Kauno technologijos universitetas
Informatikos fakultetas

Intelektikos pagrindai (P176B101)

Antro laboratorinio darbo ataskaita

Lukas Kuzmickas IFF-1/6

Studentas

Arnas Nakrošis

Dėstytojas

Kaunas 2024

Turinys

Paveikslų sąrašas

Lentelių sąrašas

Įvadas

Darbo užduotis:

1. Pasirinkti duomenų rinkinį ir atributus su kuriais bus sudaromas sprendimų medį;
2. Kaip sprendimų medžio išvestį pasirinkti prognozuojamą atributą;
3. Turimą duomenų rinkinį suskaidyti į apmokymo bei testavimo poaibius;
4. Suskaidyti duomenų poaibius į įvestis ir išvestis;
5. Naudojant apmokymo duomenų rinkinį, sudaryti sprendimų medį;
6. Grafiškai atvaizduoti gautą sprendimų medį;
7. Ištestuoti sudarytą sprendimų medį naudojant testavimo duomenis ir apskaičiuoti prognozavimo tikslumą/paklaidą. Nurodyti, kokią paklaidos metrika buvo skaičiuota. Taip pat klasifikavimo uždaviniui pateikti susimaišymo matricą;
8. Keičiant maksimalų medžio gylį, eksperimentiniu būdu išmatuoti skirtingų gylių (3-4 variacijos) medžių formavimo trukmė bei gautą tikslumą;
9. Naudojant tą patį apmokymo ir testavimo duomenų imties pasiskirstymą kaip ir formuojant sprendimų medį, suformuoti atsitiktinį mišką, kurį sudaro 5 medžiai;
10. Keičiant mišką sudarančių medžių kiekį [3-9], nustatyti geriausius rezultatus pateikiantį atsitiktinį mišką;
11. Palyginti pirminio sprendimų medžio ir atsitiktinio miško gautus rezultatus.

1. Duomenų rinkinio pasirinkimas ir paruošimas

Šiame laboratoriniame darbe naudojamas kitoks duomenų rinkinys pasiekiamas nuoroda: https://www.kaggle.com/code/qusaybtoush1990/wine-quality/notebook. Šiame duomenų rinkinyje pateikiamos „Vinho Verde“ vyno kokybės duomenų variantas. Duomenų rinkinys apibūdina įvairių vyne esančių cheminių medžiagų kiekį ir jų poveikį jo kokybei.

Norint panaudoti šį duomenų rinkinį pirmiausiai reikia šiek tiek jį apdoroti: pašalinti įrašus neturinčius tam tikrų atributų reikšmių bei atributus unikaliai identifikuojančius duomenis (jei tokių yra), duomenis normalizuoti, atributus suskirstyti į skaitinius ir kategorinius.

```
Attribute's "fixed acidity" unique value count: 91
Attribute's "volatile acidity" unique value count: 135
Attribute's "citric acid" unique value count: 77
Attribute's "residual sugar" unique value count: 80
Attribute's "chlorides" unique value count: 131
Attribute's "free sulfur dioxide" unique value count: 53
Attribute's "total sulfur dioxide" unique value count: 138
Attribute's "density" unique value count: 388
Attribute's "pH" unique value count: 87
Attribute's "sulphates" unique value count: 89
Attribute's "alcohol" unique value count: 61
Attribute's "quality" unique value count: 6
Attribute's "Id" unique value count: 1143
```

1 pav. Visų atributų kardinalumai

Iš 1 pav. galima pastebėti, kad šiame duomenų rinkinyje egzistuoja vienintelis kategorinis atributas – *quality*. Visi kiti atributai yra skaitiniai. Tačiau turime unikalų atributą Id, kurį privalome pašalinti. Kategorinis atributas turi labai mažą kardinalumą, todėl pasirenkame būtent jį.

