Sprendimų Medžiai

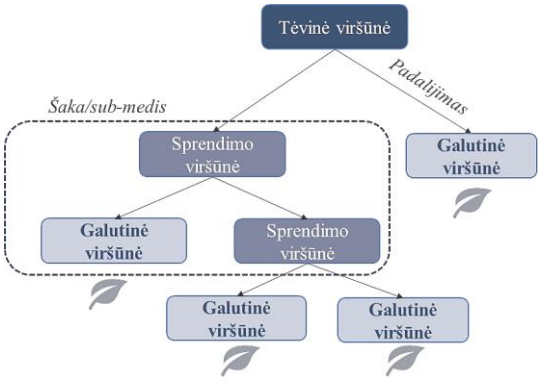
Turinys

1 Kas yra sprendimo medžiai ir kaip jie veikia

Sprendimo medžiai yra prižiūrimo (angl. *supervised*) apsimokymo tipo algoritmas, kuris yra dažniausiai naudojamas klasifikavimo uždaviniams spręsti. Kaip sprendimų medžių privalumai gali būti išskirti šie: (1) Lengva suprasti; (2) Greitas būdas duomenų analizei; (3) Laisvas duomenų tipas, nes galima apdoroti tiek skaitines, tiek kategorines reikšmes; (4) Reikalingas minimalus duomenų tvarkymas (arba kitaip valymas).

Sprendimų medžio terminologija apima 5 pagrindinių terminus (Pav 1.):

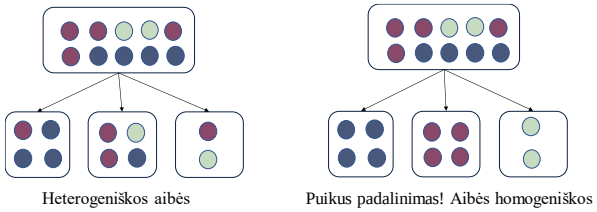
- **Tėvinė viršūnė** - aukščiausia medžio viršūnė, atvaizduojanti visą populiaciją, kuri toliau suskaidoma į dvi ar daugiau aibių, pagal tam tikras charakteristikas.
- **Padalijimas** (angl. *Splitting*) - atvaizduoja viršūnės padalinimo į dvi ar daugiau sub-viršūnių procesą.
- **Sprendimo viršūnė** (sub-viršūnė) - viršūnė, kuri padalinama į kitas sub-viršūnes.
- **Lapas** (galutinė viršūnė) - neskaidoma viršūnė. Tai reiškiasi, kad tai baigtinė viršūnė.
- **Šaka/sub-medis** - medžio tam tikra dalis.



Pav 1. Sprendimų medžių terminologijos grafinis atvaizdavimas

Sprendimų medžio konstravimo proceso tikslas – populiacijos skaidymo metu maksimizuoti aibės elementų homogeniškumą.

Kas yra homogeniškos ir heterogeniškos aibės? IT srityje, jeigu mes turime vieno tipo objektus tai sakome, kad tai homogeniška aibė, o priešingu atveju heterogeniška. Duomenų rinkinys sprendimų medžiuose yra interpretuojamas lygiai taip pat. Kaip matyti 2 paveiksle pateiktas pavyzdys, kuriame reikia padalinti visus kamuoliukus į tris grupes pagal tam tikrą atributą. Tikslas – gauti tris homogenšikas aibes pagal spalvą, tačiau kamuoliukų išskirstymą atliekant remiantis kitais juos apibudinančiais atributais. Tarkime, pavyzdžio kairėje pusėje kamuoliukai padalinti pagal svorį į tris aibes. Tačiau gautas rezultatas atsižvengiant į sprendimų medžio tikslą nėra geras, nes gautos trys heterogeniškos aibės. Štai paveikslėlio dešinėje pusėje, tarkime įvyko padalinimas į tris klases pagal atsparumą karščiui ir gautas puikus padalinimas nes pasiektas 100% homogeniškumas (grynumas).



Pav 2. Duomenų padalinimas (išskaidymas): homogeniškos ir heterogeniškos aibės

2 Sprendimų medžių sudarymo etapai

Sprendimų medžio sudarymas susideda iš trijų svarbiausių etapų:

(1) Pirmiausia, kai turimas duomenų rinkinį su kuriuo dirbsime, reikia jį išskaidyti į dvi dalis: apsimokymo imtį ir testavimo imtį. Prižiūrimo tipo mašininio mokymo algoritmai iš pradžių su vienais duomenimis turi būti apmokomi, o paskui su kitais testuojami. Paprastai apsimokymo imtis sudaro apie 70% visų duomenų, o likusioji 30 % yra paliekama testavimui.

