# **I daļa - Datu pirmapstrāde/izpēte**

Datu kopu nosaukums: “**Pima Indians Diabetes Database**”

Autors: **UCI MACHINE LEARNING**

Link: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

Es nolēmu izvēlēties datu kopu ar informāciju par pacientiem, norādot viņu medicīnisko informāciju un diabēta klātbūtni.

Šo datu kopu sākotnēji sagatavoja Diabēta un gremošanas un nieru slimību valsts institūts. Datu kopas mērķis ir diagnostiski prognozēt, vai pacientam ir vai nav diabēts, pamatojoties uz konkrētiem datu kopā iekļautajiem diagnostiskajiem mērījumiem. Šo gadījumu atlasei no lielākas datubāzes tika piemēroti vairāki ierobežojumi. Jo īpaši visi pacienti ir vismaz 21 gadu vecas sievietes, kas ir Pima indiāņu izcelsmes

Autors raksta, ka šis datu kopums ir piemērots klasifikācijai (lai norādītu, vai pacientam ir diabēts).

Persona, kas ir saistījusi darbu ar šo aktu, ir veltījusi darbu publiskai lietošanai, atsakoties no visām savām autortiesībām uz darbu visā pasaulē saskaņā ar autortiesību likumu, tostarp visām blakustiesībām, ciktāl to atļauj likums.

Ir iespēja kopēt, pārveidot, izplatīt un izpildīt darbu, pat komerciālos nolūkos, un to visu bez atļaujas pieprasīšanas.

Datu kopā ir 768 ierakstu.

Zemāk ir paradītas visi datu kopas atribūti

A screenshot of a computer

Description automatically generated with medium confidence

att.

A screenshot of a graph

Description automatically generated with low confidence

2 att.

A picture containing text, screenshot, diagram, plot

Description automatically generated

3 att.

A screenshot of a graph

Description automatically generated with low confidence

4 att.

Visi iepriekš minētie attēli vēlreiz parāda visus manas datu kopas atribūtus (attēli ir ņemti no tīmekļa vietnes, kurā es ieguvu datu kopu [Keggle.com](https://www.kaggle.com), un viņiem ir informācija par atribūtiem).

|  |  |  |  |
| --- | --- | --- | --- |
| **Atribūts** | **Paskaidrojums** | **Vērtību tips** | **Diapazons** |
| Pregnancies | Šis atribūts norāda pacienta grūtniecību skaitu. | Skaitlis | 0 - 17 |
| Glucose | Plazmas glikozes koncentrācija 2 stundu laikā pēc glikozes tolerances testa perorāli. | Skaitlis | 0 - 199 |
| BloodPreassure | Asinsspiediens (mm Hg). | Skaitlis | 0 - 122 |
| SkinThickness | Tricepsa ādas krokas biezums (mm). | Skaitlis | 0 - 99 |
| Insulin | 2 stundu seruma insulīns (mu U/ml). | Skaitlis | 0 - 846 |
| BMI | Ķermeņa masas indekss (svars kg/(augums m)^2). | Skaitlis | 0 - 66.1 |
| Age | Norāda pacientu vecumu. | Skaitlis | 21 - 81 |
| Outcome | Klase, lai norādītu, ka pacientam ir diabēts (0 - nav, 1 - ir) | Skaitlis | 0 - 1 |

Šajā datumā vienīgā klasifikācijas iespēja ir diabēta esamība vai neesamība. Līdz ar to ir divas klases. 268 ieraksti ir klasificēti kā diabēta gadījumi (1), pārējie 500 ieraksti ir klasificēti kā bez diabēta (0).

A screenshot of a computer

Description automatically generated

5 att.

5. attēlā redzams datu kopas fails csv formātā, kas atvērts programmā Excel. Attēlā redzami visi kopas atribūti, kā arī ierakstu skaits.

A screenshot of a computer

Description automatically generated with medium confidence

6 att.

Kā redzat, šajā datu kopā nav trūkstošo vērtību, un visas vērtības ir skaitliskas, tāpēc nebija jāveic nekādi papildu pasākumi, tikai jāveic datu kopas apstrāde ar Orange Continuize, lai skaitliskās vērtības iekļautu diapazonā no 0 līdz 1.

A screenshot of a computer

Description automatically generated with low confidence

7 att.

Šādi izskatās dati pēc apstrādes.

Pēc tam es devos uz izkliedes diagrammu(Scatter Plot), lai redzētu, cik lielā mērā objekti ir atdalīti viens no otra. Šis grafiks attēlo vienu atribūtu X asij un otru Y asij. Turklāt, lai informācija būtu ērtāk uztverama, es iekrāsoju laukumus izvēlētās klasifikācijas klases krāsā. Lai nebūtu jātērē daudz laika, meklējot vairākus atribūtus ar labu objektu atdalīšanu, izmantoju funkciju "Find Informative Projections", kas sniedza sarakstu ar visām informatīvajām kombinācijām, no kurām izvēlējos un nofotografēju pirmās divas.

