

**DATORZINĀTNES UN INFORMĀCIJAS TEHNOLOĢIJAS FAKULTĀTE Priekšmeta "Mākslīgā intelekta pamati(1),22/23-P"**

**2. Praktiskais darbs**

**Artūrs Adamovičs**

**2.kurss 6.grupa**

**Studenta apliecības nr. 211RDB164**

**Rīga, 2023**

Saturs

[1.Daļa- Datu pirmapstrāde/izpēte 3](#_Toc134777163)

[1.daļas secinājumi 10](#_Toc134777164)

[2.daļa – Nepārraudzītā mašīnmācīšanās 11](#_Toc134777165)

[2.Daļas secinājumi 15](#_Toc134777166)

[3. daļa – Pārraudzītā mašīnmācīšanās 15](#_Toc134777167)

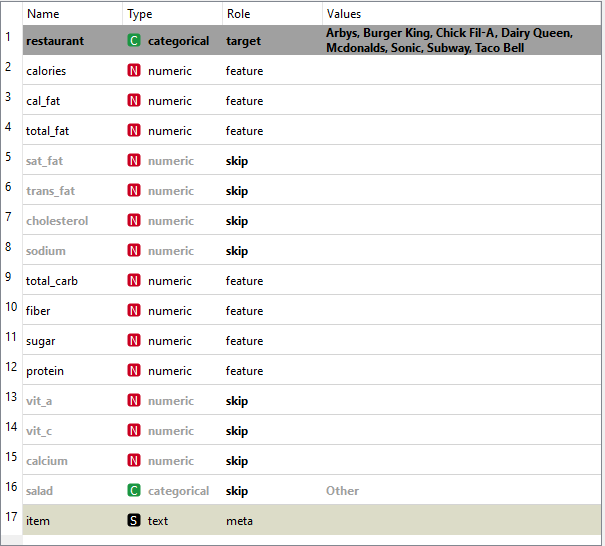
[3.daļas secinājumi 23](#_Toc134777168)

[Izmantoto avotu saraksts 24](#_Toc134777169)

# 1.Daļa- Datu pirmapstrāde/izpēte

**Datu kopas apraksts**

Esmu izvēlējies datu kopu, kas apraksta ātrās ēdināšanas edienu uztura faktiem. Datu kopas nosaukums ir “Fastfood Nutrition”. Avots: Data.world API . Šīs datu kopas izveidotājs un īpašnieks ir “Ulrik Thyge Pedersen”. Datu kopu izveidoja lietojot populārākss ātrās ēdināšanas restorānu piedāvājumus, lai informētu un parādītu patērētājiem, cik neveselīgi tas ir, pēc kā patērētājs varētu veikt informētu izvēli. Datu kopas licences nosaukums ir “Attribution 4.0 International”, kas min, ka šo ar šo datu kopu var dalīties un pārveidot pēc saviem ieskatiem, norādot atsauces uz oriģinālo darbu. Veids, kā datu kopa tika savākta diemžēl nebija minēts. Datu kopas oriģinālā saite:<https://www.kaggle.com/datasets/ulrikthygepedersen/fastfood-nutrition> .

Kopējais datu objektu skaits kopā ir 515 objekti.

1.attēls datu kopas atribūti

Datu kopā ir 8 klases: *Arbys, Burger King, Chick Fil-A, Dairy Queen, Mcdonalds, Sonic, Subway* un *Taco Bell.* Šīs klases ir zem atribūta “restaurant”, un norāda uz restorāna nosaukumu.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nr. | Nosaukums | Skaidrojums | Tips | Diapazons |
| 1. | Restaurant | Restorāna nosaukums | Kategorija | *Arbys, Burger King, Chick Fil-A,* utml. |
| 2. | Calories | Kaloriju daudzums | Skaitliska vērtība | [20;2430] |
| 3. | Cal\_fat | Kaloriju daudzums no taukiem | Skaitliska vērtība | [0;1270] |
| 4. | Total\_fat(grami) | Kopējais tauku daudzums | Skaitliska vērtība | [0;141] |
| 5. | Total\_carb(grami) | Kopējais ogļhidrātu daudzums | Skaitliska vērtība | [0;126] |
| 6. | Fiber(grami) | Šķiedrvielu daudzums | Skaitliska vērtība | [0;17] |
| 7. | Sugar(grami) | Cukura daudzums | Skaitliska vērtība | [0;87] |
| 8. | Protein(grami) | Proteīna daudzums | Skaitliska vērtība | [1;186] |
| 9. | item | Ēdiena nosaukums | Teksta vērtība | Jebkāds teksts |

