Vilniaus universitetas

Duomenų Mokslo ir Skaitmeninių Technologijų Institutas

Informatikos inžinerijos krypties doktorantų konferencija

Veiklos ataskaita už 2023 m. kovo 22d. – 2023 m. rugsėjo 26d.

Dalia BRESKUVIENĖ– Informatikos inžinerija T 007 doktorantė

Darbo vadovas – prof. habil. dr. Gintautas DZEMYDA

Doktorantūros pradžios ir pabaigos metai: 2021.12.01 – 2025.11.30

Disertacijos tema, tyrimo objektai ir tikslas

- **Preliminari disertacijos tema:**

Klasifikatoriaus (nesubalansuotos) mokymo aibės optimizavimas, siekiant geresnės klasifikavimo kokybės.

Patikslinta preliminari disertacijos tema:

Adaptyvieji mašininio mokymosi metodai, skirti nesubalansuotam duomenų su koncepcijos poslinkiu klasifikavimui

(Adaptive machine learning techniques for classifying imbalanced data with concept drift)

- **Tyrimo objektai:**

Nesubalansuotų duomenų su koncepcijos poslinkiu klasifikavimas finansinių nusikaltimų identifikavimui.

- **Tikslas:**

Sukurti arba patobulinti jau egzistuojantį mašininio mokymosi algoritmą, siekiant pagerinti klasifikavimo rezultatus nesubalansuotiems duomenims su koncepcijos poslinkiu.

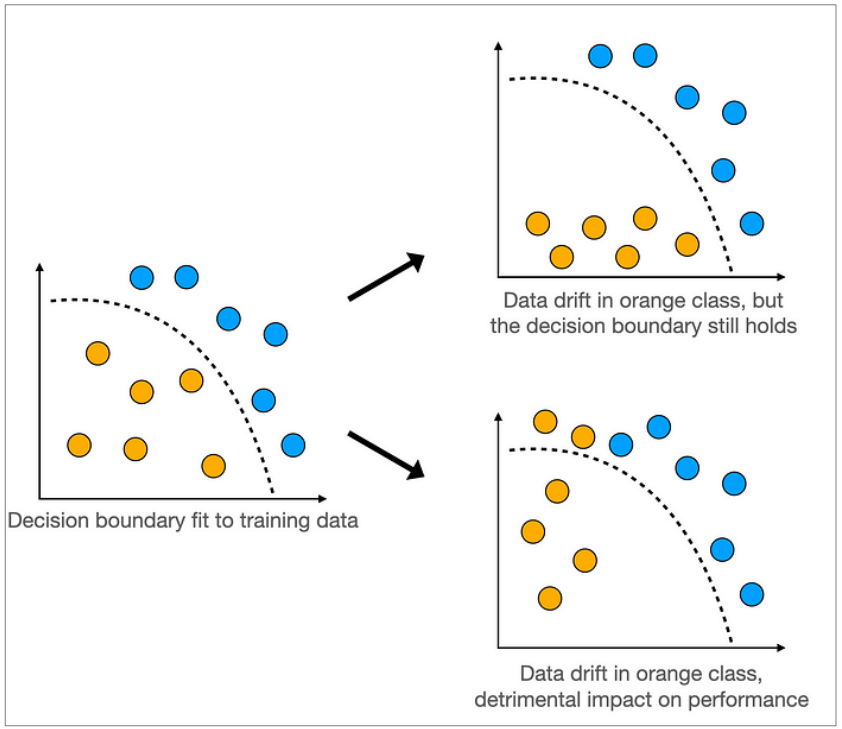
Kas yra koncepcijos poslinkis?

Koncepcijos poslinkis gali atsirasti keliose srityse. Jis gali paveikti ...

