

**Rīgas Tehniskā universitāte
Datorzinātnes un informācijas tehnoloģijas fakultāte**

Mākslīga intelekta pamati

Atskaite 2. Praktiskajam darbam

Autors: Māris Markuss Petrovs

1.grupa

211RDB311

Projekta saite: <https://github.com/RobbyM2/ML2PRDARB>

2022./2023. m.g.

SATURS

Mākslīga intelekta pamati.....	1
IEVADS	3
DATU PIRMAPSTRĀDE/IZPĒTE	4
Datu kopas apraksts	4
Datu kopas satura apraksts.....	5
Datu kopas objektu analīze.....	10
Klases atdalīšana izmantojot izkliedes diagrammas	11
Klašu atdalīšana izmantojot histogrammas	13
Pazīmju sadalījums	14
Skaitlisko rādītāju aprēķini	16
Orange rīka darbplūsmas atspoguļojums.....	17
Vizuālā atspoguļojuma analīze datu kopas objektiem	17
Datu kopas skaitlisko rādītāju analīze	18
NEPĀRRAUDZĪTĀ MAŠĪNMĀCĪŠANĀS	19
Hierarhiskā klasterizācija	19
K-vidējo algoritms	22
Nepārraudzītās mašīnmācīšanās secinājumi.....	26
PĀRRAUDZĪTĀ MAŠĪNMĀCĪŠANĀS	26
kNN algoritms	26
Nejauša meža algoritms	27
Apmācības un testēšanas rezultāti	29
Izveidoto modeļu veikspējas interpretācija un salīdzinājums.....	30
SECINĀJUMI	35
Avotu saraksts	36

IEVADS

Šī darba izpildei bija nepieciešams izvēlēties datu kopu[1] un izmantot tās apstrādei pārraudzītās un nepārraudzītās mašīnmācīšanās algoritmus.

[4]Darba mērķis ir attīstīt studentu prasmes izmantot mašīnmācīšanās algoritmus un analizēt iegūtos rezultātus. Šī darba galarezultāts ir šī atskaite par darba izpildi.

Pēc praktiskā darba nostādnes tika pieņemts lēmums izmantot tajā ieteikto Orange rīku[2].

Darba atskaite sastāv no 36 lpp; 36 attēliem un 5 tabulām.

DATU PIRMAPSTRĀDE/IZPĒTE

Datu kopas apraksts

No krātuves “Kaggle Datasets”[1] tika izvēlēta datu kopa “Amazon Prime TV Shows”. Šīs datu kopas autors ir Neelima Jauhari, kurš to izveidoja 2020. gadā, un tajā tiek attēloti visi pieejamie Amazon Prime televīzijas seriāli ar mērķi atrast visjaunākos pieejamos seriālus kā arī seriālus ar augstu reitingu.

Tabulas dati tika savākti no Amazon Prime platformas.

Datu kopa satur 404 ierakstus, un sekojošo informāciju: seriāla nosaukums(“Name of the show” vai “Title”), “*Year of release*”, kas ir gads, kurā seriāls tika izlaists vai tika pārraidīts ēterā, “*No.of seasons*” - sezonu skaits, tas nozīmē Prime platformā pieejamo seriāla sezonu skaitu, “*Language*” attiecas uz seriāla audio valodu un neņem vērā subtitru valodu; seriāla žanrs, piemēram: bērniem, drāma, asa sižeta seriāls u. c., IMDB seriāla vērtējums: lai gan daudziem TV seriāliem un bērnu seriāliem vērtējums nebija pieejams, “*Age of viewers*” ir norādīts mērķauditorijas vecums – visiem vecumiem, tātad saturs nav ierobežots kādai noteiktai vecuma grupai un to var skatīties visi skatītāji.

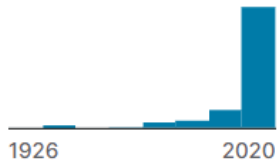
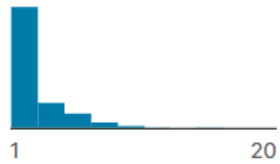
Datu kopas autors atzīmē, ka tā tika izveidota, tā kā daudziem TV seriāliem ir augsti IMDB reitingi, bet tie nav tik daudz skatīti, jo skatītāji par tiem nezina vai tie netiek daudz reklamēti, tādēļ tā veidota, lai noskaidrotu visaugstāk novērtētos seriālus katrā kategorijā vai konkrētā žanrā. Teksta daļa tika izlaista jo nav nepieciešama. Par Target tika izvēlēta “Age of viewers”.

Datu kopas satura apraksts

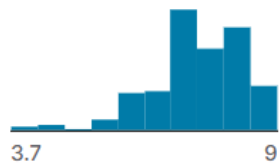
Pazīmes(atribūta) nosaukums	Vērtības	Tips
Izlaiduma gads	Gads kurā izlaists/parādīts seriāls (no 1926 līdz 2020)	Skaitlisks
Sezonu skaits	Izlaisto sezonu skaits (no 1 līdz 20)	Skaitlisks
Valoda	Seriāla audio valoda, subtitri netiek skaitīti (Angļu, Indiešu, Vācu, Franču, Ivrits, Itāļu, Japāņu, Mararšu, Nīderlandžu, Krievu, Serbu, Spāņu, Somu, Telugu)	Kategorisks
Žanrs	Filmas žanrs/žanri (Trilleris, Komēdija, Piedzīvojumu, Animācijas, Drāma, Izklaides, un šo žanru paveidi/kombinācijas)	Kategorisks
IMDb reitings	Skatītāju seriālu vērtējums (no 3,7 līdz 9,0)	Skaitlisks
Skatītāju vecums	Vecums no kura drīkst skatīties (7+,13+,16+,18+,All)	Kategorisks

1.1.tabula - pazīmes, tipi un vērtības

Piezīme: Kategorijā “Name of the show” ir 3% null vērtību, jo no 394. Rindas līdz 404. rindai tabula ir tukša.

▲ Name of the show	# Year of release	# No of seasons av...	▲ Language
Title of the show	Year in which the show went on-air	No. of seasons of the show available on Amazon Prime	Audio language of the show
[null] 3%			English 78%
The Last Ship 0%			Hindi 10%
Other (391) 97%			Other (51) 13%

1.1.att. datu kopas atribūtu apraksti 1.daļa.

▲ Language	▲ Genre	# IMDb rating	▲ Age of viewers
Audio language of the show	Category of the show	Rating	Age of the target audience
English 78%	Drama 31%		16+ 37%
Hindi 10%	Comedy 24%		18+ 20%
Other (51) 13%	Other (182) 45%		Other (175) 43%

1.2.att. datu kopas atribūtu apraksti 2.daļa.

Info 404 instances 7 features (9.8% missing values) Data has no target variable. 1 meta attribute			
Columns (Double click to edit)			
Name	Type	Role	Values
1 S.no.	N numeric	skip	
2 Year of release	N numeric	feature	
3 No of seasons available	N numeric	feature	
4 Language	C categorical	feature	Deutsch, English, French, Hebrew, Hindi, Italiano, Japanese, Marathi, Nederlands, Russian, Serbian, Spanish, Suomi, Telugu
5 Genre	C categorical	feature	Action, Action, Comedy, Adventure, Animation, Animation, Drama, Arts, Entertainment, Culture, Comedy, Comedy, Action, ...
6 IMDb rating	N numeric	feature	
7 Age of viewers	C categorical	target	7+, 13+, 16+, 18+, All
8 Name of the show	S text	skip	
Reset		Apply	

1.3.att. Orange rīkā attēlotās kopas lomas un pazīmes

Pēc dotās kopas analīzes var secināt, ka mums ir pieejamas 3 klases pēc kurām ir iespēja klasificēt seriālus (objektus), tās ir – Valoda (14 pieejamas), Žanrs(ieskaitot kombinācijas) un Skatītāju vecums (5 varianti, no visiem vecumiem līdz 18+). No tabulas analīzes tika izslēgts tikai S.no.(seriāla numurs sarakstā), jo tas nav nepieciešams strādājot Orange rīkā – numerācija jau programmā piemīt.

Genre(Žanrs)	Objektu skaits
Action	26
Action, Comedy	1
Adventure	4
Animation, Drama	1
Arts, Entertainment, Culture	12
Comedy	97
Comedy, Action	1
Comedy, Arts, Entertainment, Culture	1
Comedy, Drama	1
Documentary	3
Drama	136
Drama, Action	11
Drama, Action, Adventure	1
Drama, Action, Sci-fi	1
Drama, Action, Suspense	3
Drama, Comeddy	1
Drama, Comedy	15
Drama, Comedy, Action	1
Drama, Comedy, LGBTQ	1
Drama, Documentary	2
Drama, Fantasy	2
Drama, Horror	2
Drama, Horror, Fantasy	1
Drama, Horror, Romance, Suspense	1
Drama, Horror, Suspense	2
Drama, LGBTQ, Arts, Entertainment, Culture	1
Drama, Romance	4
Drama, Romance, Comedy	1
Drama, Sci-fi	7

Drama, Sci-fi, Suspense, Action	1
Drama, Sports	1
Drama, Suspense	12
Drama, Suspense, Action	2
Drama, Suspense, Adventure	1
Drama, Suspense, Fantasy	1
Drama, Suspense, Horror	1
Fantasy	1
Fantasy, Comedy	1
Horror	1
Kids	29
Kids, Animation	2
Sci-fi	5
Sci-fi comedy	1
Sci-fi, Action, Suspense	1
Sci-fi, Drama, Suspense	1
Sports	1
Sports, Drama	2

1.2.tabula – Atribūta “Genre” kategoriju objektu skaits

Language(Valoda)	Atribūtu skaits
Deutsch	2
English	314
French	2
Hebrew	1
Hindi	39
Italiano	7
Japanese	13
Marathi	2
Netherlands	2
Russian	1
Serbian	2
Spanish	6
Suomi	1
Telugu	1

1.3.tabula – Atribūta “Language” kategoriju objektu skaits

Age of viewers(skatītāju vecums)	Skaits
7+	28
13+	69
16+	150
18+	79
All	67

1.4.tabula – Atribūtu “Age of viewers” kategoriju objektu skaits

Datu kopas objektu analīze

Uzdevumā ir prasīts sekojošais:

ir jāizveido vismaz divas 2- vai 3-dimensiju izkliedes diagrammas, kas ilustrē klases atdalāmību, balstoties uz dažādām pazīmēm;

