



Kauno technologijos universitetas

Informatikos fakultetas

Hibridinis finansinių operacijų gretinimo algoritmas ir jo tyrimas

Baigiamasis magistro studijų projektas

Paulius Litvinas

Projekto autorius

Prof. Rytis Maskeliūnas

Vadovas

Kaunas, 2025



Kauno technologijos universitetas

Informatikos fakultetas

Hibridinis finansinių operacijų gretinimo algoritmas ir jo tyrimas

Baigiamasis magistro studijų projektas

Programų sistemų inžinerija (6211BX011)

Paulius Litvinas

Projekto autorius

Prof. Rytis Maskeliūnas

Vadovas

Prof. Robertas Damaševičius

Recenzentas

Kaunas, 2025



Kauno technologijos universitetas

Informatikos fakultetas

Paulius Litvinas

Hibridinis finansinių operacijų gretinimo algoritmas ir jo tyrimas

Akademinio sąžiningumo deklaracija

Patvirtinu, kad:

1. baigiamąjį projektą parengiau savarankiškai ir sąžiningai, nepažeisdama(s) kitų asmenų autoriaus ar kitų teisių, laikydamasi(s) Lietuvos Respublikos autorių teisių ir gretutinių teisių įstatymo nuostatų, Kauno technologijos universiteto (toliau – Universitetas) intelektualinės nuosavybės valdymo ir perdavimo nuostatų bei Universiteto akademinės etikos kodekse nustatytų etikos reikalavimų;
2. baigiamajame projekte visi pateikti duomenys ir tyrimų rezultatai yra teisingi ir gauti teisėtai, nei viena šio projekto dalis nėra plagijuota nuo jokių spausdintinių ar elektroninių šaltinių, visos baigiamojo projekto tekste pateiktos citatos ir nuorodos yra nurodytos literatūros sąrašė;
3. įstatymų nenumatytų piniginių sumų už baigiamąjį projektą ar jo dalis niekam nesu mokėjęs (-usi);
4. suprantu, kad išaiškėjus nesąžiningumo ar kitų asmenų teisių pažeidimo faktui, man bus taikomos akademinės nuobaudos pagal Universitete galiojančią tvarką ir būsiu pašalinta(s) iš Universiteto, o baigiamasis projektas gali būti pateiktas Akademinės etikos ir procedūrų kontrolieriaus tarnybai nagrinėjant galimą akademinės etikos pažeidimą.

Paulius Litvinas

Patvirtinta elektroniniu būdu



Kauno technologijos universitetas

Informatikos fakultetas

Baigiamojo magistro projekto užduotis

Projekto tema Hibridinis finansinių operacijų gretinimo algoritmas ir jo tyrimas

Reikalavimai ir sąlygos:

Funkciniai reikalavimai:

1. Sistema turi suteikti naudotojo sąsają finansų vadovams, skirta inicijuoti mokėjimų gretinimo užklausas, įskaitant atitinkamų banko sąskaitų pasirinkimą.
2. Sistema turi apdoroti banko išrašų ir klientų/tiekėjų duomenis iš „Business Central“ modulių, juos paruošiant gretinimui su taisyklėmis grindžiamais ir išoriniais metodais.
3. Sistema turi apdoroti gautus mašininio mokymo modelio rezultatus ir juos atvaizduoti Business Central sąsajoje.
4. Sistema turi suteikti finansų vadovams ir konsultantams galimybę peržiūrėti gretinimo istoriją, įskaitant praėjusių gretinimų rezultatus.
5. Sistema turi leisti konsultantams valdyti integracijos konfigūracijos nustatymus, įskaitant duomenų pateikimo dažnumą, modelio mokymo intervalus ir API kredencijų konfigūracijas.

Nefunkciniai reikalavimai:

1. Naujasis funkcionalumas turi veikti lietuvių ir anglų kalbomis
2. Sistema turi teikti aiškius ir lengvai suprantamus klaidos pranešimus naudotojams.
3. Bendras taisyklėmis grindžiamų metodų, išorinių modelių bei pačio meta-modelio veikimas galutiniam banko išrašų gretinimui neturėtų užtrukti ilgiau 30 minučių.
4. Sistema turi būti integruota su mašininio mokymosi meta-modeliu.
5. Tik konsultantai ir paskirti kliento įmonės naudotojai turi turėti galimybes keisti integracijos konfigūraciją.

Kiti reikalavimai:

Nėra

Vadovas / Vadovė

Prof. Rytis Maskeliūnas

Litvinas Paulius. Hibridinis finansinių operacijų gretinimo algoritmas ir jo tyrimas. Magistro baigiamasis projektas vadovas Prof. Rytis Maskeliūnas; Kauno technologijos universitetas, informatikos fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Programų sistemos.

Reikšminiai žodžiai: bankinių operacijų gretinimas, algoritmo parinkimas, hibridinis atpažinimo algoritmas, mašininis mokymas, taisyklėmis grindžiami metodai

Kaunas, 2025. 72 p.

Santrauka

Įmonių apskaitos kasdienėje praktikoje, tikslus ir greitas dokumentų gretinimas su bankų pateikiamais mokėjimų duomenimis verslo valdymo sistemose išlieka reikšmingas iššūkis. Skirtingų bankų išrašų struktūrų ir formatų įvairovė, priklausanti nuo šalies teisinių ir techninių reikalavimų, neleidžia pasiekti aukštos automatizacijos tradiciniais taisyklėmis grindžiamais metodais – kiekvienam klientui tenka kurti ir adaptuoti atskirus taisyklių rinkinius, o tai reikalauja daug darbuotojų laiko ir specialių žinių.