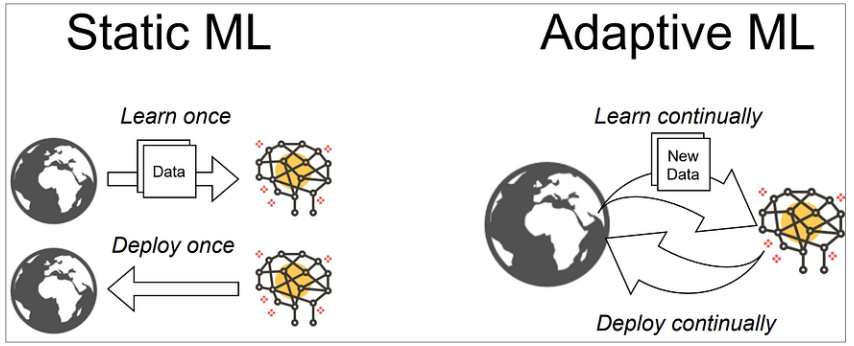
... pagrindinį duomenų pasiskirstymą
- ...vartotojų elgesio pasikeitimas.

... etikečių (label) pasiskirstymą
- ... apnuodijimo išpuoliai. Šių atakų metu įsilaužėlis gali pasiekti sistemos mokymo aibę ir įterpti arba pakeisti vieną ar kelis stebėjimus, kad paveiktų modelio mokymo procesą..

... posteriorinį pasiskirstymą
- kai pasikeičia sukčiavimo scenarijus.



https://towardsdatascience.com/dont-let-your-model-s-quality-drift-away-53d2f7899c09



https://medium.com/continual-ai/towards-adaptive-ai-with-continual-learning-f493fd0d698

Vilniaus universitetas

Finansinių sukčiavimų aptikimo iššūkiai:

- klasių nesubalansuotumas
- koncepcijos poslinkis
- įspėjimo-grįžtamojo ryšio sąveika
- imties atrankos paklaida
- skirtingos sukčiavimo schemos
- skirtingų duomenų šaltinių panaudojimas
- apdorojimas realiuoju laiku
- modelio rezultatų paaiškinamumas
- priešpriešinis mašinų mokymasis (Adversarial Learning)

Nesubalansuotų duomenų su koncepcijos poslinkiu klasifikavimas:

Dėl būdingo klasių nesubalansuotumo, kai viena klasė pasitaiko gerokai rečiau nei kita (-os), modeliai gali būti šališki daugumos klasės atžvilgiu. Koncepcijos poslinkis dar labiau padidina šią problemą, nes duomenų pasiskirstymo pokyčiai gali neproporcingai paveikti mažumos klasę.

Tyrimo uždaviniai

- Identifikuoti aktualias mokslines problemas, kylančias uždaviniuose, susijusiuose su nesubalansuotos mokymo aibės su koncepcijos poslinkiu klasifikavimu;
- Identifikuoti tinkamus metodus nesubalansuotos mokymo aibės su koncepcijos poslinkiu klasifikavimui
- Identifikuoti tinkamus metodus nesubalansuotos mokymo aibės su koncepcijos poslinkiu balansavimui;
- Sukurti arba patobulinti algoritmą nesubalansuotos mokymo aibės su koncepcijos poslinkiu klasifikavimui;
- Pritaikyti sukurtą arba patobulintą metodą nesubalansuotiems duomenims su koncepcijos poslinkiu ir atlikti gautų duomenų analizę, rezultatų apibendrinimą, išvadų parengimą.

2021–2025m.

Vilniaus universitetas

STUDIJŲ PLANAS IR JO VYKDYMO SUVESTINĖ

Studijų metai	Egzaminai [1]		Dalyvavimas konferencijose [2]				Publikacijos [3]					
			Tarptautinėse		Nacionalinėse		Su citav. rodikliu			Be citav. rodiklio		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė [4]	Planas	Įvykdyta	Būklė [4]
I (2021/2022)	3	3			1	1						
II (2022/2023)	1	1	1	1		1		2	Publikuota Pateikta	1	1	Publikuota
III (2023/2024)			1				1					
IV (2024/2025)							1					
Iš viso:	4	4	2	1	1	2	2	2		1	1	

