**RĪGAS TEHNISKĀ UNIVERSITĀTE**

Datorzinātnes un informācijas tehnoloģijas fakultāte

Vadības informācijas tehnoloģijas katedra

**Megija Krista Supe**

bakalaura studiju programmas

students, stud. apl. nr. 201RDB101

<https://github.com/megijasupe/machinelearning>

**Mašīnmācīšanās algoritmu pielietojums**

**2. praktiskais darbs**

RĪGA 2023

1. Datu pirmapstrāde un izpēte

1.1 Informācija par izvēlēto datu kopu

Darba izpildei tika izmantota sarkanā vīna kvalitātes datu kopa, kura atbilst darba ietvaros esošajiem nosacījumiem. Šī datu kopa tika iegūta UC Irvine mašīnmācīšanās krātuvē. Datu kopas nosaukums ir “Wine quality red”. Šī datu kopa saitē ir iesniegta 2009. gadā, un to ir iesniedzis Paulo Cortez. (Cortez, Cerdeira, Almeida, & Matos, 2009.)

Izvēlētā datu kopa ir dati par portugāļu sarkanā vīna variantiem, kas ir konkrēti “Vinho Verde” vīni. Datu kopā ir pieejama informācija par vīna kvalitātes mērījumiem un kāda ir to vērtību ietekme uz vīna garšu. Konkrēti šī kopa ir parocīga, lai veiktu mašīnmācīšanās uzdevumus, piemēram, regresijas vai klasifikācijas uzdevumus, jo ir iespējams noteikt vīna kvalitāti, atsaucoties uz datiem, kas ir doti par vīna īpašībām. Datu kopa ir pieejama publiski jebkuram lietotājam, taču ar noteikumu, ka, izmantojot šo datu kopu, ir obligāti jāpievieno atsauces tipa informācija, ka tieši tiek izmantota konkrētā datu kopa, šis paskaidro licencēšanas nosacījumus kopai. (Cortez, Cerdeira, Almeida, & Matos, 2009.)

Datu kopas datu iegūšanai tika atrasta kopa, kas atbilst šī darba nosacījumiem, piemēram, jābūt pieejamiem vismaz 200 datu objektiem, datu kopai jāsatur klašu iezīmes un citi nosacījumi, kas ir aprakstīti darba nosacījumos. Šī datu kopa ir pieejama csv formātā. Datu kopas iegūšanai tika lejupielādēta datu mape un izvēlēta konkrēti kopa par sarkanvīniem, kaut gan bija iespēja izvēlēties arī baltvīna klasifikāciju. Datu kopas veidotāji datus ieguva, pamatojoties uz fizikāli ķīmisko un sensoro testu veikšanu. Sensorie testi balstās uz cilvēka resursiem. (Cortez, Cardeira, & Almeida, 2009.)

Datu kopā ir 1599 datu objekti un 12 pazīmes jeb atribūti, kuri raksturo sarkanvīna fizikālās un ķīmiskās īpašības. Sekojoši tiks minēts datu kopas atribūtu atspoguļojums (Cortez, Cerdeira, Almeida, & Matos, 2009.)

1. Folskābes saturs (fixed acidity) – kvantitatīvie dati, lai noteiktu skābes daudzumu vīnā.
2. Gaistošā skābe (volatile acidity) – kvantitatīvie dati, lai noteiktu kaitīgo skābju daudzumu.
3. Citronskābes saturs (citric acid) – kvantitatīvie dati, kas ļauj regulēt pH līmeni.
4. Atlikušā cukura daudzums (Residual sugar) – kvantitatīvie dati, kas nosaka atlikušā cukura daudzumu pēc fermentācijas procesa.
5. Hlorīdu saturs (chlorides) – kvantitatīvie dati, kas nosaka sāļu daudzumu.
6. Brīvais sēra dioksīds (free sulfur dioxide) – kvantitatīvie dati, kas nosaka brīvo sēru dioksīda daudzumu.
7. Kopējais sēra dioksīds (total sulfur dioxide) – kvantitatīvie dati, kas nosaka kopējā sēru dioksīda daudzumu.
8. Blīvums (density) – kvantitatīvie dati, lai noteiktu masas attiecību pret tilpumu.
9. pH līmenis (pH) – kvantitatīvie dati, lai noteiktu skābuma līmeni vīnā.
10. Sulfātu daudzums (sulphates) – kvantitatīvie dati, kas apskaidro sēra dioksīda daudzumu vīnā.
11. Alkohola saturs (alcohol) – kvantitatīvie dati, lai noteiktu alkohola procentuālā daudzumu vīnā.
12. Kvalitāte (quality) – kvalitatīvie dati, vīna kvalitātes vērtējums, kas attiecināms skalā no 0 līdz 10. (THEFULLSTORYEDUCATIONTEAM, 2021.)