(2) Sprendimu medžio atveju, su pasirinkta apmokymo imtimi yra sukonstruojamas pilnas medis;

(3) Tikrinamas sukonstruoto medžio tinkamumas (paprastai tikslumas) su naujais, nematytais duomenimis, kurie sudaro testavimo imtį ir todėl šis etapas vadinamas *testavimu*.

Sprendimų medžio sudarymo pavyzdys „Žinduolis“.

2.1 Duomenų išskaidymas

Paimkime pavyzdį ir pradėkime nuo pirmo etapo. Pirmasis etapas turint duomenis juos padalinti į atskiras dvi grupes: apmokymo imtį ir testavimo imtį. Pavyzdžio duomenų rinkinį sudaro 20 elementų, todėl apmokymo imčiai paimta 14 duomenų, o testavimo imtį sudaro paskutiniai 6 elementai.

1 lentelė. Duomenų rinkinio pavyzdys. Išskaidymas į apmokymo ir testavimo imtis.

Apmokymo imtis:

Pavadinimas	Kūno temperatūra	Maitinimosi tipas	Keturios kojos	Žiemoja	Klasės tipas. Ar žinduolis?
Karvė	šiltakraujis	žolėdis	Taip	Ne	Taip
Šuo	šiltakraujis	visaėdis	Taip	Ne	Taip
Šikšnosparnis	šiltakraujis	visaėdis	Ne	Taip	Taip
Mėlynasis Banginis	šiltakraujis	mėsėdis	Ne	Ne	Taip
Krokodilas	šaltakraujis	mėsėdis	Taip	Taip	Ne
Komodo varanas	šaltakraujis	mėsėdis	Taip	Ne	Ne
Žaltys	šaltakraujis	mėsėdis	Ne	Taip	Ne
Lašiša	šaltakraujis	visaėdis	Ne	Ne	Ne
Erelis	šiltakraujis	mėsėdis	Ne	Ne	Ne
Egzotinė žuvis gupija	šaltakraujis	visaėdis	Ne	Ne	Ne
Meška	šiltakraujis	visaėdis	Taip	Taip	Taip
Ožka	šiltakraujis	žolėdis	Taip	Ne	Taip
Liūtas	šiltakraujis	mėsėdis	Taip	Ne	Taip
Jūros ūdra	šiltakraujis	visaėdis	Taip	Ne	Taip

Testavimo imtis:

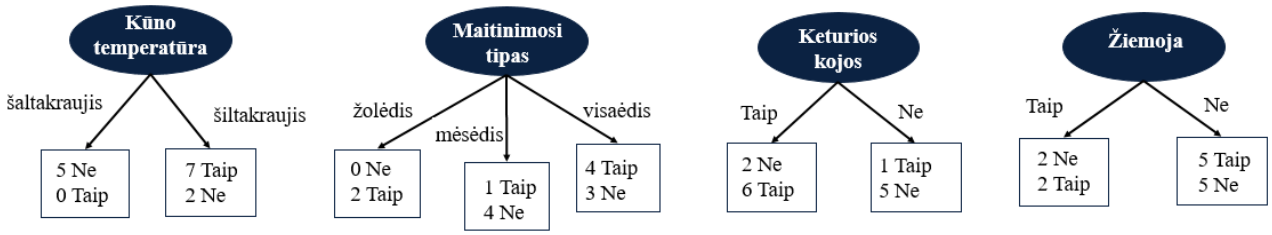
Banginis	šiltakraujis	mėsėdis	Ne	Ne	Taip
Žmogus	šiltakraujis	visaėdis	Ne	Ne	Taip
Avis	šiltakraujis	žolėdis	Taip	Ne	Taip
Kengūra	šiltakraujis	žolėdis	Ne	Ne	Taip
Gandras	šiltakraujis	mėsėdis	Ne	Ne	Ne
Varlė	šaltakraujis	mėsėdis	Taip	Taip	Ne

2.2 Sprendimų medžio sudarymas pagal apmokymo duomenų imtį

Pavyzdyje, duomenų rinkinys, tai informacija apie įvairias gyvas būtybes (gali būti gyvūnai, vabzdžiai ir pan.) nurodant jų kūno temperatūrą, maitinimosi tipą, keturių kojų egzistavimą (taip ar ne) ir žiemojimo būdą (žiemoja reiškiasi miega šaltuoju sezonu, nežiemoja tai nemiega). Tikslo aibė yra paskutinis stulpelis, kuris nusako ar gyva būtybė yra žinduolis ar ne. Akivaizdu, jog tai klasifikavimo uždavinys, kuriame pagal pateiktus atributus (charakteristikas) reikia pasakyti kokio tipo yra gyva būtybė. Norint pradėti konstruoti medį reikia iš visų atributų išrinkti patį tinkamiausią medžio viršūnei (Pav 3.). Pats tinkamiausias atributas yra tas, kurio duomenų padalinimo rezultatas – maksimalus homogeniškumas.

