**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\_2prakt\_Final**

**Rīga, 2022**

**SATURS**

[1. DATU KOPAS APRAKSTS 3](#_Toc104303578)

[2. NEPĀRRAUDZĪTA MAŠĪNMĀCĪŠANĀS 13](#_Toc104303579)

[2.1. Hierarhiskā klasterizācija 13](#_Toc104303580)

[2.2. K-vidējo algoritms 19](#_Toc104303581)

[3. PĀRRAUDZĪTA MAŠĪNMĀCĪŠANĀS 24](#_Toc104303582)

[3.1. Darbs ar Orange rīku 24](#_Toc104303583)

[3.2. Darbības ar testa kopu 30](#_Toc104303584)

[SECINĀJUMI 31](#_Toc104303585)

# DATU KOPAS APRAKSTS

**Datu kopas apraksts**

Datu kopa ir pieejama sekojošā linkā: <https://www.kaggle.com/datasets/brendan45774/test-file>

Saite uz GitHub projektu: https://github.com/jekmiz/MI\_2prakt\_Final

Datu kopas autors: Brenda N. Datu kopa ir pieejama Kaggle mājaslapā. Šīs datu kopas veidotāja veidoja šo datu kopu ar mērķi analizēt un vizualizēt datus. Tā ir testa kopa, jo ne visi dati ir pareizi, kā norādīja pati kopas autore. Darba autore savā darbā izmantoja Titanika datu kopu. Testa datu kopa aptuveni apkopo informāciju par to, kādi bija Titanika pasažieri, vai viņi nomira, vai izdzīvoja, kādi viņiem bija biļešu numuri utt.

Datu kopa jau bija pieejama .csv formātā, kā rezultātā to nebija nepieciešams transformēt.

Datu kopas licencēšanas nosacījumi: datu kopa ir brīvi pieejama mācību nolūkos. Kopas autore nesniedza informāciju par licencēšanas aspektiem.

Veids, kā datu kopa savākta: datu kopa ir savākta, izmantojot publicēto informāciju par Titānika katastrofā bojāgājušajiem, kā arī aptuveniem datiem, kurus var iegūt. Informācijas sadaļā kopas autore ir norādījusi, ka aptuveni 70% no testa kopas datiem ir ticami.

**Datu kopas satura apraksts**

Pirms aprakstīt datu kopas saturu darba autorei ir nepieciešams pievērst uzmanību noteiktām kolonnām, kuras tika izdzēstas, piemēram, kolonna Cabin.

**Chart

Description automatically generated with medium confidence**

1.1.attēls. Izdzēstā kolonna

Text

Description automatically generated

1.2.attēls. Aizpildīto un neaizpildīto datu sadalījums

Tā kā lielākā daļa datu iztrūka (un nebija iespējas tos atrast internetā), tad darba autore izvēlējās to izdzēst.

Līdzīgā veidā, lai nedzēstu kolonnas, informācija tika papildināta – daļēji izmantojot randomizētās vērtības, daļēji – izmantojot internetā atrasto informāciju par Titānika pasažieriem.

Calendar

Description automatically generated with medium confidence

1.2.attēls. Iztrūkstošie un papildinātie dati

1.2.attēlā ir atspoguļots sākotnējais variants, kā izskatījās dati pirms tie tika papildināti. Tā kā bija obligāti tie jāpapildina, tad izmantojot interneta avotus, kā arī randomizētās vērtības (ja datus nevarēja atrast), darba autore papildināja iztrūkstošos datus.

Dati tika nedaudz pārveidoti tā, lai ar tiem varētu vieglāk strādāt. Tika izdzēstas liekās vērtības, kā arī papildinātas iztrūkstošās.

Table

Description automatically generated

1.3.attēls. Pārveidotie un papildinātie dati

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 informācijas par to, cik daudz ir datu objektu, kā arī, ka ir 9 pazīmes un nav izvēlēta mērķa pazīme, pēc kuras arī tiks veikta datu analīze.

