Vilniaus universitetas
Matematikos ir informatikos fakultetas
Duomenų mokslo ir skaitmeninių technologijų institutas

# Duomenų klasifikavimas

prof. dr. Olga Kurasova
Olga.Kurasova@mif.vu.lt

# Duomenų klasifikavimas

- Pagal turimus duomenis, kurių klasės yra žinomos, **reikia sukurti mechanizmą** (klasifikatorių), kuris gebėtų priskirti klases duomenims, kuriems jos nėra žinomos.
- Duomenims klasifikuoti taikomi **įvairūs klasifikavimo metodai**: Naive Bayes, k artimiausių kaimynų, atraminių vektorių, klasifikavimo medžių ir kt.
- Dirbtiniai neuroniniai tinklai taip pat yra plačiai **naudojami duomenims klasifikuoti**.
- Net **vienas neuronas geba** spręsti nesudėtingus klasifikavimo uždavinius.

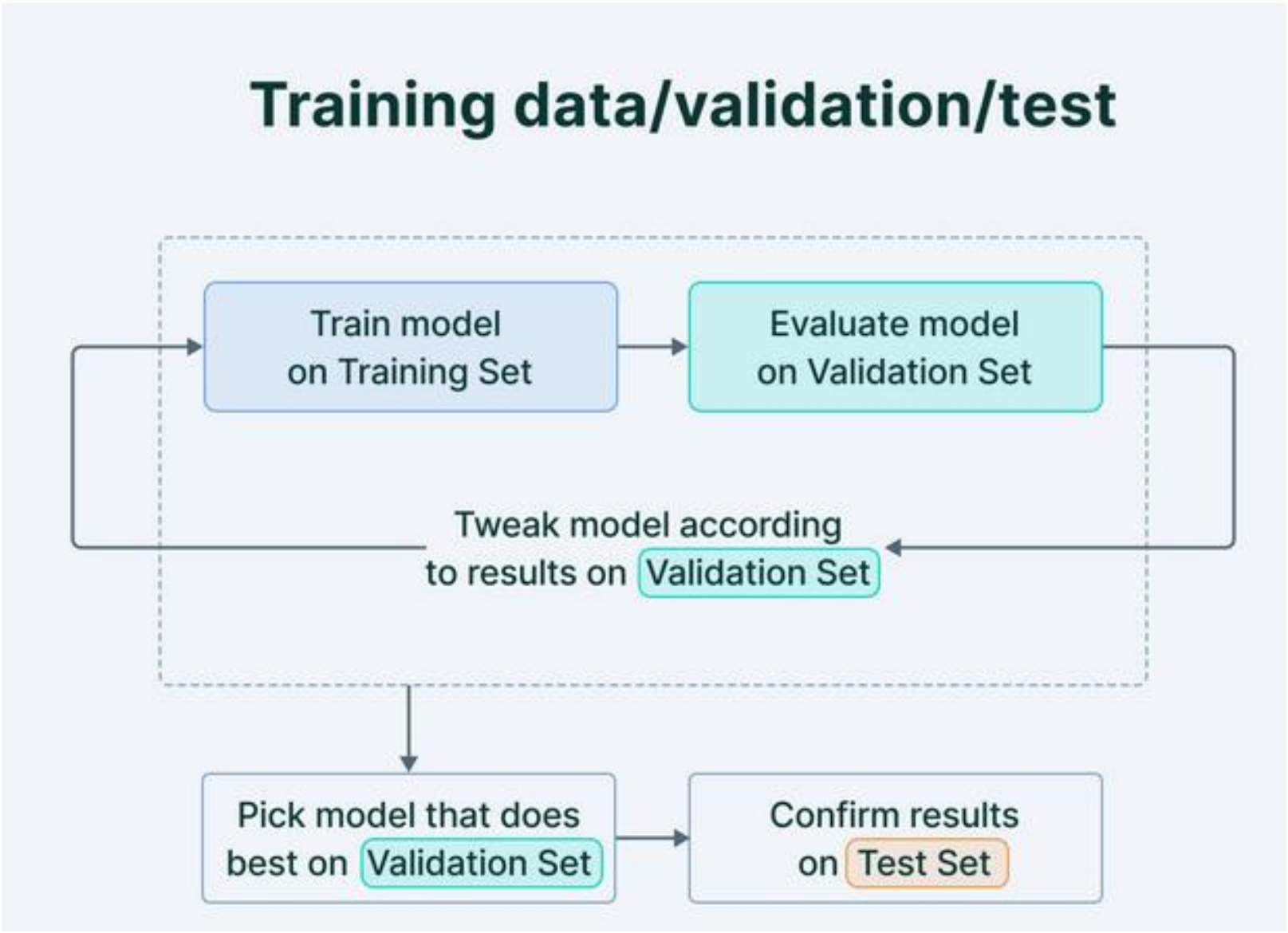
# Duomenų klasifikavimas

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | **Klasė** |
|---|---|---|---|---|---|
| $X_1$ | 85 | 0,001 | 24,1 | 1025 | sveikas |
| $X_2$ | 77 | 0,002 | 21,3 | 2036 | sveikas |
| $X_3$ | 68 | 0,015 | 35,8 | 1059 | sveikas |
| … | … | … | … | … | … |
| $X_{101}$ | 101 | 0,001 | 22,4 | 3011 | serga |
| $X_{102}$ | 95 | 0,001 | 28,0 | 2645 | serga |
| … | … | … | … | … | … |
| $X_{201}$ | 86 | 0,002 | 30,1 | 2987 | **???** |
| $X_{202}$ | 72 | 0,010 | 19,5 | 1259 | **???** |