Pirmasis duomenų aibės atributas yra „pavadinimas“, kuris yra unikalus tekstas, todėl tikrai nėra tinkamas medžio konstravimui. Aplamai atributai, kurių reikšmės unikalios mašininiame mokyme yra nereikalingi.

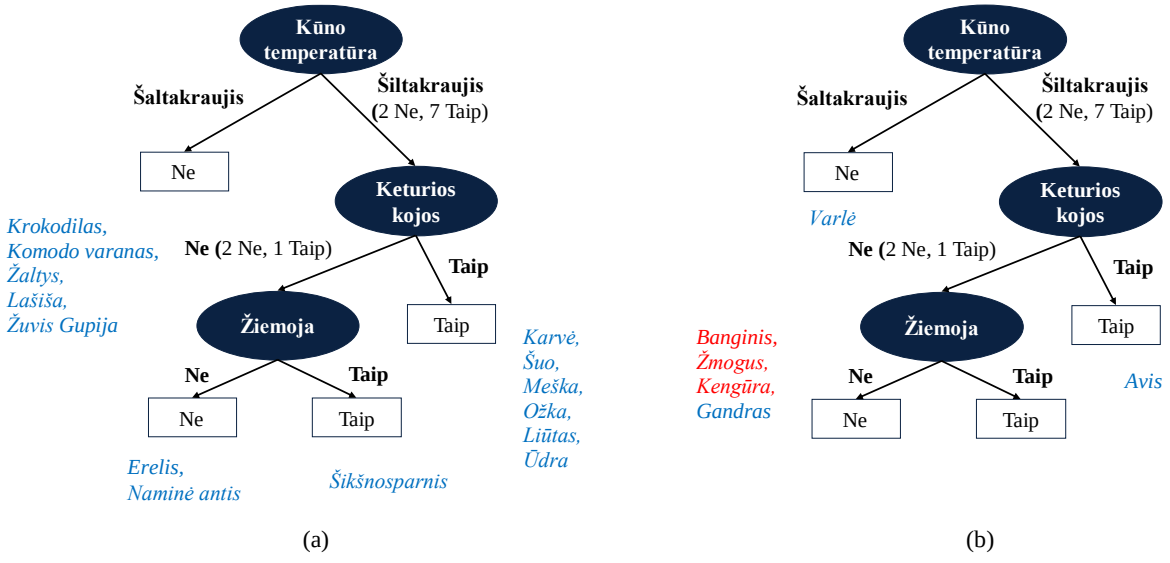
Antrasis atributas – „kūno temperatūra“. Kūno temperatūra gali įgyti vieną iš dviejų reikšmių: „šiltakraujis“ arba „šaltakraujis“. Jeigu gyva būtybė yra šaltarkraujis gyvis, tai žiūrint į paskutinį stulpelį matyti, kad jis yra niekada nebus žinduolis. Jeigu gyvūnas šiltakraujis, tai iš devynių tokių gyvūnų 7 yra žinduoliai ir 2 ne žinduoliai, nes *erelis* ir *naminė antis* nors ir šiltakraujai tačiau nėra žinduoliai. Išskaidymas pagal kūno temperatūrą yra ganėtinai tikslus, nes yra tik 2 klaidos iš 14 duomenų. Klaida yra gautoje atsakymo aibėje mažesnis skaičius. Jeigu turime dvi atsakymų klases, kaip pavyzdžiui atsakymas (4,1) tai klaida yra 1. Priešingu atveju, jeigu turime daugiau klasių, visi mažesni skaičiai už didžiausiąjį yra klaidos. Pavyzdžiui atsakymas (3,1,0,6), tai klaidų kiekis yra 4.



Pav 3. Pavyzdžio „Žinduolis“ duomenų rinkinio išskaidymo galimybės pagal keturis atributus.

Kitas atributas – tai „maitinimosi tipas“, kuris šiuo atveju gali įgyti tris reikšmes: žolėdis, visaėdis ir mėsėdis. Jeigu žolėdis, tai visais atvejais gauname, kad jis yra žinduolis. Jeigu mėsėdis, tai remiantis pateiktais duomenimis matyti, kad dažniausiai tai yra ne žinduolis, nes iš 5 mėsėdžių 4 atvejai, kad šis nėra žinduolis. Jeigu šis yra visaėdis tai gaunama stipriai heterogeniška aibė, kurią sudaro 4 žinduoliai ir 3 ne žinduoliai.

Atributas „žiemoja“, nusako ar gyva būtybė žiemoja ar ne. Kaip matyti iš duomenų tai yra visai netinkamas atributas medžio viršūnei, nes gaunamos atsakymų aibės yra heterogeniškos (Pav 3.).