Ir jāizveido vismaz 2 histogrammas, kas parāda klašu atdalīšanu, pamatojoties uz interesējošām pazīmēm;

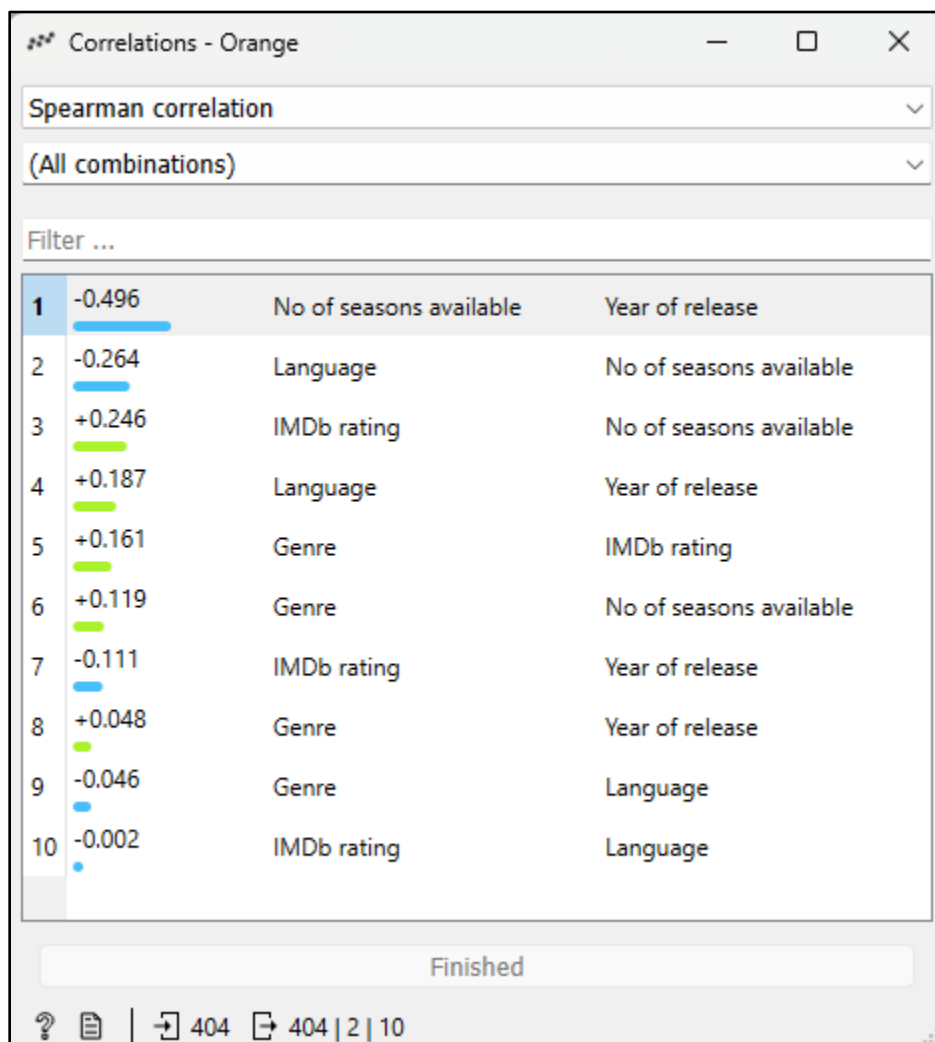
Ir jāatspoguļo 2 interesējošo pazīmju sadalījums;

Ir jāaprēķina statistiskie rādītāji.

Tālāk tiek parādītas šīs prasības.

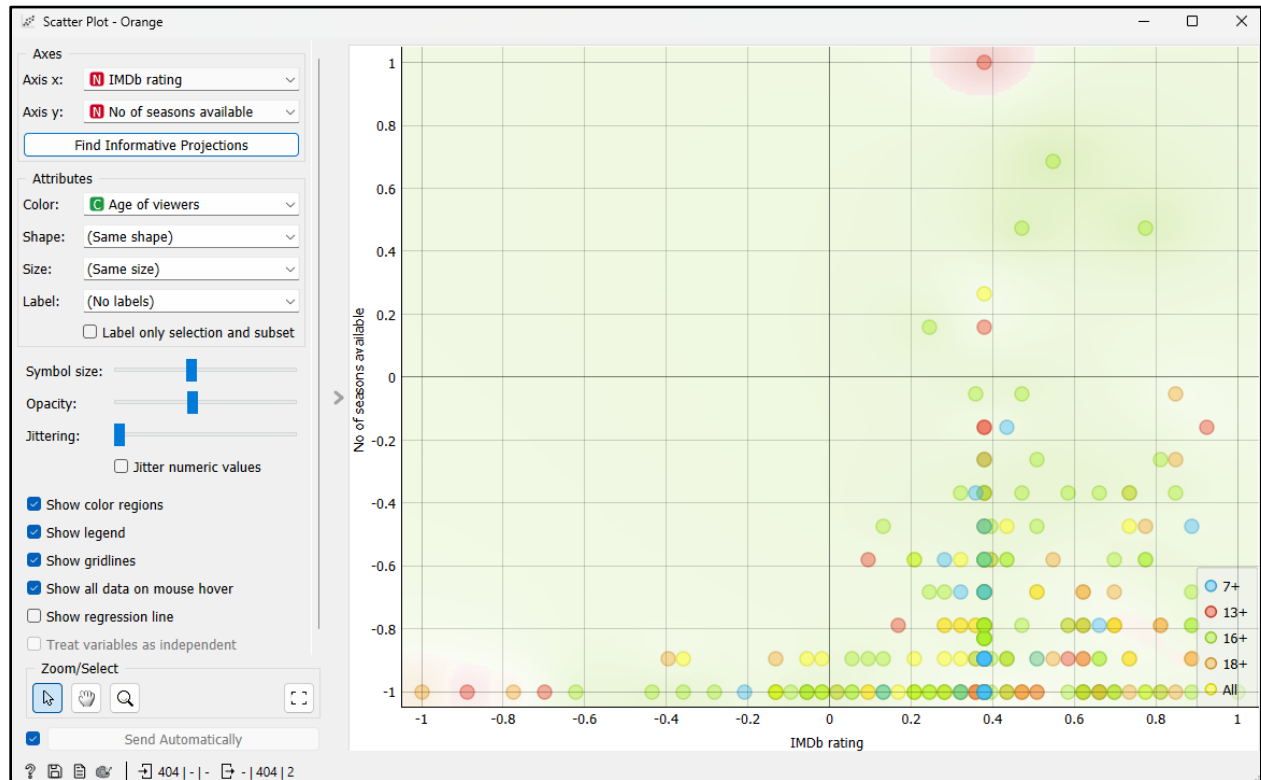
Klases atdalīšana izmantojot izkliedes diagrammas

Lai saprastu kādi atribūti (pazīmes) dos visprecīzāko klases atdalāmību attēlojumā, mēs izmantosim “Corelations” rīku, kas dos informāciju par pazīmju savstarpēju korelāciju, un kuriem tā ir visaugstākā.

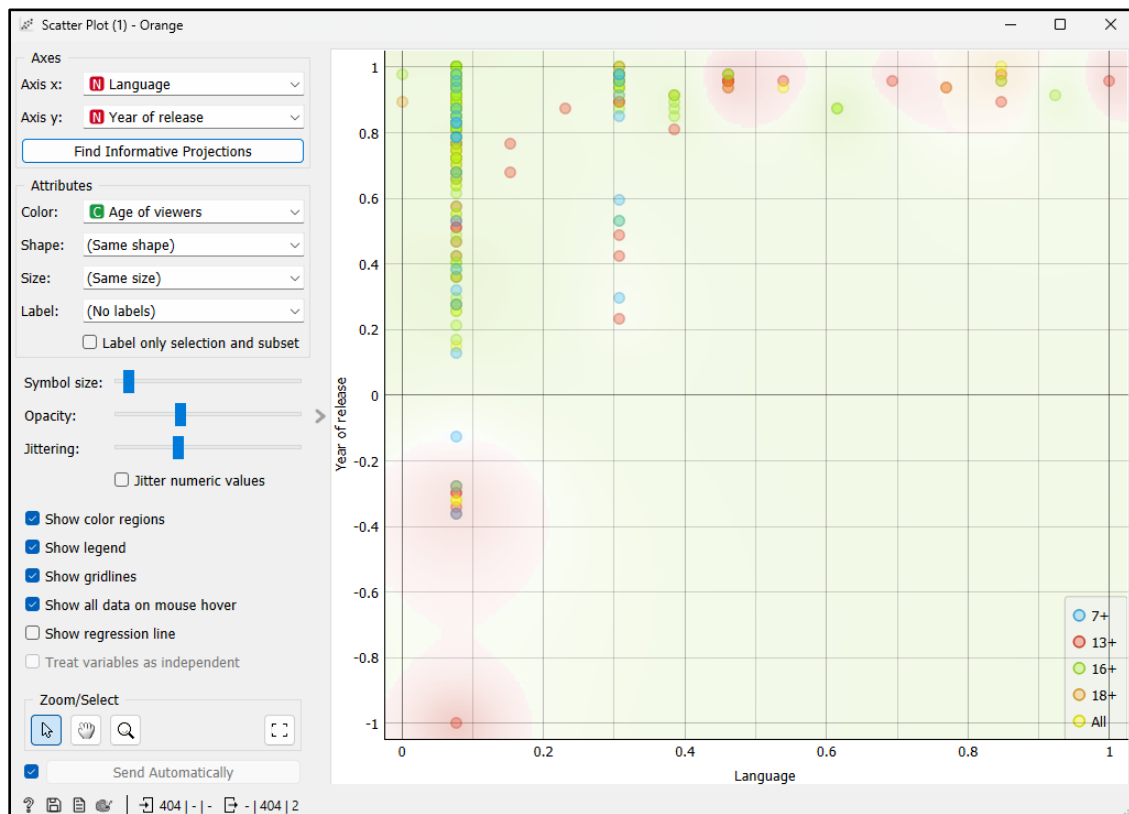


1.4.att. – Datu kopu pazīmju korelācija

Vislielākā korelācija ir “IMBd rating” un “No of seasons aviable”, un otrajā vietā ir “Language” un “Year of release”, tādēļ izkliedes diagrammas būs apskatītas izmantojot šos datus.

