



Kauno technologijos universitetas

Informatikos fakultetas

P176B101 Intelektikos pagrindai

Laboratorinis darbas Nr. 2

IFF-1/9 Martynas Kuliešius

Studentas

dėst. Nečiūnas Audrius

Dėstytojas

KAUNAS, 2024

Turinys

Įvadas.....	3
Duomenų rinkinys	3
Funkcinių reikalavimų vykdymas	5
Duomenų rinkinio išvestis. Apmokymo ir testavimo poaibiai.	5
Sprendimų medis	5
Sprendimų medžio testavimas	7
Atsitiktinis miškas	9
Palyginimas	11
Išvados.....	11

Išvadas

Šiame laboratoriniame darbe nagrinėjami sprendimų medžio modeliavimo ir atsitiktinių miškų taikymo metodai ir jų taikymas duomenų analizėje. Sprendimų medžiai pasižymi savo galia bei plačiai naudojamas mašininio mokymosi metodas, leidžiantis prognozuoti kintamuosius bei rezultatus remiantis įvesties kintamųjų reikšmėmis. Kita vertus, atsitiktiniai medžiai yra modeliavimo metodas, kuris apjungia daug sprendimų medžių ir pateikia prognozes pagal šių medžių vidurkį arba daugumos balsavimą.

Šio laboratorinio darbo atlikimo sėkmei reikia surasti duomenų rinkinį, kurio atributai yra susiję viens su kitu. Kitaip nebus įmanoma teisingai įvykdyti kai kurių funkcinių reikalavimų.

Funkciniai darbo reikalavimai:

1. Pasirinkite duomenų rinkinį kuriam sudarysite sprendimų medį.
2. Kaip sprendimų medžio išvestį pasirinkite prognozuojamą atributą (Patariama pasirinkti kategorinį kintamąjį, kurio kardinalumas yra nuo 4 iki 10).
3. Turimą duomenų rinkinį suskaidykite į apmokymo bei testavimo poaibius. Apmokymo aibė turi būti didesnė nei testavimo.
4. Suskaidykite duomenų 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 atvaizduokite gautą sprendimų medį. Atvaizdavimui galima naudoti slearn ir graphviz arba kitas bibliotekas. Jei sudarytas sprendimų medis yra labai didelis, kad pavyktų įskaitomai pateikti ataskaitoje - įkelkite failą atskirai ir ataskaitoje pateikite tik medžio fragmentą ir komentarus apie gautą struktūrą.
7. Ištestuokite sudaryta sprendimų medį naudojant testavimo duomenis ir apskaičiuokite prognozavimo tikslumą/paklaidą. Nurodykite kokią paklaidos metriką skaičiuojate (pvz., MAE, MSE ir t.t.). Taip pat klasifikavimo uždaviniui pateikite susimaišymo (angl. confusion) matricą.
8. Keičiant maksimalų medžio gylį, eksperimentiniu būdu išmatuokite skirtingų gylių (3-4 variacijos) medžių formavimo trukmę bei gaunamą tikslumą, t.y. medžio auginimas stabdomas nuo tam tikro gylio. Rezultatus pateikite ataskaitoje.
9. Naudojant tą patį apmokymo ir testavimo duomenų imties pasiskirstymą kaip ir formuojant sprendimų medį, suformuokite atsitiktinį mišką kurį sudaro 5 medžiai. Ataskaitoje pateiktų jų skirtumus. Maksimalus medžio gylis - gylis užfiksuotas eksperimento metu, kuris pateikė geriausius rezultatus.
10. Keičiant mišką sudarančių medžių kiekį [3-9], nustatykite geriausius rezultatus pateikiantį atsitiktinį mišką.
11. Palyginkite pirminio sprendimų medžio ir atsitiktinio miško gautus rezultatus.

Duomenų rinkinys

Šiam laboratoriniui darbui pasirinktas duomenų rinkinys turėjo turėti sąryšius tarp atributų, iš kurių vienas atributas turi turėti kardinalumą tarp 4 ir 10. Tokių duomenų rinkinį nebuvo lengva rasti. Po ilgų paieškų, radau raudonojo vyno kokybės duomenų rinkinį, kur vyno kokybę vertina nuo 0 iki 10. Šį vyno rezultatą ir naudosime ir laikysime kaip išvestinio atributo kardinalumą. Kiti atributai sudarys testavimo ir apmokymo poaibius, kurie bus naudojami ištestuoti ir apmokyti sprendimų medį.

Pasirinktas rinkinys: <https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009>

Pasirinktą duomenų rinkinį sudaro 1159 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 sulfur dioxide
7. Total sulfur dioxide
8. Density
9. pH
10. Sulphates
11. Alcohol
12. Quality – Išvedimo atributas, kurio iškosime naudodami sudarytą modelį.

Funkcinių reikalavimų vykdymas

Duomenų rinkinio išvestis. Apmokymo ir testavimo poaibiai.

Duomenų rinkinio sprendimų medžių išvestis bus parinktas atributas „Quality“. Šio atributo kardinalumas 10. Toks kardinalumo lygis atitinka funkcinis reikalavimus. Naudojant sprendimų medį bandysime gauti šį atributą atsižvelgiant į kitus atributus.

Įvesties poaibiui naudosime 20% duomenų rinkinio elementų.

Likusius 80% elementų, naudosime apmokymui.