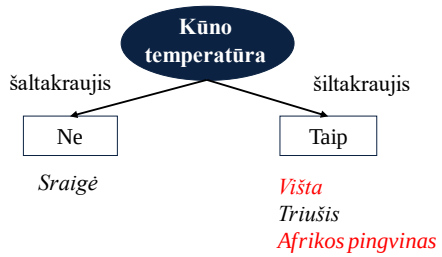
Pav 4. Sprendimų medžio taikymas apmokymo (a) ir testavimo duomenims (b)

Tarkime medžio viršūnei pasirinkome atributą „kūno temperatūra“, nes kai gyva būtybė yra šaltakraujis, tai visada bus ne žinduolis (4a pav.). Tačiau šiltakraujų atveju gaunamas atsakymas nėra visai homogeniškas: 2 ne žinduoliai ir 7 žinduoliai, todėl skaidymą galima tęsti toliau pasirenkant kitą atributą. Ar tęsti medžio konstravimą ar ne, priklauso nuo eksperto arba tam tikrų iš anksto nustatytų taisyklių. Pažiūrėjus į duomenis matosi, kad jeigu gyva būtybė turi keturias kojas tai jis visada yra žinduolis. Todėl sekantis atributas skaidymui pasirinktas „keturios kojos“ (4a pav.). Jeigu šis neturi keturių kojų, tai matyti jog ne žiemojantys, tokie kaip *erelis* ir *naminė antis* nėra žinduoliai, o žiemojantis yra. Taigi kaip paskutinis atributas pasirenkamas „žiemoja“, kuriuo remiantis suklasifikuojami paskutiniai likę duomenys (gyvos būtybės). Medžio konstravimas yra baigtas ir gaunamas pilnas medis, sukonstruotas remiantis mokimosi duomenų imtimi ir pateikia nulinę paklaidą, kas reiškiasi, kad visi duomenys yra teisingai suklasifikuoti (4a pav.). Tačiau labai svarbu įvertinti, kaip šis medis klasifikuoja testavimo duomenis, kurie yra nauji ir nematyti spendimų medžio algoritmui.

2.3 Sprendimų medžio testavimas su naujais duomenimis

Testavimo imtį sudaro 6 elementai (1 lentelė). *Banginis* yra šiltakraujis, bet neturi keturių kojų ir nežiemoja, todėl pagal medį gaunasi kad jis ne žinduolis (4b pav.). Dėl tų pačių priežasčių ir *žmogus* gaunasi jog nėra žinduolis. *Kengūra* yra šiltakraujis, tačiau skaitosi jog turi dvi kojas ir dvi rankas arba kituose šaltiniuose teigiama, jog turi 5 kojas. Bet kuriuo atveju, laikome jog tai nėra keturios kojos ir ji nežiemoja todėl kengūra ne žinduolis. Visi trys yra suklasifikuoti neteisingai, nes iš tikro jie visi yra žinduoliai (4b pav.). *Avis* yra šiltakraujis ir su keturiomis kojomis, todėl priskiriama žinduolių klasei kas yra teisingas atsakymas. *Gandras* ir *varlė* yra teisingai suklasifikuoti, nes priskiriami ne žinduolių klasei. Bendrai klasifikavimo rezultatas pateikė tik 50% tikslumą. Tai tas pats kas spėlioti atsitiktinai ar bus žinduolis, ar ne žinduolis. Kodėl taip atsitiko? Tai gali nulemti keletas priežasčių, tačiau labai dažnai žemas testavimo rezultatas gaunasi, todėl kad sprendimų medis per daug prisitaikė prie apsimokymo duomenų ir įvyko taip vadinamas *persimokinimas*. Kartais duomenų yra permažai ir spendimų medis apsimoko nepakankamai gerai kad galėtų suformuoti tikslius, duomenis apibendrinančius bruožus. Reikėtų prisiminti, jog tikslas yra gauti kuo geresnius rezultatus su testavimo duomenų imtimi, o ne tik su apmokymo imtimi.

Tarkime, jeigu atliktume klasifikavimą tik palikus vieną atributą „kūno temperatūra“. Tokiu atveju, apsimokymo paklaida nebūtų 0% (4a pav.), nes *naminė antis* ir *erelis* būtų suklasifikuoti neteisingai (jie būtų suklasifikuoti kaip žinduoliai) ir todėl gautume 14,3%. Tačiau, testavimo imčiai gautume tik vieną klaidingai suklasifikuotą kaip žinduolį *gandrą* ir bendrai gautume 16,7% paklaidą. Iš pirmo žvilgsnio atrodo, jog rastas puikus sprendimas, bet reikia suprasti, kad toks atsakymas gautas naudojant konkrečius testavimo duomenis (1 lentelė). Kas būtų jeigu testavimo imtis būtų kitokia?



Pav 5. Klasifikavimas naudojant vieną „kūno temperatūros“ atributą.