2022/2023 – II pusmetis

Vilniaus universitetas

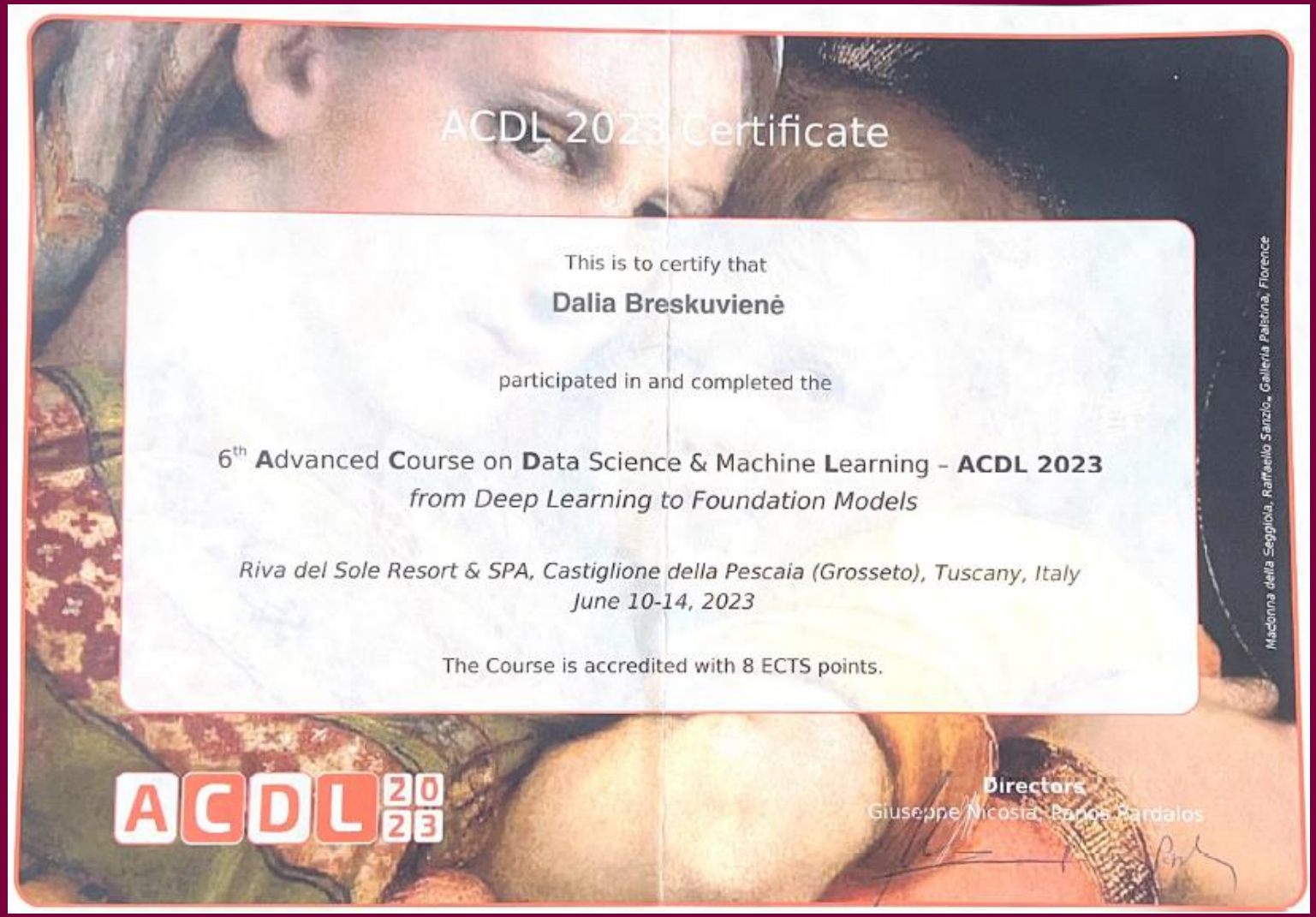
Publikacijos

Publikacijos 2022/2023 (II pusmetis)			
Planas	Įvykdyta	Būklė	Publikacijos tipas
	D. Breskuvienė, G. Dzemyda, “Categorical Feature Encoding Techniques for Improved Classifier Performance when Dealing with Imbalanced Data of Fraudulent Transactions”, INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL, vol. 18, no. 3, Art. no. 3, May 2023, doi: 10.15837/ijccc.2023.3.5433.	Publikuota	Su cituojamumo rodiklio: Impact Factor Q3 Citation Indicator Q2
	D. Breskuvienė, G. Dzemyda, “Enhancing Fraud Detection in the Digital Economy: FID-SOM - Feature Selection for Imbalanced Data”, ADVANCED ENGINEERING INFORMATICS	Pateikta	Su cituojamumo rodiklio: Impact Factor Q1 Citation Indicator Q1