A screenshot of a computer

Description automatically generated with medium confidence

8 att. Vecums – Glikoze kombinācija

A screenshot of a computer screen

Description automatically generated with medium confidence

9 att. Ķermeņa masas indekss – Glikoze kombinācija

Pēc tam es devos uz histogrammu(Bar plot) un izvēlējos 2 grafikus (2 atribūti), kas labi atspoguļo datu kopu.

A red and blue lines

Description automatically generated with low confidence

10 att. Asinsspiediena histogramma

A picture containing text, screenshot, plot, line

Description automatically generated

11 att. Ķermeņa masas indeksa histogramma

Pēc tam dodieties uz sadaļu Izplatīšana (Distribution). Tas var sniegt detalizētu informāciju par pazīmju sadalījumu, tostarp par to, kuras vērtības ir visizplatītākās un cik bieži tās sastopamas. Rīkniņš parādīs histogrammu, vizualizējot pazīmes vērtību sadalījumu un izceļot bieži sastopamās vērtības.

A screenshot of a computer

Description automatically generated with medium confidence

12 att.

A screenshot of a graph

Description automatically generated with medium confidence

13 att.

Nākamais uzdevums ir aprēķināt statistikas datus. Orange vidē ir statistikas rīks(feauture statistics). Ar to es varu atrast uzdevuma mediānu un dispersiju.

A screenshot of a computer

Description automatically generated with medium confidence

14 att.

A screenshot of a computer

Description automatically generated with medium confidence

15 att.

## **Secinājumi**

**Vai klases datu kopā ir līdzsvarotas, vai dominē viena klase (vai vairākas klases)?**

Būtībā manā datu kopā ir tikai 2 klases, tomēr viena no tām dominē, proti, cilvēki bez diabēta. 500 pret 268.

**Vai datu vizuālais atspoguļojums ļauj redzēt datu struktūru?**

Izkliedes diagrammā īsti neatspoguļo struktūru. Tā kā daži viena objekta punkti atrodas starp citiem objektiem, ir grūti noteikt struktūru. No manis izvēlētajām histogrammām (10. un 11. attēls) var redzēt datu grupēšanu (viens klases tips atrodas blakus citiem klases objektiem).

Īpaši KMI histogramma. No tās var redzēt, ka cilvēkiem bez KMI vidēji ir zemāks KMI nekā diabēta slimniekiem.

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

Aplūkojot divas izkliedes diagrammas. Pirmais grafiks (8. attēls) sniedz vislabāko grupējumu saskaņā ar citiem grafikiem. Pārējos grafikos punkti ir pārāk tuvu viens otram, tāpēc ir grūti precīzi noteikt klasi. Arī 8. attēla grafikā daži punkti pārklājas, bet šeit lielākā daļa objektu ir atdalīti un ir iespējams skaidri nodalīt grupas.

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

Grupēšana pēc atribūtiem vecums un glikoze diezgan labi atdala objektus vienu no otra. Tas nozīmē, ka tie ir normālā attālumā, lai gan joprojām ir situācijas, kad daži objekti "pārklājas" viens otram, bet šajā grupēšanā tas ir vismazāk ticams.

**Secinājumi, kas izriet no statistisko rādītāju** **analīzes.**

Tā kā esmu normalizējis ievades datus, visu atribūtu maksimālās un minimālās vērtības ir vienādas (min. - 0, maks. - 1).

Maksimālā dispersija (Dispersion) ir novērota insulīna parametram (1,44), bet mazākā - ĶMI parametram (0,25). Tas liecina, ka insulīna raksturlieluma vērtība variē visvairāk no visiem raksturlielumiem. Savukārt ĶMI atribūta vērtība svārstās vismazāk.

Kā redzams, lielākajai daļai pētījumā iesaistīto cilvēku nebija diabēta.

Asinsspiediena sadalījums ir ļoti līdzīgs normālajam sadalījumam.

Datu kopā bieži sastopama insulīna atribūta 0 vērtība. Es izlasīju vairākus komentārus un nonācu pie secinājuma, ka 0 nenozīmē, ka insulīna līmenis patiešām ir 0 (tas tā nevar būt), bet ka tas ir ļoti zems.

Arī bieža ādas tievuma vērtība arī ir 0. Tas var liecināt par to, ka cilvēkiem nav veikts ādas biezuma tests un no tā rodas 0 vērtība.