Pavyzdžiui reiktų suklasifikuoti keturias gyvas būtybes: *vištą, triušį, sraigę* ir *Afrikos pingviną*. Atlikus klasifikavimą pagal vieną atributą „kūno temperatūrą“ gaunamas tik 50% tikslumas, nes *višta* ir *pingvinas* suklasifikuoti neteisingai - kaip žinduoliai ir tik *sraigė* bei *triušis* yra teisingai suklasifikuoti (Pav 5.). Todėl akivaizdu, kad klasifikavimas pagal vieną atributą yra nepakankamas ir labai jautrus duomenų imčiai. Yra keli būdai kaip tokią problemą galime spręsti, tačiau bet kuriuo atveju turi būti pakankama duomenų imtis. Mašininio mokymo algoritmai gali apdoroti milžiniškus kiekius duomenų, todėl visais atvejais rekomenduojama gauti maksimalų jų kiekį ir paskui pagal poreikį juos susimažinti paliekant pačius reikalingiausius.

2.4 Tikslesnio sprendimo paieška

Sprendimų medžių teorijoje vienas iš tokių problemų sprendimo būdų - remtis ne vieno medžio sprendimu, bet sukurti daug sprendimų medžių ir lesti jiems balsuoti, kuri klasė ar reikšmė yra populiariausia. Tokiu būdu populiariausias atsakymas ir būtų galutinis sprendimas. Būtent čia atsirado termino **Atsitiktinis miškas** idėja. Atsitiktinis miškas – aibė atsitiktinai sugeneruotų sprendimų medžių, kurie yra originalaus sprendimų medžio „kopijos“.

Kaip sukurti tas medžio kopijas? Kopijos identiškos originaliam sprendimų medžiui yra netinkamos. Atsitiktinis miškas susideda iš sprendimų medžių, kurie yra skirtingi. Yra keletas būdų kaip galima sukurti medžio kopijas ir taip sudaryti atsitiktinį mišką. Vienas iš populiariausių būdų yra iš originalaus duomenų rinkinio sukurti naujus duomenų rinkinius ir sukonstruoti sprendimų medžius pagal skirtingus atributų rinkinius. Tuomet atliekamas testavimas su visais miško medžiais ir galutinis sprendimas priimamas balsavimo būdu.

Naujų duomenų rinkinių sudarymas **savirankos** metodu (angl. *Bootstrap method*) tai naujų duomenų rinkinių sukūrimas iš originalaus (2 lentelė), kur naujų rinkinių ir originalaus rinkinio elementų skaičius yra vienodas tačiau tam tikri įrašai gali kartotis (3 lentelė). Taip yra todėl, kad kuriant naują duomenų rinkinį kiekvienas elementas iš originalaus duomenų rinkinio yra nukopijuojamas ir padedamas atgal ir todėl sekantį kartą renkantis atsitiktinai elementą vėl gali būti pasirinktas tas pats.

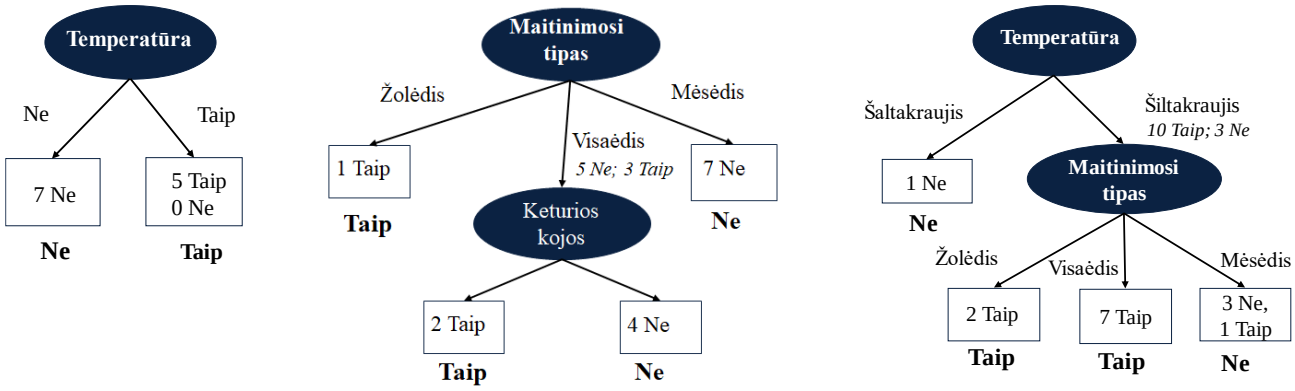
2 lentelė. Originalaus duomenų rinkinio pavyzdys