Galiausiai reikia duomenis normalizuoti. Tai reiškia, kad visų atributų reikšmės bus perskaičiuotos į intervalą $[min; max]$. Buvo pasirinktas intervalas $[0; 1]$. Normalizavimui naudota formule matoma (1):

$$x_{new} = \frac{x_{old} - \min(x_{old})}{\max(x_{old}) - \min(x_{old})} \cdot (1 - 0) + 0 = \frac{x_{old} - \min(x_{old})}{\max(x_{old}) - \min(x_{old})} \quad (1)$$

Pritaikius (1) matomą formulę, gauto normalizuoto duomenų rinkinio fragmentas matomas 2 pav.:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	0.247788	0.397260	0.00	0.068493	0.106845	0.149254	0.098940	0.567548	0.606299	0.137725	0.153846	0.4
1	0.283186	0.520548	0.00	0.116438	0.143573	0.358209	0.215548	0.494126	0.362205	0.209581	0.215385	0.4
2	0.283186	0.438356	0.04	0.095890	0.133556	0.208955	0.169611	0.508811	0.409449	0.191617	0.215385	0.4
3	0.584071	0.109589	0.56	0.068493	0.105175	0.238806	0.190813	0.582232	0.330709	0.149701	0.215385	0.6
4	0.247788	0.397260	0.00	0.068493	0.106845	0.149254	0.098940	0.567548	0.606299	0.137725	0.153846	0.4
...	...	...	...	...	...	...	...	...	...	...	...	...
1138	0.150442	0.267123	0.13	0.095890	0.106845	0.417910	0.120141	0.416300	0.535433	0.251497	0.400000	0.6
1139	0.194690	0.342466	0.08	0.068493	0.093489	0.402985	0.113074	0.472834	0.535433	0.293413	0.169231	0.6
1140	0.141593	0.328767	0.08	0.075342	0.130217	0.462687	0.134276	0.354626	0.559055	0.149701	0.323077	0.4
1141	0.115044	0.294521	0.10	0.089041	0.083472	0.567164	0.159011	0.370778	0.614173	0.257485	0.430769	0.6
1142	0.115044	0.359589	0.12	0.075342	0.105175	0.462687	0.134276	0.396476	0.653543	0.227545	0.276923	0.4

2 pav. Normalizuoto duomenų rinkinio fragmentas.

Norint sudaryti sprendimų medį, reikia pasirinkti jo išvestį – prognozuojamą atributą. Kadangi prognozuojamas atributas turėtų būti kategorinis, o naudojamas duomenų rinkinys turi tik vieną kategorinį atributą – *quality*, tai jis ir buvo pasirinktas, kaip sprendimų medžio išvestis.

Kadangi duomenų rinkinyje nėra tiek daug įrašų (1143) ir norima gauti neblogą prognozę, tai buvo pasirinkta duomenis skaidyti dažnai praktikoje naudojamu principu: 70% duomenų bus skiriami apmokymui, o 30% – testavimui. Duomenys į apmokymo bei testavimo poaibius bus skirstomi atsitiktinai, tačiau taip, kad visų skirtingų atributo *quality* reikšmių kiekiai procentais šiuose poaibiuose būtų kuo artimesni visų skirtingų atributo *quality* reikšmių kiekiams procentais visame duomenų rinkinyje. Tai bus daroma norint išvengti situacijos, kai apmokymo poaibyje nėra pakankamai duomenų reikalingų teisingai priimti sprendimus apie duomenis testavimo poaibyje. Taip tikimasi pagerinti sprendimų medžio prognozių tikslumą.

Atributai, kurie bus naudojami, kaip sprendimų medžio įvestys:

- fixed acidity;
- volatile acidity;
- citric acid;
- residual sugar;
- chlorides;
- free sulfur dioxide;
- total sulfur dioxide;
- density;
- pH;

