YouTube vaizdo įrašų populiarumo vertinimas naudojant viešai prieinamus metaduomenis

Gabrielė Ruminavičiūtė

Vilniaus Universitetas, Matematikos ir informatikos fakultetas,
Naugarduko g. 24, LT-03225 Vilnius
gabriele.ruminaviciute@mif.stud.vu.lt

Santrauka. YouTube vaizdo įrašų duomenų augimo tempai kelia daug problemų sprendžiant vaizdo įrašų populiarumo vertinimo ir klasifikavimo uždavinius. Vis dar nėra nusistovėjusios metodologijos, kokiais kriterijais remiantis, turi būti apibrėžiamas įrašų populiarumas. Šiame darbe analizuojami YouTube duomenų programos programavimo sąsajos pagalba surinkti vaizdo įrašų metaduomenys ir išvestiniai jų parametrai. Įvertinus reikšmingus vidurkių skirtumus tarp populiarumo grupių statistiniais testais, atliktas vaizdo įrašų populiarumo grupių klasifikavimas naudojant tris mašininio mokymosi metodus: atraminių vektorių, atsitiktinio miško ir daugialypės logistinės regresijos klasifikatorius. Atlikta algoritmų lyginamoji analizė bei atrinkti populiarumo lygį tiksliausiai klasifikuojantys požymių rinkiniai.

Raktiniai žodžiai: YouTube, populiarumas, klasifikavimas, atraminių vektorių klasifikatorius, atsitiktinis miškas, daugialypė logistinė regresija, mašininis mokymasis, statistinė analizė

1 Įvadas

YouTube tinklalapis yra viena populiariausių vaizdo įrašų platinimo platformų, turinčių virš dviejų milijardų naudotojų, žiūrinčių vaizdo įrašų turinį virš milijardo valandų kiekvieną dieną. Taip pat kiekvieną minutę į šį tinklalapį yra įkeliama apie 500 valandų vaizdo įrašų turinio [1]. Neabejotina, kad toks greitai augantis vaizdo įrašų duomenų kiekis sudaro vis sunkesnes galimybes vaizdo įrašus surūšiuoti ir atskirti, kurie vaizdo įrašai yra populiarūs bei nustatyti, kokiais kriterijais remiantis turi būti apibrėžiamas populiarumas. Taip pat nėra aiškiai apibrėžtos populiarumo sąvokos - kiekvienas tyrėjas tyrime populiarumo apibrėžimą pateikia savaip. Tiek vaizdo įrašo kūrėjams, tiek YouTube kompanijai informacija apie įrašų populiarumą yra svarbi, ka-

DOI: https://doi.org/10.15388/LMITT.2021.10

dangi abi šalys uždirba iš reklamų – kuo vaizdo įrašas populiaresnis ir pasiekia daugiau auditorijos, tuo daugiau pinigų uždirbama iš tame vaizdo įraše leidžiamos reklamos. Ši informacija reikalinga ir įvairioms reklamos kompanijoms – populiariuose vaizdo įrašuose rodoma reklama matoma daugiau kartų gali padidinti įmonės pelną.

Viena iš pagrindinių metrikų, pagal kurią vaizdo įrašas yra laikomas populiariu yra vaizdo įrašo peržiūros. Kita vertus, populiarumas ir vaizdo įrašų peržiūrų skaičius nėra sinonimai. Labai svarbus kriterijus populiarumui nusakyti yra ir laikas. Dviejų vaizdo įrašų klasifikavimas į tą pačią populiarumo grupę, turint tą patį įrašų peržiūrų skaičių, bet skirtingą įrašų gyvavimo viešoje erdvėje trukmę, būtų neteisingas. Dėl šios priežasties, populiarumas gali būti apibrėžtas skirtingomis kategorijomis (nepopuliarūs, populiarūs, labai populiarūs ir t.t.) atsižvelgiant į vaizdo įrašo peržiūrų skaičių bei gyvavimo trukmę. Tačiau kyla ir kitas klausimas - kokie kiti papildomi požymiai gali daryti įtaką vaizdo įrašo populiarumui?

Šio darbo tikslas yra klasifikuoti YouTube vaizdo įrašus pagal populiarumo grupes, naudojant viešai prieinamus metaduomenis bei atlikti įvairių klasifikavimo algoritmų lyginamąją analizę.

