapų skaičius: 3

Tikslumas: 85%

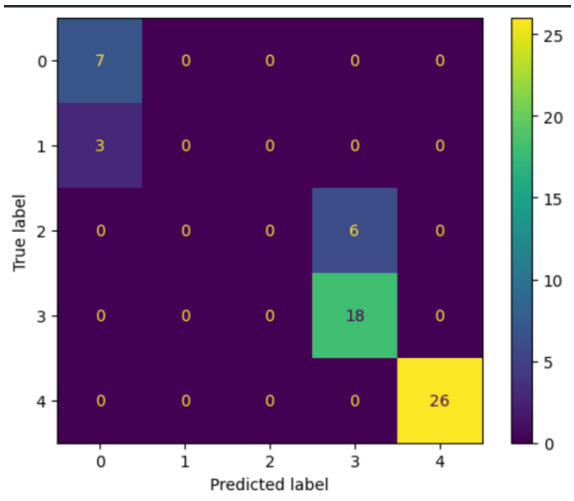
Trukmė: 0.2s



Lyginami tikri duomenys su atspėtais:

	Actual	Predicted
95	drugX	drugX
15	drugY	drugY
30	drugX	drugX
158	drugC	drugC
128	drugY	drugY

Sumaišymo matrica:



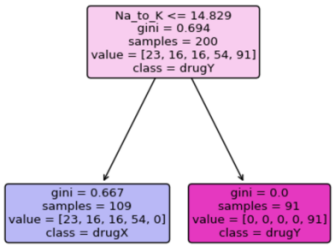
2.1. Medis (4)

Medžio gylis: 1

Lapų skaičius: 2

Tikslumas: 70%

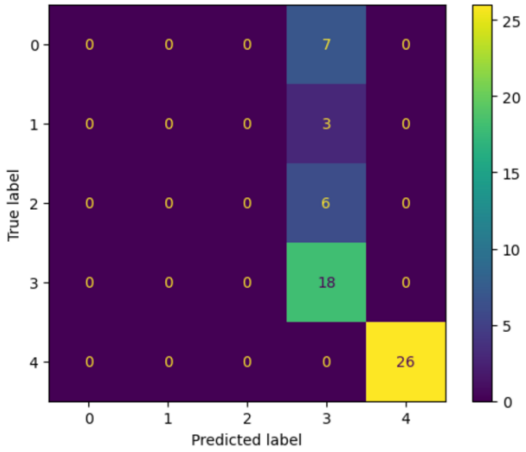
Trukmė: 0.1s



Lyginami tikri duomenys su atspėtais:

	Actual	Predicted
95	drugX	drugX
15	drugY	drugY
30	drugX	drugX
158	drugC	drugX
128	drugY	drugY

Sumaišymo matrica:



3. Atsitiktinis miškas

3.1. Pasiruošimas darbui

Sklearn bibliotekos funkcijos „**RandomForestClassifier**“ keletas pagrindinių kintamųjų:

- **n_estimators** – medžių skaičius
- **criterion** – algoritmas medžio dalinimo kokybės vertinimui (GINI, ENTROPY, LOG_LOSS)
- **max_depth** – maksimalus medžio gylis

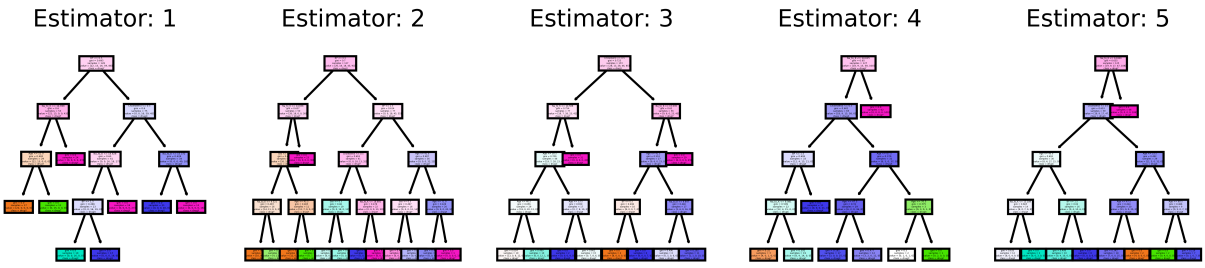
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