1.tabula atribūtu klasificēšana

Datu objektu skaits, kas pieder katrai klasei-

* “Arbys”- 55 datu objekti
* “Burger King”- 70 datu objekti
* “Chick Fil-a”- 27 datu objekti
* “Dairy Queen”- 42 datu objekti
* “Mcdonalds”- 57 datu objekti
* “Sonic”- 53 datu objekti
* “Subway”- 96 datu objekti
* “Taco Bell”-115 datu objekti

**Izkliedes diagrammas**

Attēls, kurā ir teksts, ekrānuzņēmums, programmatūra, multivides programmatūra

Apraksts ģenerēts automātiski2.attēls kalorijas pret taukiem

Pēc 2.attēlā parādītās izkliedes diagrammas var redzēt, ka ātrās ēdināšanas restorāns “Subway” savā produktu klāstā piedāvā produktus ar salīdzinoši zemu kaloriju daudzumu, kā arī mazu tauku daudzumu, tikmēr restorānam “Mcdonalds” ir pāris produkti ar vislielāko kaloriju un tauku daudzumu.

Attēls, kurā ir teksts, ekrānuzņēmums, programmatūra, datora ikona

Apraksts ģenerēts automātiskiPēc 3.attēlā parādītās izkliedes diagrammas var secināt, ka tomēr jāuzmanās izvēloties restorāna “Subway” produktus, jo tomēr no visiem ātrās ēdināšanas restorāniem “Subway” produktos ir nesamērīgi daudz cukura, bet ir arī produkti restorāna klāstā, kuriem cukura daudzums nav tik liels, kas nozīmē, kad izvēloties “Subway” produktus jāpievērš uzmanība uz cukura daudzumu.

3.attēls tauku saturs pret cukuru

Attēls, kurā ir ekrānuzņēmums, teksts, skice, diagramma

Apraksts ģenerēts automātiski**Histogrammas**

4.attēls histogramma 1

4.attēla histogrammā varam spriest, ka “Taco Bell” piedāvājumā ir vislielākais produktu klāsts, kurā ir salīdzinoši mazs tauku daudzums(pārējo restorānu perspektīvā).

Attēls, kurā ir ekrānuzņēmums, teksts, programmatūra, diagramma

Apraksts ģenerēts automātiski5.attēla histogrammā es novēroju, ka restorānam “Mcdonalds” ir diezgan liels klāsts ar vidēja-liela proteīna daudzuma produktiem.

5.attēls histogramma 2

Attēls, kurā ir ekrānuzņēmums, teksts, diagramma, skice

Apraksts ģenerēts automātiski6.attēlā redzama vidējā, mediāna, modas, dispersijas, minimālās un maksimālās vērtības starp visiem datu kopas atribūtiem.

6.attēls statistiskie rādītāji

Attēls, kurā ir teksts, ekrānuzņēmums, taisnstūris

Apraksts ģenerēts automātiski

7.attēls datu kopas fragments

7.attēlā esmu uztaisījis ekrānuzņēmumu no datu kopas faila struktūru.

# 1.daļas secinājumi

* Vai klases datu kopā ir līdzsvarotas, vai dominē viena klase (vai vairākas klases)?  
  Nē, izpētot datu kopu, dominē “Taco Bell” un “Subway” klase ar 115 un 96 datu objektiem.
* Vai datu vizuālais atspoguļojums ļauj redzēt datu struktūru?   
  Brīžiem ir diezgan grūti atšķirt kurai klasei pieder datu objekts. Novēroju, ka starp šīm ēstuvēm ir diezgan līdzīgas vērtības, jo visas šīs ēstuves ir neveselīgas, un ja pievienotu kādu veselīgu restorānu, tad uzreiz varētu atšķirt vizuālajā attēlojumā.
* Cik datu grupējumus ir iespējams identificēt, pētot datu vizuālo atspoguļojumu?  
  Pētot attēlus vel vienu reizi, varu uzreiz viegli identificēt 6 no 8 datu grupējumiem.
* Vai identificētie datu grupējumi atrodas tuvu viens otram vai tālu viens no otra?  
  Visos gadījumos datu grupējumi atrodas tuvu viens otram.