Konkrētajā datu kopā ir viena klasifikācija un tā šajā kontekstā ir sarkanvīna kvalitāte, kā iepriekš minēts, tad tā ir skalā no 0 līdz 10, taču konkrētajā datu kopā vērtību diapazons ir no 3 līdz 8.

Atsaucoties uz (Miranda, 2017.) datiem, kur tiek klasificēti vīni pēc to kvalitātes, darba ietvaros vērtība “kvalitāte” tika veidota ar Orange “Create class” rīku, kur izmantotās vērtības no 3 līdz 8, tika norādītas šādi:

* Failure (izgāzšanās) – 1-2
* Very Poor (ļoti slikts) – 3
* Poor (slikts) – 4
* Below Average (zem vidējā) –5
* Average (vidējais) - 6
* Good (labs) – 7
* Very good (ļoti labs) – 8
* Great (lielisks) – 9
* Outstanding (izcils) - 10

Taču jāpiemin, ka vērtības, kas ir skalā no 1-2 un 9-10 netiek izmantotas, jo datu kopā šādu vērtību “quality” kolonnas ietvaros nav.

1.1. tabulā ir atspoguļoti atribūti, to nozīme datu kopā, vērtību tipi un to diapazona. Lai atrastu atribūtu diapazonu tika izmantota Microsoft excel formula, kas nosaka minimālo un maksimālo vērtību konkrētā kolonā. Formula minimālajai vērtībai izskatās šādi: “=min(L2:L1600)”, kur “L2” un “L1600” ir konkrētās iezīmētās kolonnas, kurās vēlas noteikt minimālo vērtību, tas pats tika darīts ar maksimālo vērtību, formulas sākums tika mainīts uz ““=max(L2:L1600)”. Atribūtu vērtības tika noteikts ar pārskatīšanu, darba autors vizuāli aplūkoja Microsoft excel failā redzamās vērtības atbilstoši katrai kolonnai un noteica tipu. Savukārt, 1.1.attēlā ir redzams datu faila struktūras fragments, kurā ir redzamas visas kolonnas un 20 vērtības.

**1.1.tabula**

**Atribūtu nozīme, vērtību tips un diapazona**

|  |  |  |  |
| --- | --- | --- | --- |
| **Atribūta nosaukums** | **Atribūta nozīme** | **Atribūta vērtību tips** | **Atribūta datu diapazona** |
| Folskābes saturs | Skābes daudzums vīnā | Skaitlis | 5-45184 |
| Gaistošā skābe | Kaitīgo skābju daudzums | Skaitlis | 0,12-45292 |
| Citronskābes saturs | Citronskābes daudzums | Skaitlis | 0 - 1 |
| Atlikušā cukura daudzums | Atlikušo cukuru daudzums | Skaitlis | 0,9-45748 |
| Hlorīdu saturs | Sāļu daudzums | Skaitlis | 0,012 - 0.611 |
| Brīvais sēra dioksīds | Brīvo sēra dioksīdu daudzums | Skaitlis | 1 - 45051 |
| Kopējais sēra dioksīds | Kopējā sēru dioksīda daudzums | Skaitlis | 0,99007-1,00369 |
| Blīvums | Masas attiecība pret tilpumu | Skaitlis | 0,99007 – 1,00369 |
| pH līmenis | Skābuma līmenis vīnā | Skaitlis | 3-47178 |
| Sulfātu daudzums | Sēra dioksīda daudzums vīnā | Skaitlis | 0,33 - 46753 |
| Alkohola saturs | Alkohola procentuālais daudzums | Skaitlis | 9-45901 |
| Vīna kvalitāte | Kvalitātes mērījums | Kategorisks | 3-8 |