Šiame darbe pristatomas hibridinio banko operacijų gretinimo algoritmo, integruoto į „Microsoft Dynamics Business Central“ sistemos „Bankfeed“ produktą, projektavimas, įgyvendinimas ir vertinimas. Pagrindinis indėlis – algoritmų parinkimo (angl. algorithm selection) meta-modelis, kuris, pasinaudodamas istoriniais gretinimo duomenimis, dinamiškai parenka optimaliausią metodą – ar tai būtų taisyklėmis pagrįstas metodas, ar mašininio mokymosi metodas - kiekvienai gaunamai operacijai. Atlikto tyrimo metu buvo nustatytas tiksliausiai atpažinimo metodus renkantis mašininio mokymo modelis, įvertintas jo tikslumas bei bandomų spėjimų apimtis. Aprašoma modelio mokymo pritaikymo verslo sričiai eiga, kai įvedami svoriai, siekiant padidinti tikslumą, bandomų spėjimų kaina. Galiausiai aprašomas atliktas eksperimentas naudojant keleto apskaitos įmonių realius duomenis, siekiant įvertinti modelio veikimą praktikoje, įvertinamas sukurto sprendimo įnešamų pokių statistinis reikšmingumas.

Litvinas Paulius. Analysis and Implementation of a Hybrid Algorithm for Financial Transaction Reconciliation. Master's Final Degree Project supervisor Prof. Rytis Maskeliūnas; Informatics faculty, Kaunas University of Technology.

Study field and area (study field group): Program systems.

Keywords: bank transaction reconciliation, algorithm selection, hybrid matching algorithm, machine learning, rule-based methods

Kaunas, 2025. 72.

Summary

In the daily practice of corporate accounting departments, precisely and rapidly matching purchase and sales invoices stored in enterprise resource planning (ERP) systems with bank payment records remains a significant challenge. The heterogeneity of bank statement formats—driven by varied legal and technical requirements across jurisdictions—prevents full automation via rule-based methods, since each client demands custom rule sets that are time-consuming to create and maintain.

This thesis presents the design, implementation, and evaluation of a hybrid bank transaction reconciliation algorithm integrated into „Microsoft Dynamics Business Central“ system’s Bankfeed module. The core contribution is an algorithm-selection meta-model that, leveraging historical matching data, dynamically chooses the optimal reconciliation method – whether rule-based or machine-learning-based, for each incoming transaction. During the analysis part, we identified the machine-learning model variant that most accurately predicts the best matching approach, measured its precision and coverage, and introduced weighting schemes to balance accuracy against computational cost. Finally, we report on a field trial assessing practical performance and statistically validate the improvements in matching accuracy and operational efficiency compared to the legacy rule-based system. This hybrid approach hopes to significantly reduce manual configuration effort and admin cost, while boosting the day-to-day effectiveness of accounting staff in ERP environments.

Turinys

Lentelių sąrašas	9
Paveikslų sąrašas	10
Santrumpų ir terminų sąrašas	11
Įvadas.....	12
1. Analitinė dalis	14
1.1. Mašininio mokymo metodai, tinkami algoritmų parinkimo užduotims spręsti	14
1.1.1. Sprendimų medžiai.....	14
1.1.2. Ansamblio algoritmai	14
1.1.3. Modelių tikslumo vertinimo metrikos	16
1.2. Taisyklėmis grindžiamų sistemų ir mašininio mokymosi panaudojimas finansinių operacijų gretinimui	17
1.3. Statistinės analizės metodai	18
1.3.1. ANOVA.....	19
1.3.2. Efekto dydis η^2	19
1.3.3. Tukey HSD	20
1.3.4. Porinis T-testas	20
1.3.5. Shapiro-Wilk	20
1.3.6. Friedman.....	21
1.3.7. Wilcoxon	21
1.4. Pasirinktų technologijų pagrindimas	21
1.4.1. Python programavimo kalba.....	21
1.4.2. Azure Machine Learning Studio	22
1.4.3. Azure Functions.....	22
1.5. Egzistuojančių rinkoje sprendimų palyginimas.....	22
1.5.1. Xero	22
1.5.2. Intuit QuickBooks	23
1.5.3. BlackLine	24
1.5.4. Trintech (Cadency).....	24
2. Hibridinio finansinių operacijų gretinimo algoritmo projektinė dalis.....	27
2.1.1. Esama padėtis	27
2.1.2. Veiklos kontekstas.....	27
2.1.3. Sistemos panaudojimo atvejų modelis	28
2.2. Funkciniai reikalavimai	31
2.3. Nefunkciniai reikalavimai	32
2.4. Sistemos išdėstymo vaizdas	34
2.5. Sistemos architektūra.....	34
3. Tyrimo dalis	43
3.1. Tyrimo tikslas.....	43
3.2. Tyrimo eiga	43
3.2.1. Tinkamų mokymui duomenų paruošimas ir apdorojimas	43
3.2.2. Geriausio mašininio mokymo modelio parinkimo procesas	46
3.2.3. Mokymo pritaikymas teikti prioritetą aukštam tikslumui	50
3.3. Tyrimo rezultatai	52
3.3.1. Hibridinis gretinimo algoritmas veikia tiksliau už vien taisyklėmis grįstą metodą	52