# 2.daļa – Nepārraudzītā mašīnmācīšanās

**K-vidējo algoritms**

Attēls, kurā ir teksts, ekrānuzņēmums, displejs, programmatūra

Apraksts ģenerēts automātiski8.attēlā varam redzēt K-vidējā algoritma hiperparametrus:

8.attēls K-vidējo vērtības

* Klasteru skaitu, to izvēlamies 5.
* Normalizēt kolonnas, to iezīmējam, lai izņemtu ārā datu objektus ar tukšām vērtībām.
* Inicializācijas metode, testēju abus, bet rezultātu manā gadījumā nemainīja.
* Maksimālo iterāciju skaits, to atstājam 300, testēju palielinot un pazeminot skaitli, bet rezultātu nemainija.

Pie lauka *Silhouette Scores* izvēlamies klasteru skaitu, kura vērtība ir tuvāk pie 1, kas manā variantā ir 2 klasteru skaits.

Attēls, kurā ir teksts, ekrānuzņēmums, programmatūra, grafikas programmatūra

Apraksts ģenerēts automātiski

9.attēls 2 klasteri

9.attēlā varam redzēt, ka datu kopas esošās klases divos klasteros samērā labi atdalāmas, bet tiklīdz izvēlamies 3 klasterus, kā 10.attēlā uzreiz jau paliek slikti atdalāms.

Attēls, kurā ir programmatūra, ekrānuzņēmums, multivides programmatūra, datora ikona

Apraksts ģenerēts automātiski

10.attēls 3 klasteri

**hierarhiskā klasterizācija**

Attēls, kurā ir teksts, ekrānuzņēmums, fonts, cipars

Apraksts ģenerēts automātiski

11.attēls hierarhiskā klasterizācija

11.attēlā redzami mani hiperparametru iestatījumi, attāluma metodi izvēlējos Eiklīda ar datu normalizāciju.

Attēls, kurā ir teksts, diagramma, ekrānuzņēmums, skice

Apraksts ģenerēts automātiski

12.attēls atdalošā līnija 25.1%

Attēls, kurā ir teksts, diagramma, ekrānuzņēmums, plāns

Apraksts ģenerēts automātiski

13.attēls atdalošā līnija 28.4%

Attēls, kurā ir teksts, ekrānuzņēmums, displejs, programmatūra

Apraksts ģenerēts automātiski

14.attēls atdalošā līnija 20.2%

12.,13. un 14.attēlā var secināt, ka šo datu kopu ar hierarhisko klasterizāciju ir slikti atdalāmas klases. Eksperimentēju kustinot atdalošo līniju, kā ari mainīju iestatījumus, un pētīju klasteru skaitu un to saturu. Pēc eksperimentēšanas ir skaidri redzams, ka sadalot vairākos klasteros dažādas klases joprojām ir saliktas kopā vienā klasterā.

# 2.Daļas secinājumi

Veicot eksperimentēšanu ar abiem algoritmiem var secināt, ka ar hierarhisko klasterizāciju klases ir slikti atdalāmas, bet ar K-vidējo algoritmu klases ir samērā labi atdalāmas, ja klasteru skaits ir tikai 2, līdz ko klasteru skaits ir lielāks par 2, tad klases paliek slikti atdalāmas.

# 3. daļa – Pārraudzītā mašīnmācīšanās

Attēls, kurā ir teksts, ekrānuzņēmums, cipars, fonts

Apraksts ģenerēts automātiski Priekš praktiskā darba 3.daļas es izvēlējos KNN un SVM pārraudzītās mašīnmācīšanās algoritmus. Šos divus algoritmus es izvēlējos, jo tie man visvairāk ieinteresēja un likās, ka es šo algoritmu darbības principu vislabāk sapratu.

15.attēls data sampler

Algoritmu apmācībai es izvēlējos 70% procentus, un priekš testēšanas atstāju 30%. Atstāju pārējos iestatījumus pēc noklusējuma, kas attēlots 15.attēlā.