Attēls, kurā ir teksts, cipars, ekrānuzņēmums, fonts

Apraksts ģenerēts automātiski

1.1.attēls Datu faila struktūras fragments

* 1. . Datu kopas vizuālais atspoguļojums un statistiskie rādītāji

Nodaļas ietvaros tika izveidotas divas 3-dimensiju izkliedes diagrammas Orange rīkā ar “Scatter plot” funkciju. Tika izveidota viena 3-dimensiju izkliedes diagramma, kas nosaka alkohola un sulfītu ietekmi uz vīna kvalitāti, un otra, kas nosaka blīvuma un sulfātu ietekmi uz vīna kvalitāti. Atsaucoties uz 1.2 un 1.3 attēliem, var teikt to, ka identificētie datu grupējumi abos gadījumos atrodas diezgan tuvu viens otram, to ietekmē tas kādus atribūtus izvēlās kā mainīgos uz x un y asīm. 1.2 attēlā veidotā izkliedes diagramma ir vairāk pārskatāma, un ir redzams tas, ka tomēr ir atdalāmi datu grupējumi. Taču, lielākoties šajā datu kopā datu grupējumi nav tik labi atdalāmi.

1. Ja alkohola daudzums (y ass) un sulfātu daudzums (x ass) (kvalitāte - z ass) vīnā ir zems, tad lielākoties dominē zem vidējas kvalitātes vīns.
2. Attēls, kurā ir ekrānuzņēmums, teksts, programmatūra, multivides programmatūra

   Apraksts ģenerēts automātiskiJa sulfātu daudzums (x ass) ir zems, bet alkohola daudzums (y ass) ir liels, tad dominē vidējas un labas kvalitātes vīns.

1.2. att. Alkohola un sulfātu ietekme pret kategoriju.

1. Ja blīvuma daudzums (x ass) ir liels, bet sulfātu daudzums (y ass) ir zems, tad dominē zem vidējās kvalitātes vīns.
2. Attēls, kurā ir ekrānuzņēmums, teksts, programmatūra, multivides programmatūra

   Apraksts ģenerēts automātiskiJa blīvuma lielums un sulfātu daudzums ir viena līmenī, tad dominē vairākas vīna kvalitātes klasifikācijas.

1.3.att. Blīvuma un sulfātu ietekme pret kategoriju.

Attēls, kurā ir teksts, ekrānuzņēmums, programmatūra, displejs

Apraksts ģenerēts automātiskiAr Orange rīka “Distributions” palīdzību tika noskaidrots tas, ka šī datu kopā nav līdzsvarota, jo 1.4. attēlā var spilgti redzēt to, ka konkrēti dominē divas klases, kas ir zem vidējā vīna kvalitāte (681 ieraksts, kas ir 42,59% no visiem ierakstiem) un vidēja vīna kvalitāte (638 ieraksti, kas ir 39,9% no visiem ierakstiem). Nākamais lielākais pārsvars ir klasei labs vīns (199 ieraksti, kas ir 12,45% no visiem ierakstiem), sekojoši slikts (53 ieraksti, kas ir 3,31% no visiem ierakstiem), ļoti labs (18 ieraksti, kas ir 1,13% no visiem ierakstiem) un ļoti slikts (10 ieraksti, kas ir 0,63% no visiem ierakstiem).

1.4. att. Klašu līdzsvars

Attēls, kurā ir teksts, ekrānuzņēmums, programmatūra, diagramma

Apraksts ģenerēts automātiski1.5. un 1.6 attēlos var redzēt to, ka dažādās klasēs esošie datu objekti ir skaidri atdalāmi un var redzēt datu struktūru, tas ir, stabiņu diagrammā dati nepārklājas.