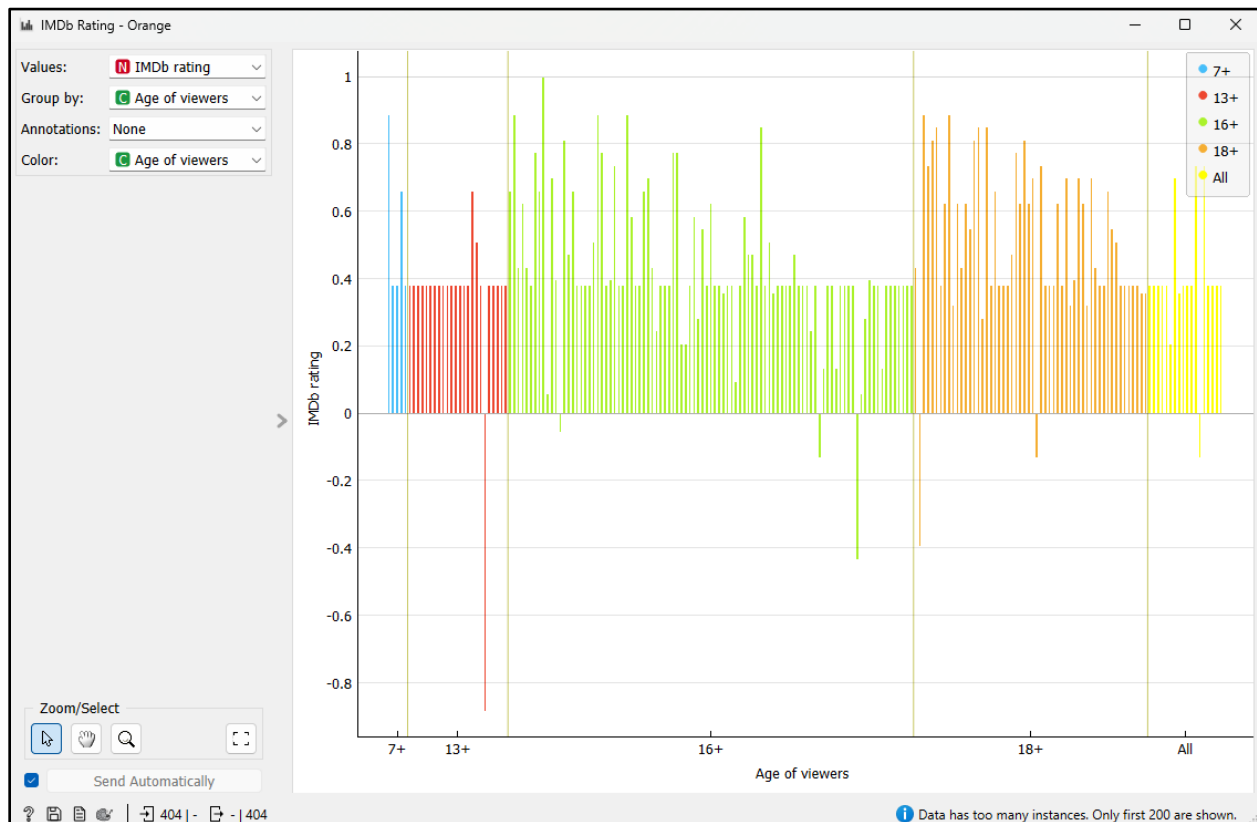


1.5.att. – “IMBd rating” un “No of seasons aviable” izkliedes diagramma

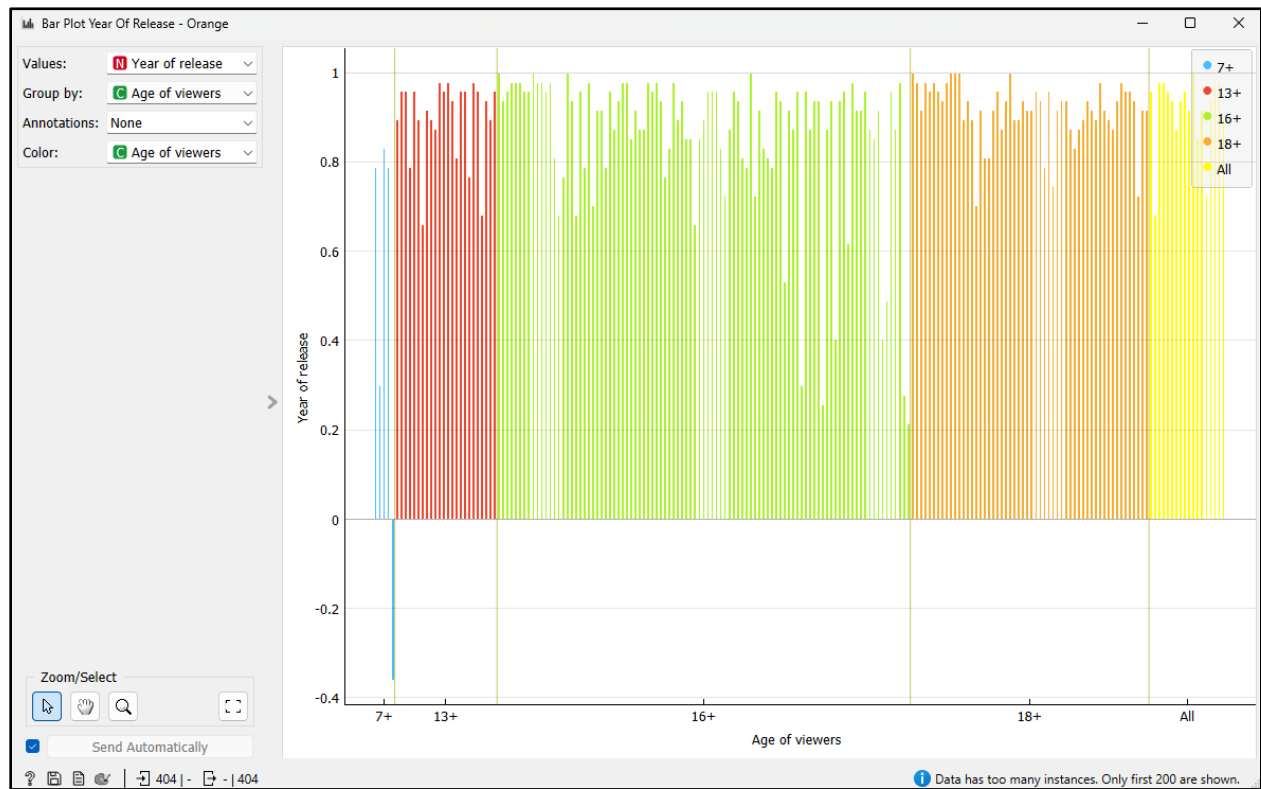


1.6.att. - “Language” un “Year of release” izkliedes diagramma

Klašu atdalīšana izmantojot histogrammas

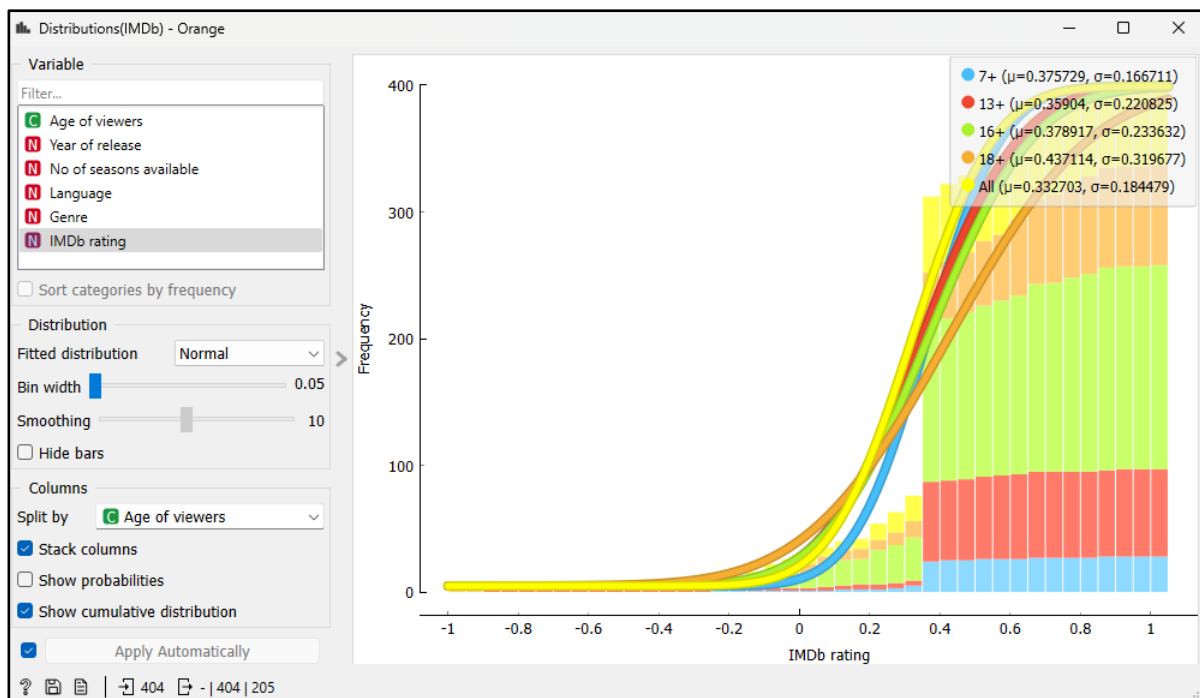


1.7.att. - Histogramma izmantojot “IMDb rating” vērtības

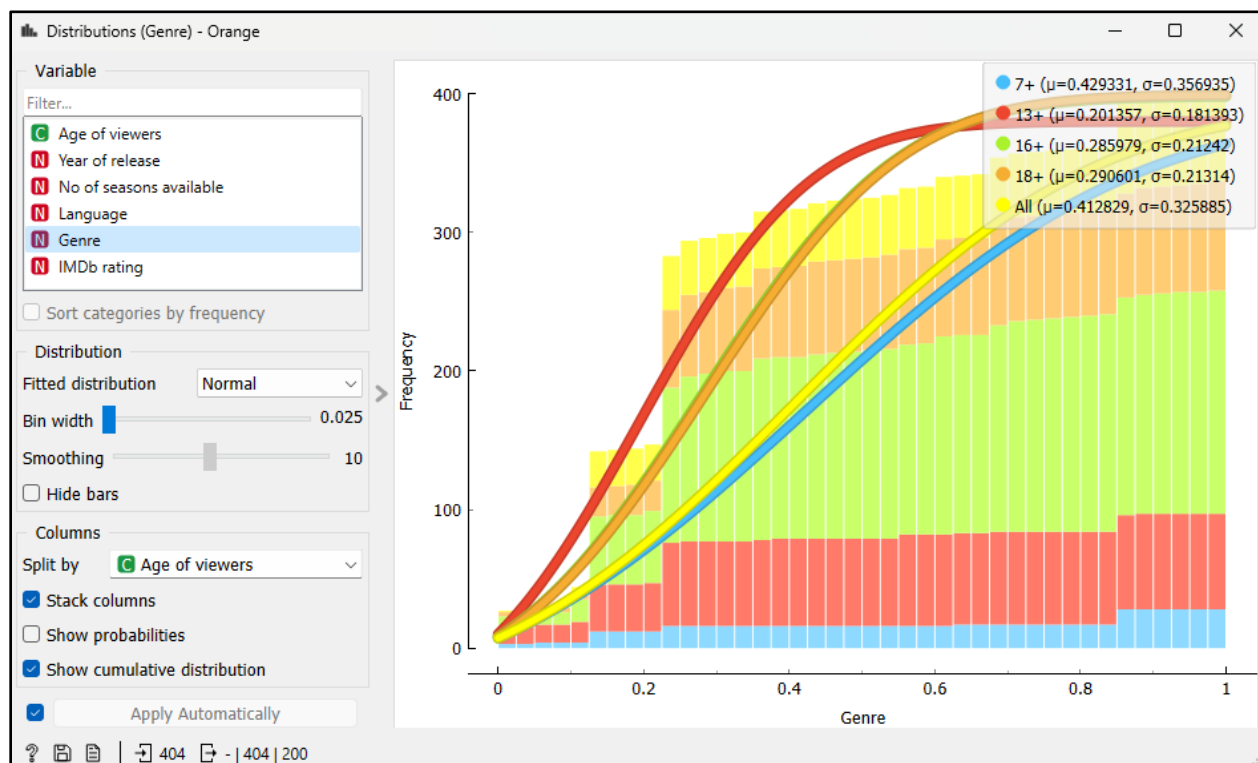


1.8.att. - Histogramma izmantojot “Year of release” vērtības

Pazīmju sadalījums



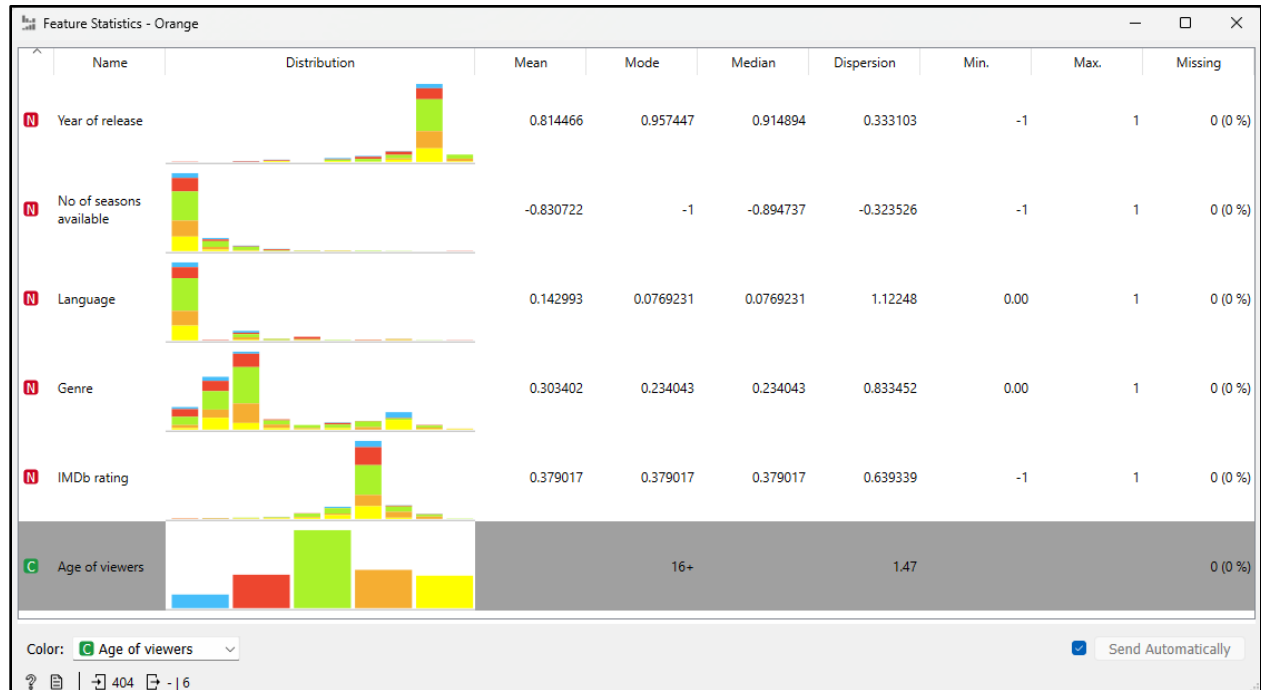
1.9.att. – “IMDb rating” pazīmes sadalījums



1.10.att. - “Genre” pazīmes sadalījums

Skaitlisko rādītāju aprēķini

Orange rīkā “Feature statistics” ir iespējams apskatīt automātiski aprēķinātās vidējās vērtības un dispersiju.



1.11.att. – “Feature statistics” vērtības

Atribūts(Pazīme)	Vid.vērt.	Moda	Mediāna	Dispersija	Min.vērt	Maks.vērt.
Year of release	0.0814466	0.957447	0.914894	0.333103	-1	1
No of seasons available	-0.830722	-1	-0.894737	-0.323526	-1	1
Language	0.142993	0.0769231	0.0769231 1	1.12248	0.00	1
Genre	0.303402	0.234043	0.234043	0.833452	0.00	1
IMDb rating	0.379017	0.379017	0.379017	0.639339	-1	1
Age of viewers	-	16+	-		-	-

1.5.tabula - Dati no “Feature statistics”

The flowchart illustrates a data science workflow, starting from data input and processing, moving through feature engineering and model training, and finally leading to model evaluation and predictions.

Data Input and Processing:

- Data** (central node) branches into:
 - Data Tabula (nemănită)** (Data Table)
 - Data Kopă** (Data Copy)
 - Impute** (Imputation)
 - Continueze** (Continue)
 - Data Info** (Data Information)
 - Data Sampler** (Data Sampling)
 - Data Sample** (Data Sample)

Feature Engineering and Analysis:

- Data** branches into:
 - Feature Statistics** (Feature Statistics)
 - Distributions (Genre)** (Distributions by Genre)
 - Distributions (IMDb)** (Distributions by IMDb Rating)