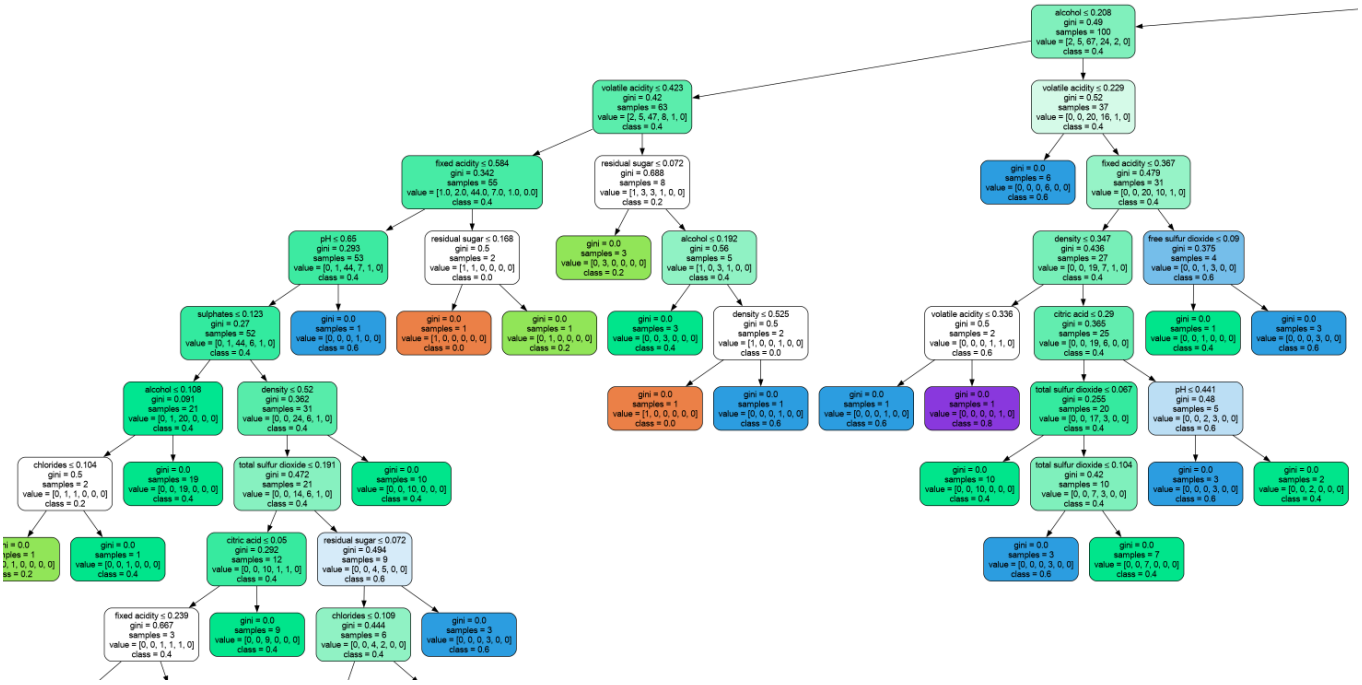
- sulphates;
- alcohol.

Atributai, kurie bus naudojami, kaip sprendimų medžio išvestys:

- quality.

2. Sprendimų medžio sudarymas ir testavimas

Sprendimo medžiui sudaryti naudota „python“ programavimo kalbos biblioteka „sklearn“. Medžiui suformuoti buvo pasirinkta naudoti, gini koeficientą skaičiuojantį, CART algoritmą. Kadangi gautas medis yra pakankamai didelis, tai 3 pav. pateiktas tik jo fragmentas:



3 pav. CART algoritmu sudaryto sprendimų medžio fragmentas.

