Paveikslėlis, kuriame yra juodas, tamsa

Automatiškai sugeneruotas aprašymas

**Kauno technologijos universitetas**

Informatikos fakultetas

**P176B101 Intelektikos pagrindai**

**Laboratorinis darbas Nr. 2**

|  |  |
| --- | --- |
|  |  |
| **Nedas Liaudanskis IFF-1/9**  Studentas | (parašas) (data) |
|  |  |
| **dėst. Nečiūnas Audrius**  Dėstytojas | (parašas) (data) |
|  |  |

**KAUNAS, 2024**

Turinys

[1. Įvadas 3](#_Toc161772963)

[2. Duomenų rinkinys 3](#_Toc161772964)

[3. Funkcinių reikalavimų vykdymas 5](#_Toc161772965)

[Duomenų rinkinio išvestis. Apmokymo ir testavimo poaibiai. 5](#_Toc161772966)

[Sprendinių medis 5](#_Toc161772967)

[Sprendinių medžio testavimas 7](#_Toc161772968)

[Atsitiktinis miškas 8](#_Toc161772969)

[Palyginimas 9](#_Toc161772970)

[4. Išvados 10](#_Toc161772971)

1. Įvadas

Šiame laboratoriniame darbe nagrinėsime sprendimų medžio modeliavimo ir atsitiktinių miškų metodų taikymą duomenų analizėje. Sprendimų medžiai yra galingas ir plačiai naudojamas mašininio mokymosi metodas, kuris leidžia prognozuoti kintamuosius, remiantis įvesties kintamųjų reikšmėmis. Atsitiktiniai miškai, kita vertus, yra modeliavimo metodas, kuris kombinuoja daugybę sprendimų medžių ir pateikia prognozes, remdamasis jų vidurkiu ar daugumos balsavimu.

Norint pradėti šį laboratorinį darbą, reikia susirasti duomenų rinkinį, kurio atraibtai yra susiję vienas tarp kito. Kitaip nebus įmanoma įvykdyti kai kurių funkcinių reikalavimų pateiktų šiam laboratoriniu darbui.

Funkciniai darbo reikalavimai:

1. Pasirinkti duomenų rinkinį, kuriam bus sudaromas sprendinių medis.
2. Pasirinkti sprendinių medžio išvestį. Atributą rinkinio kurį bandysime gauti išmokę modelį iš kitų atributų reikšmių. Rekomenduojamas, jog atributas tūrėtų kardinaluma nuo 4 iki 10.
3. Suskaidyti turimą duomenų rinkinį į apmokymo, bei testavimo poaibius.
4. Suskaidyti duomenų rinkinio poaibius į įvestis ir išvestis.
5. Naudojant apmokymo duomenų rinkinį, sudarykite sprendimų medį. Galime rinktis iš keleto algoritmų ID3, C4.5, CART ir pan. Nuo to priklauso kokius indeksus/metodus naudosite medžio dalijimui (pvz., Gini, Gain ir t.t.). Žinoti koks yra skirtumas tarp šių algoritmų ir dalijimo indeksų.
6. Grafiškai atvaizduoti gautą sprendinių medį.
7. Ištestuoti sudarytą sprendinių medį. Suskaičiuoti apmokyto modelio tikslumą ir paklaidą. Nurodyti paklaidos skaičiavimams naudotas metrikas. Taip pat pateikti sumaišymo matricą.
8. Keičiant maksimalų medžio gylį, eksperimentiniu būdu išmatuokite skirtingų gylių (3-4 variacijos) medžių formavimo trukmę bei gaunamą tikslumą.
9. Naudojant tuos pačius testavimo ir apmokymo duomenis, suformuoti atsitiktinį mišką, kurį sudaro 5 medžiai. Pateikti jų skirtumus. Maksimalus medžio gylis - gylis užfiksuotas eksperimento metu, kuris pateikė geriausius rezultatus.
10. Keičiant mišką sudarančių medžių kiekį, nustatyti geriausią rezultatą pateikiantį atsitiktinį mišką.
11. Palyginti sprendinių medžio rezultatus su atsitiktinio miško gautais rezultatais.
12. Duomenų rinkinys

Šiam laboratoriniai darbui pasirinktas duomenų rinkinys turėjo tūrėti sąryšius tarp atributų iš kurių vienas atributas turi kardinalumą nuo 4 iki 10. Tokį duomenų rinkinį surasti buvo nelengva. Po daug ieškojimo radau duomenų rinkinį, kuris nusako vyno kokybę, balais nuo 0 iki 10. Todėl šiuos balų skaičius ir laikysime kaip išvestinio atributo kardinalumą. Kiti atributai sudarys testavimo ir apmokymo poaibius, kurie bus naudojami ištestuoti ir apmokyti sprendinių medį.