 - Bar Plot Year Of Release** (Bar Plot of Year of Release)
 - Scatter Plot (1)** (Scatter Plot 1)
 - Scatter Plot** (Scatter Plot)
 - Correlations** (Correlations)
 - k-Means** (k-Means Clustering)
 - Silhouette Plot** (Silhouette Plot)
 - Distances** (Distances)
 - Hierarchical Clustering** (Hierarchical Clustering)
 - Confusion Matrix** (Confusion Matrix)
 - Predictions** (Predictions)
 - Neural Network** (Neural Network)
 - Random Forest** (Random Forest)
 - kNN** (k-Nearest Neighbors)

Model Training and Evaluation:

- Data Sample** leads to **Test and Score Learner** (Test and Score Learner).
- Test and Score Learner** leads to **Learner** (Learner).
- Learner** leads to **Predictions** (Predictions).

Vizuālā atspoguļojuma analīze datu kopas objektiem

Nē, pastāv liela atšķirība starp klasēm, piemēram ir dažas klases, kurās piemīt vairāk par 100 objektiem, bet ir ļoti liela daļa klašu kur objektu skaits ir vienāds ar 1. Tas ir noticis filmu žanra aprakstīšanas īpašību dēļ, kā jau tika minēts, šajās klasēs ir daudzas, kurām ir kopīgi žanri, bet tādēļ ka citās to ir vairāk, tās tiek ierakstītas atsevišķi. Šādu problēmu, iespējams, varētu risināt saīsinot un vienkāršojot dalījumu uz klasēm, bet šobrīd tāds uzdevums nepastāv.

Vai datu vizuālais atspoguļojums ļauj redzēt datu struktūru?

Jā, datu ir pietiekami maz lai skaidri atšķirtu pat bez pietuvināšanas izkliedes diagrammās kopējo datu struktūru un tas pats attiecas uz pazīmju sadalījuma tabulām. Tomēr histogrammās šis process ir daudz grūtāks, dati atrodas diezgan tuvu vizuāli viens pie otra.

Cik datu grupējums ir iespējams identificēt, pētot datu vizuālo atspoguļojumu?

Izkliedes diagrammās var saskatīt dažus grupējumus ja izmanto grupējumu iekrāsošanu, tomēr tie ir pārāk mazi lai dotu kādu noderīgu informāciju.

Vai identificētie datu grupējumi atrodas tuvu viens otram vai tālu viens no otra?

Faktiski tie atrodas tik tuvu viens pie otra, ka to nošķiršana rada lielas grūtības.

