Vilniaus
universitetas

Privačios informacijos išsaugojimas taikant dirbtinio intelekto technologijas

Paulius Milmantas

Įvadas

Mašininis mokymas yra dirbtinio intelekto sritis, kuri pasitelkia statistinius algoritmus, kad apibrėžtų duomenų generavimo mechanizmą, ar egzistuojančius sąryšius, priklausomybes.

Pagrindinė problema

Turint sukurtą modelį, neturi būti galima atgaminti duomenų, pagal kuriuos jis buvo mokomas, bei negali būti identifikuoti asmenys. [1]

Problemos pavyzdys: teksto atpažinimo modelis.

Gali būti atskleisti privatūs duomenys.

Vilniaus universitetas

Modelių duomenų lyginimas

$$atvirumas(s[r])_\theta = \log_2|r| - \log_2 rangas_\theta(s[r])$$

Naudojama teorijoje, dėl sunkiai apskaičiuojamo rango. [2]

s – duomenų rinkinys.
r ∈ R, parenkamas atsitiktinai.

$$atvirumas(s[r])_\theta = -\log_2 \int_0^{Px_\theta(s[r])} \rho(x)dx$$

Dėl grafinės interpretacijos naudojama praktikoje. [2]

Px – logaritminis entropijos matas.
s[r] entropija yra ρ(.) pasiskirstymo distribucijos.

Pasiūlyta tyrimo metodika

$$DMDK = \sum_{n=0}^{m} \left(\sum_{k=0}^{h} \left(max_{\epsilon} \left(\left(|\epsilon| + D_{eilut.:n,stulp.:k} \right) : \epsilon \in R \right) \right) / h \right) / m$$

$\mathbf{DMDK}$ – Didžiausias galimas duomenų nuokrypis.

$\boldsymbol{D_{eilut.:n,stulp.:k}}$ - duomenys n eilutėje ir k stulpelyje.

ϵ - ieškomas didžiausias galimas kintamasis, su kuriuo modelis nepakeičia išvesties rezultatų.

m - duomenų eilučių skaičius.

h - parametrų skaičius (stulpeliai).

Metrikos validacija (1)

Pagal KMI ir gimdymų skaičių prognozuojama, ar moteris serga cukriniu diabetu.

- Kai modelio DMDK yra mažas – gauti duomenys eksperimente buvo artimi pradiniams duomenims.
- Kai modelio DMDK yra didelis – nepavyko gauti panašių duomenų.

1 lentelė. Sugalvotos reikšmės modelio tikrinimui.

KMI	Gimdymų skaičius
25	3
25	2

2 lentelė. Tikri modelio duomenys.