**II daļa – Nepārraudzītā mašīnmācīšanās**

A screenshot of a computer

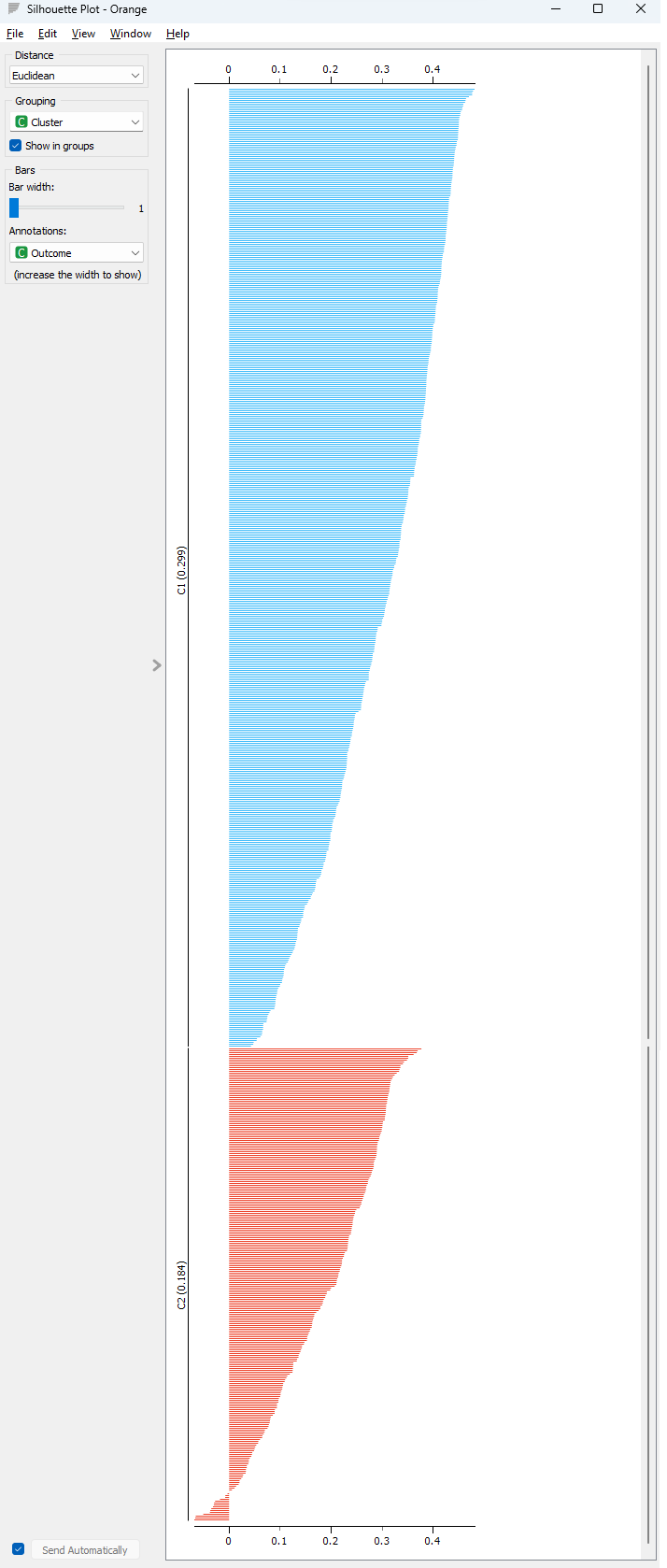
Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with low confidence

# **III daļa – Pārraudzītā mašīnmācīšanās**

|  |  |  |
| --- | --- | --- |
|  | 0(nav diab.) | 1(ir diab.) |
| Sakumā | 500 | 268 |
| Mācīšanai(80%) | 400 | 214 |
| Testam(20%) | 100 | 54 |

Prettolerantai mašīnmācīšanai nolēmu izvēlēties 2 algoritmus (kNN un Gradient Boosting), un 3 obligātais algoritms bija neironu tīkli.

A screenshot of a computer

Description automatically generated with medium confidence

## **kNN algoritms**

k-tuvākie kaimiņi (kNN) ir mašīnmācīšanās algoritms, ko izmanto, lai klasificētu datus, pamatojoties uz to tuvumu jau zināmiem datiem. Vienkārši sakot, kNN algoritms atrod k tuvāko kaimiņu jauniem datiem un klasificē tos atbilstoši klasei, kas ir visbiežāk sastopama starp šiem k tuvākajiem kaimiņiem.

Lai klasificētu datus, izmantojot kNN algoritmu, ir jānosaka, kā katram jaunam novērojumam tiks noteikti tuvākie kaimiņi. To parasti dara, mērot attālumu, piemēram, izmantojot Eiklīda attālumu.