Datu objektu skaits: 418

Kopai ir 10 pazīmes, nav tukšo vērtību.

Klase: Survived (0 = nav izdzīvojis, 1 = izdzīvojis);

Graphical user interface, text, application

Description automatically generated

1.4.attēls. .csv faila ielādēšana Orange rīkā

Pazīmju (atribūtu) skaits un to atspoguļojums:

* PassengerId – pasažiera ID;
* Pclass – pasažiera klase (1 = Pirmā klase, 2 = Otrā klase, 3 = Trešā klase);
* Name – pasažiera vārds;
* Sex – pasažiera dzimums (male, female);
* Age – pasažiera vecums;
* SibSp – radinieku skaits;
* Parch;
* Ticket – biļetes Nr.
* Fare – biļetes summa;
* Embarked – sākuma punkts (C - Cherbourg Q = Queenstown S = Southampton).

Kā mērķis tika izvēlēts Survived, saskaņā ar ko tika veikta analīze. Tam ir vairākas vērtības:

* 0 = nav izdzīvojis;
* 1 = izdzīvojis

Ielādētos datus var redzēt arī Data Table.

Graphical user interface, table

Description automatically generated

1.5.attēls. Data Table ar datiem

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

Diagram

Description automatically generated

1.6.attēls. Pievienotās programmas daļas

Pirmais, ko darba autore vēlējās uzzināt, ir datu statistiskā analīze, kura ir atspoguļota 1.7. un 1.8.attēlos.

Chart

Description automatically generated with low confidence

1.7.attēls. Statistikas dati

1.8. attēlā var redzēt mirušo (neidzīvojušo) un izdzīvojušo sadalījumu.

A picture containing chart

Description automatically generated

1.8.attēls. Statistikas dati

To pašu var redzēt, izmantojot Distributions iespēju. 1.9.attēlā ir parādīta informācija par to, kā sadalās izdzīvojušo un mirušo rezultāti. Lielākā daļa pasažieru nav izdzīvojuši.

Graphical user interface, chart

Description automatically generated

1.9.attēls. Izdzīvojušo un mirušo sadalījums

1.9.attēlā var precīzāk aplūkot izdzīvojušo un mirušo sadalījumu. 1.10.attēlā, kuru var redzēt zemāk mirušo un izdzīvojušo sadalījumu atkarībā no vecuma. Var redzēt arī ļoti interesantu faktu, ka viens no pasažieriem, kura vecums bija virs 70, ir izdzīvojis katastrofas laikā. Tāpat sadalījums ļauj novērtēt, ka lielākā daļa izdzīvojušo ir vecumā no 20 līdz 40 gadiem. Diemžēl, katastrofā gāja bojā arī daudz bērnu vecumā no aptuveni 5 gadiem līdz 10 gadiem.

Chart, histogram

Description automatically generated

1.10.attēls. Izdzīvojušo un mirušo sadalījums atkarībā no vecuma

Chart, bar chart

Description automatically generated

1.11.attēls. Izdzīvojušo un mirušo sadalījums atkarībā no ceļojuma sākuma punkta

1.11.attēlā ir atspoguļoti dati par izdzīvojušo un mirušo sadalījumu atkarībā no ceļojuma sākuma punktiem. Lielākā daļa pasažieru, kuri neizdzīvoja, ceļoja no S jeb Southampton.

Chart, scatter chart

Description automatically generated

1.12.attēls. Scatter plot (Pasažieru ID un vecums)

1.12.attēlā ir atspoguļota informācija par pasažieru ID un viņu vecuma sadalījumu. Tas bija viens no grafikiem, kur var skaidri redzēt klašu sadalījumu.