KMI	Gimdymų skaičius
26	1
22	2

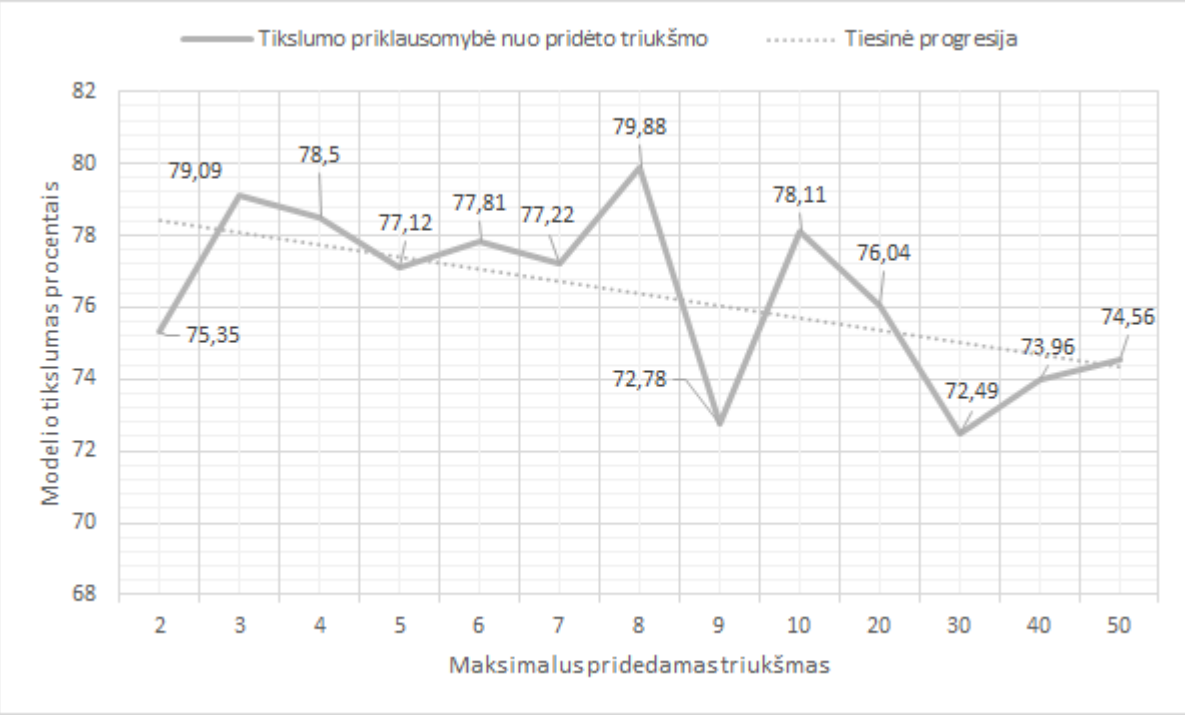
Metrikos validacija (2)

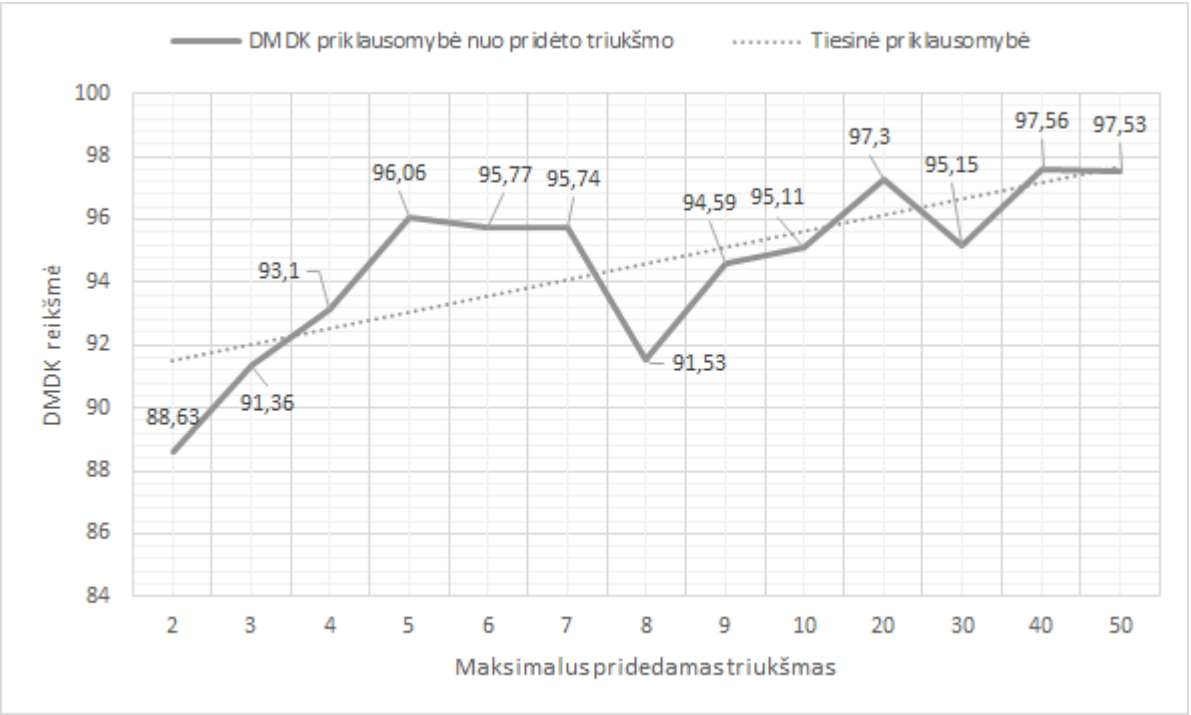
- Eilučių duomenys buvo padauginti iš N. Sukūrus kelis naujus modelius su skirtingais N, DMDK reikšmė išlieka panaši.