2 Literatūros apžvalga

YouTube vaizdo įrašų populiarumo klausimas nagrinėjant papildomus metaduomenis nėra lengva užduotis. Daugelis svarbių metrikų, kurios iš dalies galėtų pasakyti kokie požymiai lėmė konkretaus vaizdo įrašo populiarumą, nėra viešai prieinami, todėl publikuoti tyrimai atlikti panašia tema taip pat yra labai riboti. Kita vertus, vaizdo įrašų populiarumas įtraukiant įvairius metaduomenis yra plačiai nagrinėjama tema. Vienas iš klasifikatorių, naudojamas YouTube vaizdo įrašų populiarumo prognozavimui, yra atraminių vektorių klasifikatorius (angl. Support Vector Machine, SVM). [2] pateiktas tyrimas parodė, kad naudojant vizualinius vaizdo įrašų požymius (vaizdo įrašų trukmė, kadrų skaičius, rezoliucija, spalvos, veido ir teksto pasirodymas kadruose, miniatiūros kokybė ir t.t) galima nuspėti vaizdo įrašo populiarumą, o norint pagerinti šio klasifikatoriaus prognozavimo tikslumą, dinaminiai požymiai, tokie kaip peržiūrų, „patinka“, pasidalijimų ir komentarų skaičius taip pat turėtų būti įtraukti. Naudojant įvairius mašininio mokymosi metodus, nustatyta, jog penki požymiai daro įtaką vaizdo įrašų

populiarumui – peržiūros pirmąją dieną, kanalo prenumeratorių skaičius, miniatiūros kontrastas, paieškos sistemos „Google“ paspaudimai ir raktinių žodžių kiekis [3]. Kita vertus, populiarumo prognozavimas yra labai svarbi užduotis ankstyvajame vaizdo įrašo gyvavimo laikotarpyje, todėl [4] naudojama daugialypės tiesinės regresijos modeliai, skirti nuspėti ateities vaizdo įrašo peržiūras, naudojant praeities reikšmes. Vaizdo įrašų populiarumas yra vertinamas ir neturint apie vaizdo įrašus jokios istorijos arba dinaminių požymių. Tokiam populiarumo prognozavimui naudojami sudėtingesni mašininio mokymosi klasifikatoriai [5]: klasifikatorių junginys (angl. Ensemble of Classifiers) tiksliau nustato populiarumo klases negu individualūs Naiviojo Bajeso, SVM, logistinės regresijos, neuroninių tinklų ir atsitiktinio miško (angl. Random Forest, RF) klasifikatoriai.

Nagrinėjant mokslinę literatūrą pastebėta, kad vaizdo įrašų metaduomenų požymių rinkiniai skiriasi, todėl apibendrinti gautus rezultatus sudėtinga. Metaduomenų rinkiniai yra labai įvairūs jie rūšiuojasi į dinaminius – kintančius laike – požymius (peržiūrų, komentarų, „patinka“, „nepatinka“ skaičius), miniatiūrų kokybės vertinimo požymių grupę (miniatiūros kokybė, šviesumas, išsiliejimas ir kontrastas), ir tekstinio turinio kokybės vertinimo požymių grupes (simbolių skaičius pavadinime, aprašymas, raktinių žodžių kiekis). Todėl analizė gali būti atliekama naudojant bendrus požymių rinkinius, arba pasirenkant tik vieną tikslinę požymių grupę, pvz. [6] naudojami tekstiniai metaduomenys (aprašymas ir pavadinimas).

3 Duomenys

Darbe naudojami YouTube video įrašų platformos duomenys, surinkti YouTube Data API (angl. Application Programming Interface) – viešai prieinama YouTube duomenų programos programavimo sąsaja. Iš surinktų metaduomenų sukonstruoti papildomi kintamieji, kurie naudojami darbe (žr. 1 lentelę).

Taigi, iš viso turima 14 kintamųjų, kurių 12 - kiekybiniai, o likę 2 - kategoriniai. Visą duomenų imtį sudaro virš 130 tūkstančių vaizdo įrašų, patalpintų 629-iuose kanaluose. 2-oje lentelėje pateikta duomenų rinkinio apibendrinimas.