Graphical user interface, application, Word

Description automatically generated

1.13.attēls. Scatter plot (Klase un vecums)

Chart, scatter chart

Description automatically generated

1.14.attēls. Scatter plot (pasažiera ID un biļetes cena)

Chart, scatter chart

Description automatically generated

1.15.attēls. Scatter plot (Vecums un biļetes cena)

A picture containing graphical user interface

Description automatically generated

1.16.attēls. Scatter plot (Vecums un klase)

**Secinājumi:**

1. Klašu datu kopas nav līdzsvarotas.
2. Datu vizuālais atspoguļojums neļauj līdz galam redzēt datu struktūru, jo dati pārklājas, kas traucē atdalīt vienu informāciju no otras. Datu objekti, kuri pieder dažādām klasēm, nav skaidri atdalāmi.
3. Atdalāmi datu grupējumi – klases, ceļojuma sākums, dzimums.
4. Identificētie datu grupējumi atrodas tuvu viens otram, kas apgrūtina datu analīzi.

# NEPĀRRAUDZĪTA MAŠĪNMĀCĪŠANĀS

Diagram

Description automatically generated

2.1.attēls. Nepārraudzītas mašīnmācīšanās

## Hierarhiskā klasterizācija

Klasterizācijas rezultātā var grupēt datus pēc pazīmēm. Klasterus katrā hierarhijas līmenī veido, apvienojot klasterus nākamajā zemākajā līmenī. Viszemākajā hierarhijas līmenī katram datu objektam ir savs klasteris. Visaugstākajā hierarhijas līmenī ir tikai viens klasteris, kurā ietilpst visi datu objekti.

Graphical user interface, application, Word

Description automatically generated

2.2.attēls. Hierarhiskās klasterizācijas piemērs

Secinājumi:

1. Lai izveidotu šos klasterus, distances netika normalizētas, kā rezultātā arī rezultāts sanāca liels un nepārskatāms.
2. Eksperimenta rezultātā sanāca liels klasteru daudzums.

Chart

Description automatically generated

2.2.attēls. Hierarhiskās klasterizācijas piemērs (single)

Secinājumi:

1. Lai izveidotu šo klasterizāciju, tika ņemtas single (vienas) vērtības, kā rezultātā klasteru daudzums sanāca lielāks.
2. Lai izveidotu šos klasterus, distances tika normalizētas, tomēr arī šajā gadījumā rezultāts sanāca liels.
3. Datu objektu pārklāšanās nenotiek, tomēr mazāki klasteri ir grūtāk pārskatāmi.
4. Normāla un pārskatāma klasteru veidošanās ir tikai līdz noteiktam līmenim, pēc tam sākas “putra” no mazākiem klasteriem, kas apgrūtina pārskatīšanu.

Diagram

Description automatically generated

2.3.attēls. Hierarhiskās klasterizācijas piemērs (average)

Secinājumi:

1. Lai izveidotu šo klasterizāciju, tika ņemtas vidējās vērtības.
2. Lai izveidotu šos klasterus, distances tika normalizētas, tomēr arī šajā gadījumā rezultāts sanāca liels.
3. Datu objektu pārklāšanās nenotiek, tomēr mazāki klasteri ir grūtāk pārskatāmi.
4. Normāla un pārskatāma klasteru veidošanās ir tikai līdz noteiktam līmenim, pēc tam sākas “putra” no mazākiem klasteriem, kas apgrūtina pārskatīšanu.
5. Grupējot datus pēc klases, klasteri ir daudz labāk pārskatāmi, salīdzinot ar iepriekšējo, kad parametrs bija “Survived” un netika ņemtas vidējās vērtības.

Chart

Description automatically generated

2.4.attēls. Hierarhiskās klasterizācijas piemērs (Weighted)

Secinājumi:

1. Lai izveidotu šo klasterizāciju, tika ņemtas weighted vērtības.
2. Klasteris ir pārskatāmāks, salīdzinot ar iepriekšējiem trim piemēriem, kuri tika aplūkoti augstāk.
3. Datu objektu pārklāšanās nenotiek, kā arī mazāki datu klasteri ir vieglāk uztverami un pārskatāmi.