- Tarkime, kad pirmas modelis turi viena parametrą - KMI. Pagal šį parametrą, modelis prognozuoja, ar žmogus serga cukriniu diabetu ar ne. Modelio tikslumas yra 54%, jis visą laiką prognozuoja, kad žmogus serga cukriniu diabetu. DMDK reikšmė artėja link begalybės.

Vilniaus universitetas

Pridėto triukšmo tyrimas

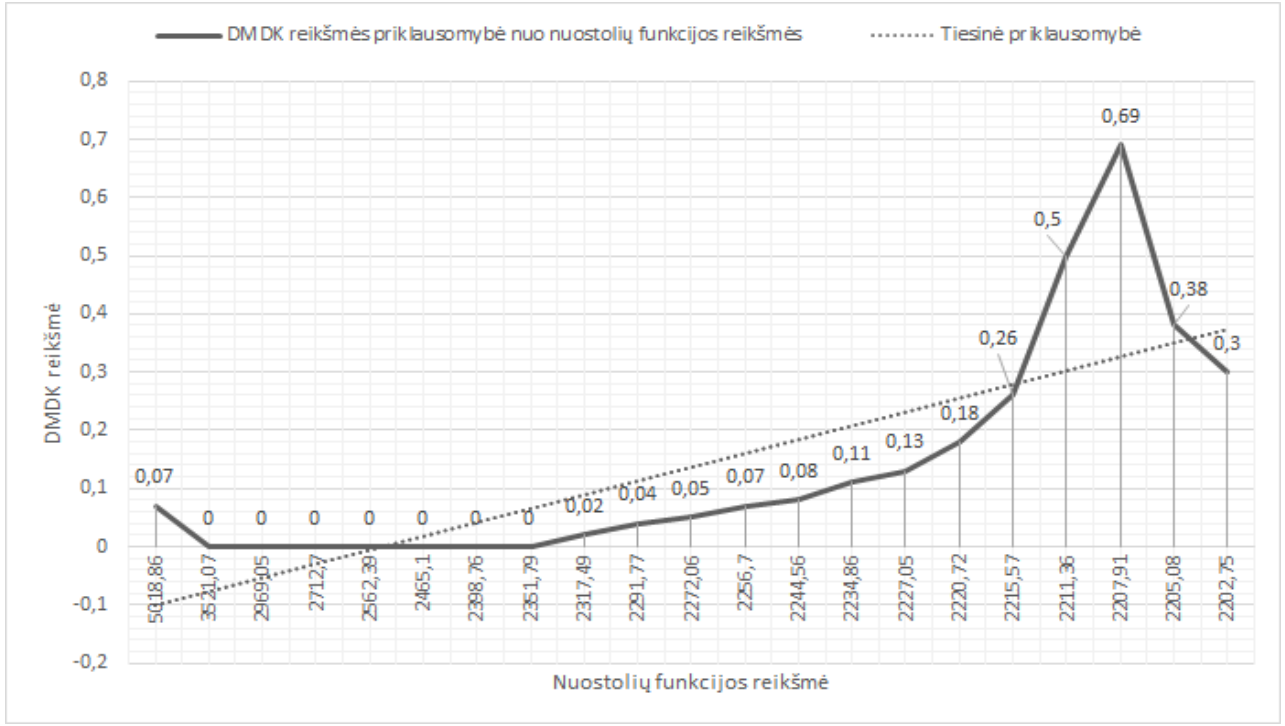


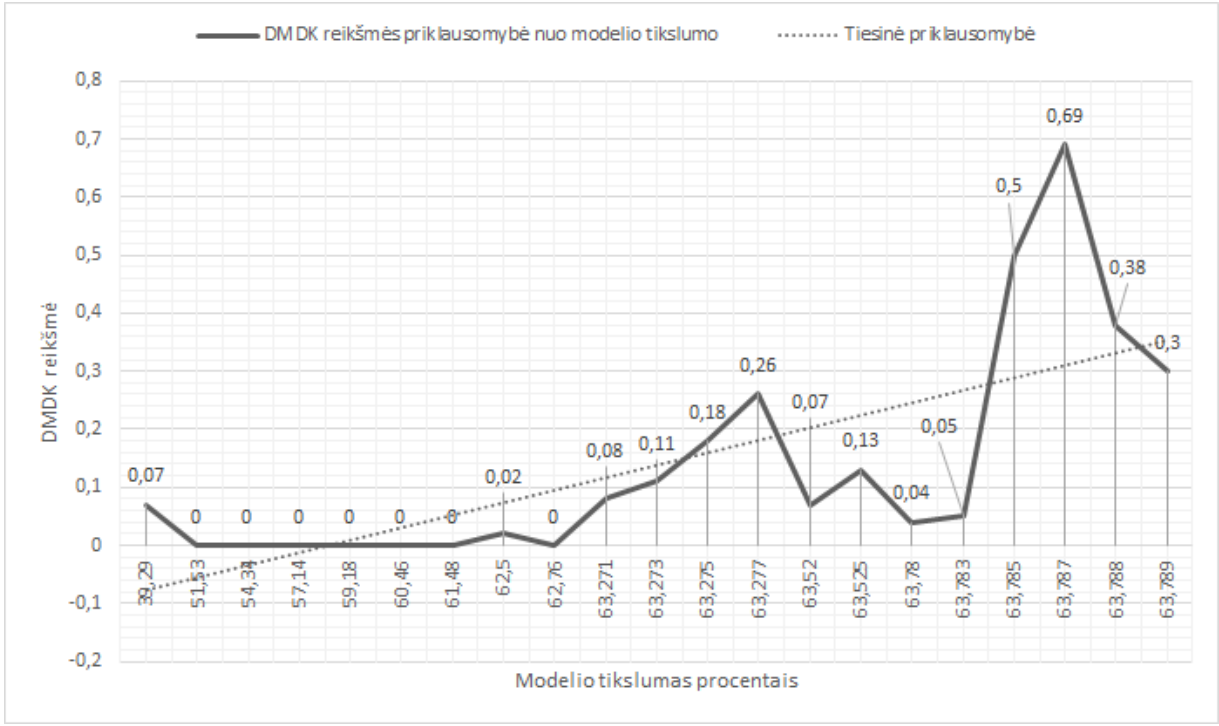


Pagal Spearman koreliacijos tikrinimo metodą, koreliacija yra 0.6978022, **vidutinio stiprumo statistinis ryšys**, p-reikšmė 3.661e-08 < 0.05 - **statistiškai reikšminga**.