1 lentelė. Duomenų aprašymas

Apibūdinimas	Tipas	Galimos reikšmės (imties plotis)
Kanalo prenumeratorių skaičius	Kiekybinis	[0, 54800000]
Vaizdo įrašo amžius dienomis	Kiekybinis	[1, 5569]
Vidutinis vaizdo įrašo peržiūrų skaičius per dieną	Kiekybinis	[0, 3543465]
Vidutinis "patinka" paspaudimų skaičius per dieną	Kiekybinis	[0, 207945,46]
Vidutinis "nepatinka" paspaudimų skaičius per dieną	Kiekybinis	[0, 24109,057]
Vidutinis komentarų skaičius per dieną	Kiekybinis	[0, 20660,382]
Egzistuoja vaizdo įrašo aprašymas	Dichotominis	1 arba 0
Simbolių pavadinime skaičius	Kiekybinis	[1, 100]
Raktinių žodžių skaičius	Kiekybinis	[0, 115]
Miniatiūros kokybė	Kiekybinis	[0,0028; 99,9595]
Miniatiūros išsiliejimas	Kiekybinis	[0, 42075]
Miniatiūros kontrastas	Kiekybinis	[0, 124,72]
Miniatiūros šviesumas	Kiekybinis	[0, 255]
Populiarumo grupė	Nominalus	0, 1, 2, 3

2 lentelė. Duomenų rinkinio apibendrinimas

Vaizdo įrašai	132321
Kanalai	629
Vidutinis vaizdo įrašo amžius (dienos)	1089
Vidutinis peržiūrų skaičius (vienam vaizdo įrašui)	4064744

Visa duomenų imtis yra suskirstyta į keturias populiarumo grupes, naudojant vidutinio vaizdo įrašo peržiūrų skaičiaus per dieną kvartilius tam, kad išvengti nesubalansuotos duomenų imties, jog kiekviena grupė duomenyse nepasitaikytų dažniau negu kita. Pirmąją grupę – nepopuliarų – sudaro vaizdo įrašai, kurių vidutinės peržiūros yra ne didesnės negu 48.6. Antrosios – vidutinio populiarumo – grupės vidutinės peržiūros per dieną patenka tarp (48,6012; 494,3954]. Populiarių vaizdo įrašų vidutinės peržiūros per dieną (trečiosios grupės) patenka į intervalą (494;3954, 3209;4048]. Visi likusieji vaizdo įrašai sudaro ketvirtąją – labai populiarių – grupę, kurių vidutinės peržiūros per dieną yra didesnės už 3209,4048. Tačiau tam, kad išvengti tiesinio

ryšio tarp populiarumo grupių ir vaizdo įrašo amžiaus bei peržiūrų skaičiaus, pastarieji kintamieji į mašininio mokymosi algoritmus nėra įtraukiami.

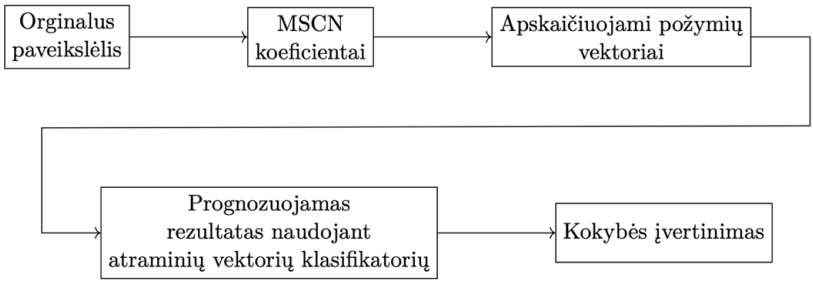
4 Tyrimo metodologija

Metaduomenų konstravimas

Pagrindinių kintamųjų, paaiškinančių vaizdo įrašo žiūrovų įsitraukimą, apskaičiavimas:

- vidutinės peržiūros per dieną: santykis visų vaizdo įrašo peržiūrų ir vaizdo įrašo amžiaus (skirtumas tarp vaizdo įrašo duomenų nusiskaitymo dienos ir vaizdo įrašo publikavimo viešai dienos);
- vidutinis komentarų skaičius per dieną: santykis visų komentarų ir vaizdo įrašo amžiaus;
- vidutinis „patinka“ paspaudimų skaičius per dieną: santykis visų „patinka“ paspaudimų skaičiaus ir vaizdo įrašo amžiaus;
- vidutinis „nepatinka“ paspaudimų skaičius per dieną: santykis visų „nepatinka“ paspaudimų skaičiaus ir vaizdo įrašo amžiaus;

Miniatiūrų įvertinimui fiksuojami kokybės, išsiliejimo, kontrasto ir šviesumo požymiai. Miniatiūros kokybė įvertinta naudojant aklojo/be nuorodų vaizdų erdvinės kokybės vertinimo metodika (angl. Blind/Referenceless Image Spatial Quality Evaluator, BRISQUE) [7]. Šis kokybės vertinimas remiasi algoritmu (žr. 1 pav.)