2022/2023 – II pusmetis

Dalyvavimas tarptautinėse vasaros mokyklos

Vilniaus universitetas



Mokslinių tyrimų ir disertacijos rengimo planas:

	Darbo pavadinimas	Atlikimo terminai
1.	Mokslinių tyrimų disertacijos tema apžvalga ir analizė (Lietuvoje ir užsienyje):	
	1.1. Disertacijos tyrimo objekto detalizavimas.	2021 m. gruodžio mėn. – 2022 m. vasario mėn.
	1.2. Atlikti būdų klasifikatorių veikimo optimizavimui analitinę apžvalgą.	2021 m. gruodžio mėn. – 2022 m. spalio mėn.
	1.3. Nustatyti (identifikuoti) mokslines problemas, kylančias uždaviniuose, susijusiuose su klasifikavimo kokybės optimizavimu, o taip pat ir naudojant giliuosius neuroninius tinklus.	2022 m. kovo mėn. – 2022 m. spalio mėn.
	1.4. Tyrimo tikslo suformavimas.	2022 m. kovo mėn. – 2022 m. spalio mėn.
2.	Mokslinio tyrimo vykdymas:	
	2.1. Tyrimo metodikos sudarymas:	
	2.1.1. Tyrimo metodikos iškeltiems uždaviniams spręsti parinkimas;	2022 m. kovo mėn. – 2022 m. spalio mėn.
	2.1.2. Teorinio ir empirinio tyrimų suplanavimas pagal pasirinktą metodiką.	2022 m. kovo mėn. – 2022 m. spalio mėn.
	2.2. Teorinis tyrimas:	
	2.2.1. Klasifikatorių efektyvumo galimybių tyrimas optimizuojant mokymo aibės taškų parinkimą.	2022 m. lapkričio mėn. – 2023 m. spalio mėn.
	2.2.2. Giliųjų neuroninių tinklų panaudojimo galimybių optimaliai mokymo aibei rasti tyrimas.	2022 m. lapkričio mėn. – 2023 m. spalio mėn.
	2.3. Empirinis tyrimas:	
	2.3.1. Sudarytų metodų pritaikymas praktinių uždavinių sprendimui.	2024 m. kovo mėn. – 2024 m. spalio mėn.
	2.3.2. Gautų duomenų analizė, rezultatų apibendrinimas, išvadų parengimas.	2024 m. spalio mėn. –2025 m. vasario mėn.
3.	Atskirų daktaro disertacijos dalių (tyrimo metodikos, rezultatų, ginamų teiginių, išvadų, ir kt.) parengimas:	
	3.1. Tikslų, uždavinių, tyrimo metodikos, ginamųjų teiginių patikslinimas;	2024 m. spalio mėn. – 2025 m. vasario mėn.
	3.2. Analitinės disertacijos dalies parengimas;	2024 m. kovo mėn. – 2025 m. rugpjūčio mėn.
	3.3. Teorinės disertacijos dalies parengimas;	2024 m. kovo mėn. – 2025 m. rugpjūčio mėn.
	3.4. Eksperimentinės disertacijos dalies parengimas;	2024 m. kovo mėn. – 2025 m. rugpjūčio mėn.
	3.5. Bendrųjų išvadų formulavimas.	2024 m. kovo mėn. – 2025 m. rugpjūčio mėn.
4.	Daktaro disertacijos parengimas ir svarstymas padalinyje	2025 m. rugsėjo mėn.
5	Daktaro disertacijos gynimas	2025 m. lapkričio mėn.

Vilniaus universitetas

Atlikti darbai

- Išbandyti skirtingi kategorinių duomenų kodavimo metodai nesubalansuotiems duomenims.
  - James-Stein ir WOE kodavimo būdai parodė geriausius rezultatus iš visų įvertintų kodavimo technikų.
  - CatBoost kodavimo būdas gali būti netinkamas nesubalansuotiems duomenų rinkiniams.
  - Tyrimo rezultatai publikuoti „International Journal of Computers Communications & Control“ žurnale su Impact Factor Q3.