Diagram

Description automatically generated

2.5.attēls. Hierarhiskās klasterizācijas piemērs (Complete)

Secinājumi:

1. Lai izveidotu šo klasterizāciju, tika ņemtas complete vērtības.
2. Klasteris ir pārskatāmāks, salīdzinot ar pirmajiem trim piemēriem, kuri tika aplūkoti augstāk.
3. Datu objektu pārklāšanās nenotiek, kā arī mazāki datu klasteri ir vieglāk uztverami un pārskatāmi.

## K-vidējo algoritms

Pirmais, kam darba autore vēlētos vērst uzmanību, ir tas, ka algoritms ir salīdzinoši vienkāršs un to ir viegli lietot. Pēc k-Means var redzēt, cik daudz klasterus algoritmam būtu jāizveido. Tas nozīmē, ka K ir šī algoritma hiperparametrs.

Var redzēt, ka vislabākais rezultāts ir 4 klasteri. Tomēr darba autorei ir nepieciešams vērst uzmanību uz to, ka kopumā tiks veikti 5 dažādi eksperimenti ar dažādu klasteru skaitu.

Graphical user interface

Description automatically generated

2.6.attēls. Silueta koeficienti

Darba autorei jau pašā sākumā ir jānorāda, ka silueta koeficienti ir tuvāki nullei nekā tiecas uz 1. Jo tuvāk 1, jo labāk, jo labāk klasteri ir atdalāmi. Ja vērtības ir tuvāk 0, tad tas norāda, ka sadalījums starp klasteriem nav visai nozīmīgs. Ja vērtējumi nāk ar – zīmi, tad objekti nepareizi tiek attiecināti uz klasteriem.

Vispirms eksperiments tika veikts ar 4 klasteriem.

Graphical user interface, application

Description automatically generated

2.7.attēls. 4 klasteri

Darba autore vēlējās uzzināt, kā izskatīsies algoritma darbības rezultāts, ja tiks izmantoti 3 klasteri. Gadījumā ar 3 klasteriem var redzēt, ka parādās daži vērtējumi ar – zīmi, kas nozīmē, ka objekti tika nepareizi attiecināti pret klasteriem.

Chart, bar chart

Description automatically generated

2.8.attēls. 3 klasteri

Izvērtējot 2.7. un 2.8.attēlos sniegto informāciju, darba autorei ir jānorāda, ka neliels klasteru skaits padara datus pārskatāmākus, tomēr, pēc darba autores domām, pārskatāmāks variants ir ar 4 klasteriem. Nākamais eksperiments – algoritma darbības rezultāts ar 5 klasteriem. Ir jānorāda, ka palielinot klasteru skaitu, dati var palikt mazāk pārskatāmi.

Chart, bar chart

Description automatically generated

2.9.attēls. 5 klasteri

Gadījumā ar 5 klasteriem var redzēt, kā ir parādīts 2.9.attēlā, ka parādās daži vērtējumi ar – zīmi, kas nozīmē, ka objekti tika nepareizi attiecināti pret klasteriem.

A picture containing bar chart

Description automatically generated

2.10.attēls. 6 klasteri

Gadījumā ar 6 klasteriem var redzēt, ka parādās daži vērtējumi ar – zīmi, kas nozīmē, ka objekti tika nepareizi attiecināti pret klasteriem.

Graphical user interface

Description automatically generated with medium confidence

2.11.attēls. 8 klasteri

2.11.attēlā ir atspoguļots rezultāts ar 8 klasteriem. Klasteri ir mazāk pārskatāmi un sliktāk uztverami, kā arī tam parādās vērtējumi ar – zīmi, kas nozīmē, ka datu objekti netika pareizi attiecināti uz noteikto klasteri.