1.5.att. Datu objektu iedalījums pēc kopējā sēra dioksīda daudzuma

Attēls, kurā ir programmatūra, grafikas programmatūra, multivides programmatūra, skice

Apraksts ģenerēts automātiski

1.6.att. Datu objektu iedalījums pēc pH līmeņa

Ar Orange rīku un “Feature statistics” tika veikti secinājumi par statistiskajiem rādītājiem. Tika iegūti rādītāji, kas redzami 1.7.attēlā. Vidējā vērtība pH līmenim ir 3,31, kā zināms, tad to aprēķina saskaitot visas esošās ph līmeņa vērtības un, izdalot ar to, skaitu. pH līmeņa mediāna ir 3,31, ja datu kopā ir nepāra skaitlis ar vērtībām, šajā datu kopā ir 1599 vērtības, kas ir nepāra skaitlis, tātad pie 1559 pieskaita 1 un sadala ar divi, tas ir, 800. 800 vērtība pH līmeņa kolonnā ir 3,29. Pēc “Feature statistics” rezultātiem tā sanāk 3,31, kā rezultātā ir radusies neprecizitāte. 1.7.attēlā ir redzama vērtība dispersija, kas šajā gadījumā ir 4,6%, tas nozīmē to, ka datu kopa ir viendabīga.



1.7.att. Statistiskie rādītāji

2. Nepārraudzītās mašinmācīšanās algoritmi

Nodaļas ietvaros tika apskatīti divi nepārraudzītas mašīnmācīšanās algoritmi: hierarhiskā klasterizācija un K-vidējo algoritms. Lai darbotos Orange rīkā ar šiem algoritmiem, hierarhiskajai klasterizācijai Orange rīkā tika izmantoti: “Distance” un “Hierarchical Clustering” rīki:

1. “Distance” logā tika izvēlēta distance starp rindām (distance between rows) un sadaļā, kas aprēķina attālumu starp datu kopas kolonnām vai rindām (distance metric) tika izvēlēti parametri:

* Leņķis kosinuss (cosine), kas ir starp diviem vektoriem.
* Eiklīds (euclidean), kas ir attālums starp diviem punktiem.
* Manhetena (manhattan), kas ir absolūto atšķirību summa visiem atribūtiem.

2. “Hierarchical Clustering” logā sadaļā savienojums (linkage) tika izvēlēti sadalījuma veidi:

* Pēc pilnīga sadalījuma (complete linkage), kas aprēķina attālumu starp klasteru visattālākajiem elementiem.
* Pēc viena sadalījuma (single linkage), kas nosaka attālumu starp tuvākajiem divu klasteru elementiem.
* Pēc vidējā sadalījuma (average linkage), kas nosaka vidējo attālums starp divu klasteru elementiem.

3. K-vidējo algoritmu noteikšanai tika izmantoti Orange iebūvētie rīki “k-Means” un “Silhouette Plot”. “k-Means” logā tika izvēlēti 3 hiperparametri, tie ir:

* Sadaļā klasteru skaits (number of clusters), kas šajā gadījumā būs 5
* Tie tika izvēlēts klasteru diapazons no 2-8, kas nosaka klasterizācijas rezultātus izvēlētajam klasteru diapazonam.
* Trešais hiperparametrs ir inicializācija (initialization), šoreiz tika izvēlēta inicializācija pēc Kmeans++ (Initialize with Kmeans++), kas nozīmē, ka pirmais centrs tiek izvēlēts nejauši, taču nākamie tiek izvēlēti no atlikušajiem punktiem ar varbūtību.

Sekojoši tika veikti uzdevumi, atsaucoties uz iepriekš iegūto informāciju Orange Data Mining dokumentācijā. (OrangeDataMining, 2015.)