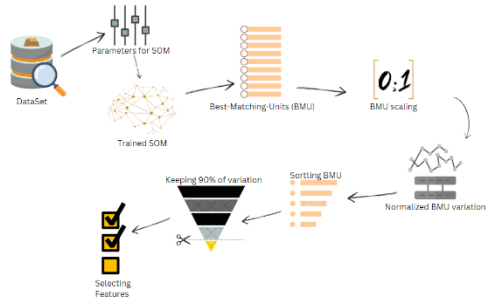
- Ištirta nesubalansuotų duomenų požymių atrinkimo metodai ir pasiūlytas naujas metodas požymių atrinkimui.
  - Metodo gerumui įvertinti naudojamos F1, MCC, G-Mean, AUCPR, and AUCROC.
  - Tirti trys mašininio mokymosi algoritmai - XGBoost, CatBoost, and Random forest.
  - Metodo sėkmė buvo vertinama skaičiuojant, kiek kartų metodas buvo pasirinktas kaip geriausias. Siūlomas FID-SOM metodas pademonstravo sėkmės rodiklį, kuris siekė 51 %. Šis pasiekimas yra reikšmingas ne tik dėl to, kad jis gali būti lygiavertis esamiems metodams ar net juos pranokti, bet ir dėl to, kad rodo jo inovacinį potencialą.
  - Tyrimo rezultatai įteikti „Advanced Engineering Informatics“ žurnalui su Impact Factor Q1.

Publikacijos

Vilniaus universitetas

FID-SOM: Inovatyvus požymių atrinkimo metodas

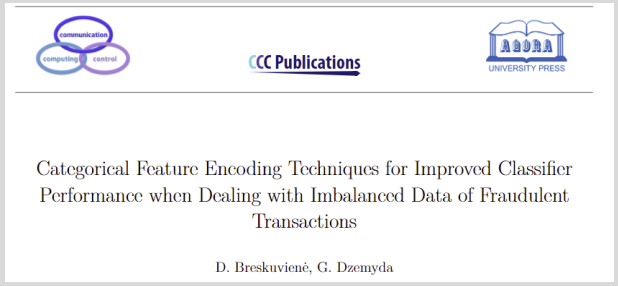
(2022-2023 II pusmetis)



Kategorinių požymių kodavimo poveikis nesubalansuotiems duomenims

(2022-2023 II pusmetis)

- James-Stein
- Weight of Evidence (WOE)



Categorical Feature Encoding Techniques for Improved Classifier Performance when Dealing with Imbalanced Data of Fraudulent Transactions

D. Breskuvienė, G. Dzemyda

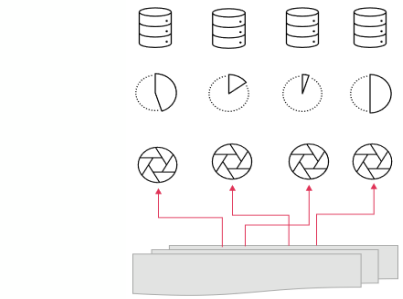
Klasterizacijos pagrindu veikiantis klasifikavmo metodas

(2022-2023 I pusmetis)

Imbalanced Data Classification Approach Based on Clustered Training Set

Dalia Breskuvienė, Gintautas Dzemyda

Institute of Data Science and Digital Technologies, Vilnius University
Akademijos str. 4, Vilnius LT-08412, Lithuania
dalia.breskuviene@gmail.com, gintautas.dzemyda@mif.vu.lt
http://www.mii.lt