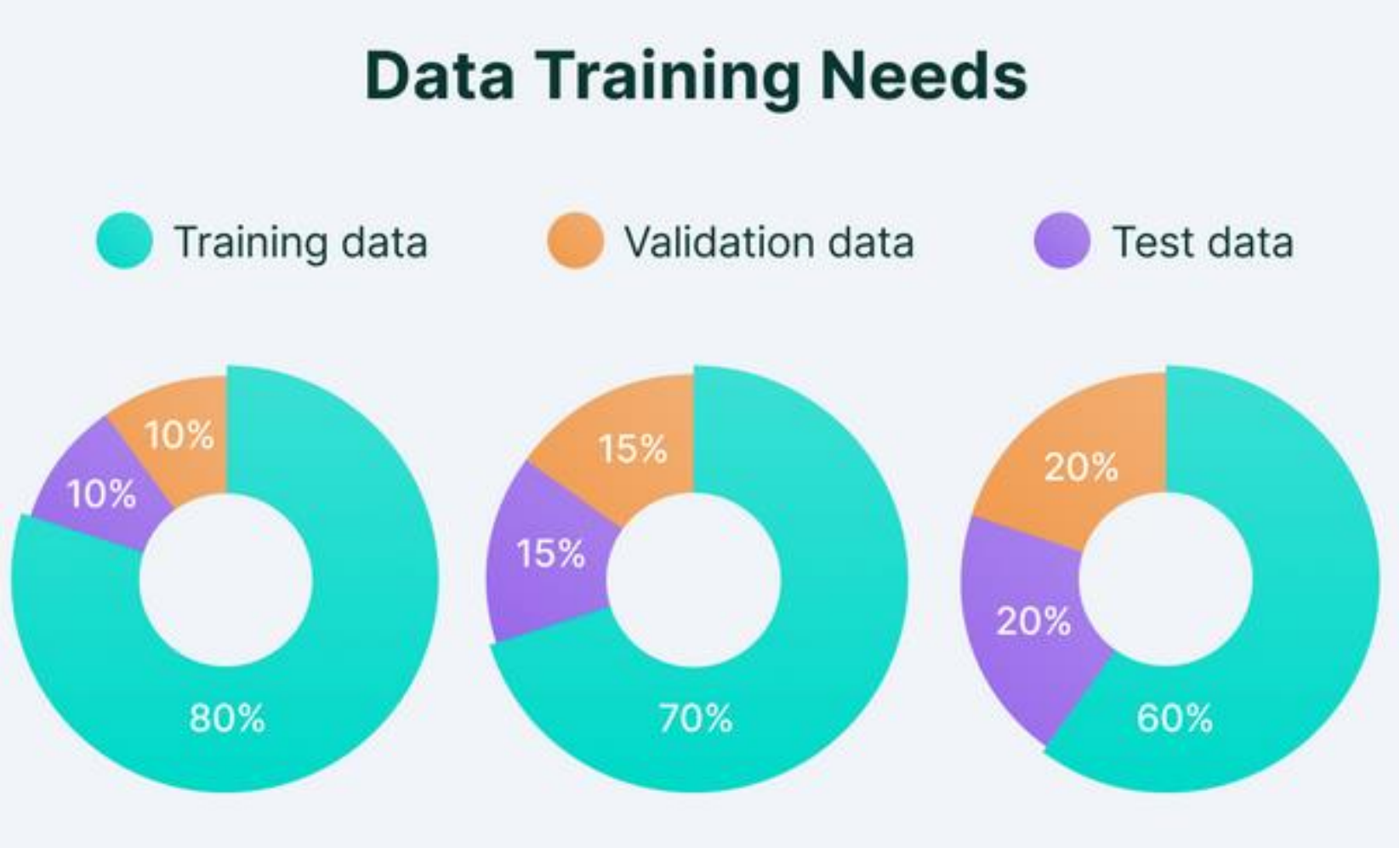
# Duomenys klasifikavimui

Sprendžiant **klasifikavimo** uždavinius išskiriami **trijų tipų duomenys**:

- **mokymo duomenys** naudojami klasifikatoriui sukurti,
- **validavimo duomenys** naudojami patikrinti išmokymo klasifikuoti lygį, rezultatai gali būti naudojami hiperparametrų parinkimui ar mokymo stabdymui.
- **testavimo duomenys** naudojami pabaigoje patikrinti išmokymo klasifikuoti lygį.
- **nauji duomenys**, kurių klasės nėra žinomas, bet taikant sukurtą klasifikatorių jos yra nustatomos.




https://www.v7labs.com/blog/train-validation-test-set




https://www.v7labs.com/blog/train-validation-test-set

# Klasifikavimo tikslumo matai

- Klasifikatorius **turi būti išmokytas** taip, kad gebėtų gerai klasifikuoti duomenys, kurių klasės nėra žinomos.
- Vadinasi reikia turėti to **išmokymo įvertinimo matus**.
- Klasifikavimo tikslumui nustatyti dažniausiai vertinami šie matai:
  - **bendras klasifikavimo tikslumas** (*accuracy*),
  - **atkūrimas** (*recall*), **jautrumas** (*sensitivity*),
  - **preciziškumas** (*precision*),
  - **specifiškumas** (*specificity*),
  - **F1 matas**.