Pasirinktas rinkinys: <https://www.kaggle.com/datasets/yasserh/wine-quality-dataset>

Pasirinktą duomenų rinkinį sudaro 1143 įrašai. Rinkinys turi 13 atributų, vienas iš jų yra [Id] atributas, šis atributas neturės reikšmės medžio formavime, tai suformuoti medį naudosime 12 atributų.

Duomenų rinkinio atributai:

1. Fixed acidity
2. Volatile acidity
3. Citric acid
4. Residual sugar
5. Chlorides
6. Free sulfor dioxide
7. Total sulfor dioxide
8. Density
9. pH
10. Sulphates
11. Alcochol
12. Quality – Išvedimo atributas, kurio iškosime naudodami sudarytą modelį.
13. Id – Indentifikavimo atributas, kuris nebus naudojamas sudaryti sprendinių medį.
14. Funkcinių reikalavimų vykdymas

Duomenų rinkinio išvestis. Apmokymo ir testavimo poaibiai.

Duomenų rinkinio sprendimų medžių išvestis yra parinktas atributas „Quality“, šis atributas turi kardinalumą 10. Toks kardinalumas puikai tinka į funkcinius reikalavimus. „Quality“ atributą sudaro skaičiai nuo (0 iki 10) šie skaičiai parodo vyno kokybę. Naudodami sprendimų medį bandysime gauti šį atributą atsižvelgiant į kitus atributus.

Įvesties poaibiui naudosime 20% duomenų rinkinio elementų.

Likusius 80% elementų, naudosime apmokymui.

