

# Privačios informacijos išsaugojimas taikant dirbtinio intelekto technologijas

Paulius Milmantas

Darbo vadovas: asist. dr. Linas Petkevičius

Vilniaus Universitetas  
Matematikos ir informatikos fakultetas

Bakalauro darbo gynimas

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

**Darbo tikslas** - ištirti ir palyginti privatumą saugančius dirbtinio intelekto algoritmus pagal jų saugumą, našumą ir panaudojamumą, bei pateikti rekomendacijas.

Darbo tikslui įgyvendinti, iškelti šie **uždaviniai**:

- 1 Išanalizuoti esamus algoritmus pagal jų saugumą ir panaudojamumą.
- 2 Identifikuoti kriterijus, kurių pagalba galima įvertinti privatumo išsaugojimą, bei palyginti algoritmus tarpusavyje.
- 3 Ištirti kurie algoritmai yra realizuoti ir realizuoti dalį algoritmų, kurie nėra atvirai prieinami.
- 4 Palyginti algoritmus pagal našumą ir pateikti rekomendacijas.

- 1 Turint sukurtą modelį, neturi būti galima atgaminti duomenų, pagal kuriuos jis buvo mokomas, bei negali būti identifikuoti asmenys [1].
- 2 Trečios šalys neturi matyti įvedamų duomenų. Tai gali būti tinklo saugumo spragos, duomenų surinkimo aplikacijų spragos ir t.t. . .
- 3 Modelio išvesties neturi matyti asmenys, kuriems šie duomenys nepriklauso.
- 4 Sukurtas modelis negali būti niekieno pasisavintas.

$$atvirumas(s[r])_{\theta} = \log_2 |r| - \log_2 rangas_{\theta}(s[r]) \quad (1)$$

Naudojama teorijoje, dėl sunkiai apskaičiuojamo rango [2].

**s** - duomenų rinkinys.

**r**  $\in \mathcal{N}$ , parenkamas atsitiktinai.

$$atvirumas(s[r])_{\theta} = -\log_2 \int_0^{P_x(s[r])} \rho(x) dx \quad (2)$$

Dėl grafinės interpretacijos naudojama praktikoje.

**P<sub>x</sub>** - logaritminis entropijos matas.

**s[r]** - entropija yra  $\rho(.)$  pasiskirstymo distribucijos.

$$DMDK = \sum_{n=0}^m \left( \sum_{k=0}^h (\max_{\epsilon} (|\epsilon| + D_{n,k}) : \epsilon \in R, \text{modelis}(|\epsilon| + D_{n,k}) = \text{modelis}(D_{n,k})) / h \right) / m \quad (3)$$

**DMDK** - Didžiausias modelio duomenų nuokrypis.

**D<sub>eilut:n, stulp:k</sub>** - 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).

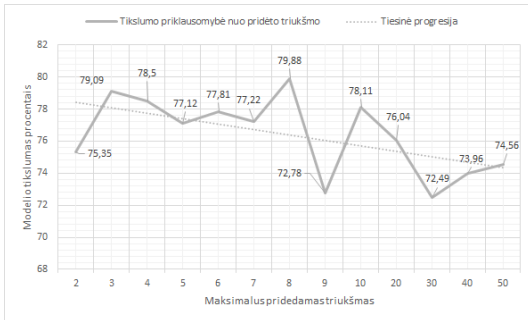
Pagal KMI ir gimdymų skaičių prognozuojama, ar moteris serga cukriniu diabetu.

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

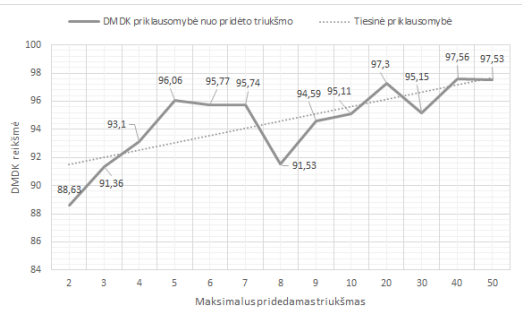


- Eilučių duomenys buvo padauginti iš N. Sukūrus kelis naujus modelius su skirtingais N, DMDK reikšmė išlieka panaši.
- Tarkime, kad pirmas modelis turi vieną parametą - KMI. Pagal šį parametą, modelis prognozuoja, ar žmogus serga cukriniu diabetu ar ne. Modelio tikslumas yra 54%, jis visą laiką prognozuoja, kad žmogus serga cukriniu diabetu. DMDK reikšmė artėja link begalybės.