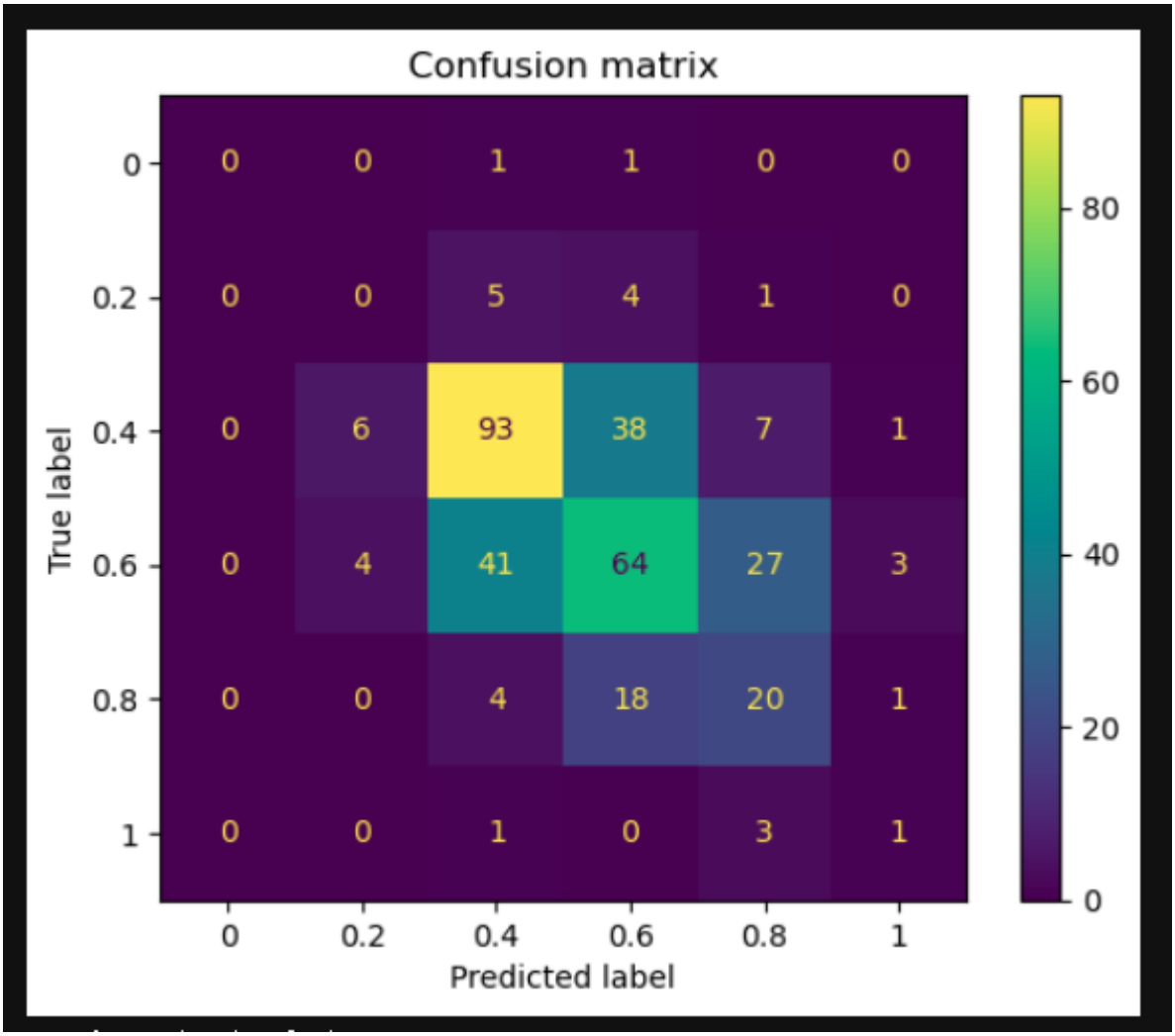
Iš 3 pav. pateikto sprendimų medžio fragmento galima pastebėti, kad medžio mazgai yra skaidomi tol, kol pasiekiamas maksimalus atsakymų aibės homogeniškumas (atsakymų aibėje yra duomenys, kurių atributo *quality* reikšmė sutampa). Tai ir nulemia sprendimų medžio dydį. Gali būti, kad taip sudarytas medis labiau pritaikytas teisingai prognozuoti atsakymą pagal mokymosi poaibio duomenis, nes apsimokymo poaibis išskaidomas į labai mažus poaibius, turinčius vos po kelis elementus, ir pagal juos nusprendžiama išvesties reikšmė.