4. Eksperimentinė dalis	58
4.1. Eksperimento tikslas.....	58
4.2. Eksperimento eiga	58
4.2.1. Eksperimento atlikimo procesas.....	58
4.2.2. Duomenų surinkimas.....	60
4.3. Modelio tikslumo vertinimas naudojant realių įmonių duomenis.....	61
4.4. Eksperimento rezultatų statistinė analizė	63
4.4.1. ANOVA rezultatai ir efektų dydžiai.....	64
4.4.2. Tukey HSD post-hoc palyginimas.....	64
4.4.3. T-testų rezultatai	64
4.4.4. Shapiro-Wilk testo rezultatai	65
4.4.5. Neparametriniai Friedman / Wilcoxon testai	65
Išvados	67
Literatūros sąrašas	68
Priedai.....	71
1 priedas. Sujungto galutinio mokymo duomenų rinkinio laukų paaiškinimai.....	71

Lentelių sąrašas

1 lentelė.	Populiarių ansamblio metodų apibendrinimas	15
2 lentelė.	Maišos matricos pavyzdys.....	16
3 lentelė.	Klasifikatorių kokybės vertinimo metrikos	16
4 lentelė.	K-kartotinių kryžminės patikros vizualizacija.....	17
5 lentelė.	Statistinės analizės metodų apibendrinimas	18
6 lentelė.	Egzistuojančių technologijų ir kuriamo sprendimo palyginimas	26
7 lentelė.	PA1 Inicijuoti mokėjimų gretinimo procesą	29
8 lentelė.	PA2 Apdoroti duomenis gretinimui	29
9 lentelė.	PA3 Pritaikyti gretinimo metodus	29
10 lentelė.	PA4 Parinkti geriausią galutinį gretinimo metodą	29
11 lentelė.	PA5 Registruoti mokėjimų gretinimus.....	30
12 lentelė.	PA6 Peržiūrėti gretinimo istoriją.....	30
13 lentelė.	PA7 Valdyti integracijų konfigūraciją.....	30
14 lentelė.	PA8 Tikrinti gretinimo metodo parinkimo proceso statusą	30
15 lentelė.	PA9 Gražinti gretinimo rezultatus	30
16 lentelė.	Funkcinis reikalavimas 1 – naudotojo sąsaja	31
17 lentelė.	Funkcinis reikalavimas 2 – „Business Central“ duomenų apdorojimas.....	31
18 lentelė.	Funkcinis reikalavimas 3 – mašininio mokymo rezultatų apdorojimas	31
19 lentelė.	Funkcinis reikalavimas 4 – gretinimo istorijos peržiūra	31
20 lentelė.	Funkcinis reikalavimas 5 – integracijos konfigūracijos valdymas.....	32
21 lentelė.	Nefunkcinis reikalavimas 6 – funkcionalumo kalba	32
22 lentelė.	Nefunkcinis reikalavimas 7 – klaidos pranešimai	32
23 lentelė.	Nefunkcinis reikalavimas 8 – algoritmo veikimo trukmė	33
24 lentelė.	Nefunkcinis reikalavimas 9 – integracijos	33
25 lentelė.	Nefunkcinis reikalavimas 10 – integracijos konfigūracijos valdymo teisės	33
26 lentelė.	Azure Functions komponentų aprašymas.....	34
27 lentelė.	Klasifikatoriaus sudarymo taisyklės.....	46
28 lentelė.	Tyrime išbandytų mašininio mokymo modelių sąrašas	47
29 lentelė.	Apibendrintas modelių rezultatų, tikslumo, rizikos įvertinimo, mokymo laiko sąrašas .	48
30 lentelė.	Mašininio mokymo modelių rezultatai, lyginant su atpažinimo metodais.....	53
31 lentelė.	Tyrimo metu atliktos 10 k kryžminės patikros rezultatai	55
32 lentelė.	Atpažinimo metodų tikslumo santrauka.....	56
33 lentelė.	Mašininio mokymo modelių rezultatų apibendrinimas	56
34 lentelė.	Eksperimento metu naudotų įmonių duomenų parametrų suvestinė.....	60
35 lentelė.	Eksperimento, atlikto ant 8 įmonių realių duomenų, rezultatai.....	61
36 lentelė.	Eksperimento rezultatų vidurkių apibendrinimas.....	62
37 lentelė.	ANOVA rezultatų apibendrinimas	64
38 lentelė.	Tukey HSD rezultatų apibendrinimas	64
39 lentelė.	T-testų rezultatų apibendrinimas	65
40 lentelė.	Shapiro-Wilk rezultatų apibendrinimas.....	65
41 lentelė.	Friedman rezultatų apibendrinimas	65
42 lentelė.	Wilcoxon rezultatų apibendrinimas.....	66