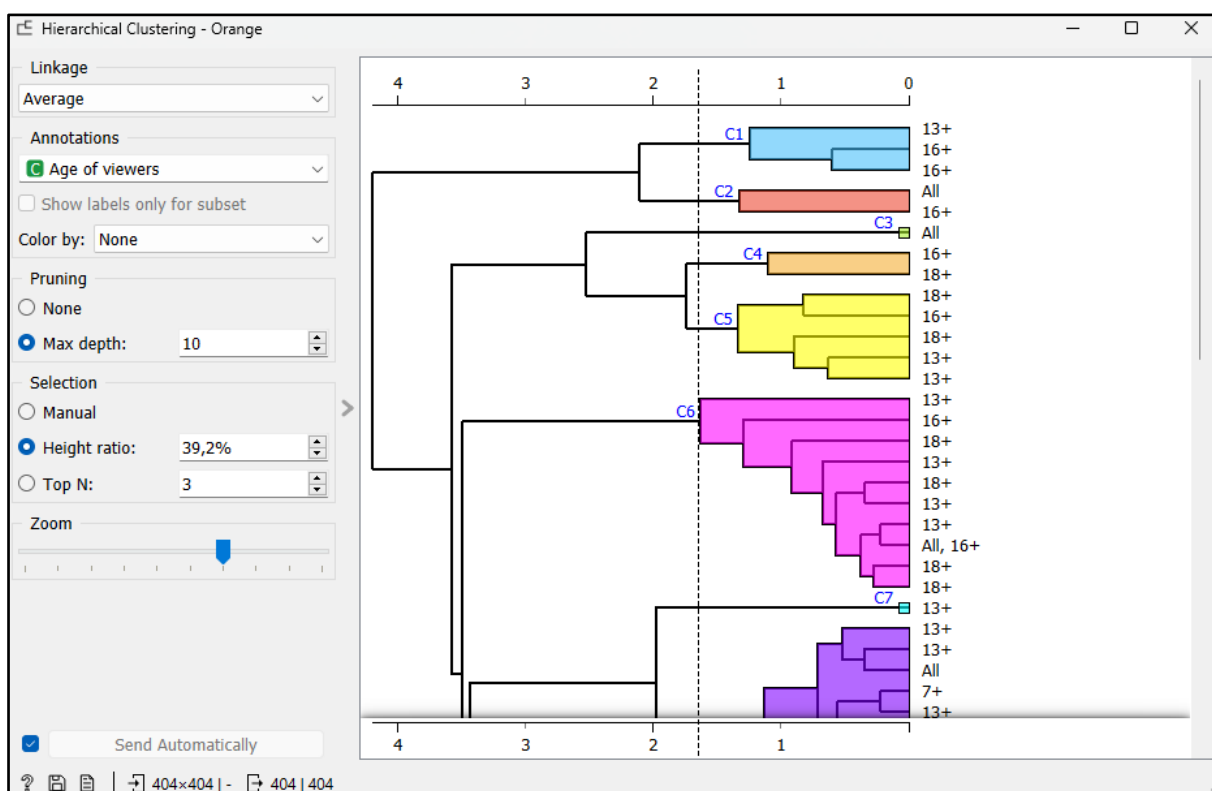
Datu kopas skaitlisko rādītāju analīze

- Nekur nebija pazaudēti datu objekti;
 - “Year of release” pazīmei ir vislielākās vidējās, modas un mediānas vērtības, bet Min. tai ir vienāds ar -1 ;
 - “No of seasons aviable” ir vienīgā, kam visas vērtības ir negatīvas;
 - “Language” pazīmei ir vismazākās vidējās, modas un mediānas vērtības, bet otrā lielākā dispersija;
 - “Age of viewers” pazīmei nav vidējās un mediānas vērtības, tomēr tā rāda Modas vērtību kā arī dispersiju.
- Tātad, Orange rīkā simboliskas vērtības no datu kopas tiek apstrādātas specifiskā veidā.

NEPĀRRAUDZĪTĀ MAŠĪNMĀCĪŠANĀS

Nepārraudzītā mašīnmācīšanās[5] paredz divu sekojošo algoritmu izmantošanu: 1) Hierarhiskā klasterizācija, kā arī 2) K-vidējo algoritms.

Hierarhiskā klasterizācija



2.13.att. Nostādījuma logs Hierarhiskās klasterizācijas rīkam

Linkage attiecas uz saistību starp klasteriem. Ir pieejami vairāki saistību veidi, piemēram, *Single*, *Average*, *Weighted*, *Complete* un *Ward*. Katrs no tiem aprēķina attālumu starp diviem klasteriem, taču izmanto atšķirīgus kritērijus, piemēram, vidējo attālumu, vislielāko attālumu vai summas kļūdas pieaugumu, lai šo attālumu noteiktu. Klasteru elementu anotēšana ir iespējama, bet tā neietekmē klasterizācijas procesu. *Pruning* var būt noderīgs, lai ierobežotu dendrogrammas

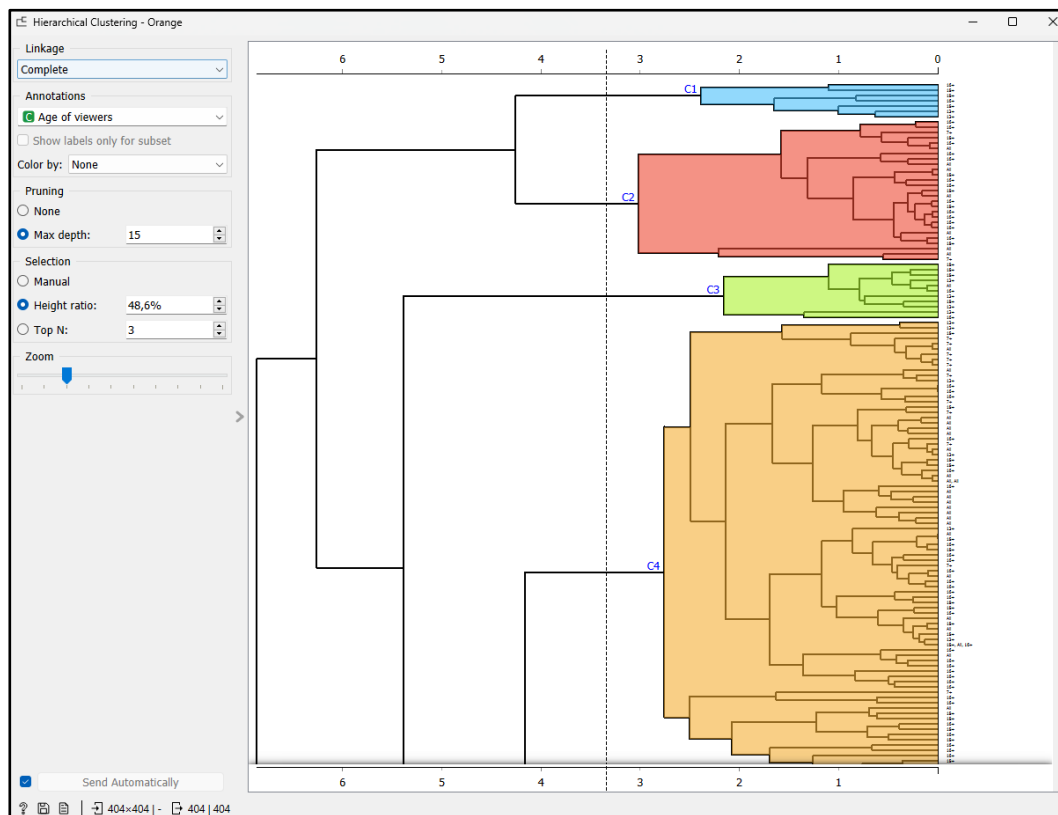
dziļumu un, tādējādi, klasteru skaitu sākumā. Klasteru izvēle var būt manuāla un ļaut lietotājam izvēlēties klasterus ar kursoru. *Height ratio*, kas ir augstuma attiecība, ir atdalošās līnijas atrašanās pret paša klastera augstumu. *Top N* ir augšējo klasteru N skaits, kas ļauj izvēlēto klasteru skaitu displejā definēt.

Tiks izdarīti 3 eksperimenti Hierarstiskās klasterizācijas ietvaros, tas tiks darīts izmantojot saites veida un atdalošās līnijas izmaiņu.



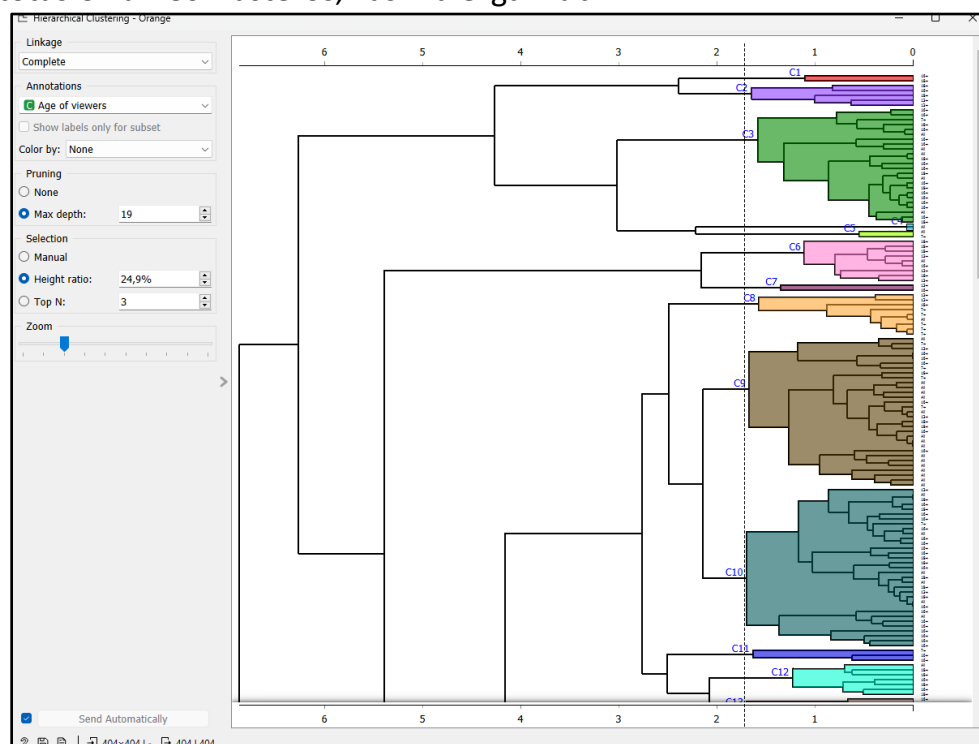
2.14.att. Hierarhiskā klasterizācija ar vidējo saiti, maksimālais dziļums ir 11, un Height ratio ir vienāds ar 79,3%

Izmantojot izvēlētos hiperparametrus, tika izvēlēti 5 klasteri. Tādēļ, ka dažādu tipu zvaigznes ir ievietotas, var secināt ka tas nav pārāk labi.