Kaip sprendimų medžio įvestį naudojant testavimo duomenų aibę gauti tokie rezultatai: 178 kartus išvesties reikšmė prognozuota teisingai, 166 kartus išvesties reikšmė prognozuota neteisingai. Suformuoto medžio prognozavimo tikslumas apskaičiuotas naudojantis (2) formule yra lygus 51,74%.

$$\frac{teisingų\ prognozių\ kiekis}{visų\ prognozių\ kiekio} \cdot 100 \quad (2)$$

Kadangi gautas prognozavimo tikslumas $> 50\%$, tai galima sakyti, kad sudarytas sprendimų medis prognozuoja išvestį šiek tiek tiksliau, nei ji būtų prognozuojama spėliojant atsitiktinai. Tikslumas greičiausiai nėra didesnis dėl persimokymo.

Atliekant prognozes gauta susimaišymo matrica matoma 4 pav.:



4 pav. Susimaišymo matrica.

Iš 4 pav. matoma, kad daugiausiai teisingų prognozių buvo padaryta apie įrašus, kurių atributo *quality* reikšmė yra lygi 0,4, o daugiausiai neteisingų prognozių buvo padaryta apie įrašus, kurių atributo *quality* reikšmė yra lygi 0,6. Apie įrašus, kurių *quality* reikšmė lygi 1, 0,2 arba 0 prognozės labai netikslios. Taip gali būti dėl to, nes pasirinktame duomenų rinkinyje yra mažai tokių įrašų.

Siekiant pagerinti sudaryto sprendimų medžio rezultatus bei išvengti persimokymo, galima

bandyti apriboti maksimalų medžio gylį. 1 lentelėje pateikti skirtingų maksimalių gylių medžių gauti rezultatai bei formavimo laikai:

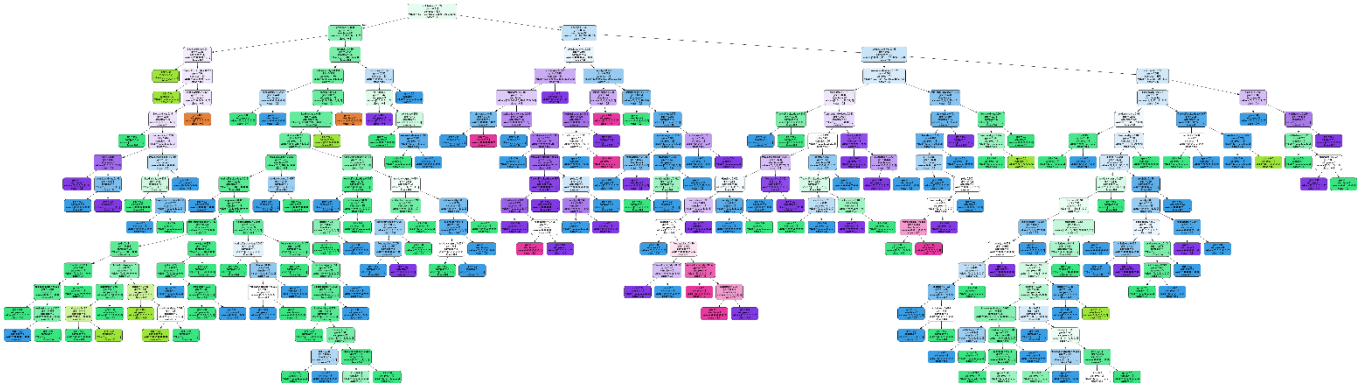
1 lentelė. Skirtingų maksimalių gylių medžių gauti rezultatai bei formavimo laikai.

Maksimalus medžio gylis	Prognozių tikslumas	Formavimo laikas
3	50,87%	0,003s
7	51,45%	0,005s
13	51,45%	0,006s
17	51,74%	0,006s