Hierarhiskā klasterizācija:

Attēls, kurā ir diagramma, taisnstūris, teksts, ekrānuzņēmums

Apraksts ģenerēts automātiski2.1.attēlā dati tiek ievākti izmantojot distances sadaļā leņķa kosinuss un hierarhiskā klasterizācijas daļā pie savienojuma tika izvēlēts pilnīgs savienojums. 2.1.attēlā nav pilnīgi redzams, taču tika sadalīts 5 klasteros, un katrā no klasteriem lielākoties dominē vidējas un zem vidējas kvalitātes vīni.

2.1.att. Dendogramma pēc N=5 izvēles

2.2.attēlā tika izmantots distances sadalījums pēc manhetana un hierarhiskā klasterizācijas daļā pie savienojuma tika izvēlēts viens sadalījums. 2.2.attēlā var redzēt to, ka izvēloties manuālu sadalījuma veidu izveidojas divi klasteri. Pirmajā klasterī iedalās tikai labas kvalitātes vīni, un otrajā klasterī lielākoties dominē vidējas un zem vidējas kvalitātes vīni, taču ir redzami arī daži labas un sliktas kvalitātes vīni, bet to ir maz. Savukārt 2.3.attēlā tika sadalīts pēc N=5 sadalījuma. Tur dendogramma tiek sadalīta 5 klasteros, 1. klasterī iedalās tikai labas kvalitātes vīni, 2. klasterī vidējas kvalitātes vīni, 3. klasterī sliktas kvalitātes vīni, 4. klasterī ir tikai zem vidējās kvalitātes vīni un 5. klasterī ir visas iekļautās kvalitātes.

Attēls, kurā ir teksts, ekrānuzņēmums, taisnstūris

Apraksts ģenerēts automātiski

Attēls, kurā ir teksts, ekrānuzņēmums, diagramma, rinda

Apraksts ģenerēts automātiski2.2.att. Dendogramma pēc manuāla sadalījuma

2.3.att. Dendogramma pēc N=5 sadalījuma

Pēdējai sadalījums tika izvēlēts sekojoši: distances sadaļā tika izvēlēts eiklīda sadalījums, un hierarhiskā klasterizācijas daļā pie savienojuma tika izvēlēts vidējais sadalījums. 2.4.attēlā var redzēt to, ka pēc N=5 sadalījuma veidojas 5 klasteri. 1. klasterī iedalās tikai labas kvalitātes vīni, 2. klasterī un 3. klasterī iedalās visu veida vīnu kvalitātes, bet 4. klasterī un 5. klasterī iedalās tikai vidējas un zem vidējas kvalitātes vīni.

Attēls, kurā ir teksts, ekrānuzņēmums, programmatūra, taisnstūris

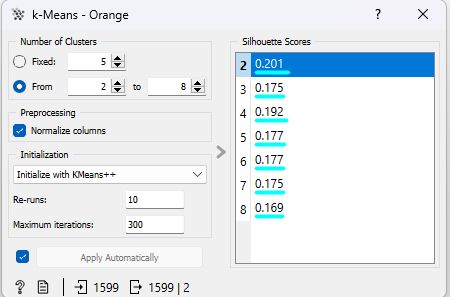
Apraksts ģenerēts automātiski

2.4.att. Dendogramma pēc N=5 iedalījuma

Pēc hierarhiskā klasterizācijas veikšanas var secināt to, ka datu kopā esošās klases ir ne pārāk labi atdalāmas, taču nav sliktākais gadījums, vairākās dendogrammās var redzēt to, ka klasteros klašu sadalījums iedalās labi, bet dažos vienā klasterī dominē visi vīna kvalitātes veidi. Tas varētu būt atkarīgs no tā, kādus parametrus izvēlās distances sadaļā un hierarhiskā klasterizācijas sadaļā. Iespējams, izvēloties citus parametrus, veidoties labāks pārskats par to, kā datu kopā esošās klases atdalās.