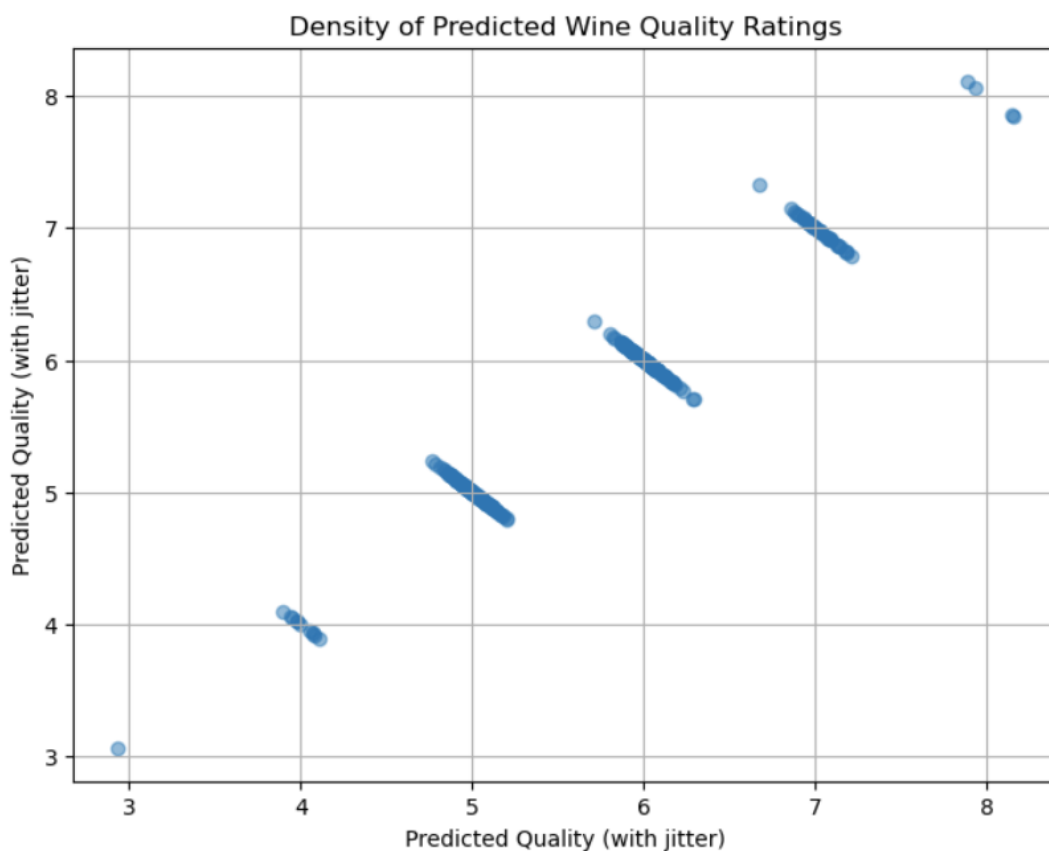


Figure 1 Duomenų rinkinio Quality atributo pasiskirstymas

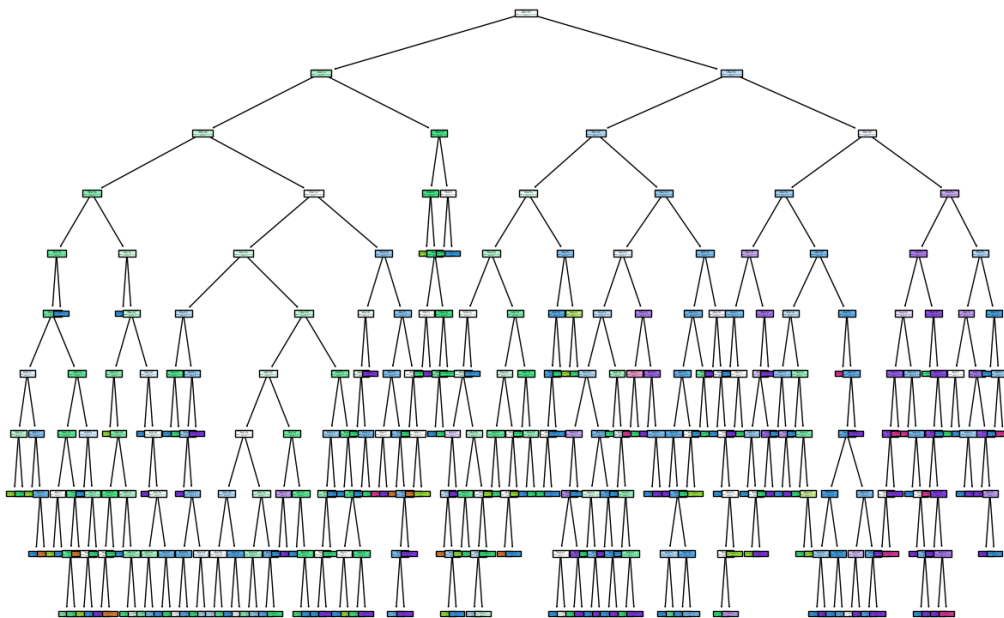
Sprendimų medis

Sprendimų medžio sudarymui naudosime sklearn Python biblioteką. Ši biblioteka naudoja CART algoritmą, kuris veikia sukurdamas binarinį medį, kuriame kiekvienas vidutinio dydžio mazgas yra priklausomybės matrica. Medžio šakos ir lapai nustato tam tikras sąlygas, pagal kurias duomenys yra suskaidomi į dvi dalis, šio pritakymo atveju, naudosime „Gini“ kriterijų. Šis

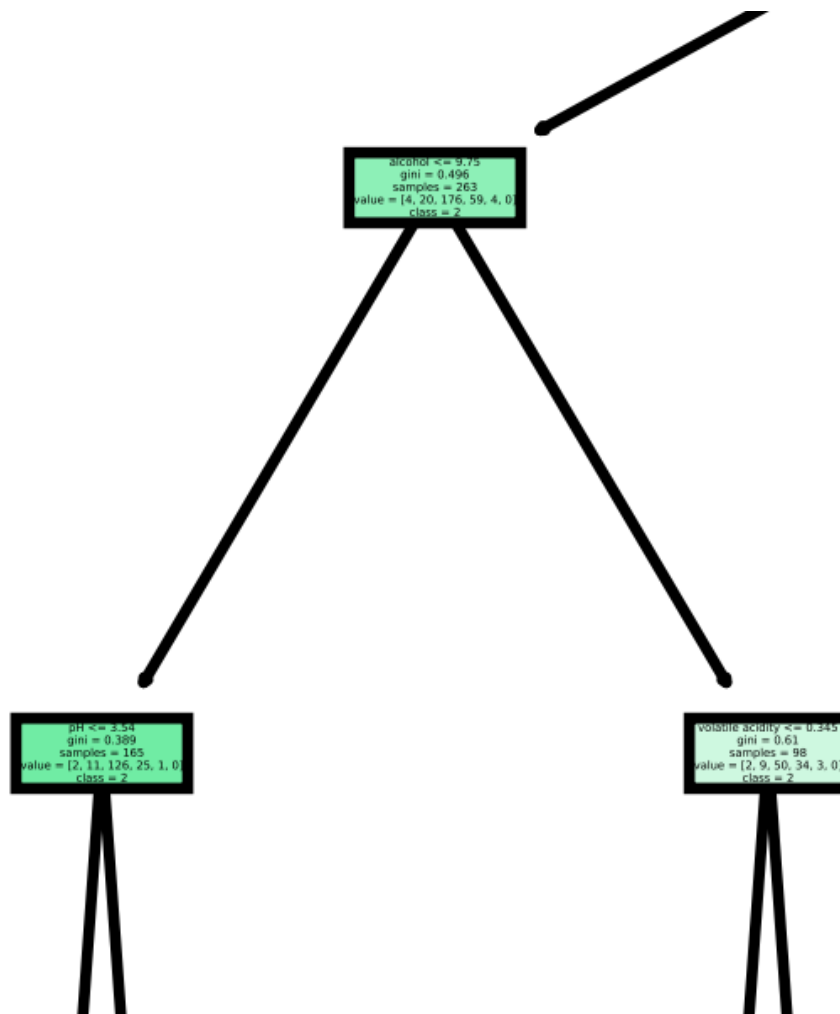
procesas kartojamas rekursyviai, kur šakų skaidymas tęsiamas, kol pasiekama sustojimo sąlyga, tokia kaip pasiektas maksimalus medžio dydis arba minimalus mazgo dydis.