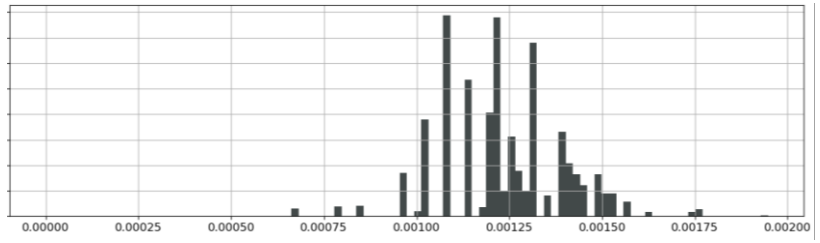


Duomenys naudojami tyrimams

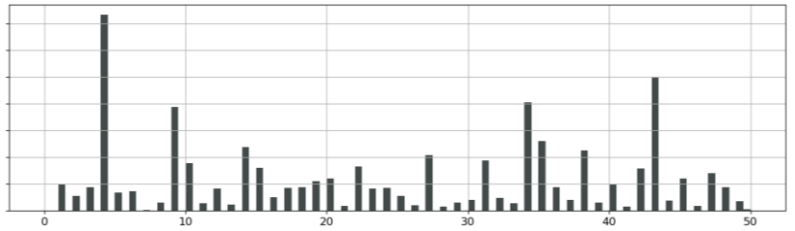
DataSet-A	
Not Fraud	99.86%
Fraud	0.14%
# of instances	3 445 553
# of features	25

DataSet-B	
Not Fraud	99.48%
Fraud	0.52%
# of instances	1 852 394
# of features	11

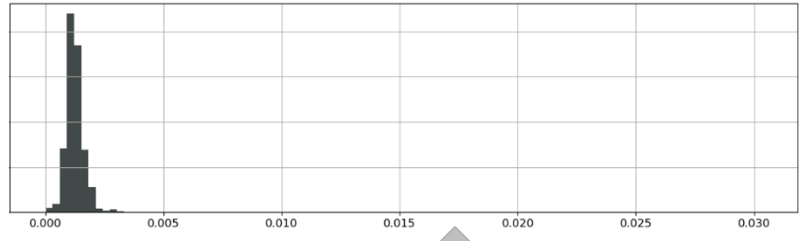
DataSet-C	
Not Fraud	99.13%
Fraud	0.87%
# of instances	161 388
# of features	257