K-vidējo algoritms:

Klasteru skaitā tika izvēlēts no 2 līdz 8 un pie inicializācijas tika izvēlēts inicializēt pēc kMeans++. “kMeans” logā, kas redzams 2.5.attēlā var redzēt to, cik klasteros būtu efektīvāk sadalīt datu objektus. Orange rīks norāda to, ka efektīvāk būt sadalīt divos klasteros, taču fiksēti ir 5 klasteri.



* 1. att. “kMeans” logs Orange rīkā

Lai izprastu šo sniegtu informāciju sekojoši tika pievienots Orange rīks “Silhouette Plot”. Šajā rīkā distance tiks izvēlēta eiklīda, grupēts pēc klasteriem un kā anotācija, tiks izvelētas vīnu kvalitātes kategorijas. 2.6.attēlā redzams tas, ka liela daļa no datu kopas objektiem ir attālu no klasteru centriem, kas nozīmē to.

Attēls, kurā ir ekrānuzņēmums, teksts, skice, rinda

Apraksts ģenerēts automātiskiAttēls, kurā ir teksts, ekrānuzņēmums, skice, diagramma

Apraksts ģenerēts automātiski

2.6.att. 'Silhouette Plot sadalījums

Kopumā no nodaļas ietvaros veiktajiem uzdevumiem var teikt to, ka, piemēram, sadalot 5 klasteros, dažviet var labi atdalīti vīnus pēc to kvalitātes, ja vēlas iegūt līdzīgas kvalitātes vīnus. Piemēram, ja vēlas iegūt tikai vidējas un zem vidējas kvalitātes vīnus, tad klasteru sadalījums ir labs, bet, ja grib iegūt klasteri, kur būs tikai un vienīgi vidējas kvalitātes vīni, tad klases nav tik labi atdalītas.

1. Pārraudzīta mašinmācīšanās

Nodaļas ietvaros tika izmantoti divi algoritmi un aprakstīta to nozīme. Darba autors izvelējās “Logistic regression” jeb loģistiskā regresija, “Random forest” jeb nejaušais mežs un “Neural network” jeb neironu tīkli. Sekojoši tiks aprakstīts pirmie divi pirms tam minētie algoritmi.

* Loģistiskā regresija – ievades ir ievades datu kopa un pirmapstrādes metodes. Izejas ir loģistiskās regresijas mācīšanās algoritms, apmācīts modelis un loģistiskās regresijas koeficienti. Loģistisko regresiju izmanto, lai aprakstītu datus un izskaidrotu attiecības starp vienu atkarīgu bināro mainīgo un vienu vai vairākiem nominālajiem, kārtas, intervāla vai attiecības līmeņa neatkarīgiem mainīgajiem. Šis algoritms tika izvēlēts, jo tas darbojas korekti, ja jūsu datu kopa ir lineāri atdalāma, kā arī to ir viegli īstenot, interpretēt un apmācīt. (Gupta, 2020.)
* Loģistiskā regresija Orange rīkā:

Tika izvēlēti divi hiperparametri, tie ir, “regularization type”, kur var izvēlēties vai nu L1, vai L2. L1 ir normas zuduma funkcija, pazīstama kā mazākās absolūtās novirzes un mazākās absolūtas kļūdas. L2 ir normas zuduma funkcija, to sauc arī par mazāko kvadrātu summu. L1 ir saukts par Lasso un L2 saukts par Ridge. Tālāk ir “strenght” parametrs, kas robežās no weak (vājš) līdz strong (stiprs). (Aditya, 2018.)

* Nejaušais mežs – ievades ir ievades datu kopa un pirmsapstrādes metodes. Izejas ir izlases meža mācīšanās algoritms un apmācīts modelis. Šis algoritms veido lēmumu koku kopu. Katrs no kokiem ir veidots, pamatojoties uz saknēšanās parauga apmācības datiem. (OrangeDataMining, 2015.)
* Nejaušais mežs Orange rīkā:

Tiks izvelēti divi hiperparametri – number of trees (koku skaits) un Number of attributes considered at each split (katrā sadalījumā izvēlēto atribūtu skaits). Koku skaits norāda, cik daudz lēmumu koki tiks iekļauti mežā. Katrā sadalījumā izvēlēto atribūtu skaits norāda to, cik atribūtu tiks patvaļīgi izveidots izskatīšanai katrā mezglā. (OrangeDataMining, 2015.)