Paveikslų sąrašas

1 pav. Sprendimų medžio pavyzdys	14
2 pav. Xero išrašų gretinimo langas	23
3 pav. Quickbooks išrašų gretinimo langas.....	23
4 pav. BlackLine išrašų gretinimo langas	24
5 pav. Trintech (Cadency) išrašų gretinimo langas.....	25
6 pav. Hibridinio finansinių operacijų gretinimo algoritmo veiklos konteksto diagrama	28
7 pav. Hibridinio finansinių operacijų gretinimo algoritmo panaudojimo atvejų diagrama.....	28
8 pav. Hibridinio finansinių operacijų gretinimo algoritmo diegimo diagrama	34
9 pav. Panaudojimo atvejo „Apdoroti duomenis gretinimui“ veiklos diagrama	35
10 pav. Panaudojimo atvejo „Apdoroti duomenis gretinimui“ sekos diagrama	36
11 pav. Panaudojimo atvejo „Parinkti geriausią galutinio gretinimo metodą“ veiklos diagrama	37
12 pav. Panaudojimo atvejo „Parinkti geriausią galutinio gretinimo metodą“ sekos diagrama	38
13 pav. Panaudojimo atvejo „Grąžinti gretinimo rezultatus“ veiklos diagrama.....	39
14 pav. Panaudojimo atvejo „Grąžinti gretinimo rezultatus“ sekos diagrama.....	39
15 pav. Panaudojimo atvejo „Tikrinti gretinimo proceso statusą“ veiklos diagrama	40
16 pav. Panaudojimo atvejo „Tikrinti gretinimo proceso statusą“ sekos diagrama.....	40
17 pav. Panaudojimo atvejo „Pritaikyti gretinimo metodus“ veiklos diagrama	41
18 pav. Panaudojimo atvejo „Pritaikyti gretinimo metodus“ sekos diagrama.....	42
19 pav. Apibendrinta banko išrašų importo į Business Central schema	43
20 pav. Apibendrinta Business Central duomenų surinkimo ir apdorojimo schema.....	44
21 pav. Apibendrinta duomenų apdorojimo Azure Functions sistemoje schema.....	45
22 pav. Apibendrinta funkcijos <i>process_training_dataframe()</i> schema.....	46
23 pav. Tyrimo išbandytų prižiūravimo mokymo algoritmų rezultatų maišos matricos	49
24 pav. Tyrimo išbandytų prižiūravimo mokymo algoritmų svarbiausių požymių grafikai	50
25 pav. Tyrimo išbandytų prižiūravimo mokymo algoritmų slenksčių vizualizacijos	52
26 pav. POST užklausa, siunčianti įmonės ZIP archyvą modelio mokymui	58
27 pav. Mokymo ir patalpavimo procesas Azure ML aplinkoje	58
28 pav. Detalesnis Azure ML mokymo procesas su metrikomis.....	59
29 pav. GET užklausa, skirta rezultatų gavimui iš Azure ML.....	59
30 pav. Naujų duomenų spėjimo procesas Azure ML platformoje	59
31 pav. Gaunamų apdorotų rezultatų iš Azure ML pavyzdys.....	60
32 pav. XGBoost modelio, vieno iš medžių pavyzdys	60
33 pav. Baudos vertės pasiskirstymas pagal metodą	62
34 pav. Tikslumo pasiskirstymas pagal metodą.....	63
35 pav. Metodų baudos vertės išsidėstymas per įmones.....	63

Santrumpų ir terminų sąrašas

Santrumpos:

BC (angl. *Business Central*) – „Microsoft Dynamics 365 Business Central“ verslo valdymo sistema, kurioje integruotas „Bankfeed“ produktas finansinių operacijų gretinimui.

API (angl. *Application Programming Interface*) – sąsaja, leidžianti perduoti duomenis tarp „Business Central“ ir išorinių sistemų, naudojama finansinių duomenų apsikeitimui ir išorinių gretinimo metodų integracijai.

ML (angl. *Machine Learning*) - mašininis mokymas. Sistemoje naudojamas mokytas MM modelis, skirtas automatiškai atpažinti finansinių operacijų gretinimo dėsningumus ir pateikti sprendimus, kokią sąskaitą reikia gretinti su tam tikru mokėjimu.

ERP (angl. *Enterprise Resource Planning*) – verslo valdymo sistema. Tokios sistemos skaitmeniniu būdu integruotos į įmonės, įstaigos ar organizacijos kasdienes veiklas, leidžia įmonėms valdyti pirkimus, pardavimus, gamybą, sandėlio procesus vienoje sistemoje.

Terminai:

Finansinių operacijų gretinimas – procesas, kurio metu banko išrašų operacijos susiejamos su atitinkamais dokumentais ar sąskaitomis verslo valdymo sistemoje.

Hibridinis algoritmas – algoritmas, kuris naudoja keletą skirtingų metodų (šiuo atveju taisyklėmis grįstus metodus kartu su mašininio mokymosi modeliais), siekiant optimaliai atpažinti ir gretinti finansines operacijas.

Meta-modelis – mašininio mokymosi modelis, kuris derina taisyklėmis grįstų metodų ir MM modelių rezultatus, siekdamas atrinkti geriausią metodą tam tikros banko operacijos susiejimui su sąskaita ar apskaitos dokumentu.

Taisyklių rinkinys – taisyklių sistema, kuri atliekama pirmame finansinių operacijų gretinimo etape, siekiant greitai atpažinti atitikimus, pvz., pagal banko sąskaitą ar dokumento numerį.

Įvadas

Problemos aktualumas

Įmonių apskaitininkai, naudojantys verslo valdymo sistemas (VVS, angl. *enterprise resource planning, ERP*), kasdien susiduria su užduotimi, kai sistemose esantys dokumentai, tokie kaip pirkimo, pardavimo sąskaitos faktūros, turi būti sugretinami su bankinių mokėjimų duomenimis, nurodančiais, jog dokumentai buvo mokėti. Šis gretinimas leidžia įmonėms stebėti klientų mokumą, užfiksuoti vėluojančius mokėjimus, nustatyti skolas, taip pat atlikti finansinę analizę bei auditą. Nors rinkoje egzistuojančios sistemos dažnai turi automatizuotus sprendimus, visiško automatizavimo įgyvendinimą riboja bankinių duomenų heterogeniškumas – skirtingų bankų išrašai būna nevienodos struktūros, formato ir detalumo, tai priklauso nuo šalies teisinių bei techninių reikalavimų.