1 pav. BRISQUE algoritmas

Miniatiūros išsiliejimas skaičiuojamas naudojant Laplaso operatorių [8]. Laplaso operatorius naudojamas briaunų nustatymui, nes pabrėžia atvaizdo regionų staigų kitimą, todėl aukštesnės dispersijos paveiksliukai turi

daugiau staigių kitimų, o tai reiškia didesnį skaičių briaunų. Kuo daugiau nustatoma paveikslėlyje kraštinių, tuo paveikslėlis yra ryškesnis ir mažiau išsiliejęs.

Paveikslėlio kontrastas ir šviesumas apskaičiuojamas atitinkamai pikselio intensyvumo ir pikselio šviesumo vidurkio kvadratine šaknimi [9]:

$$Kontrastas = \sqrt{\frac{1}{MN}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}(I(i,j)-\bar{I})^2},$$

$$Šviesumas = \sqrt{\frac{1}{MN}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}(\delta(i,j)-\bar{\delta})^2},$$

$i \in \{0, 1, ..., M - 1\}$; $j \in \{0, 1,..., N - 1\}$; M ir N miniatiūros aukštis ir plotis atitinkamai, I pikselio intensyvumas, $\bar{I}$ – pikselių intensyvumo vidurkis, δ – pikselio šviesumas, $\bar{\delta}$ – pikselių šviesumo vidurkis.

Statistiniai testai

Skirtumams tarp vidurkių populiarumo grupėse įvertinti naudojamas MANOVA testas. MANOVA testą galima naudoti esant tam tikroms sąlygoms: priklausomi kintamieji turi būti normaliai pasiskirstę grupėse (daugiamatis normalumas), dispersijų homogeniškumas bei tiesiškumas tarp visų priklausomų kintamųjų porų. Tikrinama nulinė hipotezė:

$$\begin{cases} H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = \boldsymbol{\mu}_4 \\ H_1: \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j \text{ nors vienai porai } (i,j), \end{cases}$$

čia $\boldsymbol{\mu}'_s = (\mu_{s1}, \mu s_2, ..., \mu_{sk})$ – grupės $s \in \{1, 2, 3, 4\}$ vidurkių vektorius, $k \in \{1, 2, ..., p\}$ – kintamasis.

Tyrime skaičiuojama Wilks MANOVA testo statistika [10]. Tyrimo duomenims daroma prielaida, kad daugiamačio normalumo ir tiesiškumo prielaidos yra tenkinamos.

Vienmatė ANOVA naudojama ištirti kiekvieną priklausomą kintamąjį. Kadangi dispersijos homogeniškumo prielaidą duomenys pažeidė (tikrinama Levene testu [11]), naudojamas ANOVA Welch kriterijus [12]. Post-hoc analizėje poriniai palyginimai atlikti Games-Howell kriterijumi [13].

Taikomi klasifikatoriai

1. Atraminių vektorių klasifikatorius, naudojant radialinį branduolį. Pagrindinė klasifikatoriaus idėja yra apmokymo duomenų projektavimas į bendrą požymių erdvę ir hiperplokštumos sukūrimas, kuri atskiria skirtingas požymių klases. Radialinio branduolio SVM klasifikatoriui parinktas C – parametras, apibrėžiantis tolerantiškumą klaidingam klasifikavimui ir γ – parametras, atsakingas už sprendimo ribos sklandumą ir kontroliuoja modelio dispersiją [14]. Prieš atliekant SVM, duomenys yra standartizuojami. Kintamųjų standartizacija yra būtina SVM klasifikatoriui, kadangi tokiu būdu visų kintamųjų rėžiai turės panašią įtaką skaičiuojant atstumus hiperplokštumos konstravimui.
2. Atsitiktinis miškas – klasės formuojamos apjungiant sprendimus daugumos balsu iš sprendimų medžių, sukurtų sujungiant skirtingus duomenų poaibius iš visos duomenų aibės naudojant atsitiktinai parinkus požymių poaibius iš požymių aibės. Atlikus atsitiktinio miško klasifikatorių, įvertinta požymių svarba, remiantis Gini indekso:

$$I_G(p) = \sum_{i=1}^{J} p_i(1 - p_i),$$

čia $i \in \{1, 2,..., J\}$, J – imtyje esančių kategorijų skaičius, o p_i yra i-tosios kategorijos proporcija imtyje, pokyčiu, kuomet medyje atliekamas atskyrimas remiantis tam tikru požymiu. Požymio svarbumas yra visų medžių sprendimų vidurkis, tai reiškia, kad požymio svarba yra vidutinė tarp visų medžių [14]. RF parinkti šie parametrai: sprendimų medžių kiekis miške, kiekvieno padalijimo metu atsitiktinai atrinktų požymių skaičius, maksimalaus lygių skaičius kiekviename sprendimų medyje, minimalus duomenų kiekis taškų, reikalingas toliau skaidyti medžio lapams bei minimalus duomenų taškų kiekis lapui.
3. Daugialypės logistinės regresijos modelis (angl. Multinomial Logistic Regression, MLR) – dvinarės logistinės regresijos apibendrinimas, kuomet priklausomas kintamasis įgyja daugiau nei dvi skirtingas reikšmes. Kategorijų priskyrimui yra skaičiuojamos visų galimų Y kategorijų įgijimo tikimybės ir Y prognozuojame tą reikšmę, kurios įgijimo tikimybė yra didžiausia. Matematinis modelis yra nusakomas tikimybėmis:

$$P(Y_i = j|x_i) = \frac{\exp\{z_j\}}{1 + \sum_{h=1}^{J-1} \exp\{z_h\}},$$

čia $j \in \{1, 2, ..., J\}$ – imtyje esančių kategorijų skaičius, $z_j = \alpha_j + \beta_j x_i$, x_i – požymiai. Kaip ir atsitiktiniame miške, taip ir daugialypėje logistinėje regresijoje galima apskaičiuoti požymių svarbą. Ji apskaičiuojama sudedant gautas kiekvieno kintamojo *z* reikšmes kiekvienoje kategorijoje [14]. Prieš atliekant MLR, duomenys yra standartizuojami bei kiekybiniams priklausomiems kintamiesiems pritaikoma kvadratinės šaknies transformacija, kadangi kintamieji yra stipriai asimetriški.

Klasifikatorių tikslumo vertinimui naudojamas k-dalių kryžminis patikrinimas (angl. k-fold Cross Validation), duomenų imtis yra padalinama į lygiai $k = 10$ dalių.

Klasifikatorių įvertinimas

Klasifikavimo algoritmų įvertinimui naudojamos trys metrikos:

$$Bendras\ klasifikavimo\ tikslumas = \frac{Teisingi\ spėjimai}{Visi\ spėjimai}$$ (angl. general accuracy),

$$jautrumas\ (klasė = J) = \frac{TT}{TT + KN}$$ (angl. recall),

$$tikslumas\ (klasė = J) = \frac{TT}{TT + KT}$$ (angl. precision),

čia TT – visi klasės *J* objektai priskiriami klasei *J* , KT – visi ne klasės *J* objektai priskiriami klasei *J,* TN – ne klasės *J* objektai nepriskiriami klasei *J*, KN – visi klasės *J* objektai nepriskiriami klasei *J*.

5 Eksperimentinis tyrimas

Statistinės analizės rezultatai

Priklausomų kokybinių kintamųjų vidurkiai grupėse skiriasi (MANOVA: $F(3, 127) = 3801{,}1$; $\Lambda = 0{,}35$; p – reikšmė = $2{,}2 \times 10^{-16}$). Tolesnės vienmatės ANOVA rezultatai, naudojant Welch kriterijų, pateikti 3 lentelėje. Taigi, yra statistiškai reikšmingi skirtumai visuose kintamuosiuose tarp vaizdo įrašų populiarumo grupių.