Attēls, kurā ir teksts, programmatūra, grafikas programmatūra, multivides programmatūra

Apraksts ģenerēts automātiskiModelim tika pievienots rīks “Data sampler”, kurš nosaka kopējo datu kopās esošo datu objektu skaitu. Testa datu kopai tika izvēlēti 70% (1120 datu objekti), taču apmācību datu kopai 30% (479 datu objekti). Sekojoši testa un apmācības datiem tika pievienots “distributions” rīks, lai vizuāli redzētu to, cik datu objektu no katras klases ir iekļauts apmācību un testa datu kopās, to var redzēt 3.1. un 3.2. attēlā.

Attēls, kurā ir teksts, programmatūra, grafikas programmatūra, multivides programmatūra

Apraksts ģenerēts automātiski3.1.att. Distributions testa datiem

3.2.att. Distributions apmācības datiem

Testa datiem:

* Ļoti slikts vīns – 10 datu objekti, kas ir 0,89%.
* Slikts vīns – 31 datu objekts, kas ir 2,77%.
* Zem vidējās kvalitātes - 483 datu objekts, kas ir 43,12%.
* Vidējā kvalitāte - 444 datu objekts, kas ir 39,64%.
* Labs vīns - 139 datu objekts, kas ir 12,41%.
* Ļoti labs vīns - 13 datu objekts, kas ir 1,16%.

Apmācības datiem:

* Ļoti slikts vīns – 0 datu objekti, kas ir 0%.
* Slikts vīns – 22 datu objekts, kas ir 4,59%.
* Zem vidējās kvalitātes - 198 datu objekts, kas ir 41,34%.
* Vidējā kvalitāte - 194 datu objekts, kas ir 40,50%.
* Labs vīns - 60 datu objekts, kas ir 12,53%.
* Ļoti labs vīns - 5 datu objekts, kas ir 1,04%.

Lai veiktu darbības ar “Logistic Regression”, “Random Forest” un “Neural Network”, zemāk dotajās tabulās ir noradīts, kādi parametri tiks izmantoti šajos logos un rezultāti pēc parametru pielietošanas.

3.1.tabula

Loģistiskā regresijas ievades informācija

|  |  |  |
| --- | --- | --- |
| NR. | Regularization  type | Strength (C) |
| 1. | Lasso (L1) | 5 |
| 2. | Ridge (L2) | 200 |

3.2.tabula

Loģistiskā regresijas rezultāti

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| NR. | AUC | CA | F1 | Precision | Recall |
| 1. | 0,754 | 0,588 | 0,548 | 0,571 | 0,588 |
| 2. | 0,759 | 0,588 | 0,571 | 0,569 | 0,588 |

3.3.tabula

Nejaušā meža ievades informācija

|  |  |  |
| --- | --- | --- |
| NR. | Number of trees | Number of attributes |
| 1. | 5 | 10 |
| 2. | 100 | 5 |

3.4.tabula

Nejaušā meža izvades rezultāti

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| NR. | AUC | CA | F1 | Precision | Recall |
| 1. | 0,720 | 0,580 | 0,569 | 0,565 | 0,580 |
| 2. | 0,802 | 0,653 | 0,637 | 0,635 | 0,653 |

3.5.tabula

Neironu tīklu ievades informācija

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| NR. | Naurons in h.l. | Activision | Solver | Regularization |
| 1. | 1 | Logistic | Adam | 0,0001 |
| 2. | 1000 | Relu | SGD | 0,01 |

3.6.tabula

Neironu tīklu izvades rezultāti

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| NR. | AUC | CA | F1 | Precision | Recall |
| 1. | 0,722 | 0,471 | 0,387 | 0,487 | 0,471 |
| 2. | 0,733 | 0,592 | 0,563 | 0,578 | 0,592 |

Lai analizētu rezultātus, kas tika iegūti, sākuma tika noskaidrots, kas ir konkrētās vērtības - AUC, CA, F1, Precision, Recall (OrangeDataMining, 2015.)