„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.

Gauto medžio gylis: 10.



pav. 1 Visas sprendimų medis.



pav. 2 Priartinta sprendimų medžio dalis.

Sprendimų medžio testavimas

Suformavus sprendimų medį bei naudojant bibliotekas buvo apskaičiuota sprendimų medžio tikslumas bei suformuota sumaišymo matrica (pav.3), kuri 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. Tai yra regresijos modelio vertinimo įvertis, skirtas įvertinti, kiek vidutiniškai modelio prognozės skiriasi nuo tikrųjų stebėtų reikšmių. Jis parodo kiek modelis vidutiniškai nukrypsta nuo reikiamos reikšmės.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

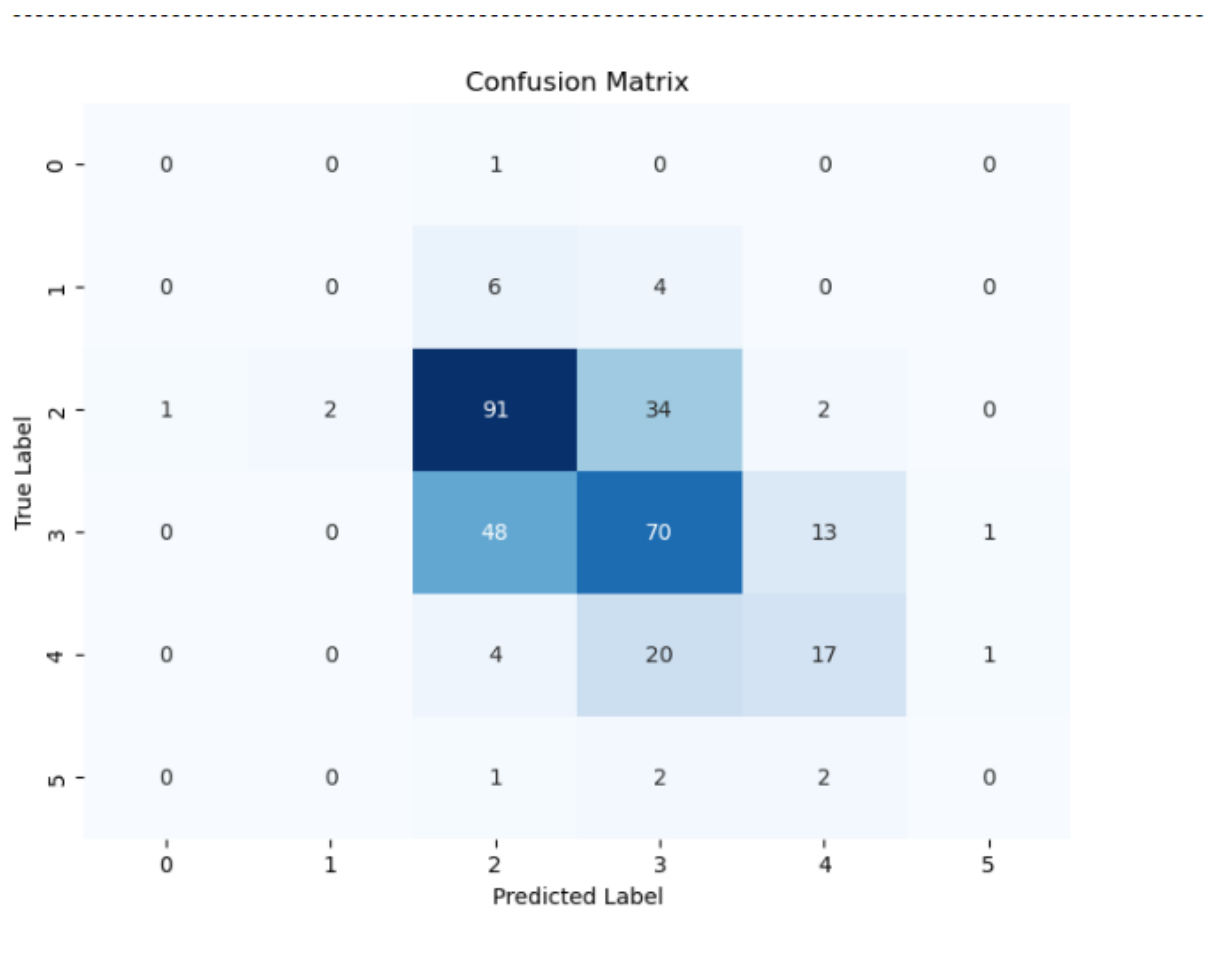
Čia, y_i yra stebėtos reikšmės, \hat{y}_i yra modelio prognozuotos reikšmės, n yra duomenų skaičius.

Testuojant pastebėjau, kad mano duomenų rinkinys turi gan nemažą kiekį sukoncentruotų dydžių, kurie kelė sunkumus sprendimo medžio apmokymui, nes didžioji dalis išeinančio atributo buvo paskirstyta tarp nedidelių intervalų. Todėl mano sprendimų medis negavo daug įvairovės ir gautas tikslumas tesiekia 56%, o paklaida yra 0,5. Tai yra sąlyginai didelė paklaida, kai žinome, kad mano duomenų rinkinyje pilnas 1 keičia atributo reikšmę. Testavimo pabaigoje suvokiau, kad šis duomenų modelis kogeru nebuvo pats geriausias šiai užduočiai.

Gauti rezultatai:

Accuracy: 0.55625

MAE: 0.496875



pav. 3 Sumaišymo matrica

Po testavimų, sudariau kelis medžius skirtinguose gyliuose ir kiekvienam iš šių medžių buvo apskaičiuotas tikslumo balas bei sudarymo laikas.

Gauti rezultatai:

```
Max Depth: 3, Accuracy: 0.528125, Time: 0.004413127899169922 seconds
Max Depth: 4, Accuracy: 0.53125, Time: 0.004058837890625 seconds
Max Depth: 5, Accuracy: 0.559375, Time: 0.004914283752441406 seconds
Max Depth: 6, Accuracy: 0.5375, Time: 0.005736827850341797 seconds
```

