**Datorzinātnes un informācijas tehnoloģijas fakultāte**

Datorvadības, automātikas un datortehnikas institūts

A picture containing clipart

Description automatically generated

# Mākslīgā intelekta pamati

Referāts

**2.praktiskā darba atskaite**

**Jekaterina Leitarte**

**Stud.apl.Nr. 181RDB219**

**Saite uz GitHub projektu: https://github.com/jekmiz/MI\_Orange\_2023**

**Rīga, 2023**

**SATURS**

[1. DATU KOPAS APRAKSTS 3](#_Toc134748580)

[2. NEPĀRRAUDZĪTA MAŠĪNMĀCĪŠANĀS 13](#_Toc134748581)

[2.1. Hierarhiska klasterizācija 13](#_Toc134748582)

[2.2. K-vidējo algoritms 18](#_Toc134748583)

[SECINĀJUMI 22](#_Toc134748584)

# DATU KOPAS APRAKSTS

Datu kopa ir pieejama: <https://www.kaggle.com/datasets/utkarshx27/breast-cancer-dataset-used-royston-and-altman>

Saite uz GitHub projektu: https://github.com/jekmiz/MI\_Orange\_2023

Datu kopas autors: Utkarsh Singh

Šīs datu kopas veidotāja veidoja šo datu kopu ar mērķi analizēt, kā arī trenēties datu vizualizācijā. Šī datu kopa satur pacientu ierakstus no 1984.gada līdz 1989.gada pētījuma, ko veica Vācijas Krūts vēža pētījumu grupa. Tajā piedalījās 720 pacienti ar krūts vēzi. Datu kopā ir saglabāti 686 pacienti ar pilnīgiem datiem un prognostiskajiem mainīgajiem.

Pēc datu sākotnējās sagatavošanas .csv fails tika ielādēts Orange rīkā. Pēc datu ielādēšanas var redzēt, kā izskatās ielādēti dati tam speciāli paredzētajā rīkā. Uzreiz tiks parādīta informācija par to, cik daudz ir datu objektu, kā arī , ka ir 12 pazīmes un nav izvēlēta mērķa pazīme, pēc kuras tiks veikta datu analīze.

**Datu kopas satura apraksts**

Darba autorei ir jānorāda, ka dati netika papildināti no darba autores puses.

A screenshot of a table

Description automatically generated with low confidence

1.1.attēls. Datu kopa

Dati .csv formātā tika ielādēti Orange rīkā.

Kā mērķa pazīmi darba autore ir norādījusi grade (pakāpe).

Datu objektu skaits 686 objekti.

Kopai ir 12 pazīmes, nav tukšo vērtību, ko var redzēt 1.1.attēlā.

**A screenshot of a computer

Description automatically generated**

1.2.attēls. Faila ielādēšana Orange rīkā

Pazīmju (atribūtu) skaits un to atspoguļojums:

* Pid – personas ID;
* Age – vecums;
* Meno – vai ir menopauze (menopauzes status, 0 = pre-menopauze, 1 = pēcmenopauze)
* Size – audzēja lielums, mm;
* Grade – audzēja pakāpe;
* Nodes – cik limfmezgli ir skarti;
* Pgr – progesterona receptori (fmol/l);
* Er – estrogena receptori (fmol/l);
* Hormon – hormonālā terapija (0 = nav, 1 = ir);
* Rfstime – izdzīvošanas rādītājs pēc remisijas;
* Status – 0 = dzīve bez atkārtota gadījuma, 1 = atkārtotais gadījums vai nāve.

Kā mērķis tika izvēlēts audzēja pakāpe (Grade), saskaņā ar ko tika veikta analīze. Tam ir vairākas vērtības. Ielādētos datus var redzēt datu tabulā.

A screenshot of a computer

Description automatically generated with medium confidence

1.3.attēls. Datu tabula

Pēc datu ielādes un sagatavošanas darba autore izvēlējās datu analīzei Feature Statistics, Distributions, Scatter Plot, ko piedāvā Orange rīks, lai varētu izdarīt secinājumus par izliedes diagrammām, histogrāmmām un sadalījumu par datu kopas klašu atdalāmību.

A diagram of data flow

Description automatically generated with low confidence

1.4.attēls. Pievienotās programmas daļas

Pirmais, ko darba autore izmantoja, ir Feature Statistics, kura ir atspoguļota 1.4.attēlā.