Augant bankinių duomenų kiekiui ir automatizavimo poreikiui finansinėse apskaitos sistemose, vis aktualesniu tampa efektyvus bankinių operacijų gretinimo procesas. Tradiciniai, taisyklėmis grindžiami metodai reikalauja daug laiko bei specialių žinių kiekvienam klientui atskirai. Srities specialistai, tam tikram klientui sukuria taisyklių sistemą, pagal kliento specifiką, naudojančią tokias taisykles, kaip „Regex“ atpažinimas, filtrai istoriniams įrašams ar pan., kad rastų tinkamiausią sąskaitą tam tikrai banko operacijai. Toks procesas yra itin imlus laikui, kadangi kiekvienas klientas turi savo specifiškumą.

Šiai problemai spręsti buvo pradėti kurti metodai, reikalaujantys minimalaus konsultantų indėlio, tačiau vis tiek pateikiantys pasiūlymus, tinkančius gretinti, tik ne taip tiksliai. Papildomai, pradėti kurti mašininio mokymosi algoritmai, kurie taip pat rodo nemažą potencialą pakeisti taisyklėmis grindžiamus metodus, tačiau jie taip pat turi savų apribojimų specifinėse situacijose. Todėl ypač aktualu ieškoti būdų, kaip šiuos metodus veiksmingai derinti tarpusavyje, išnaudojant kiekvieno stiprybes.

Turint taisyklėmis grindžiamus metodus, veikiančius vienose situacijose, ir mašininio mokymosi algoritmus, veikiančius gerai kitose situacijose – iškilo poreikis spręsti algoritmo parinkimo (angl. *algorithm selection*) problemą, sukuriant procesą, parenkantį geriausią gretinimo metodą tam tikram banko išrašui, panaudojant istorinius gretinimo įrašus, kaip kontekstą mokytis.

Darbo objektas

Darbo objektas – bankinių operacijų gretinimo metodai „Microsoft Dynamics Business Central“ aplinkoje, koncentruojantis į „Bankfeed“ produkto funkcionalumo analizę ir išplėtimą, panaudojant taisyklėmis pagrįstus ir mašininio mokymosi metodus.

Darbo tikslas

Suefektvinti bankinių operacijų gretinimo procesą sukuriant hibridinį algoritmą, gebantį dinamiškai parinkti optimaliausią bankinių operacijų gretinimo metodą, taip sumažinant administravimo kaštus ir padidinant apskaitos darbuotojų efektyvumą kasdienėje veikloje.

Darbo uždaviniai

1. Išanalizuoti esamus taisyklėmis grindžiamus ir mašininio mokymosi metodus, naudojamus bankinėms operacijoms gretinti;

2. suprojektuoti ir sukurti algoritmų parinkimo (angl. *algorithm selection*) meta-modelį, gebantį dinamiškai pasirinkti tinkamiausią gretinimo metodą konkrečiai banko operacijai;
3. realizuoti siūlomą metodą realioje darbinėje aplinkoje;
4. įvertinti siūlomo hibridinio algoritmo veiksmingumą, jį lyginant su esamu taisyklėmis grindžiamo metodo veiksmingumu.

Darbo struktūra

Darbą sudaro keturios pagrindinės dalys: pirmoje dalyje nagrinėjama su tyrimo problema susijusi literatūra ir esami metodai, antroje – pristatomas projektas, sistemos reikalavimai bei architektūra, trečioje – pateikiama tiriamoji dalis, aprašanti algoritmo sudarymą, tyrimo eigą ir rezultatus, o ketvirtoje dalyje aprašoma eksperimentinė dalis, kur geriausiai pasirodęs algoritmas pritaikomas praktiškai, formuluojamos išvados.

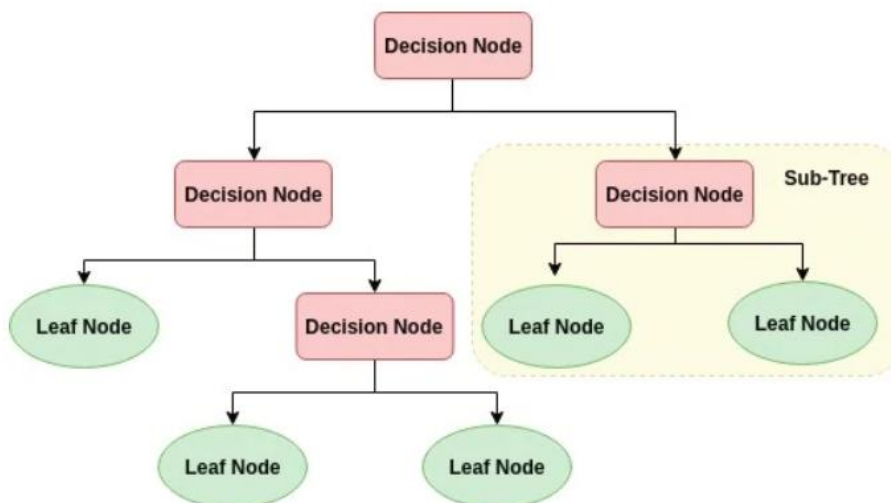
1. Analitinė dalis

Visi mokymo algoritmai skirtingose situacijose veikia skirtingu tikslumu, kaip NFL teorema (angl. *no-free-lunch theorem*) teigia – nėra vieno algoritmo, pranokstančio kitus visose srityse. Sujungiant keletą algoritmų, galima pagerinti bendrą klasifikatoriaus tikslumą, kur teoriškai sujungiant begalinį skaičių klasifikatorių būtų galima priartėti prie optimalaus Bajeso klasifikatoriaus (angl. *Bayes classifier*) [1]. Mašiniame mokyme sprendimų sujungimo uždavinys užima svarbų vaidmenį didinant prognozių tikslumą bei užtikrinant patikimumą, taip sprendžiant vieną iš esminių šios srities iššūkių [2].