Attēls, kurā ir teksts, ekrānuzņēmums, cipars, fonts

Apraksts ģenerēts automātiskiAttēls, kurā ir teksts, ekrānuzņēmums, fonts, cipars

Apraksts ģenerēts automātiski16.attēlā redzams KNN algoritma hiperparametri, kurus es atstāju pēc noklusējuma priekš pirmā eksperimenta.

17.attēls Neironu tīkli 1

16.attēls Knn iestatījumi 1

17.attēlā redzami neironu tīkla hiperparametru, kurus arī atstāju pēc noklusējuma priekš pirmā eksperimenta. Turpmāk ari 18.attēlā, kur redzami SVM algoritma hiperparametri ,atstāšu pēc noklusējuma.

Attēls, kurā ir teksts, ekrānuzņēmums, cipars, fonts

Apraksts ģenerēts automātiski

18.attēls SVM 1

Attēls, kurā ir teksts, ekrānuzņēmums, programmatūra, displejs

Apraksts ģenerēts automātiski

19.attēls testu rezultāts

Attēls, kurā ir teksts, ekrānuzņēmums, fonts, cipars

Apraksts ģenerēts automātiski19.attēlā ir redzami pašreizējie algoritmu testa rezultāti, var secināt, ka neironu tīkls pagaidām uzrāda vislabāko rezultātu.

20.attēls kNN 2

20.attēlā pamainīju kaimiņu skaitu uz 6, un Metriku uz *Manhattan*.

Attēls, kurā ir teksts, ekrānuzņēmums, cipars, fonts

Apraksts ģenerēts automātiski

21.attēls neironu tīkls 2

21.attēlā pamainīju neironus slēptajā slāni uz 100,100, un aktivizāciju uz *Identity*, kā arī nomainīju iterāciju skaitu uz 500.

Attēls, kurā ir teksts, ekrānuzņēmums, cipars, fonts

Apraksts ģenerēts automātiski22.attēla ir redzams, ka nomainīju SVM hipermarametrus: *Kernel* nomainīts uz *Polynomial*, iterāciju limits ir paaugstināts uz 200.

22.attēls SVM 2

Attēls, kurā ir teksts, ekrānuzņēmums, displejs, programmatūra

Apraksts ģenerēts automātiski

23.attēls eksperiments 2

23.attēlā var apskatīt otrā eksperimenta rezultātu. Neironu tīkla algoritms nav īpāši uzlabojies, bet var uzreiz pamanīt, ka kNN un SVM testa rezultāti ir nedaudz uzlabojušies.

Attēls, kurā ir teksts, ekrānuzņēmums, fonts, cipars

Apraksts ģenerēts automātiski

24.attēls kNN 3

24.attēlā redzams, ka kNN algoritma kaimiņu skaitu pazemināju uz 4, un metriku nomainīju uz *Chebyshev*.

Attēls, kurā ir teksts, ekrānuzņēmums, cipars, fonts

Apraksts ģenerēts automātiskiAttēls, kurā ir teksts, ekrānuzņēmums, cipars, fonts

Apraksts ģenerēts automātiski25.attēlā pamainīju neironus slēptajā slāni uz 100,100,100 un aktivizāciju uz *relu*, kā arī nomainīju iterāciju skaitu uz 1000.

26.attēls SVM 3

25.attēls Neironu tīkls 3

26.attēlā redzams, ka nomainīju hiperparametrus: *Cost* uz 2 un *Regression loss epsilon* uz 0.3

# Attēls, kurā ir teksts, ekrānuzņēmums, displejs, programmatūra Apraksts ģenerēts automātiski3.daļas secinājumi

27.attēls eksperiments 3

27.attēlā redzams trešā eksperimenta rezultāti. KNN algoritma rezultāti ir pasliktinājušies, vislabākie rezultāti priekš kNN algoritma bija otrajā eksperimentā. Neironu tīkla rezultāti ir vislabākie trešajā eksperimentā, kā arī KVM algoritma rezultāti ir bijuši vislabākie trešajā eksperimentā.

# Izmantoto avotu saraksts

1. https://www.kaggle.com/datasets/ulrikthygepedersen/fastfood-nutrition

2. <https://estudijas.rtu.lv/course/view.php?id=252548>