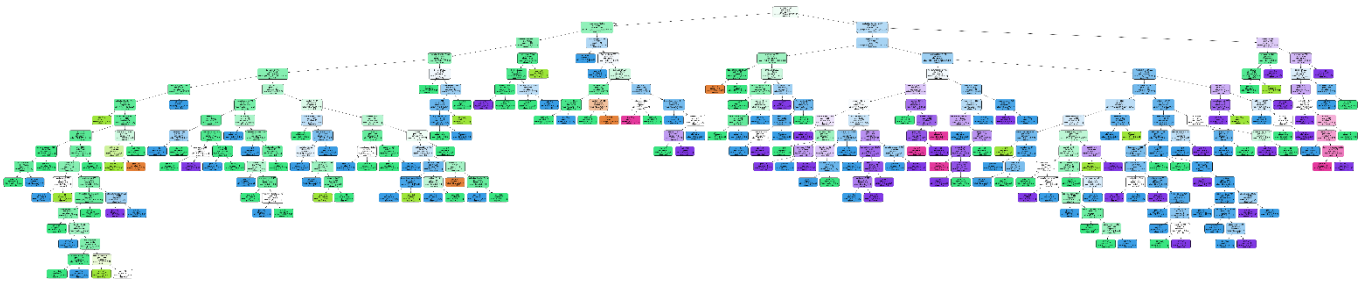
Iš 1 lentelės galima pastebėti, kad medis kurio maksimalus gylis yra 17, gauna geresni rezultatą negu orginalus medis, sutaupydamas skaičiavimo resursų.
Vadinasi maksimalaus sprendimų medžio gylio apribojimas pasiteisino.

3. Atsitiktinio miško sudarymas ir testavimas

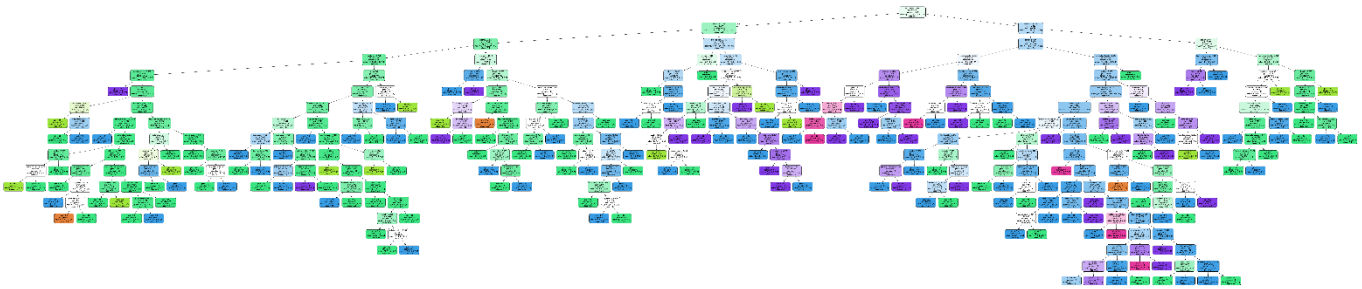
Atsitiktiniam miškui sudaryti taip pat buvo naudota „python“ programavimo kalbos biblioteka „sklearn“. Atsitiktinio miško sudarymui pritaikytas CART algoritmas bei duomenų saviranka (angl. bootstraping). Suformuotas atsitiktinis miškas turintis 5 medžius, kurių maksimalus gylis yra 17 (su šiuo gyliu gauti geriausi prieš tai sudaryto sprendimų medžio rezultatai). Gauto atsitiktinio miško medžiai yra matomi 5-9 pav.:



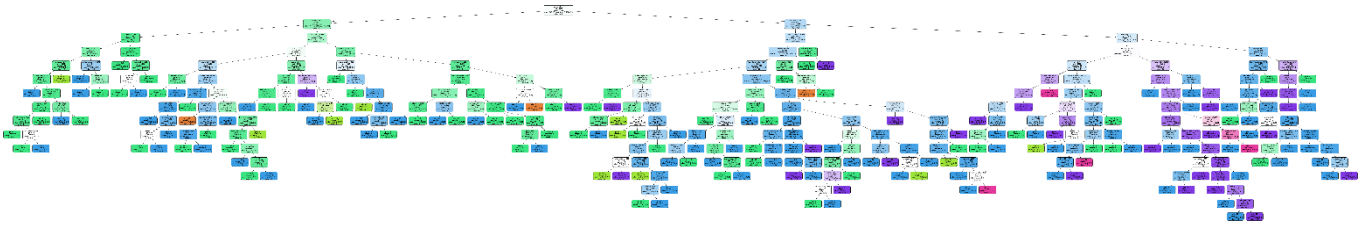
5 pav. Pirmas gauto atsitiktinio miško medis.