1.1. Mašininio mokymo metodai, tinkami algoritmų parinkimo uždutims spręsti

1.1.1. Sprendimų medžiai

Sprendimų medžiai priskiriami prie logika grindžiamų algoritmų, skirtų prižiūrimam klasifikavimui (angl. *supervised classification*) [3]. Dažniausiai jie naudojami atvaizduoti taisyklėms, apibūdinančioms hierarchinį duomenų rinkinio padalijimą. Tai reiškia, kad sprendimų medžiai gali būti taikomi ne tik klasifikavimo uždaviniuose, bet ir kitose duomenų mokslo uždutyse, tokiose kaip duomenų aprašymas ar apibendrinimas. Šiame darbe dėmesys bus skiriamas sprendimų medžio taikymui klasifikavimo uždaviniuose. Sprendimų medžiai klasifikuoja įrašus, rūšiuojant juos pagal išmokstamus kriterijus. Pradedama nuo šakninio mazgo, pasirenkant tokį požymį, kuris geriausiai padalintų duomenų rinkinį į dvi galimas klases. Mazgai reprezentuoja atskyrimo kriterijus, o šakos – reikšmes, atskirtas pagal šiuos kriterijus. Galutinis rezultatas – tai taisyklių, aprašančių požymių reikšmes, rinkinys, išdėstytas (pradedant nuo šaknies mazgo) nuo labiausiai iki mažiausiai skiriančio požymio. Sprendimų medžiai (žr. 1 pav.) lengvai vizualizuojami ir gali būti siejami su taisyklėmis grįžtais metodais, kadangi juos lengva suprasti [4].



1 pav. Sprendimų medžio pavyzdys

1.1.2. Ansamblio algoritmai

Ansamblio algoritmai (angl. *ensemble algorithms*) – tai metodai, kai naudojami ne vieno, o kelių modelių spėjimai, galiausiai imant didžiausią / mažiausią / suvidurkintą / populiariausią (priklausomai nuo taikymo) reikšmę.

Klasifikavimo užduotyse, skirtingi klasifikatoriai gali būti kuriami per skirtingus mokymo inicializavimo būdus, išmokant skirtingus duomenų rinkinius, arba išrenkant skirtingus požymių rinkinius modelių mokymui. Ansamblio metodai [5], [6] generuoja kelis bazinius mokymo modelius ir juos sujungia į vieną sudėtinį spėjimą, skirtą naujų duomenų klasifikavimui. Šis mokymo būdas remiasi įvairovės (angl. *diversity*) principu [7], kur du klasifikatoriai yra vadinami skirtingais, jeigu jie grąžina skirtingai klaidingus spėjimus nematytiems duomenims. Išnaudojant įvairovę, ansamblio mokymasis pagerina klasifikavimo tikslumą, lyginant su bet kuriuo individualiu klasifikatoriumi, kai pateikiamas toks pat duomenų rinkinys. Ansamblio mokyme, klasifikatoriai ansamblyje gali būti mokomi nepriklausomai, kaip „bagging“ metode [6], arba nuosekliai, kaip „boosting“ [5] tipo metoduose. Ansamblio mokymo metodų efektyvumas stipriai priklauso nuo silpniausių spėjimų grąžinamų rezultatų įvairovės, kadangi jie padidina galimų grąžinti sprendimų aibę bei sustiprina ansamblio apibendrinimo galimybes. Tai pasiekama agreguojant spėjimus, naudojant strategijas, tokias kaip balsavimas (angl. *voting*), ar skaičiuojant svertinį vidurkį (angl. *weighted averaging*).

Istoriškai yra pateikta daug pasiūlymų spėjimų agregavimo metodikoms, kurie galėtų pasiekti geresnių rezultatų nei pavieniai algoritmai, derinant kelias to paties, arba skirtingų spėjimų iteracijas. Tarp jų, du plačiai paplitę metodai su gerai žinomomis įgyvendinimo priemonėmis, pagrįstomis sprendimų medžiais, yra „boosting“ [4] ir „bagging“ [3]. Esminis skirtumas tarp šių dviejų metodų yra tai, kaip tarpusavyje susijusios generuojamos spėjimų iteracijos. „Bagging“ metode, mokymo rinkiniai (angl. *bootstrap sets*) yra atsitiktinai parenkami ir naudojami tam, kad būtų sugeneruota daug nepriklausomų spėjimų. „Boosting“ atveju, kiekviename žingsnyje naudojama visa mokymosi duomenų aibė, o nauja spėjimų iteracija priklauso nuo ankstesnių, bandant ištaisyti trūkumus, paveldėtus iš ankstesnių kartų [8]. Abiem atvejais, galutinis prognozavimo rezultatas gaunamas kombinuojant visų generuotų spėjimų išvestis. „Bagging“ atveju, pagrindinis rezultatas yra paklaidos sumažėjimas, atsiradęs dėl atsitiktinių duomenų rinkinių kiekvieno mokymo metu, o „boosting“ atveju – pagrindinis galutinės prognozės privalumas yra klaidingo klasifikavimo klaidos rodiklio sumažinimas [9]. Pagrindinės minėtų „boosting“ ir „bagging“ metodų realizacijos, naudojančios sprendimų medžius, yra atitinkamai „XGBoost“ [10] ir atsitiktiniai miškai (angl. *Random Forests*) [11]. Populiarių ansamblio metodų apibendrinimas [1] pateikiamas žemiau (žr. 1 lent.):