Pavadinimas	Kūno temperatūra	....	Klasės tipas. Ar žinduolis?
Karvė	šiltakraujis	...	Taip
Šuo	šiltakraujis	...	Taip
Šikšnosparnis	šiltakraujis		Taip
Banginis	šiltakraujis		Taip
Krokodilas	šaltakraujis		Ne
Komodo varanas	šaltakraujis		Ne
Žaltys	šaltakraujis		Ne
Lašiša	šaltakraujis	...	Ne

3 lentelė. Duomenų rinkinio kopijos pavyzdys

Pavadinimas	Kūno temperatūra	....	Klasės tipas. Ar žinduolis?
Šuo	šiltakraujis	...	Taip
Šuo	šiltakraujis	...	Taip
Lašiša	šaltakraujis		Ne
Banginis	šiltakraujis		Taip
Krokodilas	šaltakraujis		Ne
Banginis	šiltakraujis		Taip
Krokodilas	šaltakraujis		Ne
Žaltys	šaltakraujis		Ne

Kai sukuriamos duomenų kopijos, yra konstruojami sprendimų medžiai, tačiau parenkant tik dalį atributų. Tarkime, iš originalaus, 14 elementų duomenų rinkinio, buvo sukurti trys duomenų rinkiniai (kopija I, kopija II ir kopija III), kuriuose sprendimų medžių sudarymui panaudoti du atributai. Tarkime I-*ajai* kopijai „kūno temperatūra” ir “žiemojimo“ atributai. II-*ajai* kopijai medis konstruojamas su dviem kitais atributais: „maitinimosi tipas“ ir „keturios kojos“. Ir III-*ajai* kopijai medis konstruojamas naudojant „kūno temperatūros“ atributą ir „maitinimosi tipą“. Atsitiktinį mišką sudarantys sprendimų medžiai yra pateikti Pav 7.



(a) (b) (c)

Pav 6. Atsitiktinis miškas sudarytas iš trijų sprendimų medžių.

Kaip matyti iš Pav. 6a. pakanka vieno atributo kad būtų priimamas galutinis sprendimas ir nebelieka duomenų, kuriuos būtų galima klasifikuoti įtraukiant „žiemojimo“ atributą. Kituose dvejuose sprendimų medžiuose

panaudojami abu atributai medžio konstravimui. Trečiojo medžio (Pav. 6c) atveju gaunama 1 atsakymo klaida. Galutinis sprendimas naudojant sudarytą atsitiktinį mišką yra visų trijų sprendimų medžių „balsavimo“ rezultatas. Testavimo etape kiekvienam rinkinio elementui „ar žinduolis?“ yra pateikiami trys atsakymai gauti naudojant tris sprendimų medžius. Galiausiai išrenkamas dažniausiai pasikartojęs atsakymas (4 lentelė).

4 lentelė. Atsitiktinio miško galutinis sprendimas pateiktai duomenų imčiai

Pavadinimas	I Medis	II Medis	III Medis	Galutinis sprendimas
Banginis	Taip	Ne	Ne	Ne (klaidingas)
Žmogus	Taip	Ne	Taip	Taip
Avis	Taip	Taip	Taip	Taip
Kengūra	Taip	Ne	Ne	Ne
Gandras	Ne	Ne	Ne	Ne

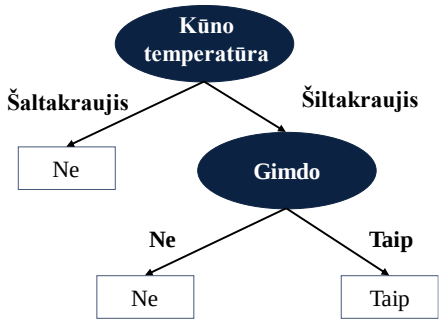
Prisiminkime testavimo duomenų imtį (1 lentelė), kuriai naudojant pirminį sprendimų medį buvo gautas 50% klasifikavimo tikslumas (Pav 4b). Taikant atsitiktinį mišką, gauname tik vieną klaidą ir tikslumas išauga iki 83,3% (4 lentelė).

Kitas būdas kaip surasti tikslesnį sprendimą yra neinformatyvių atributų pašalinimas ir naujų įtraukimas, kurie leistų daug tiksliau klasifikuoti. Paanalizavus duomenų rinkinį matyti, jog „maitinimosi“ tipas ir „žiemojimas“ yra žemo informatyvumo atributai t.y., nuo jų reikšmės atsakymo reikšme priklauso labai mažai arba beveik nepriklauso (5 lentelė). Šiuos atributus būtų galima pakeisti naujais, informatyvesniais atributais tokiais kaip pavyzdžiui, žymėjimas „ar gyvūnas gimdo“ arba „išorė“, kuri tarkime gali būti plunksnos, kailis, žvynai, kiautas ir pan. (5 lentelė).