Atsitiktinis miškas

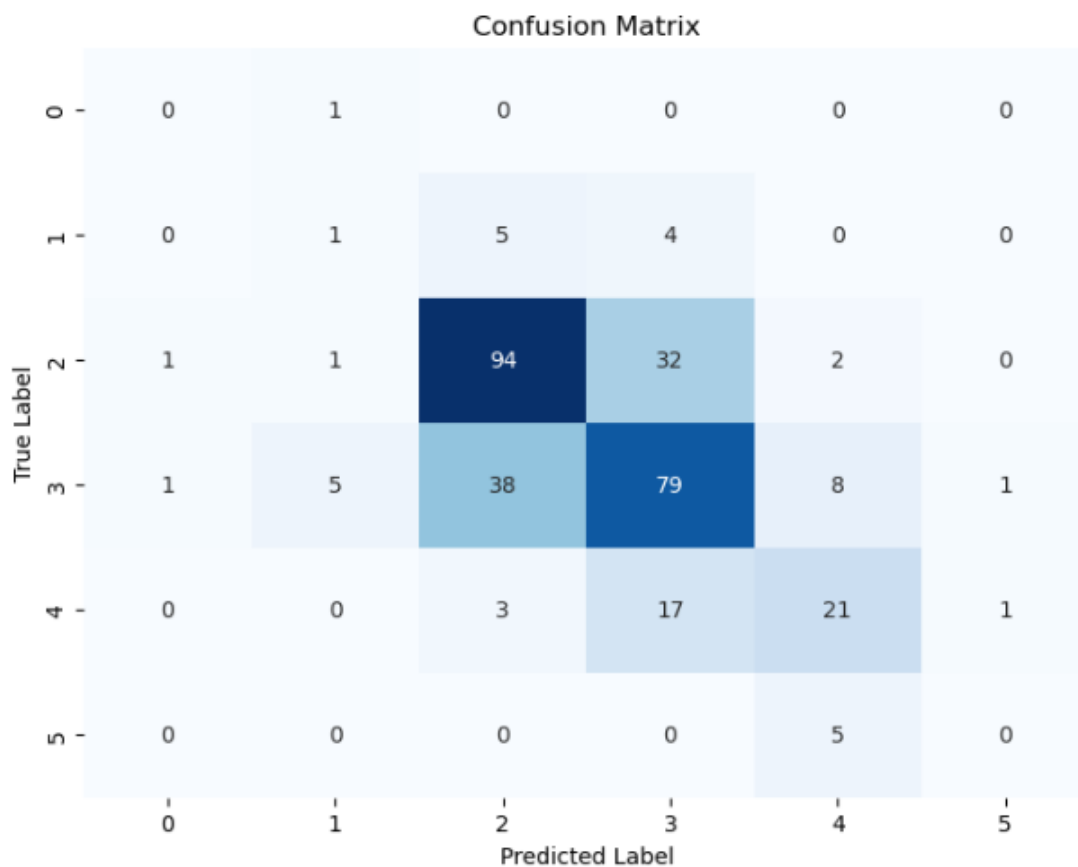
Atsitiktinis miškas yra mašininio mokymosi algoritmas, kuris naudoja daugybę sprendimų medžių. Kiekvienas sprendimų medis yra mokomas atsitiktinai pasirinktų poaibių iš pradinio duomenų rinkinio ir tuo pačiu kiekvienas medis balsuoja dėl prognozės. Galutinis sprendimas yra sudaromas remiantis daugumos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

Pagal funkcinius reikalavimus reikia sudaryti atsitiktinį mišką, kurį sudaro 5 sprendimo medžiai. Šiuos visus suskaičiuoti, su gautu galutiniu sprendimų medžiu atlikti testavimus.

Gauti rezultatai rodo, kad atsitiktinis miškas gaunasi šiek tiek tikslesnis negu sprendimų medis. Atsitiktinis miškas pasiekė 61% tiklumą bei 0,45 paklaidą. Šie dydžiai neitin žymiai skiriasi nuo paprasto sprendimo medžių miško, kurio tikslumas siekė 56%.

Gauti rezultatai:

```
Number of Trees: 5, Accuracy: 0.609375
MAE: 0.446875
```



pav. 4 Atsitiktinio miško sumaišymo matrica.

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

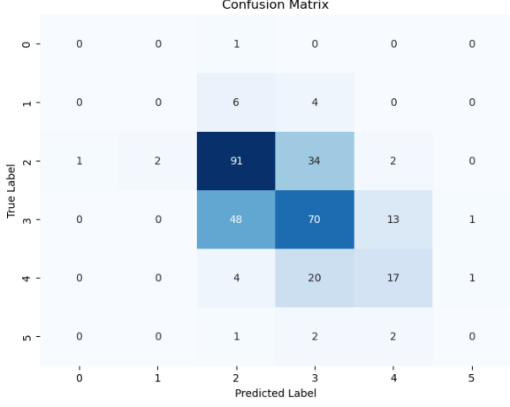
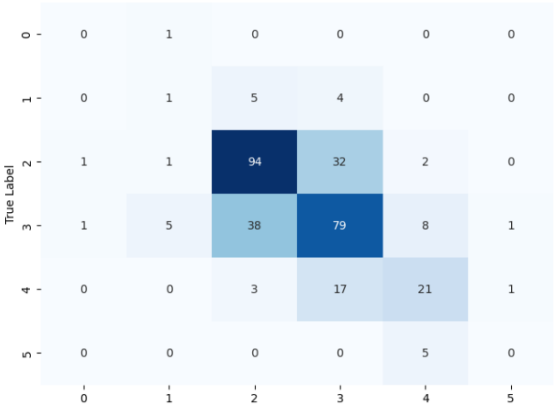
Gauti rezultatai:

Number of Trees: 3, Accuracy: 0.540625, MAE: 0.55625
 Number of Trees: 4, Accuracy: 0.60625, MAE: 0.45625
 Number of Trees: 5, Accuracy: 0.609375, MAE: 0.446875
 Number of Trees: 6, Accuracy: 0.61875, MAE: 0.4125
 Number of Trees: 7, Accuracy: 0.60625, MAE: 0.4375
 Number of Trees: 8, Accuracy: 0.634375, MAE: 0.396875
 Number of Trees: 9, Accuracy: 0.646875, MAE: 0.3875

Kaip matome pagal gautus rezultatus, kuo daugiau medžių naudojame miško kūrimui, tuo tikslesni rezultatai patampa, toliau reikėtų patestuoti iki kokios ribos tikslumas bei MAE įvertis nustoja tobulėti. Tačiau šiuo atveju, geriausius rezultatus pavyko gauti su 9 medžiais miške.

Palyginimas

Sugeneravus sprendimų medį ir atsitiktinį mišką, galima lengvai juos palyginti bei nustatyti, kuris iš jų yra pranašesnis metodas. Nors ir paprastas sprendimų medis turėjo apie 56% tikslumą, kuris praktiškai yra kaip monetos metimas, bet šis tikslumas priklauso nuo trūkstamų žemesnių reikšmių duomenų rinkinyje, tačiau su ta pačia problema susiduria ir atsitiktinis miškas, bet dėl didesnio medžių kiekio atsitiktinis miškas geba geriau išgryninti rezultatus ir pasiekti didesnę tikslumą su mažesne paklaida. Kaip matome rezultatuose pavaizduotuose žemiau, sprendimų medžio tikslumas tebuvo 56% ir turėjo didelę 0,49 paklaidą, tuo tarpu 9 medžių atsitiktinis miškas turėjo 65% tikslumą, bei 0,39 paklaidą. Skirtumas tarp šių dviejų metodų yra ganėtinai didelis ir parodo, kad atsitiktinio miško metodas yra daug patikimesnis kai yra norima nustatyti tikslumą ir gauti geresnį/tikslesnį atsakymą.