A screenshot of a computer

Description automatically generated with medium confidence

1.5.attēls. Statistikas dati

Zemāk var redzēt vecuma sadalījumu (sk.1.6.att.).

A screenshot of a computer

Description automatically generated with medium confidence

1.6.attēls. Vecuma sadalījums

A screenshot of a computer

Description automatically generated with medium confidence

1.7.attēls. Pakāpes (stadijas) sadalījums

Distributions iespēja. 1.8.attēlā ir parādīta informācija par to, kā dalās dati. Distributions funkcija sniedz iespēju aplūkot tās pašas iespējas, tikai detalizētāk. Piemēram, 1.8.attēlā ir parādīta informācija, kā sadalās audzēju izmēri.

A screenshot of a computer

Description automatically generated with medium confidence

1.8.attēls. Audzēju izmēru sadalījums

1.9.attēlā var aplūkot arī vecuma sadalījumu.

A screenshot of a computer

Description automatically generated with medium confidence

1.9.attēls. Vecuma sadalījums

Tāpat var aplūkot skarto limfmezglu skaitu (sk.1.10.att.).

A screenshot of a computer

Description automatically generated with medium confidence

1.10.attēls. Skarto limfmezglu skaits

1.11.attēlā var redzēt sadalījumu, kur 0 ir krūts audzēja pacienti remisijā, bet 1 – pacienti bez remisijas, ar atkārtotiem audzēja gadījumiem vai letāls gadījums. Tāpat var aplūkot, cik daudz šiem pacientiem bija skarto limfmezglu.

A screenshot of a computer

Description automatically generated with medium confidence

1.11.attēls. Sadalījums atkarībā no skarto limfmezglu un izdzīvošanas

1.11.attēlā ir parādīts vecuma un pacientu statusa, kuriem 0 ir krūts audzēja pacienti remisijā, bet 1 – pacienti bez remisijas, ar atkārtotiem audzēja gadījumiem vai letāls gadījums, sadalījums.

A screenshot of a computer

Description automatically generated with medium confidence

1.12.attēls. Vecuma un statusa sadalījums

Tāpat darba autore ir analizējusi rezultātus, izmantojot Scatter Plot funkcionalitāti (sk.1.13.att.).

A screenshot of a computer screen

Description automatically generated with medium confidence

1.13.attēls. Scatter plot

1.14.attēlā ir atspoguļots x asī – vecums, y asī – audzēja izmērs.

A screenshot of a computer screen

Description automatically generated with low confidence

1.14.attēls. Vecuma un audzēja izmēra sadalījums

1.15.attēlā ir atspoguļots x asī – vecums, bet y asī – skarto limfmezglu skaits.

A screenshot of a computer

Description automatically generated with medium confidence

1.15.attēls. Vecuma un skarto limfmezglu sadalījums

**Secinājumi:**

1. Klašu datu kopas nav līdzsvarotas.
2. Datu vizuālais atspoguļojums, pēc darba autores domām, neļauj līdz galam redzēt datu struktūru, jo dati pārklājas. Datu objekti, kuri pieder dažādām klasēm, nav skaidri atdalāmi. Datu grupējumi atrodas tuvu viens otram, kas apgrūtina datu analīzi.

# NEPĀRRAUDZĪTA MAŠĪNMĀCĪŠANĀS

Tika sagatavota darba virsma Orange rīkā.

A picture containing text, diagram, screenshot, circle

Description automatically generated

2.1.attēls. Sagatavota darba virsma

## Hierarhiska klasterizācija

Klasterizācijas rezultātā var grupēt datus pēc noteiktām pazīmēm. Klasterus katrā līmenī veido, apvienojot klasterus nākamajā – zemākajā līmenī. Viszemākajā hierarhijas līmenī katram datu objektam ir savs klasteris.

A screenshot of a computer

Description automatically generated with medium confidence

2.2.attēls. Hierarhiskās klasterizācijas piemērs, izmantojot Grade

A screenshot of a computer

Description automatically generated with medium confidence

2.3.attēls. Klasterizācijas piemērs

Secinājumi:

1. Distances netika normalizētas, kā rezultātā rezultāts sanāca nepārskatāms.
2. Eksperimenta rezultātā sanāca ļoti liels klasteru daudzums.