5 lentelė. Žemo ir aukšto informatyvumo atributai

		Žemo informatyvumo atributai				Aukšto informatyvumo atributai	
Pavadinimas	**Kūno temperatūra**	**Maitinimosi tipas**	**Keturios kojos**	**Žiemoja**	**Klasės tipas. Ar žinduolis?**	**Gimdo**	**Išorė**
Višta	šiltakraujis	visaėdis	Ne	Ne	Ne	Ne	Plunksnos
Triušis	šiltakraujis	žolėdis	Taip	Ne	Taip	Taip	Kailis
Sraigė	šaltakraujis	žolėdis	Ne	Taip	Ne	Ne	Kiautas
Afrikos pingvinas	šiltakraujis	mėsėdis	Ne	Ne	Ne	Ne	Plunksnos
Tigras	šiltakraujis	mėsėdis	Taip	Ne	Taip	Taip	Kailis
Australijos tunas	šaltakraujis	mėsėdis	Ne	Ne	Ne	Ne	Žvynai

Pavyzdžiui, jeigu duomenų aibėje (1 lentelė) be pavadinimo, paliktume tik du atributus: senąjį atributą „kūno temperatūra“ ir naują atributą „gimdo“, būtų galima sukonstruoti dviejų atributų sprendimų medį, kuris leistų suklasifikuoti tiek apsimokymo, tiek testavimo duomenis su 0% paklaida. Tai rodo, kad šie atributai yra informatyvus ir stipriai koreliuoja su atsakymo atributu (Pav 7.).



Pav 7. Sprendimų medis iš dviejų informatyvių atributų.

Taigi akivaizdu, kad būtina identifikuoti pačius informatyviausius atributus, kas leidžia sukonstruoti korektišką ir optimalų sprendimų medį. Klausimas, tik kaip geriausiai identifikuojamas reikšmingiausias kintamasis ir nustatoma jo reikšmė?

3 Entropijos skaičiavimai sprendimų medžiuose

Entropija – tai vertinimo kriterijus skirtas aprašyti susimaišymo (heterogeniškumo) laipsnį. Jeigu duomenų rinkinys yra pilnai homogeniškas, tai entropija = **0**; jeigu duomenų rinkinys turi vienodą skaičių *n* kiekvienos rūšies *k* (klasės) objektų, kai k>1, tai entropija lygi **1**. Entropijos formulė pateikta žemiau:

$$-\mathrm{p}\log_2\mathrm{p} \; - \mathrm{q}\log_2\mathrm{q}$$

čia *p* ir *q* yra sėkmės ir nesėkmės tikimybė viršūnėje.

Paimkime paprastą pavydį su spalvotais kamuoliukais: 3 melsvi ir 3 violetiniai (Pav 8a.). Tarkime *p* yra melsvi kamuoliai, o *q* yra violetiniais. Reiktų atkreipti dėmesį, kad *p* ir *q* yra tikimybės, todėl *p* =3/6 ir q = 3/6.

(a) (b)

Pav 8. Pavyzdys su spalvotais kamuoliukais: (a) heterogeniškas ir (b) homogeniškas duomenų rinkinys

Įstačius *p* ir *q* reikšmes į entropijos formulę iš tiesų gauname entropijos reikšmę lygią vienetui, kas reiškia jog aibės elementai yra maksimaliai susimaišę:

$$\mathrm{E} = -\mathrm{p}\log_2\mathrm{p} \; - \mathrm{q}\log_2\mathrm{q}$$
$$\mathrm{E} = -3/6\log_2(\frac{3}{6}) \; - 3/6\log_2(\frac{3}{6})$$
$$\mathrm{E} = -0{,}5\log_2(0{,}5) \; - 0{,}5\log_2(0{,}5)$$
$$\mathrm{E} = -0{,}5(-1) - 0{,}5(-1) = \mathbf{1}$$

Pavyzdys dešinėje (Pav. 8b) atvaizduoja pilnai homogeniškos aibės atvejį, kur *p* =6/6, *q* = 0/6 . Įstačius reikšmes į formulę gauname entropijos reikšmę lygią nuliui.

$$\mathrm{E} = -\mathrm{p}\log_2\mathrm{p} \; - \mathrm{q}\log_2\mathrm{q}$$
$$\mathrm{E} = -6/6\log_2(\frac{6}{6}) \; - 0/6\log_2(\frac{0}{6})$$
$$\mathrm{E} = -1\log_2(0) \; - 0$$
$$\mathrm{E} = -1(0) - 0 = \mathbf{0}$$

Todėl prieš pradedant konstruoti sprendimų medį iš visų duomenų aibės atributų reikia išrinkti tą, kuris turi mažiausią entropijos reikšmę. Prisiminkime pavyzdį su žinduoliais (1 lentelė), kur iš visų keturių atributų reikia išrinkti patį informatyviausią ir jį pastatyti sprendimų medžio viršūnėje. Tokiame mažame pavyzdyje ir be skaičiavimų matyti, kad tai turėtų būti „kūno temperatūra“, tačiau patikrinkime ir pagrįskime tai entropijos skaičiavimais.