Sprendimų medis						Atsitiktinis miškas					
Gauti rezultatai:						Medžių kiekis: 9,					
Tikslumas: 0.556						Tikslumas: 0.647					
Paklaida: 0.497						Paklaida: 0.388					
<p>Confusion Matrix</p> 						<p>Confusion Matrix</p> 					

Išvados

Galiu teigti, jog užduotį pavyko sėkmingai įvykdyti, nors pasirinktas duomenų rinkinys nebuvo pats geriausias. Sprendimų medžio tikslumas tesiekė 56%, o atsitiktinio miško tikslumas siekė netgi 65%. Pagal šį 9% pokytį ir 0,12 skirtumą tarp MAE reikšmių skirtumo (suapvalinus iki šimtųjų) galiu teigti, jog atsitiktinio miško metodas yra tikslesnis bei geresnis

Gynimo rezultatai

Sugeneruoti duomenis, juos suklasifikuoti, pavaizduoti medžiu ir tada sprendimo kodui pritaikyti laboratoriniam darbui naudotą duomenų rinkinį

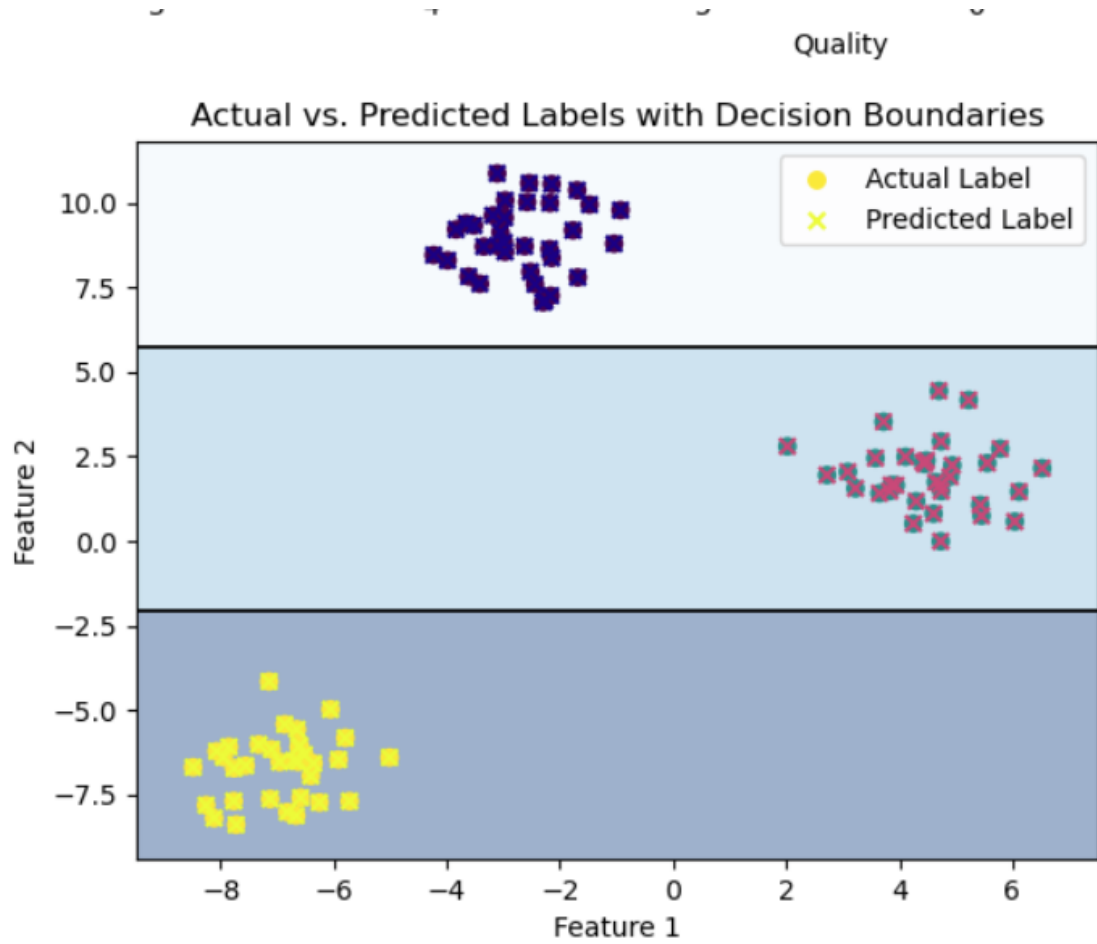


Figure 2 Spendimų medžio sugrupuoti ir atvaizduoti grafiškai

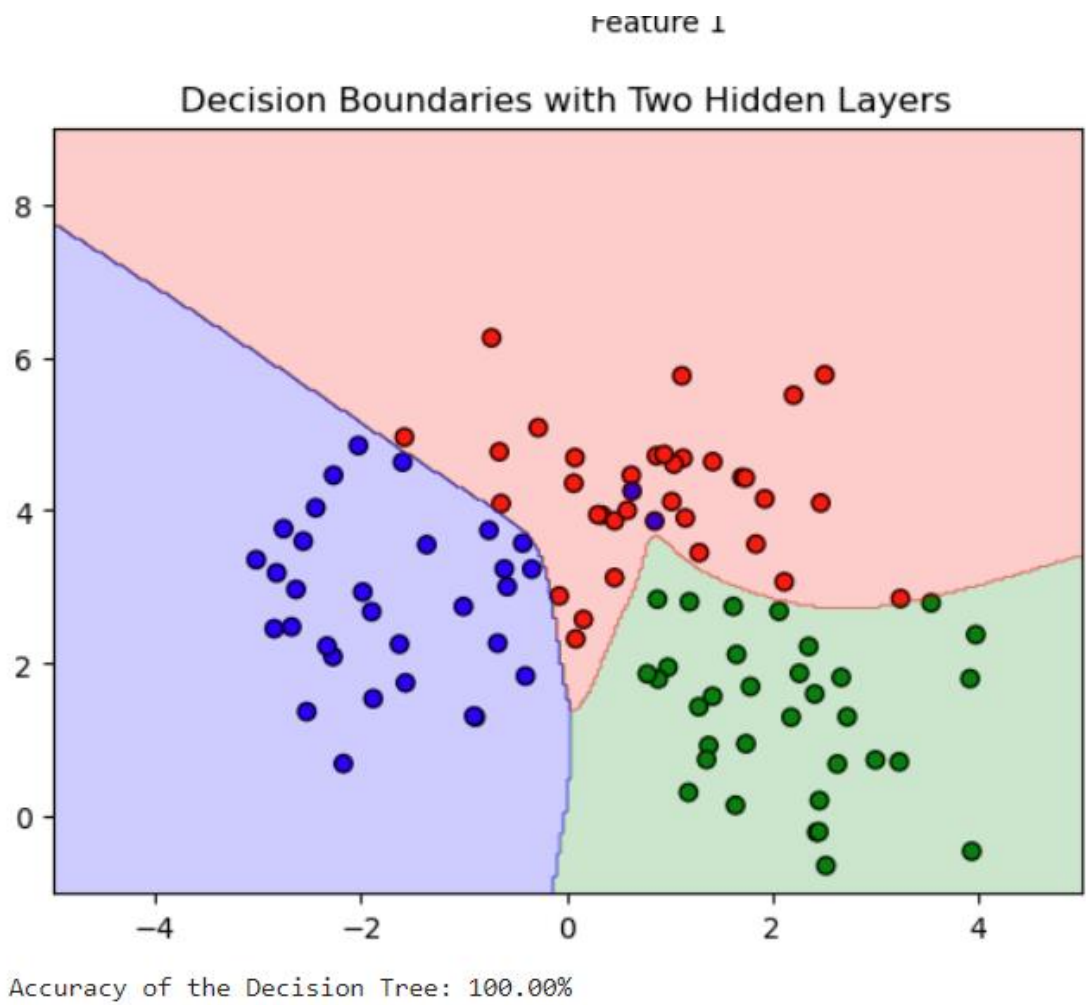
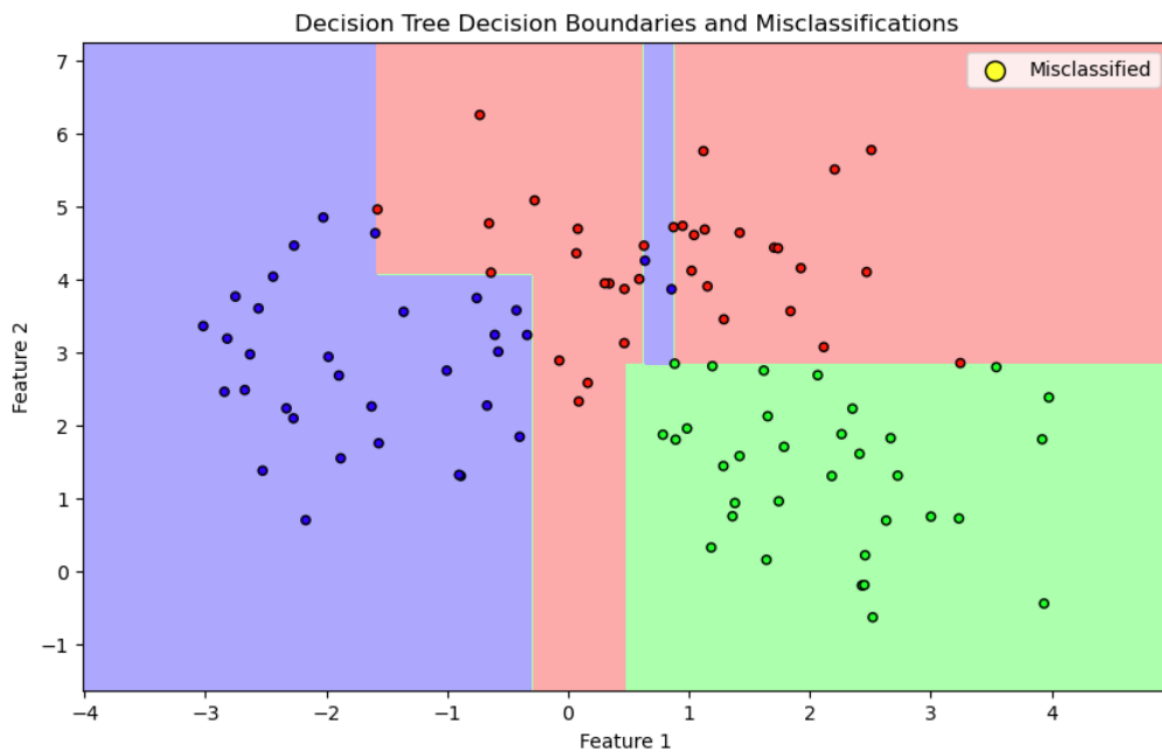


Figure 3 Sprendimų medžio rezultatai atvaizduoti scatter plot ir subrėžtos ribos



Accuracy of the Decision Tree: 100.00%

Figure 4 Atvaizduoti rezultatai parodant klaidas

Confusion Matrix:
[[34 0 0]
[0 33 0]
[0 0 33]]