2.15.att. Hierarhiskā klasterizācija ar pilnīgo saiti, maksimālais dziļums ir 15, un Height ratio ir vienāds ar 48,6%

Izmantojot izvēlētos hiperparametrus, tika izvēlēti 8 klasteri. HyperGiants tipa zvaigznes tiks ievietotas C4 un C6 klasteros, kas ir diezgan labi.

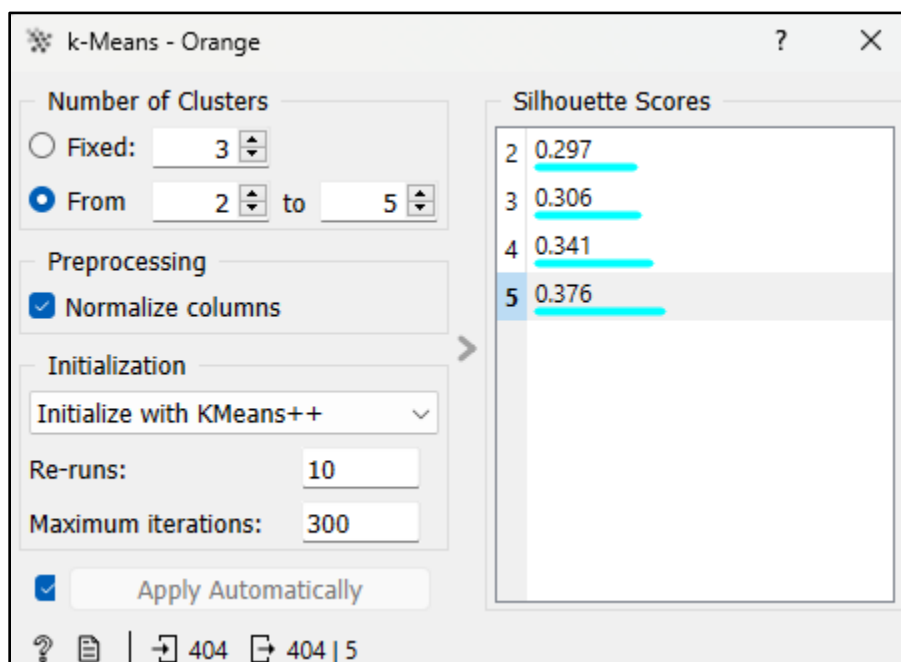


2.16.att. Hierarhiskā klasterizācija ar pilnīgo saiti, maksimālais dziļums ir 19, un Height ratio ir vienāds ar 24,9%

Šis sadalījums ir ļoti slikts, jo rodas 29 klasteri, zvaigžņu tipi tika sadalīti vienās kopās vēl pirms klasterizācijas.

Pēc doto dentogrammu analīzes var secināt, ka Hierarhiskā klasterizācija var atļaut atdalīt klase un/vai kategorijas labāk nekā to varētu k-vidējais algoritms, bet rezultātu var uzlabot.

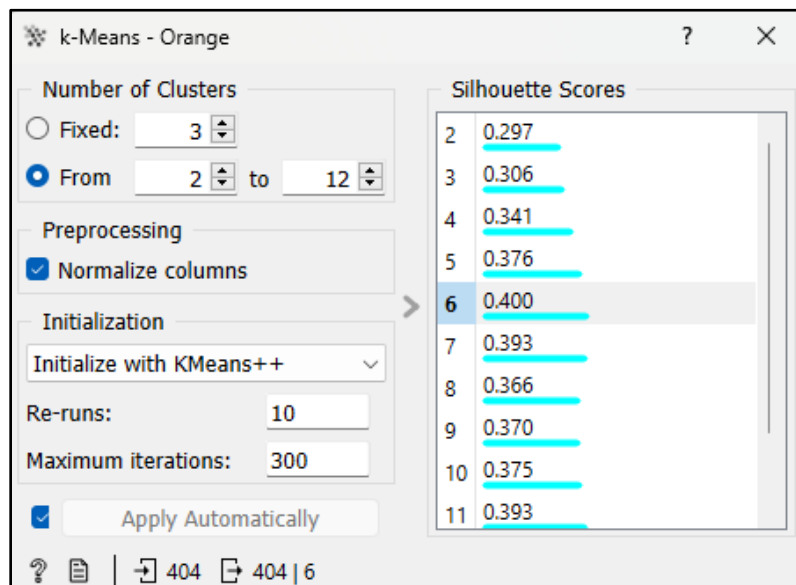
K-vidējo algoritms



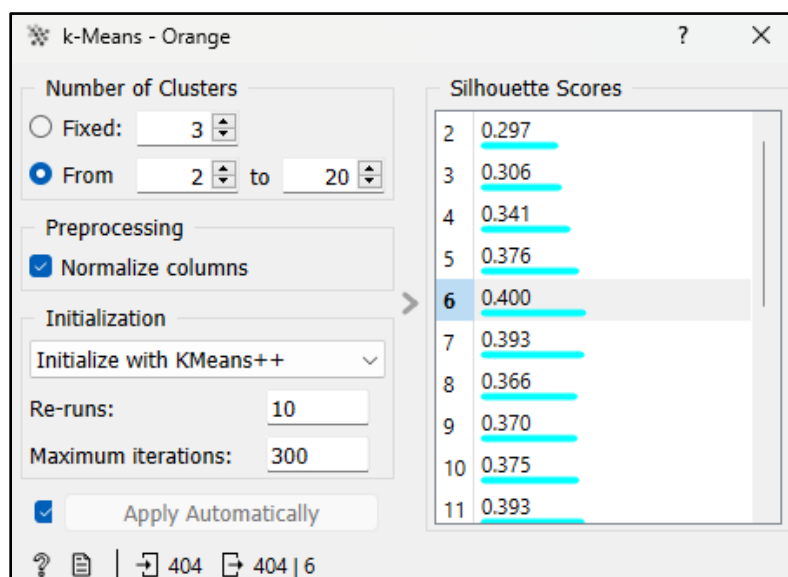
2.17.att. K-means parametru logs

Dotais attēls ir ievietots lai labāk paskaidrotu k-vidējo algoritma hiperparametrus. Šis logs ļauj lietotājam izvēlēties klasteru skaitu, kas jāatrod algoritma izpildes beigās, kā arī noteikt citus hiperparametrus, piemēram, datu iepriekšējo apstrādi un klasterizācijas inicializāciju. Ir divi veidi, kā noteikt klasteru skaitu: fiksēts skaits vai klasteru skaits, kas svārstās no N līdz K. Algoritms veic klasterizāciju katrā

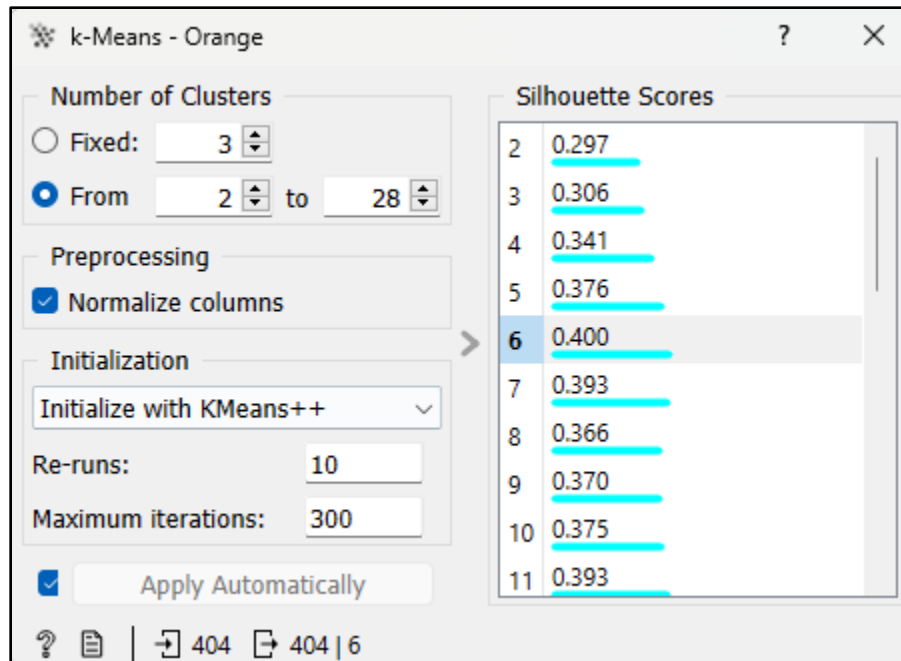
elementu skaita vērtībā no diapazona un aprēķina Silhouette Score, kas salīdzina vidējo attālumu starp elementiem vienā klasterī ar vidējo attālumu starp elementiem citos klasteros. Iepriekšēja datu apstrāde var tikt veikta, normalizējot kolonnas, lai centrētu to vidējo vērtību uz 0 un mērogotu to standarta novirzi līdz 1. Klasterizācijas inicializāciju var veikt nejauši vai, ja vēlamais klasteru skaits nav sasniegts, no vairākiem punktiem. Algoritms var tikt izpildīts vairākas reizes, sākot no nejauši izvēlētas pozīcijas, vai arī tiek ierobežots ar maksimālo iterāciju skaitu. Mainot fiksēto klasteru skaitu, tiks redzams, kā mainās Silhouette Score. Ja noteiktā maksimālā klasteru skaits ir 30, tad var noteikt, ka vislabākais klasteru skaits ir 19.