1 lentelė. Populiarių ansamblio metodų apibendrinimas

Metodas	Stiprybės	Silpnybės	Tipiniai panaudojimai
Bagging	Mažina rezultatų įvairovę, pagerina stabilumą	Mažiau efektyvus esant neobjektyviems (angl. <i>biased</i>) modeliams	Vaizdų klasifikavimas, lentelių duomenys
Boosting	Aukštas tikslumas, adaptyvus mokymasis	Linkęs į persimokymą (angl. <i>overfitting</i>), nuoseklus mokymasis	Sukčiavimų nustatymas, reitingavimo užduotys
Atsitiktiniai miškai (random forests)	Patikimumas, požymių surinkimas, plečiamumas	Reikalauja vietos daugelio medžių saugojimui	Medicininė diagnostika, nuotolinis stebėjimas
XGBoost	Efektyvus, geba tvarkytis su trūkstamais duomenimis, greitas mokymas	Didelis atminties naudojimas, mokymo sudėtingumas	Laiko eilučių prognozavimas, natūralios kalbos apdorojimas

1.1.3. Modelių tikslumo vertinimo metrikos

Klasifikavimo uždavinių tikslumas gali būti vertinamas atsižvelgiant į spėjimų rezultatus, naudojant apibendrintas metrikas. Tinkamos metrikos parinkimas yra pirmas žingsnis, siekiant įvertinti sprendimų, priimtų naudojant klasifikavimo algoritmą, pagrįstumą.

Klasifikavimo algoritmams vertinti dažnai naudojama maišos matrica [12] (angl. *confusion matrix*) (žr. 2 lent.), kuri aiškiai ir struktūruotai parodo klasifikavimo modelio prognozių teisingumą bei klaidas. Maišos matrica išskiria keturias pagrindines klasifikavimo baigtis:

- teisingi teigiami atvejai (angl. *true positive*, TP) – atvejai, kai modelio spėjimo reikšmė yra teigiama ir spėjimas teisingas;
- klaidingi teigiami atvejai (angl. *false positive*, FP) – atvejai, kai modelio spėjimo reikšmė yra teigiama ir spėjimas yra neteisingas;
- klaidingi neigiami atvejai (angl. *false negative*, FN) – atvejai, kai modelio spėjimo reikšmė yra neigiama ir spėjimas yra neteisingas;
- teisingi neigiami atvejai (angl. *true negative*, TN) – atvejai, kai modelio spėjimo reikšmė yra neigiama ir spėjimas yra teisingas.

2 lentelė. Maišos matricos pavyzdys

		Tikroji reikšmė	
		Teigiama	Neigiama
Spėta reikšmė	Teigiama	TP	FP
	Neigiama	FN	TN

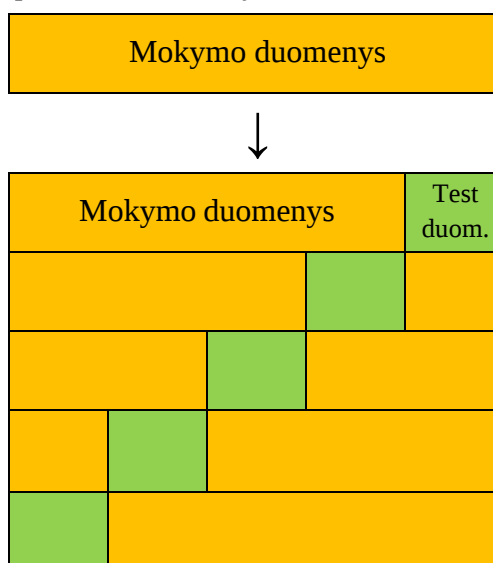
Naudojant maišos matricos rezultatus galima atlikti išsamesnį klasifikavimo kokybės įvertinimą ir išvesti kitus svarbius rodiklius, tokius kaip tikslumas (angl. *accuracy*), tikslumas (angl. *precision*), atpažinimas (angl. *recall*), F1 įvertis (angl. *F1-score*) ar kt. (žr. 3 lent.). Šie rodikliai leidžia tiksliau ir objektyviau įvertinti modelio spėjimo gebėjimus, atskleisti klasifikavimo klaidų pobūdį bei identifikuoti galimas tobulinimo kryptis.