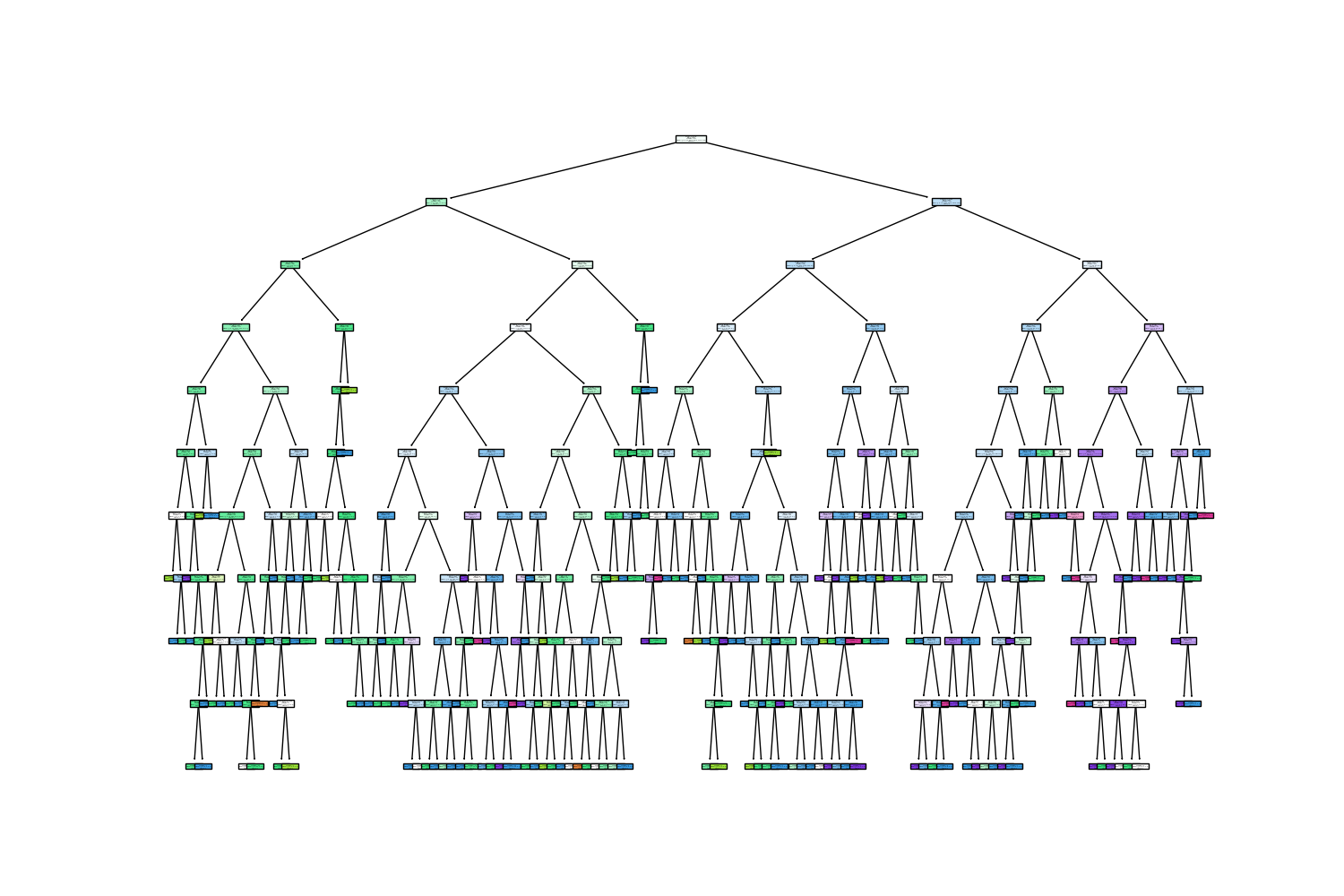
Sprendinių medis

Sudaryti sprendinių medį naudosime python biblioteką: sklearn. Ši biblioteka naudoja CART algoritmą. CART algoritmas veikia sukurdamas binarinį medį, kuriame kiekvienas vidutinio dydžio mazgas yra priklausomybės matrica. Medžio šakos ir lapai nustato tam tikras sąlygas, pagal kurias duomenys yra skaidomi į dvi dalis(mano duomenų rinkinio atveju naudoju „Gini“ kriterijų). Šis procesas kartojamas rekursyviai, kur šakų skaidymas tęsiamas, kol pasiekiama numatyta sustojimo sąlyga, pavyzdžiui, maksimalus medžio gylis ar minimalus mazgo dydis.

„Gini“ kriterijus – tai kriterijus, kuris vertina duomenų grupavimo kokybę pagal tikimybę, kad atsitiktinai pasirinktas elementas bus neteisingas. Kuo mažesnis "Gini" koeficientas, tuo geresnis skaidymas laikomas. „Gini“ apskaičiavimo formulė, kiekvienam medžio mazge:

c yra klasės skaičius, o ​ yra tikimybė, kad atsitiktinai pasirinktas elementas mazge.

Gauto medžio gylis: 10.



pav. 1 Visas sprendinių medis.

Paveikslėlis, kuriame yra linija, diagrama, tekstas, trikojis

Automatiškai sugeneruotas aprašymas

pav. 2 Priartinta sprendimų medžio dalis.

Sprendinių medžio testavimas

Suformavus sprendinių medį naudojant bibliotekas apskaičiuota buvo ir sprendinių medžio tikslumas, taip pat buvo suformuota sumaišymo matrica [pav. 3], kuria leidžia pamatyti nukrypimus skaičiavimuose.

Tikslumui naudojama paprastas tikslumo balas. Tai metrika kurioje galime lengvai rasti tikslumą.

Taip pat naudojama buvo ir MAE metrika. MAE – tai vidutinė absoliutinė paklaida. ai yra regresijos modelio vertinimo metrika, skirta įvertinti, kiek vidutiniškai modelio prognozės skiriasi nuo tikrųjų stebėtų reikšmių. Jis parodo kiek modelis vidutiniškai nukrypsta nuo reikiamos reikšmės.

Čia, yra stebėtos reikšmės, yra modelio prognozuotos reikšmės, n yra duomenų skaičius.

Testuojant buvo pastebėta, jog mano duomenų rinkinys turi, gan nemažą kiekį sukoncentruotų dydžių. Tai sukėlė sunkumų kai buvo mokamas sprendimų medis, nes didžioji dalis išeinančio atributo, buvo paskirstyta tarp nedideliu intervalu. Todėl mano sprendinių medis negavo įvairovės ir gautas tikslumas siekia tik 0.55%, o paklaida 0.5. Tai gan didelė paklaida, kai mes žinome, jog mano duomenų rinkinyje pilnas 1 keičia atributo reikšmę. Gaila, tačiau po testavimo supratau, jog šis duomenų modelis nebuvo pats geriausiai šiai užduočiai.

**Gauti rezultatai**:

**Accuracy**: 0.5502183406113537

**MAE**: 0.4978165938864629

Paveikslėlis, kuriame yra tekstas, ekrano kopija, diagrama, Stačiakampis

Automatiškai sugeneruotas aprašymas

pav. 3 Sumaišymo matrica

Po šių testavimų, buvo sudaryti medžiai skirtinguose gyliuose. Kiekvienam medžiui buvo apskaičiuotas gylis, tikslumo balas ir medžio sudarymo laikas.