3 lentelė. ANOVA Welch kriterijaus rezultatai

Kintamasis	Laisvės laipsniai, n	Laisvės laipsniai, d	F statistika	p -reikšmė
Vidutinis komentarų skaičius per dieną	3	38329	23775,41	<0,05
Vidutinis "nepatinka" paspaudimų skaičius per dieną	3	38354	27842,65	<0,05
Vidutinis "patinka" paspaudimų skaičius per dieną	3	38252	32180,52	<0,05
Kanalo prenumeratorių skaičius	3	41121	15670,81	<0,05
Simbolių pavadinime skaičius	3	49148	240,81	<0,05
Raktinių žodžių skaičius	3	49198	1472,09	<0,05
Miniatiūros kokybė	3	48920	375,04	<0,05
Miniatiūros išsiliejimas	3	49364	1122,26	<0,05
Miniatiūros šviesumas	3	48821	915,8	<0,05
Miniatiūros kontrastas	3	48913	149,81	<0,05

Atlikus porinius palyginimus naudojant Games-Howell kriterijų, nereikšmingi skirtumai tarp vidurkių yra šiuose požymiuose: simbolių pavadinime skaičiaus pirmoje ir antroje bei pirmoje ir trečioje populiarumo grupėse; miniatiūros kokybė antroje ir trečioje populiarumo grupėse; miniatiūros išsiliejime trečioje ir ketvirtoje bei miniatiūros kontraste antroje ir trečioje populiarumo grupėse. Kitur visi poriniai palyginimai, tarp populiarumo grupių reikšmingi kiekvienam rezultato kintamajam.

Klasifikavimo rezultatai

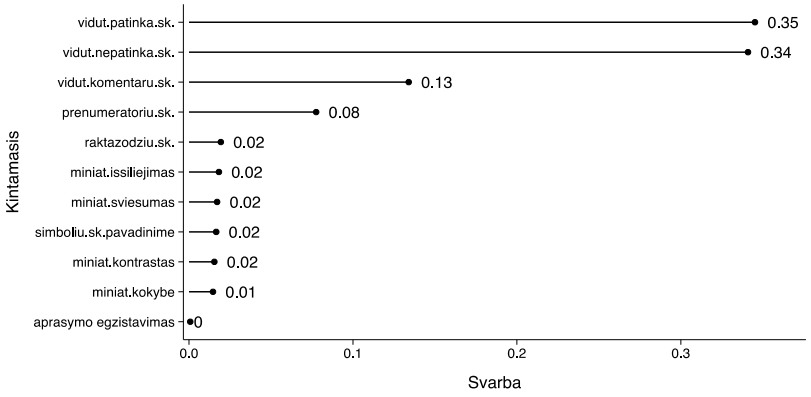
SVM, atsitiktinio miško ir daugialypės regresijos klasifikatoriams apmokyti naudojama atsitiktiniu būdu parinkta subalansuota (kiekvieną klasę sudaro vienodas skaičius vaizdo įrašų) imtis, sudaranti 70 % duomenų. Rezultatų patikrinimui naudojama 10 dalių kryžminis patikrinimas. Likusi 30 % duomenų imtis sudaro testavimo aibę, ji skirta patikrinti, kaip gerai klasifikatorius klasifikuoja nematytus duomenis. Visiems klasifikatoriams naudota ta pati duomenų aibė, sudaryta iš vienuolikos požymių: kanalo prenumeratorių skaičius, vidutinis „patinka", „nepatinka" paspaudimų skaičius per dieną, vidutinis komentarų skaičius per dieną, vaizdo įrašo aprašymo egzistavimas, simbolių pavadinime skaičius, raktinių žodžių skaičius, miniatiūros kokybė, išsiliejimas, kontrastas ir miniatiūros šviesumas.

Atraminių vektorių klasifikatoriaus parametrai: C = 1000 ir γ = 1. Pritaikant šį klasifikatorių testavimo imčiai su parinktais minėtais C ir γ parametrais, gautas 0,67 bendras klasifikavimo tikslumas klasifikuojant populiarumo grupes. Testavimo imčiai pritaikytas klasifikatorius pasiekė 0,68 bendrą klasifikavimo tikslumą. SVM klasifikatoriumi prasčiausiai identifikuojama antra klasė, o tiksliausiai ketvirta (žr. 4 lentelę).