* AUC – laukums zem ROC līknes.
* CA – klasifikācijas precizitāte jeb pareizi klasificētu piemēru proporcija.
* F1 – precizitātes un atsaukšanas harmoniskais vidējais rādītājs.
* Precision – precizitāte, kas nosaka patieso pozitīvo gadījumu proporciju starp gadījumiem, kas klasificēti kā pozitīvi.
* Recall – patieso pozitīvo attiecību īpatsvars starp visiem pozitīvajiem gadījumiem datos.

Atsaucoties uz tabulās iegūtajiem rezultātiem var teikt to, ka vislielākā AUC vērtība 0,802 bija nejaušā meža algoritma rezultātos, kur koku skaits tika izvēlēts 100 un atribūtu skaits 5, kā arī vismazākā 0,720 ir tur pat un pie tiem pašiem parametriem. Vislielākā CA vērtība 0,653 ir nejaušā meža izvades rezultātos un mazākā neironu tīklu izvades rezultātos. Vislielākā F1 vērtība 0,637 tika sasniegta nejaušā meža izvades rezultātos, taču vismazākā neironu tīklu izvades rezultātos. Vislielākā precision vērtība 0,578 tika sasniegta neironu tīklu izvades rezultātos, taču vismazākā tur pat. Vislielākā recall vērtība 0,653 tika sasniegta nejaušā meža izvades rezultātos, taču vismazākā neironu tīklu izvades rezultātos. Salīdzinot modeļus pēc AUC vērtības, vislabākā ir pie nejaušā meža algoritma, jo lielākā AUC vērtība nosaka labāko veiktspējas kvalitāti. Salīdzinot pēc CA vērtības, nejaušā meža algoritms ir vislabākais, jo tur ir sasniegta CA vērtība 0,653, kas ir vislielākā un tā nosaka klasifikācijas precizitāti jeb pareizi klasificētu piemēru proporciju. Atsaucoties uz F1, precision un recall vērtībām, pie nejaušā meža algoritma nr.2 soļa, kur tika izvēlēti tādi hiperparametri, kā koku skaits = 100 un atribūtu skaits = 5, tika iegūtas vislabākās vērtības, no tā var secināt to, ka vislabākais darba ietvaros izmantotais algoritms bija nejaušā meža. Vissliktākā veiktspēja ir neironu tīklu algoritmam, jo tam ir viszemākās vērtības CA (0,471), F1 (0,387), Precision (0,487) un Recall (0,471), taču AUC (0,722) vērtība tam ir pavisam nedaudz lielāka nekā nejaušā meža algoritmam. (OrangeDataMining, 2015.)

1. orange darbplūsma

Attēls, kurā ir teksts, diagramma, karte, rinda

Apraksts ģenerēts automātiski

4.1.att. Orange modeļa gatavais izskats

Izmantotie informācijas avoti

1. Aditya, P. (2018.). Retrieved from https://medium.com/@aditya97p/l1-and-l2-regularization-237438a9caa6

2. Cortez, P., Cerdeira, A., Almeida, F., & Matos, T. (2009.). Modeling wine preferences by data mining from physicochemical properties. Retrieved from https://archive.ics.uci.edu/ml/datasets/wine+quality

3. Gupta, R. (2020.). Retrieved from https://www.c-sharpcorner.com/article/logistic-regression/#:%7E:text=or%20explanatory)%20variables.-,Advantages,-/Features%20of%20Logistic

4. Miranda, E. (2017.). Wine Rating Systems. Retrieved from https://www.linkedin.com/pulse/wine-rating-systems-eduardo-miranda/?trk=mp-reader-card

5. OrangeDataMining. (2015.). Orange Visual Programming. Retrieved from https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/index.html

6. THEFULLSTORYEDUCATIONTEAM. (2021.). Retrieved from https://www.fullstory.com/blog/qualitative-vs-quantitative-data/