Tas arī pierāda silueta koeficienti, kas tika atspoguļots 2.6.attēlā, kā vislabākais rezultāts būs ar 4 klasteriem. Jo tuvāk 1, jo labāk, jo labāk klasteri ir atdalāmi. Silueta koeficients 4 klasteriem bija vislielākais un vistuvākais nullei. Visi pārējie koeficienti bija zemāki par to. Ja vērtības ir tuvāk 0, tad tas norāda, ka sadalījums starp klasteriem nav visai nozīmīgs. Ja vērtējumi nāk ar – zīmi, tad objekti nepareizi tiek attiecināti uz klasteriem.

# PĀRRAUDZĪTA MAŠĪNMĀCĪŠANĀS

## Darbs ar Orange rīku

Diagram

Description automatically generated

3.1.attēls. Pārraudzītas mašīnmācīšanās rīki Orange

3.1.attēlā ir atspoguļota informācija par pārraudzītas mašīnmācīšanās algoritmu rezultātu salīdzināšanu. Eksperimentāli tika mainīti hiperparametri uz 0 (neizdzīvojušie) un uz 1 (izdzīvojušie).

Algoritmu metrikas, kuras tika iegūtas:

1. CA – precizitāte;
2. F1 – F1 mērs;
3. Precision – pozitīvā prognozējošā vērtība;
4. Recall – jutīgums.

Eksperimenti, mainot hiperparametru:

1. Parametrs nav izvēlēts.

Table

Description automatically generated

3.2.attēls. Pārraudzītas mašīnmācīšanās algoritmu salīdzināšana, mainot hiperparametru

Aplūkojot CA metriku, vislabākā precizitāte ir SVM. Tas dod labāko rezultātu, jo ir tuvāks 1. Vislabākā pozitīvā prognozējošā vērtība ir SVM, jo tā ir tuvāka 1. Lielāka vērtība norāda uz labāku mašīnmācīšanās algoritma veikumu. Aplūkojot jūtīguma metriku, ir jānorāda, ka vislabākais rezultāts bija SVM algoritmam. Kā arī iepriekšējā gadījumā, lielāka vērtība norāda uz labāku mašīnmācīšanās algoritma veikumu.

1. Ir izvēlēts parametrs 0.

Aplūkojot CA metriku, vislabākā precizitāte ir SVM. Tas dod labāko rezultātu, jo tas ir tuvāks 1.

Graphical user interface, table

Description automatically generated

3.3.attēls. Pārraudzītas mašīnmācīšanās algoritmu salīdzināšana, mainot hiperparametru (neidzīvoja)

Aplūkojot CA metriku, vislabākā precizitāte ir SVM. Tas dod labāko rezultātu, jo ir tuvāks 1. Vislabākā pozitīvā prognozējošā vērtība ir SVM, jo tā ir tuvāka 1. Lielāka vērtība norāda uz labāku mašīnmācīšanās algoritma veikumu. Aplūkojot jūtīguma metriku, ir jānorāda, ka vislabākais rezultāts bija neironu tīklam. Kā arī iepriekšējā gadījumā, lielāka vērtība norāda uz labāku mašīnmācīšanās algoritma veikumu.

Graphical user interface, table

Description automatically generated

3.4.attēls. Pārraudzītas mašīnmācīšanās algoritmu salīdzināšana, mainot hiperparametru (izdzīvoja)

Aplūkojot CA metriku, vislabākā precizitāte ir SVM. Tas dod labāko rezultātu, jo ir tuvāks 1. Vislabākā pozitīvā prognozējošā vērtība ir SVM, jo tā ir tuvāka 1. Lielāka vērtība norāda uz labāku mašīnmācīšanās algoritma veikumu. Aplūkojot jūtīguma metriku, ir jānorāda, ka vislabākais rezultāts bija SVM algoritmam. Kā arī iepriekšējā gadījumā, lielāka vērtība norāda uz labāku mašīnmācīšanās algoritma veikumu.