# Klasifikavimo tikslumas

Apibrėžkime pagrindines sąvokas:

- **tikrai teigiamas** (TT) (angl. *true positive*) – objektas $X_i$ priskirtas klasei $C_j$, ir iš tiesų jis jai priklauso,
- **tikrai neigiamas** (TN) (angl. *true negative*) – objektas $X_i$ nepriskirtas klasei $C_j$, ir iš tiesų jis jai nepriklauso;
- **klaidingai teigiamas** (KT) (angl. *false positive*) – objektas $X_i$ priskirtas klasei $C_j$, bet iš tiesų jis jai nepriklauso;
- **klaidingai neigiamas** (KN) (angl. *false negative*) – objektas $X_i$ nepriskirtas klasei $C_j$, bet iš tiesų jis jai priklauso.

# Klasifikavimo matrica

Apskaičiavus šiuos įverčius, sudaroma **klasifikavimo matrica** (angl. *classification* ar *confusion matrix*)

| | | gauta klasė | |
|---|---|---|---|
| | | $C_1$ (teigiama) | $C_2$ (neigiama) |
| tikroji klasė | $C_1$ (teigiama) | tikrai teigiamas (TT) | klaidingai neigiamas (KN) |
| | $C_2$ (neigiama) | klaidingai teigiamas (KT) | tikrai neigiamas (TN) |

# Klasifikavimo matrica

Apskaičiavus šiuos įverčius, sudaroma **klasifikavimo matrica** (angl. *classification* ar *confusion matrix*) (žymėjimai anglų k.)

| | | Predicted class | |
|---|---|---|---|
| | | $C_1$ (positive) | $C_2$ (negative) |
| True class | $C_1$(positive) | true positive (TP) | false negative (FN) |
| | $C_2$(negative) | false positive (FP) | true negative (TN) |

# Klasifikavimo matricos pavyzdys

| | | gauta klasė | |
|---|---|---|---|
| | | sveikas | serga |
| tikroji klasė | sveikas | 90 | 10 |
| | serga | 5 | 80 |

# Klasifikavimo matai

**Klasifikavimo matų** reikšmės yra apskaičiuojamos pagal šias formules:

- bendras tikslumas $= \frac{TP+TN}{TP+TN+FP+FN}$
- specifiškumas $= \frac{TN}{TN+FP}$
- atkūrimas(jautrumas) $= \frac{TP}{TP+FN}$
- precižiškumas $= \frac{TP}{TP+FP}$
- $F1 = \frac{2 \times \text{precižiškumas} \times \text{atkūrimas}}{\text{precižiškumas} + \text{atkūrimas}}$

# Kryžminė patikra

- Klasifikavimo tikslumas gali priklausyti nuo to, kaip visa **duomenų aibė padalinta į mokymo ir testavimo** aibes.
- Todėl tikslinga **klasifikavimą atlikti keliems skirtingiems** tos pačios duomenų aibės mokymo ir testavimo rinkiniams ir **įvertinti vidutinį** klasifikavimo tikslumą.
- Tam tikslui dažnai naudojamas **kryžminės patikros metodas** (angl. *cross validation*).

# Kryžminė patikra

- Kryžminės patikros metu duomenų aibė yra **suskaidoma** į *q* **nesusikertančių blokų** (angl. *folds*).
- Klasifikavimo algoritmas yra **apmokomas** naudojant *q* – 1 bloko duomenis, o likusi duomenų dalis yra panaudojama algoritmui **testuoti**.
- **Fiksuojamos** klasifikavimo matų reikšmės.
- Ši procedūra atliekama *q* **kartų**, mokymui imant vis kitus *q* – 1 blokus, pabaigoje randamos klasifikavimo matų **vidutinės reikšmės**. Pagal jas vertinamas **sukurto klasifikatoriaus tikslumas**.

# Kai daug duomenų

- **Kryžminė patikra** yra daug laiko reikalaujantis procesas, kai analizuojami dideli duomenų kiekiai
- Pavyzdžiui, giliojo mokymosi procese turi būti naudojama **daug duomenų**.
- Tokiu atveju kryžminė patikra neatliekama. Po **mokymo etapo** iš karto **seka testavimas**.
- Kai testavimui naudojama daug duomenų, **daroma prielaida**, kad yra maža tikimybė, jog į testavimo duomenų rinkinį pateks tik lengvai (arba sunkiai) klasifikuojami duomenys.