**Gauti rezultatai:**

**Max Depth**: 3, **Accuracy**: 0.5807860262008734, **Time**: 0.00436711311340332 seconds

**Max Depth**: 4, **Accuracy**: 0.6157205240174672, **Time**: 0.004029989242553711 seconds

**Max Depth**: 5, **Accuracy**: 0.6069868995633187, **Time**: 0.0046846866607666016 seconds

**Max Depth**: 6, **Accuracy**: 0.5589519650655022, **Time**: 0.00526738166809082 seconds

Atsitiktinis miškas

Atsitiktinis miškas yra mašininio mokymosi algoritmas, kuris naudoja daugybę sprendimų medžių , kadangi jie sudaro "mišką". Kiekvienas sprendimų medis yra mokomas atsitiktinai pasirinktų poaibių iš pradinio duomenų rinkinio, ir kiekvienas medis balsuoja dėl prognozės. Galutinis sprendimas yra sudaromas remiantis daugumos balsavimo principu - kiekvieno medžio rezultatai yra įvertinami, ir daugumos balsas nusprendžia galutinį rezultatą.

Pagal funkcinius reikalavimus reikia sudaryti atsitiktinį mišką, kurį sudaro 5 sprendinių medžiai. Suos visus suskaičiuoti su gautu galutiniu sprendimu medžiu atlikti testavimus.

Gauti rezultatai parodo, jog atsitiktinis miškas gaunasi daug tikslesnis, palyginus su sprendinių medžiu. Atsitiktinis miškas pasiekė tikslumą: 0.65, paklaidą: 0.38. Šie dydžiai nebėra tokie didelis kaip su paprastu sprendinių mišku, kuriame tikslumas buvo tik 0.55 %.

**Gauti rezultatai:**

**Number of Trees**: 9, **Accuracy**: 0.6462882096069869

**MAE**: 0.38427947598253276

Paveikslėlis, kuriame yra tekstas, ekrano kopija, Stačiakampis, diagrama

Automatiškai sugeneruotas aprašymas

pav. 4 Atsitiktinio miško sumaišymo matrica.

Keičiant mišką sudarančių medžių kiekį, reikėjo pažiūrėti kaip skiriais tikslumas ir juos atvaizduoti. Išsirinkti iš jų geriausią.

**Gauti rezultatai:**

**Number of Trees**: 3, **Accuracy**: 0.5851528384279476, **MAE**: 0.4978165938864629

**Number of Trees**: 4, **Accuracy**: 0.62882096069869, **MAE**: 0.4148471615720524

**Number of Trees**: 5, **Accuracy**: 0.6244541484716157, **MAE**: 0.4104803493449782

**Number of Trees**: 6, **Accuracy**: 0.6375545851528385, **MAE**: 0.38427947598253276

**Number of Trees**: 7, **Accuracy**: 0.6550218340611353, **MAE**: 0.36681222707423583

**Number of Trees**: 8, **Accuracy**: 0.62882096069869, **MAE**: 0.39737991266375544

**Number of Trees**: 9, **Accuracy**: 0.6462882096069869, **MAE**: 0.38427947598253276

Kaip matome pagal gautus atsakymus, kai turime daugiau medžių gauname tikslesnius atsakymus. Iki tok kol nepasiekiame 7 medžius. Nuo 7 medžių mūsų tikslumas krenta.

Palyginimas

Padarius ir sprendinių medį ir atsitiktinį mišką, galima lengvai juos palyginti ir nustatyti, kuris iš jų yra pranašesnis algoritmas, pagal gautus rezultatus. Kaip matome paprastas sprendinių medis turi tikslumą, tik 0.55%. Šis tikslumas nėra didelis, tačiau tai yra dėl duomenų rinkinio trūkstamų žemesnių reikšmių. Tačiau tą pačia problemą, turi ir atsitiktinis miškas, tačiau kadangi miške yra naudojame 9 sprendinių medžiai, atsitiktinis miškas įgauna geresnį tikslumą. Tikslumas pasiekia, net 0.65%. Taigi galiu teigti, jog atsitiktinio miško metodas yra daug patikimesnis, kai yra norima nustatyti tikslumą ir gauti tikslesnį atsakymą.

|  |  |
| --- | --- |
| **Sprendinių medis** | **Atsitiktinis miškas** |
| **Gauti rezultatai**:  **Accuracy**: 0.5502183406113537  **MAE**: 0.4978165938864629  Paveikslėlis, kuriame yra tekstas, ekrano kopija, diagrama, Stačiakampis  Automatiškai sugeneruotas aprašymas | **Number of Trees**: 9, **Accuracy**: 0.6462882096069869  **MAE**: 0.38427947598253276  Paveikslėlis, kuriame yra tekstas, ekrano kopija, Stačiakampis, diagrama  Automatiškai sugeneruotas aprašymas |

1. Išvados

Galiu pasidžiaugti, jog užduotį pavyko įvykdyti, nors ir duomenų rinkinys nebuvo pats geriausias. Paprasto sprendinių medžio tikslumas siekė tik 55%. Atsitiktinio miško tikslumas siekė: 0.65%. Pagal šiuos dydžius galiu teigti, jog atliktinio miško metodas yra tikslesnis ir geresnis.