Entropijos reikšmės paskaičiavimas atributui susideda iš tokių trijų žingsnių: (1) paskaičiuoti entropiją kiekvienai klasei; (2) paskaičiuoti kiekvienos klasės svorinę reikšmę; (3) paskaičiuoti svorinę sumą.

Tarkime šiam pavyzdžiui į klausimą ar žinduolis ar ne, *p* yra atsakymas *Taip*, o *q* yra atsakymas *Ne*. „Kūno temperatūros“ atributas turi dvi reikšmes: *šaltakraujis* ir *šiltakraujis*, todėl *p* ir *q* reikia paskaičiuoti atskirai šaltakraujams ir šiltakraujams. Duomenų rinkinyje šaltakraujų turime 5. Reikia paskaičiuoti kiek turime *p*

šaltakraujų ir q šaltakraujų, arba kitaip sakant kiek yra šaltakraujų žinduolių ir ne žinduolių. Analogiškus skaičiavimus reikia padaryti ir su šiltakraujais, kurių yra 9.

$$p_{šaltakraujis} = \frac{0}{5} = 0, \qquad q_{šaltaraujis} = \frac{5}{5} = 1$$

$$p_{šiltakraujis} = \frac{7}{9} = 0{,}77\ , \qquad q_{šiltakraujis} = \frac{2}{9} = 0{,}22$$

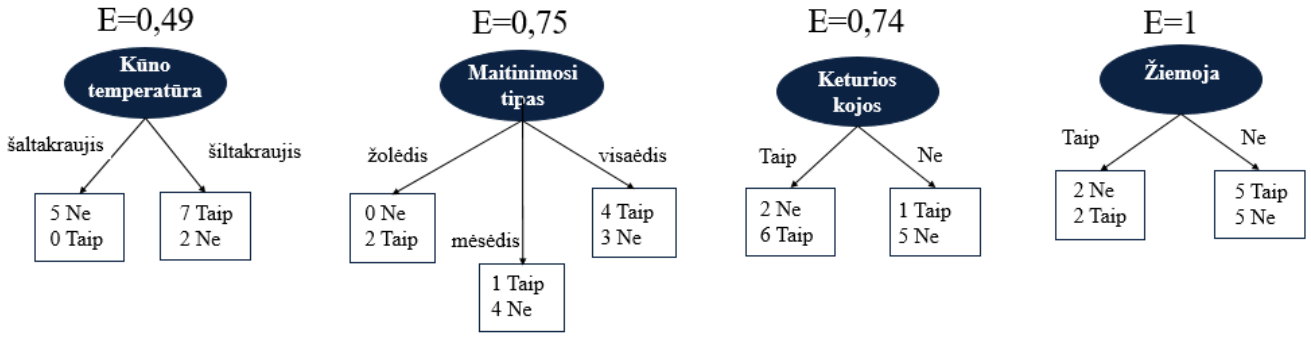
Turint p ir q abiem atvejams galime pradėti skaičiuoti entropijas šiltakraujams ir šaltakraujams pagal entropijos formulę.

$$\mathrm{E(šaltakraujis)} = -(0/5)\log_2(0/5)\ -(5/5)\log_2(5/5)=0$$

$$\mathrm{E(šiltakraujis)} = -(7/9)\log_2(7/9)\ -(2/9)\log_2(2/9)=0{,}76$$

E=(5/14)*0+ (9/14)*0,76 = **0,49**

Paskaičiavus entropijas kiekvienai atributo reikšmei (klasei) galima suskaičiuoti bendrą atributo svorinę reikšmę. Svorinė reikšmė nurodo, koks procentas yra kiekvienos klasės elementų. Entropijos reikšmė 0 šaltakraujams yra dauginama iš svorinės klasės reikšmės 5/14. Šiltakraujų klasės gauta entropijos reikšmė 0,76 dauginama iš svorinės klasės reikšmės, o tai yra 9/14. Sudėjus šias sandaugas gauname, jog „kūno temperatūros“ entropijos svorinė reikšmė yra lygi **0,49**. Paskaičiavus entropijos reikšmės visiems keturiems atributams, matyti, kad „kūno temperatūros“ atributas iš tiesų yra pats tinkamiausias atributas iš visų keturių norint pradėti konstruoti sprendimų medį, nes jo reikšmė yra pati mažiausia (Pav. 9). Toliau norint plėsti medį reiktų vėl perskaičiuoti entropijas likusiems duomenims ir taip tęsti iki tol kol nuspręsime nebeplėsti sprendimo medžio arba neliks nesuklasifikuotų duomenų.



Pav 9. Entropijos svorinės reikšmės keturiems atributams