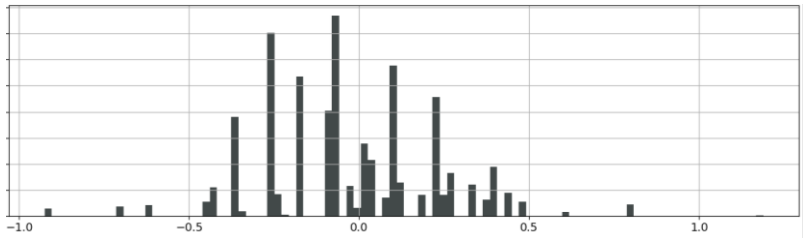
James-Stein Encoder



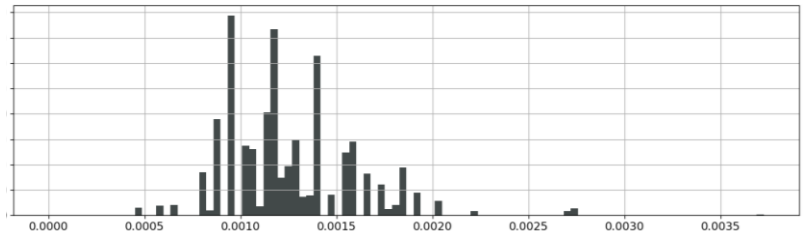
Label Encoder



CatBoost Encoder



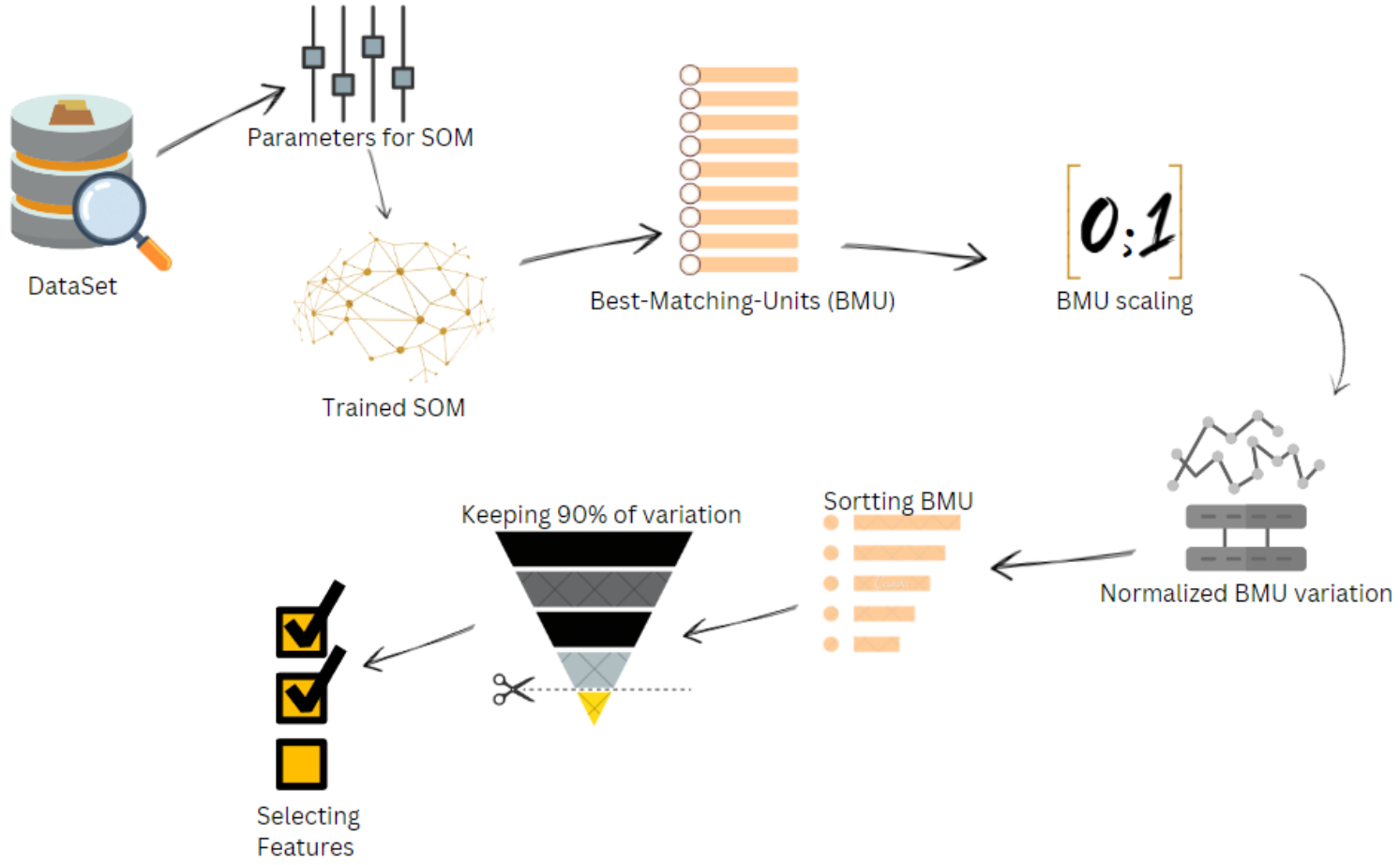
WOE Encoder



m-estimate Encoder

Kategorinių požymių kodavimo poveikis nesubalansuotiems duomenims

Rezultatai rodo, kad kodavimo metodo pasirinkimas gali turėti didelės įtakos mašininio mokymosi modelių veikimui. James-Stein ir WOE kodavimo būdai yra veiksmingiausi iš visų įvertintų kodavimo technikų. Papildomai pažymėtina, kad CatBoost kodavimo būdas gali būti netinkamas nesubalansuotiems duomenų rinkiniams. Be kita ko, naudojant tokius kodavimo būdus, kaip hashing ir One-Hot kodavimas, labai svarbu atsižvelgti į galimą duomenų rinkinio matmenų prakeiksmą (curse of dimensionality).