3.2. Atsitiktinis miškas (1)

Medžių skaičius: 5

Tikslumas: 98%

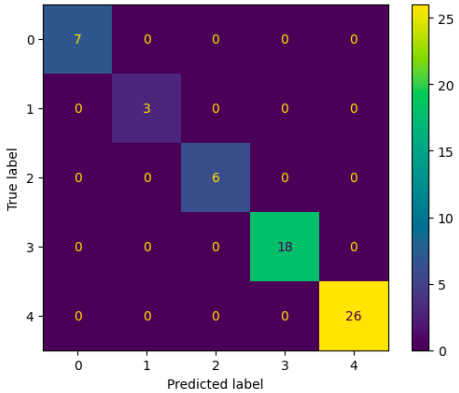
Trukmė: 2.1s



Lyginami tikri duomenys su atspėtais:

	Actual	Predicted
95	drugX	drugX
15	drugY	drugY
30	drugX	drugX
158	drugC	drugC
128	drugY	drugY

Sumaišymo matrica:



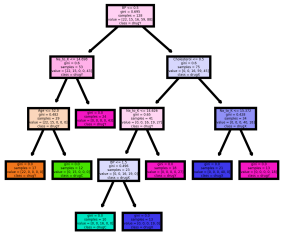
3.3. Atsitiktinis miškas (2)

Medžių skaičius: 4

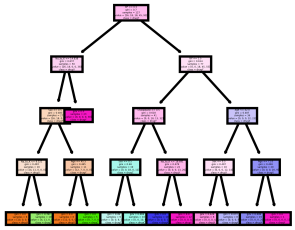
Tikslumas: 100%

Trukmė: 2.1s

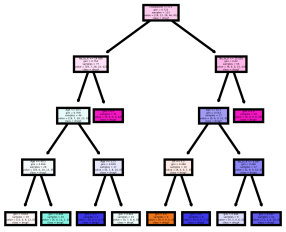
Estimator: 1



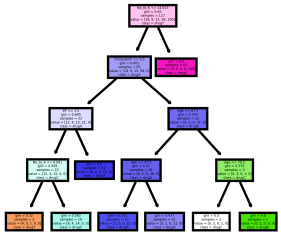
Estimator: 2



Estimator: 3



Estimator: 4



Lyginami tikri duomenys su atspėtais:

	Actual	Predicted
95	drugX	drugX
15	drugY	drugY
30	drugX	drugX
158	drugC	drugC
128	drugY	drugY

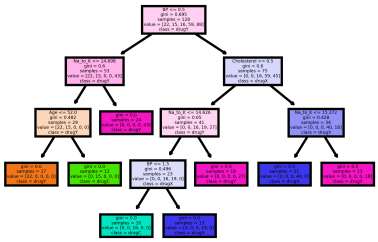
3.4. Atsitiktinis miškas (3)

Medžių skaičius: 3

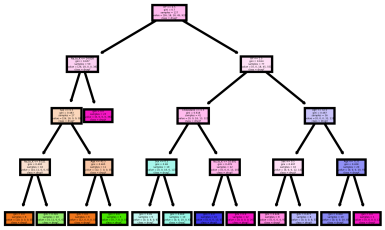
Tikslumas: 100%

Trukmė: 1.7s

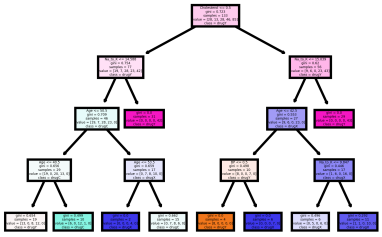
Estimator: 1



Estimator: 2



Estimator: 3



Lyginami tikri duomenys su atspėtais:

	Actual	Predicted
95	drugX	drugX
15	drugY	drugY
30	drugX	drugX
158	drugC	drugC
128	drugY	drugY

4. Palyginimas

Geriausius rezultatus pateikė medis (1).
Tikslumas 100%. Trukmė: 0.2s

Atsitiktinis miškas (3), geriausias iš miškų:
Tikslumas 100%. Trukmė 1.7s

Mano atveju norint gauti 100% tikslumą, užtenka sudaryti medį kurio gylis 4, naudojant CART algoritmą.

5. Išvados

Pagal gautuosius rezultatus matome, kad šitas duomenų rinkinys yra „paruoštas“ klasifikavimo uždaviniams.

Dažnu atveju gaunamas labai didelis tikslumas, ko realiame gyvenime negali būti.