Vilniaus universitetas

Pallier tyrimas

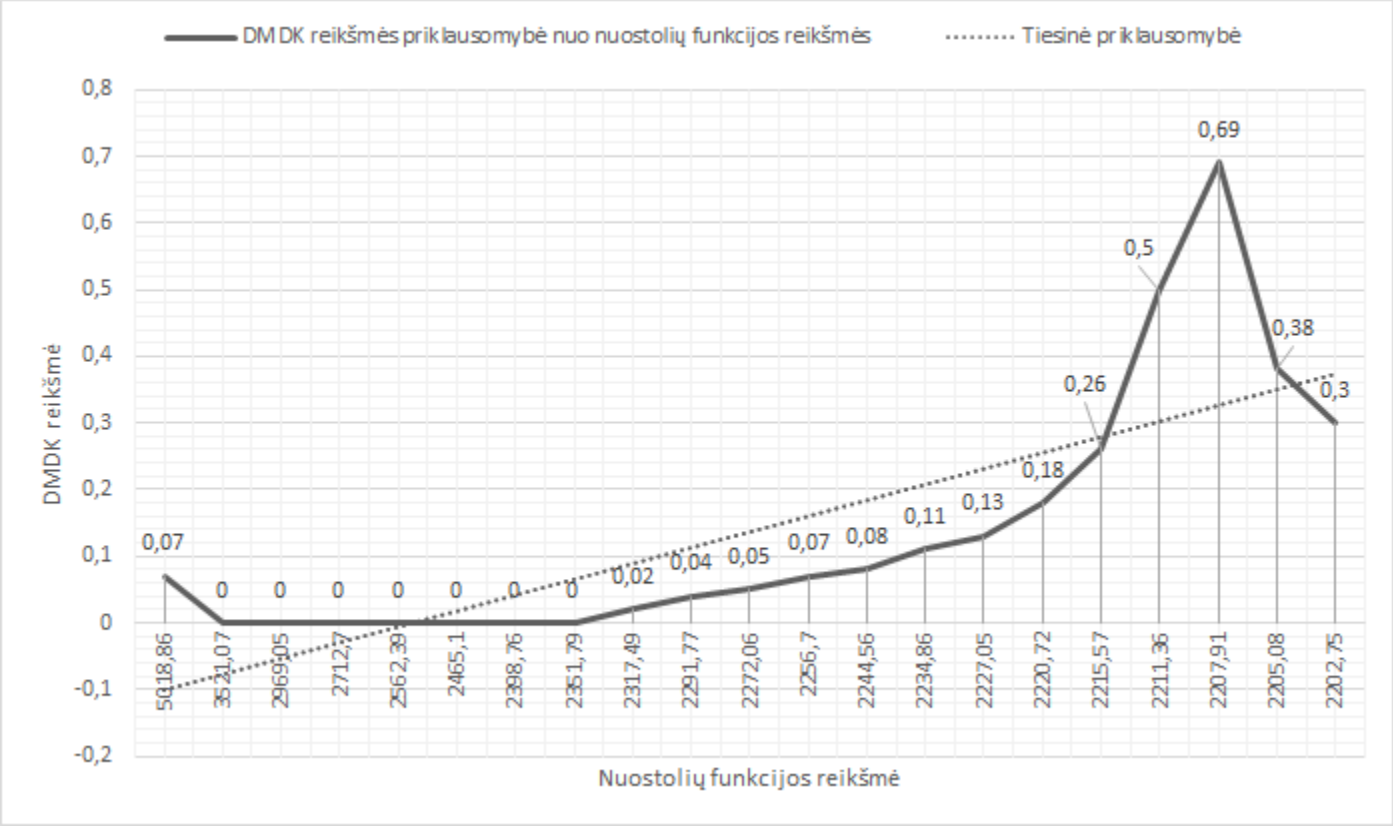




Pagal Spearman koreliacijos tikrinimo metodą, koreliacija yra 0.8969792, **stiprus statistinis ryšys**, p-reikšmė 0.01033 < 0.05 - **statistiškai reikšminga**.

Vilniaus universitetas

PyTorch neuroninio tinklo tyrimas



Išvados

- Esant aukštam modelio tikslumui, rekomenduojama naudoti homomorfinį šifravimą. Esant mažesniam, nei 70% tikslumui, rekomenduojama naudoti PyTorch neuroninius tinklus.
- Esant didesniam modelio parametrų skaičiui, PyTorch neuroniniai tinklai labiau prisimena pradinius mokymosi duomenis ir juos galima lengviau atskleisti.
- Naudojant neuroninius tinklus be homomorfinio šifravimo ir modelio tikslumui esant daugiau nei 80%, rekomenduojama pridėti triukšmą prie pradinių modelio duomenų.
- Pradinių duomenų kiekis neturi įtakos modelio duomenų saugumui.

Šaltiniai

[1] Patricia Thaine. Perfectly privacy-preserving ai, 01 2020

[2] Nicholas Carlini, Chang Liu, Ulfar Erlingsson, Jernej Kos, and Dawn Song. The secretsharer: Evaluating and testing unintended memorization in neural networks, 2019.