2.18.att. Vislabāko k-vidējā algoritma klasteru skaits (no 2 līdz 12)

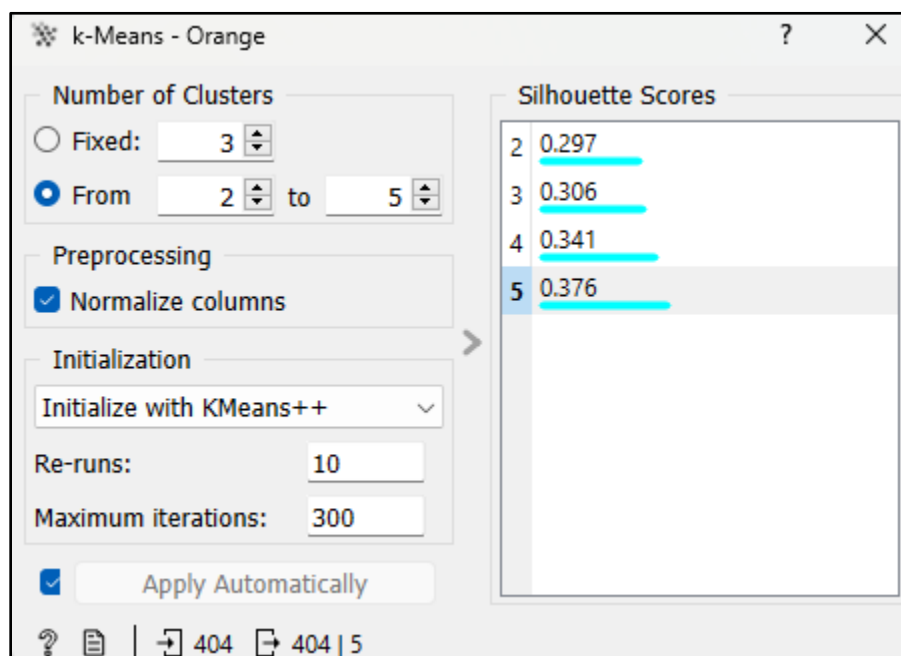


2.19.att. Vislabāko k-vidējā algoritma klasteru skaits (no 2 līdz 20)



2.20.att. Vislabāko k-vidējā algoritma klasteru skaits (no 2 līdz 28)

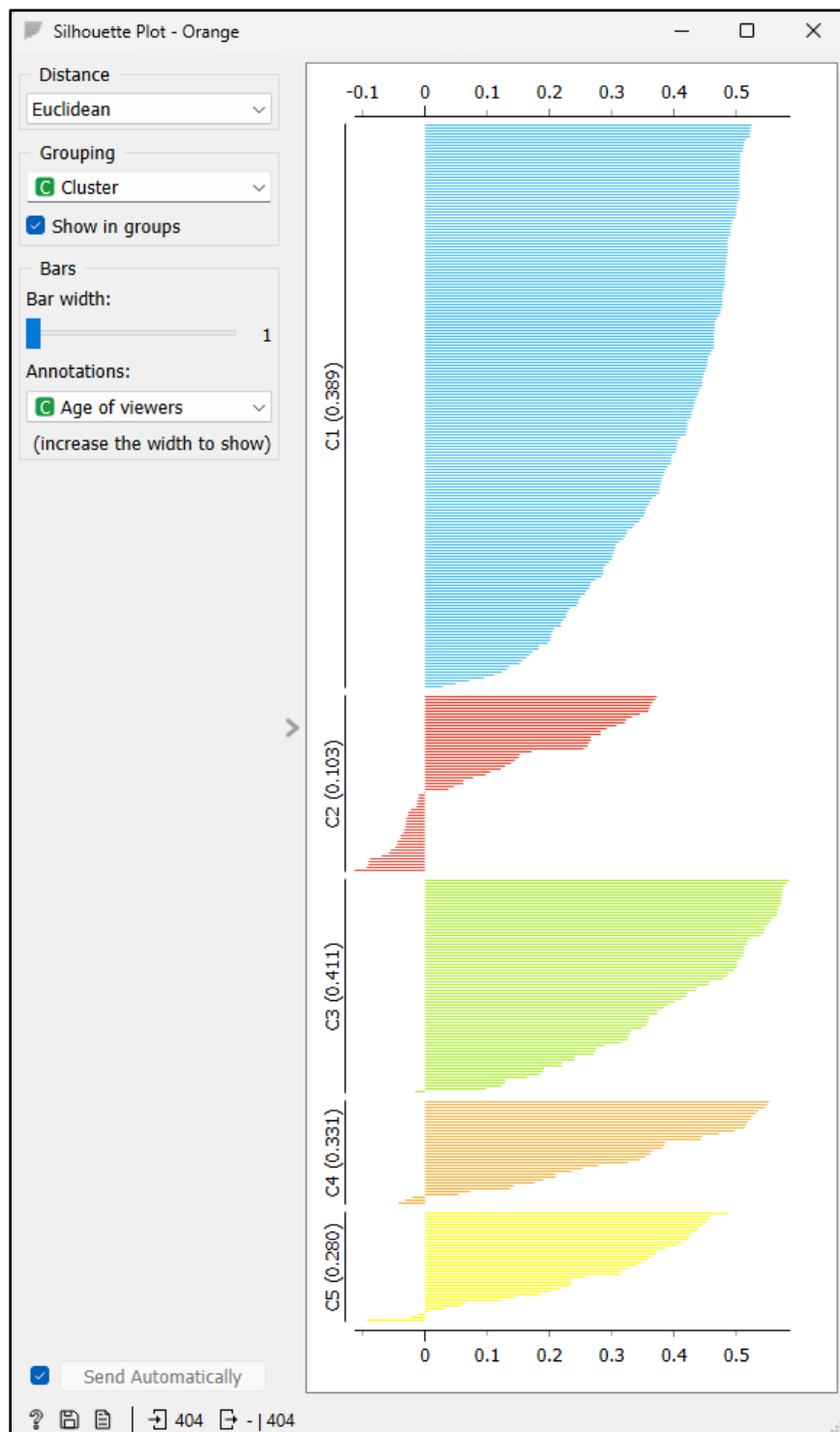
Var secināt, ka neskatoties uz klasteru skaita izmaiņām, visoptimālākais rezultāts ir pie 6, tātad, nav jēgas likt klasteru skaitu lielāku par šo.



2.21.att. Vislabāko k-vidējā algoritma klasteru skaits (no 2 līdz 5)

Izvirzītā hipotēze apstiprinājās – Ja klasteru diapazonu samazināt zemāk par 6, tad vislabākais rezultāts netiks iegūts.

Tālāk apskatīsim datu kopu izskatu ar 5 klasteriem:



2.22.att. Silhouette Plot datu kopai ar klasteriem.

Visiem klasteriem ir iegūtas praktiski pilnībā pozitīvas vērtības, tātad, dati tika sadalīti pareizajā klasterī. Visneprecīzākais izrādījās C2 klasteris, kas atbilst 13+ Vecuma ierobežojumam.

Nepārraudzītās mašīnmācīšanās secinājumi

Izpildot darbības izmantojot nepārraudzītās mašīnmācīšanās, ir redzams, ka datu kopas klases ir diezgan labi sadalītas, tātad k-videjais algoritms spēj datu objektus labi klasificēt. Hierestiskā klasterizācija arī dod pietiekami labus rezultātus. Abi nepārraudzītās mašīnmācīšanās algoritmi spēj tikt galā ar doto kopu.

PĀRRAUDZĪTĀ MAŠĪNMĀCĪŠANĀS

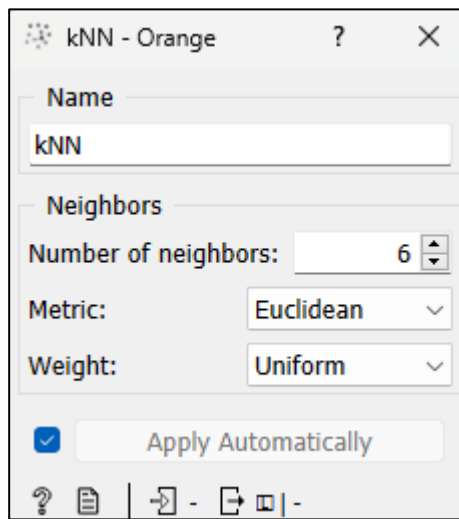
Pārraudzītā mašīnmācīšanās paredz 3 algoritmu izmantošanu[6]:

1)kNN alg., 2)Neironu tīkls un 3) Nejaušs mežs.

Ar Oracle logrīku datu kopa tika sadalīta apmācīšanas un testēšanas kopās, 90% un 20% respektīvi.

kNN algoritms

KNN algoritms ir uzraudzītās mašīnmācīšanās algoritms, kas ir piemērojams gan klasifikācijas, gan regresijas uzdevumos. Šī algoritma izvēle tika pamatota ar to, ka tas ir vienkāršs un tika apskatīts lekcijās. Algoritma darbības princips ir balstīts uz to, ka, lai noteiktu kategoriju vai klasi jaunam datu objektam, algoritms meklē datu kopā līdzīgāko kaimiņu (objektu) un piešķir jaunajam datu objektam tādu pašu kategoriju vai klasi.

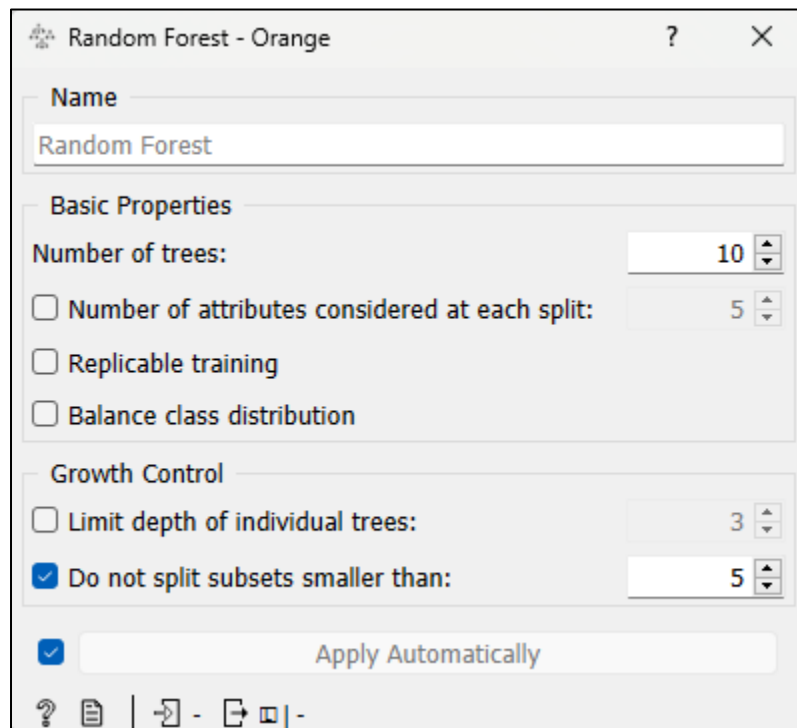


3.23.att. Nostādījuma logs kNN algoritmam

Šajā sistēmā ir iespējams definēt kategoriju vai klasi, nosakot kaimiņu skaitu (Number of neighbors), kas tiks ņemti vērā. Turklāt ir jānorāda metrika (Metric), kas tiek izmantota, lai iegūtu un aprēķinātu attālumu, kuru var definēt kā Eiklīdu (Euclidean), Manhetenas (Manhattan), Maksimālo (Maximal) vai Mahalanobisa (Mahalanobis) attālumu starp diviem punktiem. Papildus tam ir nepieciešams noteikt punktu svaru (Weight), kas var būt vienāds visiem blakus esošajiem punktiem (Uniform) vai vistuvākie punkti var ietekmēt jaunu datu objektu vairāk (Attālums).