A screenshot of a graph

Description automatically generated with medium confidence

2.4.attēls. Klasterizācijas piemērs (single)

Secinājumi:

1. Lai izveidotu šo klasterizāciju, tika ņemtas single vērtības, kā rezultātā arī klasteru daudzums sanāca lielāks.
2. Lai izveidotu šo klasteru, distances tika normalizētas.
3. Mazāki klasteri ir grūtāk pārskatāmi.

A screenshot of a computer

Description automatically generated with medium confidence

2.5.attēls. Klasterizācijas piemērs (average)

Secinājumi:

1. Lai izveidotu šo klasterizāciju, tika ņemtas vidējās vērtības.
2. Distances tika normalizētas, tomēr arī šajā gadījumā rezultāts sanāca ļoti liels.
3. Mazāki klasteri ir grūtāk pārskatāmi.
4. Grupējot datus pēc klasēm, klasteri ir labāk pārskatāmi.

## K-vidējo algoritms

Šis algoritms ir salīdzinoši vienkāršs. To ir viegli lietot. Pēc k-means var redzēt, cik daudz klasteru algoritmam būtu jāizveido. Tas nozīmē, ka K ir šī algoritma hiperparametrs. Var redzēt, ka vislabākais rezultāts ir 9 klasteri.

A screenshot of a computer

Description automatically generated with medium confidence

2.6.attēls. Silueta koeficienti

Darba autorei ir jānorāda, ka silueta koeficienti ir tuvāki nullei nekā tiecas uz 1. Jo tuvāk 1, jo labāk klasteri ir atdalāmi. Ja vērtības ir tuvāk 0, tad tas norāda, ka sadalījums starp klasteriem nav visai nozīmīgs. Vispirms eksperiments tika veikts ar 9 klasteriem.

A screenshot of a computer

Description automatically generated with medium confidence

2.7.attēls. 9 klasteri

Darba autore vēlējās uzzināt, kā izskatīsies algoritma darbības rezultāts, ja tiks izmantoti 8 klasteri.

A screenshot of a computer

Description automatically generated with medium confidence

2.8.attēls. 8 klasteri

Izvērtējot iegūto informāciju, ir jānorāda, ka neliels klasteru skaits padara datus labāk pārskatāmus. Tomēr pēc darba autores domām, ir nepieciešams turpināt eksperimentu, izmantot 4 vai 5 klasterus, kā ir parādīts zemāk.

A screenshot of a computer

Description automatically generated with medium confidence

2.9.attēls. 5 klasteri

Pārskatot rezultātus, ir jānorāda, ka, palielinot klasteru skaitu, dati var palikt mazāk pārskatāmi, ko var redzēt 2.10.attēlā. Tāpat ir jānorāda, ka, palielinoties klasteru skaitam, parādās arī vērtējumi ar – zīmi, kas norāda uz to, ka objekti netika pareizi attiecināti uz noteikto klasteri.

A screenshot of a computer

Description automatically generated with medium confidence

2.10.attēls. 10 klasteri

# SECINĀJUMI

1. Darba ietvaros tika analizēta un izvērtēta datu kopa, kas tiek darīts jau ne pirmo reizi. Tika izvēlēta laba datu kopa, uz kā var trenēties ne tikai vizualizēt datus, bet arī analizēt to.
2. Pirms darba uzsākšanas bija nepieciešams aplūkot datus, analizēt tos, kā arī sagatavot darbam ar tiem. Pirms algoritmu izmantošanas dati tika aplūkoti un ielādēti programmā, kā arī aprakstīti.
3. Pēc pirmās darba daļas pabeigšanas darba autore var izdarīt sekojošus secinājumus:

* Klašu datu kopas nav līdzsvarotas;
* Datu vizualizācija līdz galam neļauj redzēt datu struktūru, jo dati pārklājas;
* Datu pārklāšanās traucē atdalīt vienu informāciju no otras.
* Datu objekti nav skaidri atdalāmi.

1. Darbā tika izmantoti nepārraudzītas mašīnmācīšanās algoritmi – hierarhiskā klasterizācija un K-vidējo algoritms. Dati tika sadalīti klasteros. Tika veikti arī eksperimenti ar klasteriem, kā rezultātā parādījās vērtējumi ar – zīmi. Tas nozīmē, ka objekti netika pareizi attiecināti pret klasteriem.
2. Darba autorei ir jānorāda, ka Orange rīku viņa izmantoja jau otro reizi, kā rezultātā pārliecinājās, ka tas ir labs un ērts rīks datu analīzei.