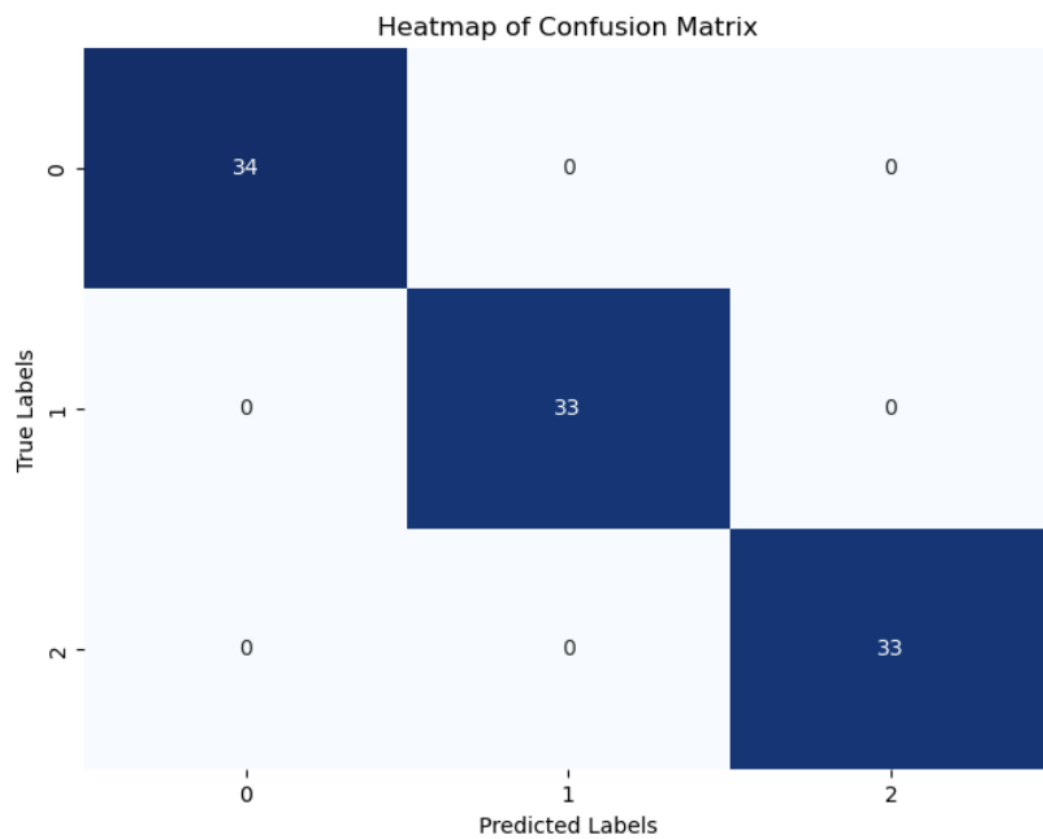


Figure 5 Sugeneruotų duomenų klasifikacijos heatmapas

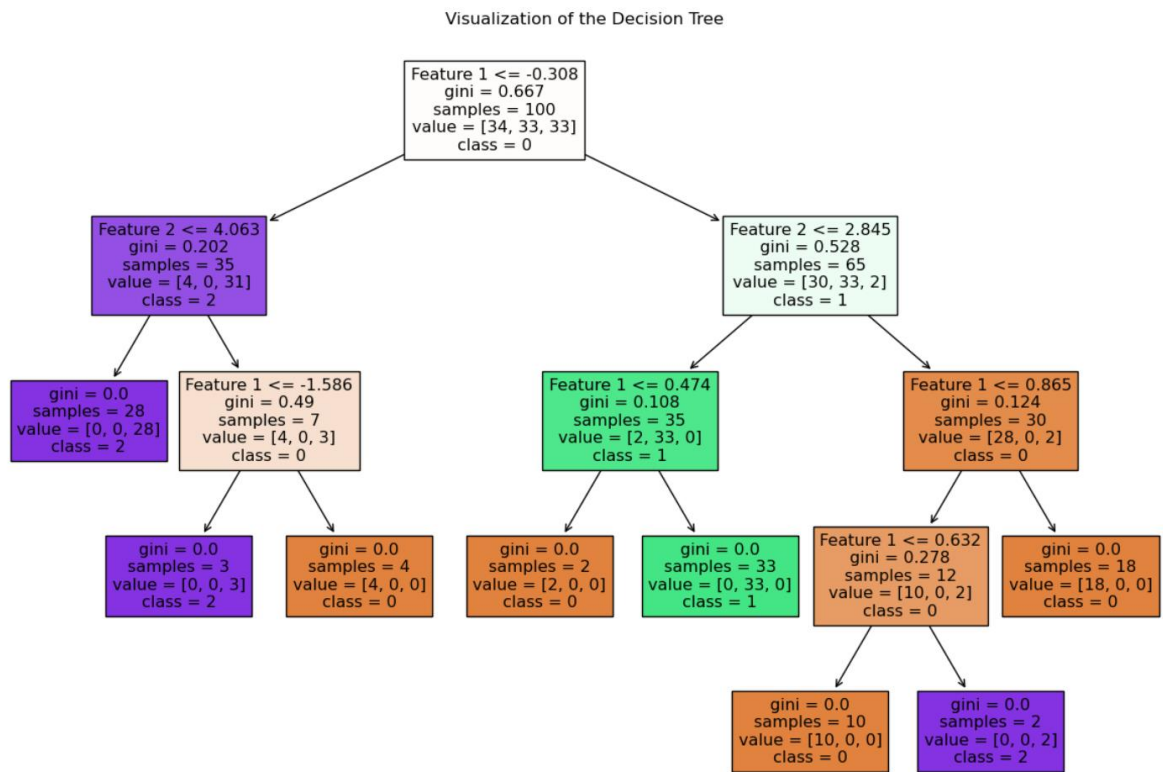
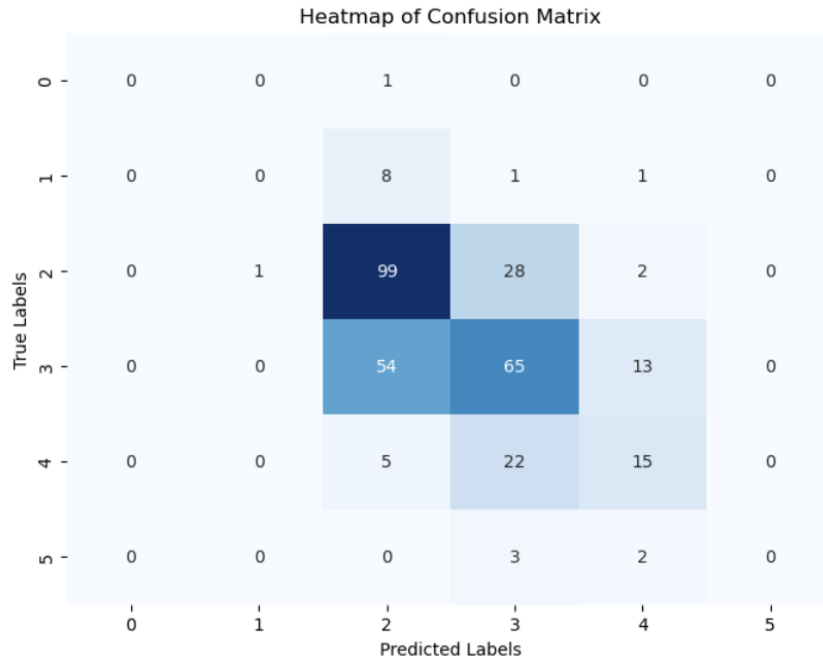


Figure 6 Sugeneruotų duomenų atvaizdavimas medžiu

Accuracy of the Decision Tree: 55.94%

Confusion Matrix:

```
[[ 0  0  1  0  0  0]
 [ 0  0  8  1  1  0]
 [ 0  1 99 28  2  0]
 [ 0  0 54 65 13  0]
 [ 0  0  5 22 15  0]
 [ 0  0  0  3  2  0]]
```



Visualization of the Decision Tree

Figure 7 Kodo panaudojimas realiems duomenims, atvaizduota klasifikavimo tikslumas heatmapu