Nejauša meža algoritms

Nejaušs mežs(Random Forest) ir algoritms, ko izmanto gan klasifikācijas, gan regresijas uzdevumos, un kas ir balstīts uz ļoti vienkāršu ideju - lēmumu koku modeli, bet ar daudziem kokiem, nevis tikai vienu, turklāt, tie ir dažādi. Tā darba princips ir tāds, ka mežā, kas sastāv no daudziem lēmumu koku modeļiem, katrs koks var piešķirt jaunam datu objektam kategoriju vai klasi. Tālāk, apkopojot visus iegūtos rezultātus no visiem kokiem, jaunam objektam tiek piešķirta tā kategorija vai klase, kurai ir vislielākā summa.



3.24.att. Nostādījuma logs Random Forest algoritmam

Bāzes nosacījumi(Basic Properties):

Meža koku skaits (Number of trees)

Atribūtu skaits, kas tiek izskatīts katrā sadalījumā (Number of attributes considered at each split)

Atkārtojama trenēšana (Replicable training) - fiksēt koka sēklas, lai nodrošinātu atkārtojamību.
Kokauguma kontroles (Growth Control) nosacījumi:

Ierobežot atsevišķu koku dziļumu (Limit depth of individual trees)

Nepadalīt apakškopas, kuru izmērs ir mazāks par norādīto limitu (Do not split subsets smaller than)

Apmācības un testēšanas rezultāti

1.Eksperiments(apmācību kopa)

Parametri: kNN – (6 kaimiņi, Euclidian, Uniform)

Random Forest – (5 koki)

Neironu tīkls (10,; Logistic; Max iter – 500)

Model	AUC	CA	F1	Precision	Recall
kNN	0.830	0.576	0.552	0.584	0.576
Random Forest	0.955	0.806	0.802	0.805	0.806
Neural Network	0.696	0.488	0.400	0.405	0.488

3.25.att. 1.eksperimenta rezultāti

2.Eksperiments(apmācību kopa)

Parametri: kNN – (12 kaimiņi, Manhattan, Uniform)

Random Forest – (10 koki)

Neironu tīkls (20,; tanh; Max iter – 1000)

Model	AUC	CA	F1	Precision	Recall
kNN	0.786	0.541	0.504	0.526	0.541
Random Forest	0.969	0.837	0.835	0.835	0.837
Neural Network	0.773	0.555	0.525	0.574	0.555

3.26.att. 2.eksperimenta rezultāti

3.Eksperiments(apmācību kopa)

Parametri: kNN – (20 kaimiņi, Euclidian, Uniform)

Random Forest – (18 koki)

Neironu tīkls (35,; Logistic; Max iter – 750)

Model	AUC	CA	F1	Precision	Recall
kNN	0.748	0.477	0.407	0.438	0.477
Random Forest	0.981	0.852	0.850	0.857	0.852
Neural Network	0.705	0.488	0.428	0.502	0.488

3.27.att. 3.eksperimenta rezultāti

Izveidoto modeļu veikspējas interpretācija un salīdzinājums

Eksperimentu veikšanas gaitā bija pārbaudīts, kādā veidā izmainās modeļu precizitāte, noderīgums un veikspēja.

kNN algoritms viszemāko precizitāti ieguva 3. Eksperimentā, tomēr mainot Metric var iegūt labāku rezultātu precizitātei. Tāpat, mainot kaimiņu skaitu tika atklāts, ka izmantojot Euclidean Metric, jo mazāks to skaits, jo lielāka precizitāte – vislabākais rezultāts pie 1 kaimiņa, tā sasniedza 96,1%, bet trešajā eksperimentā tikai 31,74%

Random Forest precizitāte palielinājās kopā ar koku skaitu – jo to vairāk, jo tuvāk tā kļuva 100%. 3. eksperimentā tā sasniedza 80,3% atzīmi

Neirona tīkla veikspēja ir atkarīga no paslēpto slāņu skaita, un mainot tikai to, izmantojot Logistic, ir grūti iegūt lielāku precizitāti nekā ar 35. Tomēr ar šo slāņu skaitu, precizitāte bija vislielākā ar ReLu algoritmu. 3. eksperimentā - 33,66%

		Predicted					Σ
		7+	13+	16+	18+	All	
Actual	7+	4	4	6	0	4	18
	13+	0	24	17	3	0	44
	16+	0	5	99	11	5	120
	18+	1	5	23	16	3	48
	All	0	7	24	2	20	53
Σ		5	45	169	32	32	283

3.28.att. 1.eksperimenta rezultāti kNN

		Predicted					Σ
		7+	13+	16+	18+	All	
Actual	7+	13	1	1	0	3	18
	13+	1	31	7	4	1	44
	16+	1	2	108	6	3	120
	18+	0	2	8	36	2	48
	All	1	2	9	2	39	53
Σ		16	38	133	48	48	283

3.29.att. 1.eksperimenta rezultāti Random Forest

		Predicted					Σ
		7+	13+	16+	18+	All	
Actual	7+	1	4	8	0	5	18
	13+	1	11	32	0	0	44
	16+	0	5	109	0	6	120
	18+	0	2	43	0	3	48
	All	0	5	31	0	17	53
Σ		2	27	223	0	31	283

3.30.att. 1.eksperimenta rezultāti Neural Network

		Predicted					Σ
		7+	13+	16+	18+	All	
Actual	7+	2	3	6	0	7	18
	13+	2	19	20	2	1	44
	16+	0	6	102	7	5	120
	18+	1	3	28	12	4	48
	All	0	4	28	3	18	53
Σ		5	35	184	24	35	283

3.31.att. 2.eksperimenta rezultāti kNN

		Predicted					Σ
		7+	13+	16+	18+	All	
Actual	7+	13	0	2	0	3	18
	13+	0	38	3	2	1	44
	16+	0	2	110	5	3	120
	18+	0	2	9	36	1	48
	All	0	3	6	2	42	53
Σ		13	45	130	45	50	283

3.32.att. 2.eksperimenta rezultāti Random Forest

		Predicted					Σ
		7+	13+	16+	18+	All	
Actual	7+	7	5	5	0	1	18
	13+	1	18	24	1	0	44
	16+	1	4	102	6	7	120
	18+	1	2	31	12	2	48
	All	1	4	29	1	18	53
Σ		11	33	191	20	28	283

3.33.att. 2.eksperimenta rezultāti Neural Network

		Predicted					Σ
		7+	13+	16+	18+	All	
Actual	7+	0	3	6	0	9	18
	13+	0	12	32	0	0	44
	16+	0	3	105	6	6	120
	18+	0	2	38	3	5	48
	All	0	1	31	1	20	53
Σ		0	21	212	10	40	283

3.34.att. 3.eksperimenta rezultāti kNN

		Predicted					Σ
		7+	13+	16+	18+	All	
Actual	7+	14	0	3	0	1	18
	13+	2	35	4	1	2	44
	16+	0	0	117	2	1	120
	18+	0	4	6	36	2	48
	All	1	4	9	1	38	53
Σ		17	43	139	40	44	283

3.35.att. 3.eksperimenta rezultāti Random Forest

		Predicted					Σ
		7+	13+	16+	18+	All	
Actual	7+	3	3	9	0	3	18
	13+	2	12	30	0	0	44
	16+	1	7	103	3	6	120
	18+	0	2	39	4	3	48
	All	1	5	31	0	16	53
Σ		7	29	212	7	28	283

3.36.att. 3.eksperimenta rezultāti Neural Network

SECINĀJUMI

Darba procesā, es ieguva praksi darbā ar Orange rīku, kas ļāva man apstrādāt datu kopas, izmantojot mašīnmācīšanās algoritmus, gan nepārraudzītiem, gan pārraudzītiem. Es dziļi iepazinos ar k-vidējo, Random Forest, kNN algoritmiem, un eksperimentēju ar hierarhisko klasterizāciju, izmantojot atbilstošo rīku. Lai novērtētu algoritmu piemērotību datu kopai un efektivitāti Orange rīka darbplūsmā, man bija nepieciešams iegūt informāciju par klasifikācijas algoritmiem. Man patika arī darbs ar Orange logrīkiem, kas ļāva man manuāli apstrādāt datus, mainīt atribūtu nosaukumus un to vērtības. Orange rīks bija ļoti ērts lielu datu apstrādei un ieskatam tajos. Darbs bija izaicinošs, bet man patika, un es ceru, ka turpmāk arī tiks dota iespēja strādāt ar lieliem datiem un datu kopām.

Avotu saraksts

“Amazon Prime Series”; skatīts 04.05.2023.

[1] <https://www.kaggle.com/datasets/dhruvjha/amazon-prime-series>

“Otrā praktiskā darba demonstrācijas piemērs”; skatīts 05.05.2023.

[2] <https://www.youtube.com/watch?v=dKURyzjh5Gc>

“List of Amazon Prime Video original films”; skatīts 05.05.2023.

[3] https://en.wikipedia.org/wiki/List_of_Amazon_Prime_Video_original_films

DATU IZPĒTE PYTHON VALODĀ UN ORANGE RĪKĀ; skatīts 05.05.2023.

[4] <https://youtu.be/bmwH3EcTBEM>

NEPĀRRAUDZĪTĀ MAŠĪNMĀCĪŠANĀS PYTHON VALODĀ UN ORANGE RĪKĀ; skatīts 05.05.2023.

[5] <https://youtu.be/ojxvlQSYLr0>

PĀRRAUDZĪTĀ MAŠĪNMĀCĪŠANĀS PYTHON VALODĀ UN ORANGE RĪKĀ; skatīts 06.05.2023.

[6] <https://youtu.be/UiGH4v3VKPc>