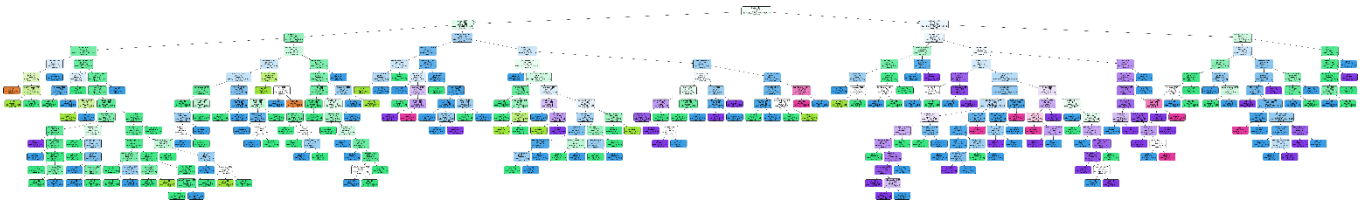
6 pav. Antras gauto atsitiktinio miško medis



7 pav. Trečias gauto atsitiktinio miško medis



8 pav. Ketvirtas gauto atsitiktinio miško medis



9 pav. Penktas gauto atsitiktinio miško medis

Iš aukščiau pateiktų paveikslėlių matoma, kad gauti atsitiktinio miško medžiai yra gana skirtingi savo sandara. Taip gali būti todėl, nes jiems sudaryti buvo naudota prieš tai minėta duomenų saviranka (angl. bootstraping). Tai daroma dėl to nes norima išgauti, kad kiekvienas medis įvertintų įvestį skirtingu būdu.

Kaip atsitiktinio miško įvestį naudojant testavimo duomenų aibę gauti tokie rezultatai: 199 kartus išvesties reikšmė prognozuota teisingai, 145 kartus išvesties reikšmė prognozuota neteisingai. Suformuoto atsitiktinio miško prognozavimo tikslumas apskaičiuotas naudojantis (2) formule yra lygus 57,85%.

Bandant pagerinti sudaryto atsitiktinio miško prognozių tikslumą, bus keičiamas mišką sudarančių medžių kiekis. Šio eksperimento rezultatai matomi 2 lentelėje:

2 lentelė. Rezultatai gauti keičiant atsitiktinio miško medžių kiekį.

Medžių skaičius	Prognozių tikslumas	Formavimo laikas
3	53,78%	0,008s
4	57,85%	0,009s
6	58,72%	0,013s
7	59,3%	0,015s
8	61,63%	0,017s
9	59,88%	0,019s

Iš 2 lentelės galima pastebėti, kad visų sudarytų miškų prognozių tikslumas yra panašus. Geriausias tikslumas išgaunamas, kai atsitiktiniame miške yra 8 medžiai. Formavimo laikas nuo medžių skaičiaus ilgėja nežymiai.

4. Sprendimų medžio ir atsitiktinio miško rezultatų palyginimas

Sprendžiant klasifikavimo užduotį atsitiktinio miško metodu išgaunamas šiek tiek geresnis prognozių tikslumas nei sprendžiant jį naudojant sprendimų medį, tačiau atsitiktinio miško formavimas užtrunka daugiau laiko, nes reikia sudaryti ne vieną medį, o daug.

Išvados

1. Sprendimų medžių bei atsitiktinių miškų prognozių tikslumas labai priklauso nuo pasirinktame duomenų rinkinyje egzistuojančių sąryšių bei pačios duomenų rinkinio kokybės. Vieni duomenų rinkiniai duoda gerus rezultatus, kiti – prastus;
2. Atsitiktinių miškų gaunami rezultatai dažniausiai bus geresni nei sprendimų medžių, tačiau jiems sudaryti reikės daugiau resursų.