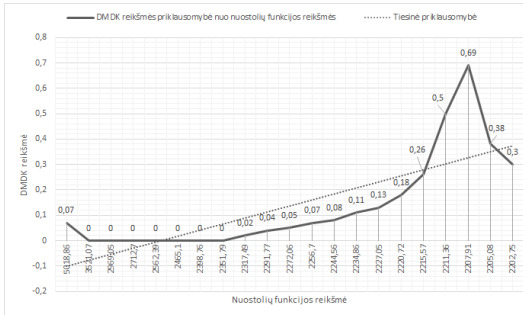
# Pridėto triukšmo tyrimas



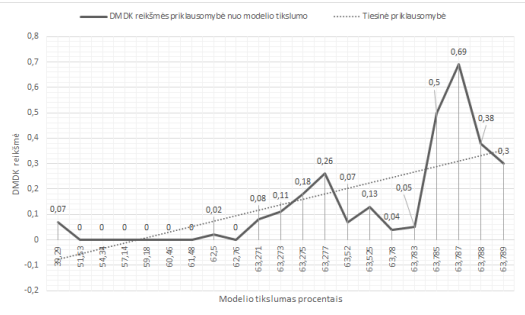
(a) Tikslumo tyrimas



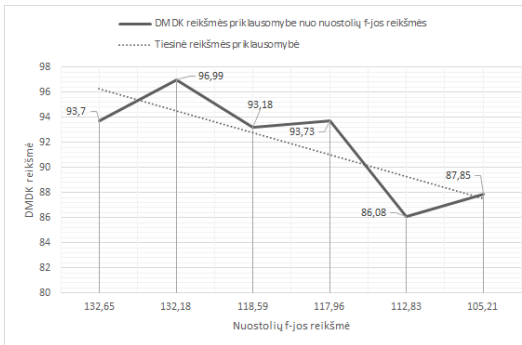
(b) DMDK tyrimas



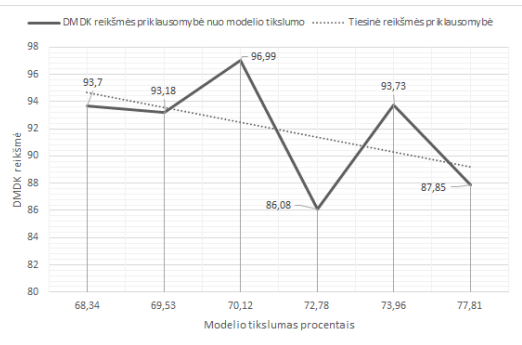
(c) DMDK tyrimas



(d) Tikslumo tyrimas



(e) Nuostolių f-jos priklausomybės tyrimas



(f) Tikslumo tyrimas

Baigiamojo darbo rezultatai buvo pristatyti nacionalinėje konferencijoje, o pranešimo tezės publikuotos konferencijos leidinyje:

**Paulius Milmantas** (2021) *Privačios informacijos išsaugojimas taikant dirbtinio intelekto technologijas*, Vilnius University Open Series, pp. 71-76. doi: 10.15388/LMITT.2021.8.

- Esant aukštam modelio tikslumui, rekomenduojama naudoti homomorfinį šifravimą. Esant mažesniui, nei 70% tikslumui, kai modelis priima  $< 20$  parametrų, rekomenduojama naudoti PyTorch karkaso neuroninius tinklus.
- Esant didesniui modelio parametrų skaičiui, PyTorch karkaso neuroniniai tinklai labiau prisimena pradinį mokymosi duomenį 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.

**[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.