FID-SOM: Inovatyvus požymių atrinkimo metodas

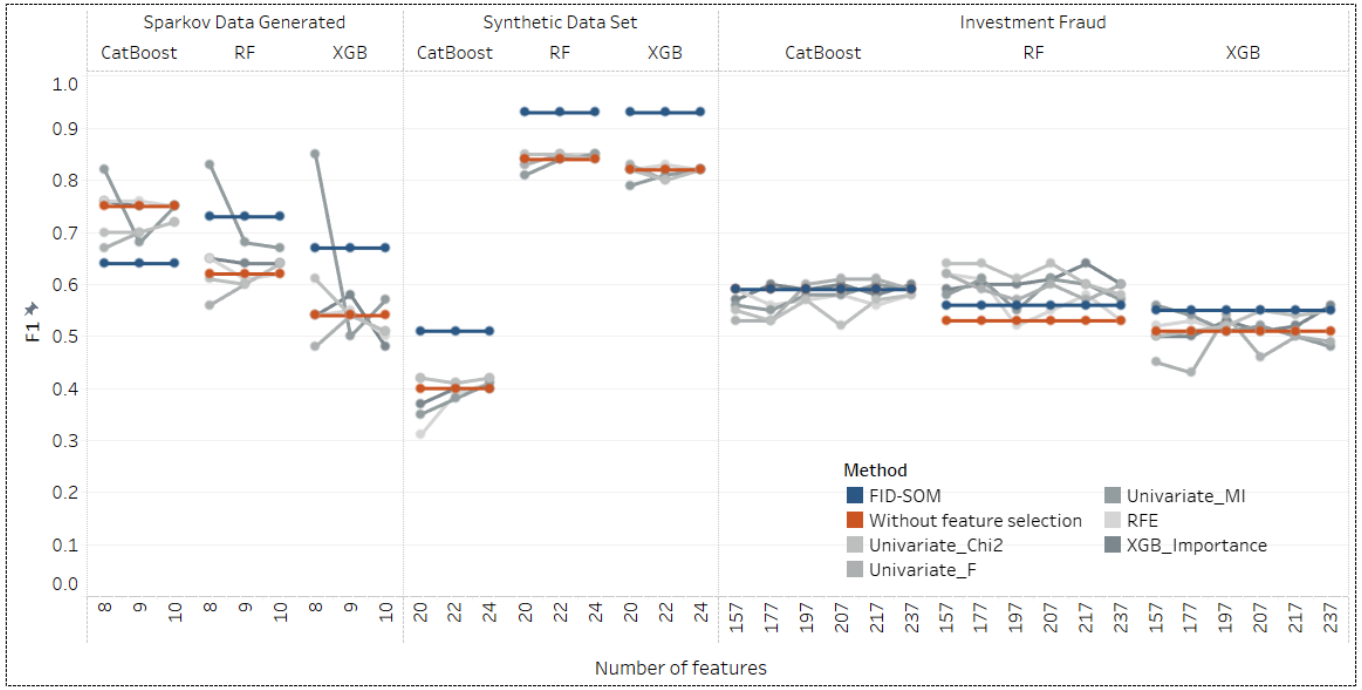
Grafinė straipsnio santrauka

Rezultatai

Vilniaus universitetas

Method	Success	Total	Percentage
Without feature selection	6	45	13%
FID-SOM	23	45	51%
Univariate_Chi2	11	45	24%
Univariate_F	10	45	22%
Univariate_MI	12	45	27%
RFE	7	45	16%
XGB_Importance	7	45	16%
Var method	0	45	0%

Method	Success	Total	Percentage
Without feature selection	34	180	19%
FID-SOM	76	180	46%
Univariate_Chi2	37	180	21%
Univariate_F	38	180	21%
Univariate_MI	49	180	27%
RFE	35	180	19%
XGB_Importance	45	180	26%



KITO PUSMEČIO DARBO PLANAS

- Ištirti jau egzistuojančius tyrimus susijusius su duomenų klasifikavimu, kai duomenys pasižymi koncepcijos poslinkiu.
- Patobulinti arba sukurti naują klasifikavimo metodą nesubalansuotiems duomenims, kurie turi koncepcijos poslinkio savybę.
- Paruošti publikaciją apie atliktus uždavinius.

Vilniaus universitetas

Ačiū už dėmesį!

Klausimai?