Table

Description automatically generated

3.5.attēls. Pārraudzīšanas mašīnmācīšanās algoritmu salīdzināšana, mainot uz Random sampling

Arī mainot no Cross validation uz Random sampling, rezultāti nedaudz pamainījās, bet tomēr vislabākie metrikas rezultāti ir SVM algoritmam. Tas dod labāko rezultātu, jo to metriku rezultāts ir tuvāks 1.

Piemēram, vislabākā pozitīvā prognozējošā vērtība ir SVM, jo tā ir tuvāka 1. Lielāka vērtība norāda uz labāku mašīnmācīšanās algoritma veikumu. Aplūkojot jūtīguma metriku, ir jānorāda, ka vislabākais rezultāts bija SVM algoritmam. Kā arī iepriekšējā gadījumā, lielāka vērtība norāda uz labāku mašīnmācīšanās algoritma veikumu.

Paredzēšana sniedz iespēju aplūkot, kāda bija katra datu objekta klase, kādas prognozes (klasifikācijas iezīmes) nodrošināja katrs algoritms. To var aplūkot 2.16.attēlā. Tas palīdz izpētīt algoritmu vērtības.

Graphical user interface, application, table, Excel

Description automatically generated

3.5.attēls. Paredzēšana

3.6.attēlā tiek salīdzināti algoritmi savā starpā.

Chart, line chart

Description automatically generated

3.6.attēls. Algoritmu darbības salīdzināšana

## Darbības ar testa kopu

Diagram

Description automatically generated

* 1. attēls. Orange izskats

# SECINĀJUMI

1. Darba ietvaros darba autore analizēja un izvērtēja datu kopu, ko viņa darīja pirmo reizi. Tika noskaidrots, ka tā ir laba datu kopa, uz kā var trenēties ne tikai vizualizēt datus par izdzīvojušajiem un mirušajiem Titanika pasažieriem, kā arī kādām klasēm viņi piederēja, bet arī ko var analizēt un uz kā var trenēties izmantot iegūtās zināšanas.
2. Pirms darba uzsākšanas bija nepieciešams aplūkot datus, sagatavot tos, lai ar tiem varētu strādāt. Pirms algoritmu izmantošanas dati tika analizēti un izvērtēti, izmantojot histogrammu, sadalījuma analīzi.
3. Pēc pirmās darba daļas pabeigšanas darba autorei ir sekojoši secinājumi: klašu datu kopas nav līdzsvarotas, datu vizuālais neļauj līdz galam redzēt datu struktūru, jo dati pārklājas, kas traucē atdalīt vienu informāciju no otras. Datu objekti, kuri pieder dažādām klasēm, nav skaidri atdalāmi. Identificētie datu grupējumi atrodas tuvu viens otram, kas apgrūtina datu analīzi.
4. Darbā tika izmantoti nepārraudzītas mašīnmācīšanās algoritmi: hierarhiskā klasterizācija un K-vidējo algoritms. K-vidējā algoritma gadījumā darba autore pateicoties koeficientiem sadalīja datu objektus klasteros. Tika veikti arī eksperimenti, kuru gaitā parādījās arī vērtības ar – zīmi, kas nozīmē, ka objekti netika pareizi attiecināti pret klasteriem. Hierarhiskās klasterizācijas gadījumā var norādīt, ka ir nepieciešams izveidot tādus klasterus, lai tie būtu pārskatāmi (arī mazākie klasteri) un ar tiem varētu strādāt.
5. Izvērtējot pārraudzītas mašīnmācīšanās algoritmus, ir jānorāda, ka vislabākos rezultātus parādīja SVM. Tam ir visaugstākā precizitāte, pozitīvā prognozējošā vērtība, jūtīguma metrika.
6. Darba autore var pieminēt, ka ar Orange rīku ir ļoti viegli, kā arī ātri un ērti strādāt. To var viegli apgūt, kā arī tas ir intuitīvi lietojams.