Kad katram jaunam novērojumam ir atrasti k tuvākie kaimiņi, kNN algoritms izmanto balsošanas metodi, lai noteiktu klasi, kurai piederēs katrs jaunais novērojums. Tas nozīmē, ka algoritms saskaita katrai klasei piederošo kaimiņu skaitu, un klase ar lielāko balsu skaitu tiek piešķirta kā jaunā novērojuma klase.

Es nolēmu izmantot šo algoritmu, jo tas šķita pietiekami vienkāršs, un mēs šo algoritmu esam aplūkojuši arī lekcijās.

Hiperparametri kNN ietver:

Parametrs k ir tuvāko kaimiņu skaits, kas tiek izmantots, lai klasificētu vai regresētu jaunos datus. Liela k

Attāluma metrika - metode attāluma mērīšanai starp novērojumiem. Visizplatītākā metode ir Eiklīda attālums, bet var izmantot arī citas metrikas, piemēram, Manhetenas.

Svari - nosaka, kā svērt attālumu starp kaimiņiem. Svērtie svari ņem vērā attālumu līdz katram kaimiņam.

### **Tests**

A screenshot of a computer

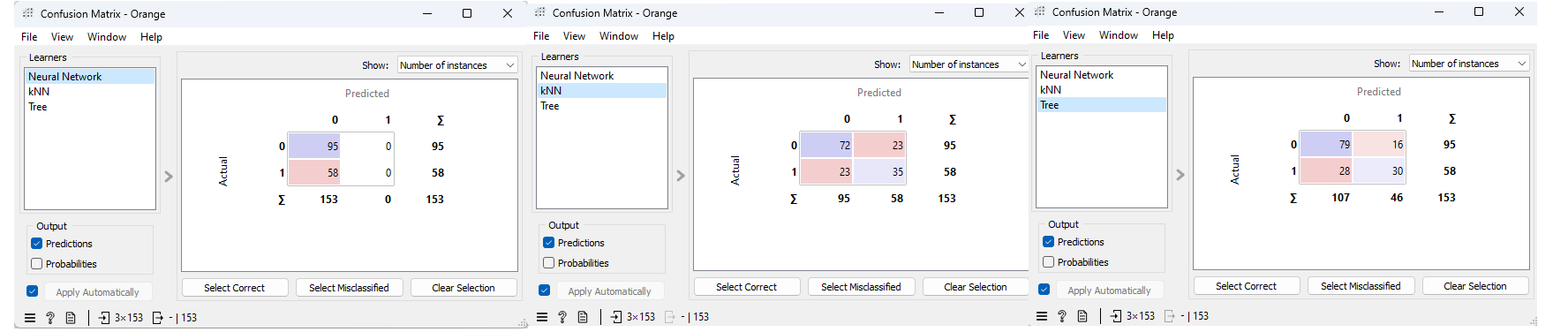
Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated



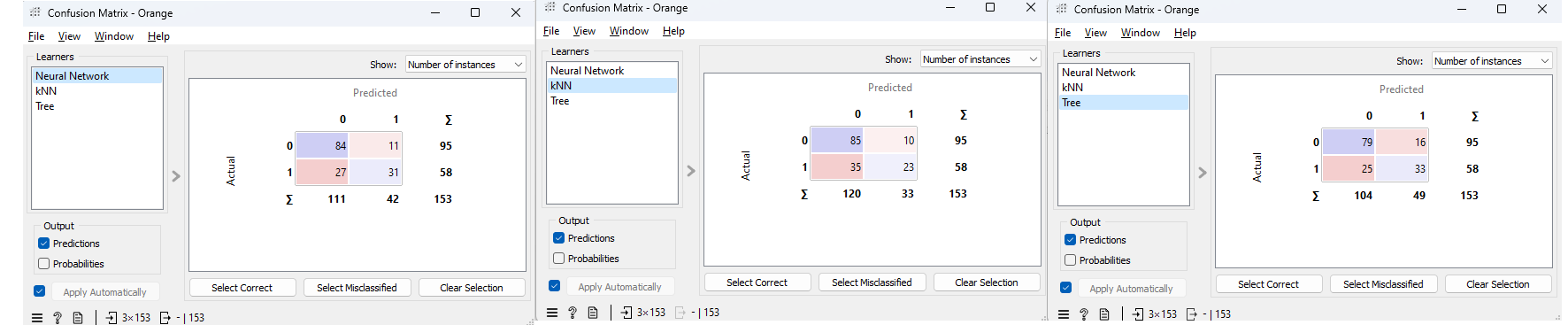
### **Tests**

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

### **Tests**

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