4 lentelė. Klasifikavimo rezultatai

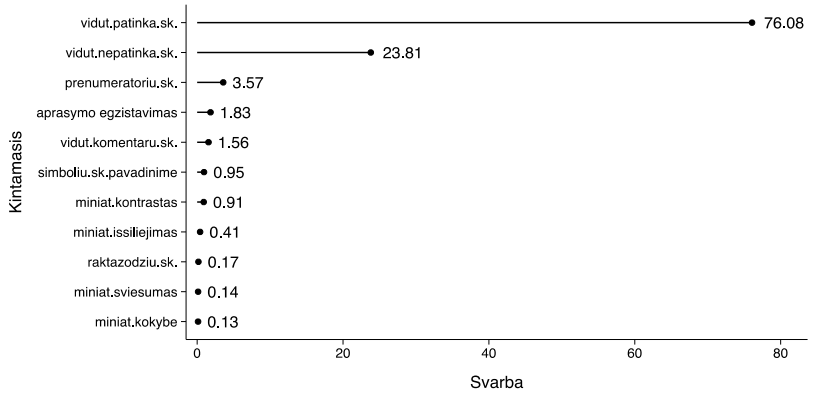
Klasė	SVM		Atsitiktinis miškas		MLR	
	Jautrumas	Tikslumas	Jautrumas	Tikslumas	Jautrumas	Tikslumas
1	0,71	0,63	0,93	0,91	0,91	0,86
2	0,57	0,54	0,83	0,83	0,77	0,77
3	0,62	0,68	0,82	0,82	0,77	0,77
4	0,81	0,89	0,90	0,92	0,86	0,91
Bendras klasifikavimo tikslumas	0,68		0,87		0,83	

Atsitiktinio miško klasifikatorius, įgyjantis aukščiausią tikslumą yra sudarytas iš 600 medžių, turintis 20 lygių kiekviename sprendimų medyje. Testavimo imties bendras klasifikavimo tikslumas yra 0,87 (4 lentelė). Įvertinus klasifikatoriaus jautrumą ir tikslumą, galima daryti išvadą, kad RF yra gana tiksliai suskirstė objektus į reikiamas klases. RF požymių svarba klasifikatoriaus bendram tikslumui matoma 2 pav. Pagal šį paveikslėlį matoma, kad vidutinis „patinka“ ir „nepatinka“ skaičius turi didžiausią svarbą.



2 pav. Atsitiktinio miško požymių svarba klasifikatoriaus bendram tikslumui

Atliekant daugialypę logistinę regresiją mokymo duomenims, gautas 0,83 bendras klasifikavimo tikslumas. Pritaikius MLR testavimo duomenims, gautas toks pats bendras klasifikavimo tikslumas kaip ir mokymo duomenims – 0,83 (žr. 4 lentelę). Nagrinėjant klasifikatoriaus jautrumą, matome, kad MLR klasifikatorius yra mažiau jautrus antros ir trečios populiarumo klasės atžvilgiu, tačiau gana tiksliai klasifikuoja kitas klases. Daugialypės regresijos požymių svarba matoma 3 pav. Kaip ir atsitiktinio miško klasifikatoriuje, taip ir čia, didžiausią svarbą turi tie patys pirmi du kintamieji.



3 pav. Daugialypės logistinės regresijos požymių svarba bendram klasifikatoriaus tikslumui

Lyginant visų atliktų klasifikatorių rezultatus testavimo imčiai, efektyviausias yra atsitiktinio miško klasifikatorius, kurio bendras klasifikatoriaus tikslumas yra 0,87. Žemiausią tikslumą turi SVM klasifikatorius – testavimo duomenims šis klasifikavimo algoritmas pasiekė 0,68 bendrą klasifikavimo tikslumą.

6 Išvados

Darbe atliktas vaizdo įrašų populiarumo tyrimas paremtas YouTube duomenų programos programavimo sąsajos pagalba surinktais vaizdo įrašų metaduomenimis ir išvestiniais jų parametrais. Įvertintas jungtinių kokybinių požymių vidurkių reikšmingas skirtumas tarp populiarumo grupių naudojant MANOVA statistinį testą ($F(3, 127) = 380{,}1$; $\Lambda = 0{,}35$; p – reikšmė = $2{,}2\times10^{-16}$)

bei nustatyti reikšmingi vidurkių skirtumai daugumoje požymių tarp kiekvienos populiarumo grupės porų naudojant Games-Howell kriterijų.