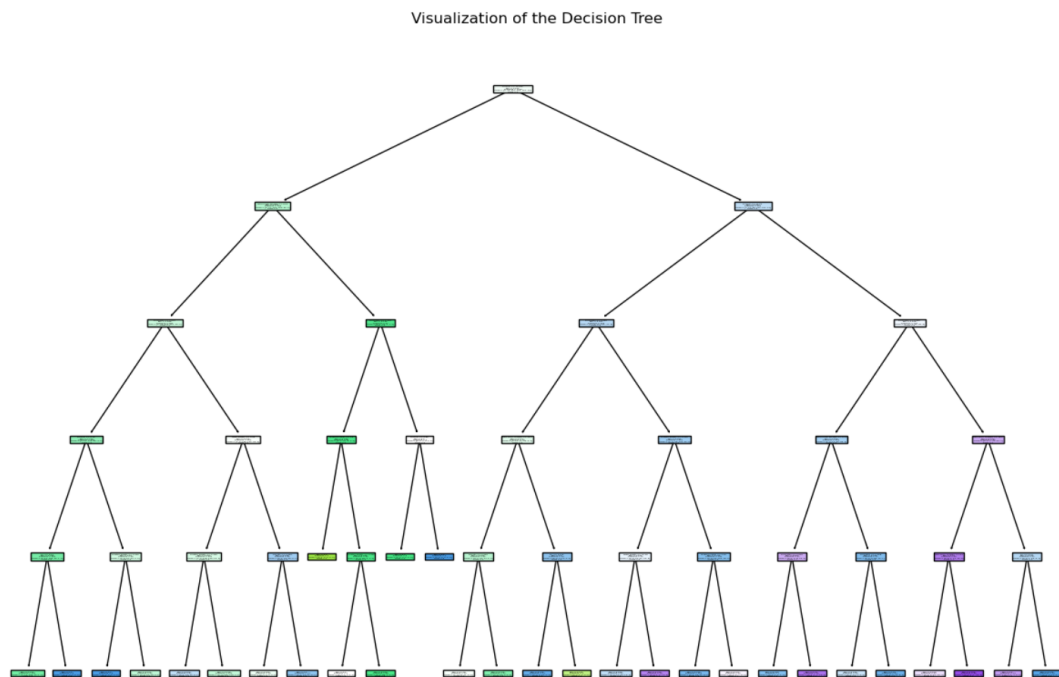


Figure 8 Kodo panaudojimas realiems duomenims atvaizduota decision tree

Accuracy of the Decision Tree: 51.25%

Confusion Matrix:

```
[[ 0  0  1  0  0  0]
 [ 0  0  5  5  0  0]
 [ 0  0 96 31  3  0]
 [ 0  0 56 57 19  0]
 [ 0  0  5 26 11  0]
 [ 0  0  0  1  4  0]]
```

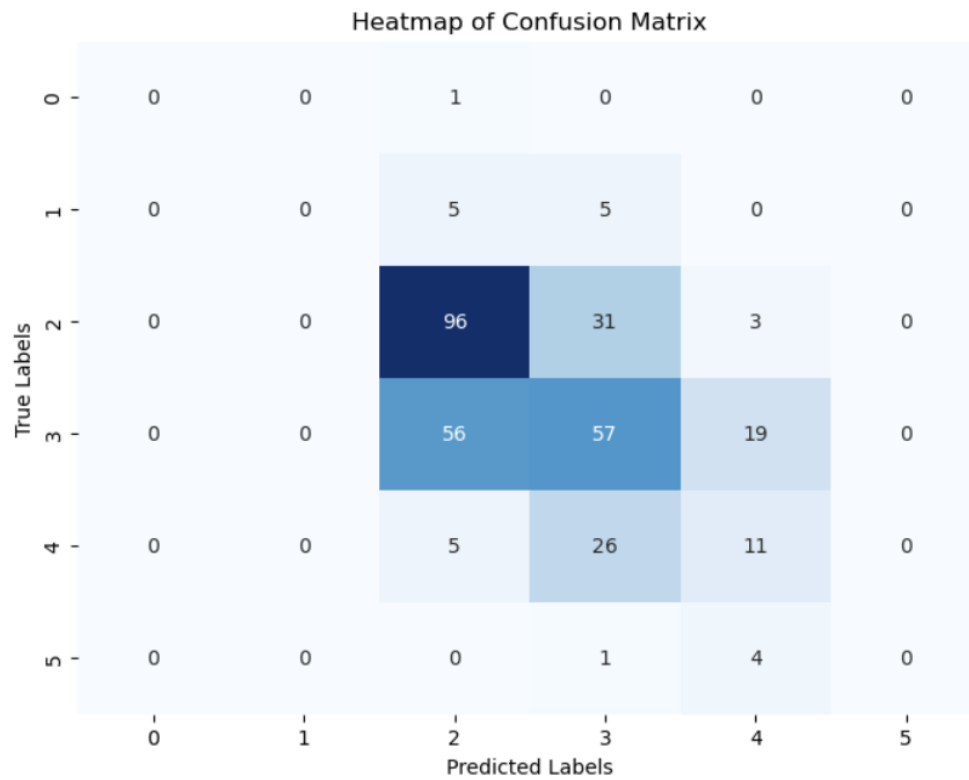


Figure 9 Klasifikavimo tikslumo atvaizdavimas heatmap

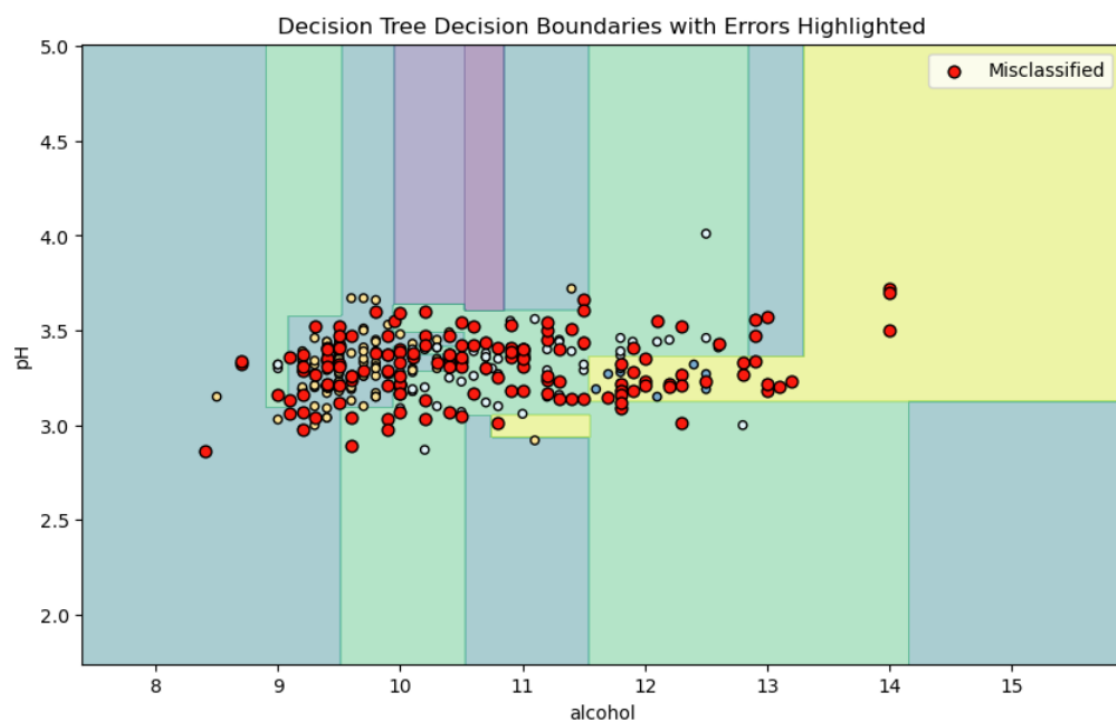


Figure 10 Atvaizduota klaidų ir klasifikavimo rezultatų scatterplot su ribom