Pasirinkus populiarumo klasių žymėjimą, naudojant vidutinio peržiūrų per dieną skaičiaus kvartilius, atliktas vaizdo įrašų klasifikavimas naudojant tris mašininio mokymosi algoritmus: atraminių vektorių klasifikatorių, atsitiktinį mišką ir daugialypę logistinę regresiją. Geriausi klasifikavimo rezultatai gauti atsitiktinio miško klasifikatoriumi (bendras klasifikavimo tikslumas yra 0.87).

Didžiausią įtaką klasifikavimo tikslumui daro dinaminiai požymiai - „patinka“ ir „nepatinka“ paspaudimų skaičius. Tačiau nagrinėti požymiai nepasako apie tai, ar vaizdo įrašų turinys atitinka kūrėjų pateiktus tekstinius duomenis (aprašymas, pavadinimas, raktiniai žodžiai), todėl tolimesniuose tyrimuose planuojama pasinaudojus natūralios kalbos apdorojimo algoritmais panaudoti įrašų turinio kontekstinę informaciją ir įvertinti jos įtaką populiarumui.

Taip pat vaizdo įrašų populiarumo klasių suskirstymas šiame darbe parinktas naudojant statistinius metodus, todėl ateities darbuose prasminga įvertinti, ar šis būdas yra efektyvus. Planuojama panaudoti grupavimo algoritmus ir įvertinti klasių susidarymo tendencijas bei ryšį su vaizdo įrašų populiarumu. Apibendrinus grupavimo rezultatus siekiama patikslinti populiarumo kriterijaus apibrėžimą ir tyrimą pakartoti naudojant tikslesnį populiarumo kriterijų. Taip pat planuojami papildomi tyrimai išplečiant klasifikatorių aibę, panaudojant rekurentinius neuroninius tinklus ir palyginti rezultatus.

Literatūra

[1] YouTube, „YouTube for Press,“ [Tinkle]. Available: https://blog.youtube/press/. [Kreiptasi 14 03 2021].

[2] T. Trzcinski ir P. Rokita, „Predicting popularity of online videosusing Support Vector Regression,“ *IEEE Transactions on Multimedia,* 10 2015.

[3] W. Hoiles, A. Aprem ir V. Krishnamurthy, „Engagement dynamics and sensitivity analysis of YouTube videos,“ *IEEE Transactions on Knowledge and Data Engineering,* t. PP, 11 2016.

[4] H. Pinto, J. Almeida ir M. Gonçalves, „Using early view patterns to predict the popularity of YouTube videos,“ *WSDM 2013 - Proceedings of the 6th ACM International Conference on Web Search and Data Mining,* 02 2013.

[5] Y.-L. Chen ir C.-L. Chang, „Early prediction of the future popularity of uploaded videos,“ *Expert Systems with Applications,* t. 133, nr. 0957-4174, pp. 59-74, 2019.

[6] G. Kalra, R. Kathuria ir A. Kumar, „YouTube Video Classification based on Title and Des-

cription Text," įtraukta *2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 2019.
[7] A. Mittal, A. Moorthy ir A. Bovik, „No-Reference Image Quality Assessment in the Spatial Domain," *EEE transactions on image processing : a publication of the IEEE Signal Processing Society,* t. 21, 2012.
[8] J. L. Pech Pacheco, G. Cristobal, J. Chamorro-Martinez ir J. Fernandez-Valdivia, „Diatom autofocusing in brightfield microscopy: A comparative study," įtraukta *Pattern Recognition, Proceedings. 15th International Conference on*, 2000.
[9] R. A. Frazor ir W. S. Geisler, „Local luminance and contrast in natural images," *Vision Research,* t. 46, pp. 1585-1598, 2006.
[10] R. A. Johnson ir D. W. Wichern, Applied Multivariate Statistical Analysis, Prentice-Hall, Inc., 2007.
[11] J. Gastwirth, Y. Gel ir W. Miao, „The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice," *Statistical Science,* t. 24, 2010.
[12] B. L. Welch, „On the Comparison of Several Mean Values: An Alternative Approach," *Biometrika,* t. 38, pp. 330-336, 1951.
[13] P. A. Games ir J. F. Howell, „Pairwise Multiple Comparison Procedures with Unequal N's and/or Variances: A Monte Carlo Study," *Journal of Educational Statistics,* t. 1, pp. 113-125, 1976.
[14] G. James, D. Witten, T. Hastie ir R. Tibshirani, An Introduction to Statistical Learning: With Applications in R, Springer Publishing Company, Incorporated, 2017.