3 lentelė. Klasifikatorių kokybės vertinimo metrikos

Metrika	Funkcija
Tikslumas	$\frac{TP + TN}{TP + FP + FN + TN}$
Klaidų dažnis	$\frac{FP + FN}{TP + FP + FN + TN}$
Atpažinimas	$\frac{TP}{TP + FN}$
Specifiškumas	$\frac{TN}{FP + TN}$
Tikslumas	$\frac{TP}{TP + FP}$
Neigiamų atpažinimų įvertis	$\frac{TN}{FN + TN}$
Neigiamų reikšmių aptikimo dažnis	$\frac{FP}{TP + FP}$
Klaidingai teigiamų atvejų dažnis	$\frac{FP}{FP + TN}$

K-kartotinių kryžminė patikra (angl. *k-fold cross validation*) [13] taikoma modelių mokymo metu, siekiant įvertinti atskirų algoritmų bei jų ansamblio tikslumą. Šis metodas atsitiktinai padalina pradinį duomenų rinkinį Y į k tarpusavyje nesikertančių ir panašaus dydžio pogrupių $Y_1, Y_2, \dots, Y_k$. Kiekvienoje iš k iteracijų, vienas pogrupis naudojamas kaip testavimo aibė, o likusieji $k-1$ pogrupiai yra apjungiami į mokymo aibę. Kryžminės patikros procesas atliekamas k kartų taip, kad kiekvienas pogrupis tiksliai vieną kartą būtų panaudotas testavimui. Dėl šios priežasties, k-kartotinių kryžminės patikros įvertinimo dispersija būna mažesnė, o gauti rezultatai – patikimesni, ypač esant ribotam duomenų kiekiui. Tačiau k-kartotinė kryžminė patikra turi ir trūkumą – didesnes skaičiavimo išlaidas, nes mokymo procesą reikia kartoti k kartų. Šiame darbe naudojama plačiai paplitusi 10-ies kartų kryžminė patikra, nes duomenų kiekis nėra labai didelis, todėl norima išnaudoti kiek įmanoma didesni jų kiekį.

4 lentelė. K-kartotinių kryžminės patikros vizualizacija



1.2. Taisyklėmis grindžiamų sistemų ir mašininio mokymosi panaudojimas finansinių operacijų gretinimui

Praktikoje, taisyklėmis grįsti metodai, tokie kaip *if/then* taisyklės, ar tam tikri iš anksto nustatyti atitikimo kriterijai, ilgą laiką buvo naudojami operacijų gretinime dėl jų skaidrumo ir lengvo suprantamumo [14]. Pavyzdžiui, taisyklė, nurodanti „jei mokėjimo išrašo suma ir data atitinka atviros sąskaitos dokumento sumą bei datą – šiuos įrašus galima gretinti“, yra lengvai suprantama, tiek žmogui, tiek kompiuteriui, ir tokio tipo taisyklės efektyviai ir dideliu tikslumu susitvarko su daugeliu paprastų atvejų. Tačiau šios sistemos susiduria su sunkumais, kai reikia prisitaikyti prie kintančių duomenų modelių ir išimčių, kurių konsultantai, ar programuotojai iš anksto nenumato.

Tradiciškai – gretinimai remiasi rankiniu darbu, arba taisyklėmis grindžiama logika, pvz. - sutampančiais likučiais, datomis, identifikaciniais numeriais [15]. Pavyzdžiui, [16] šaltinyje aprašoma, kaip daugelis organizacijų vis dar remiasi heuristiniais (taisyklėmis grindžiamais) arba rankiniais gretinimais, kurie per laiką tampa vis mažiau efektyvūs ir klaidingi, didėjant duomenų kiekiui ir skirtingumui. Šiai problemai spręsti, autoriai sukūrė „finansinių duomenų analitikos modelį“ (pagerintą atpažinimo algoritmą), kad geriau automatizuotų gretinimą. Nors tokie sprendimai duoda geresnius rezultatus – didėjant duomenims problema išlieka, ir reikalinga ieškoti kitų sprendimų jai spręsti.

Per pastaruosius keletą metų pastebima tendencija papildyti šias sistemas mašininio mokymosi atpažinimo metodais, kuriant hibridą, kuris išnaudoja taisyklių patikimumą ir mašininio mokymosi prisitaikomumą [17]. Sistemos, finansų srityje integruoja taisyklėmis grindžiamus metodus su MM, kur taisyklių sistemos tvarko pasikartojančias, struktūruotas dalis (duomenų surinkimą, perkėlimą tarp sistemų, taisyklių taikymą), o mašininio mokymosi komponentai tikrina duomenų tikslumą ir sprendžia sudėtingesnius gretinimo klausimus [18].

Konkretus hibridinės sistemos pavyzdys [19] – procesas, kuriame automatizuota sistema renka sandorius iš banko ir keityklos sistemos, o mašininio mokymosi modelis siūlo geriausiai sutampančius įrašus gretinimui. Sistema gali pirmiausia taikyti taisyklių sistemą (automatiškai sutaikant akivaizdžius atvejus, pvz., tikslas nuorodas ar tikslų sumų atitikmenis), o tada panaudoti mašininio mokymosi modelį likusiems nesugretintiems dokumentams. Mašininio mokymo algoritmai padeda sumažinti naudojamų taisyklių skaičių, paliekant svarbiausias, taip pagerinant našumą ir bendrą sistemos veikimą. Šaltinyje paminima, kad pilna automatizacija darbinėje aplinkoje nėra realizuojama, nes būtina turėti atsakingą žmogų priežiūros tikslais. Kadangi operacijų gretinimas yra jautrus ir ypač svarbus procesas – net ir hibridinėje sistemoje mašininio mokymosi algoritmai dažnai veikia kaip rekomendacijų sistema, siūlanti atitikmenis, kurias vėliau patvirtina galutinis naudotojas. Tai užtikrina audito sekos buvimą ir kontrolę, nes apskaitininkas patikrina mašinos pasiūlymus, tuo pačiu išsprendžiant atskaitomybės ir paaiškinamumo klausimus. Ilgainiui, augant pasitikėjimui modeliu, vis daugiau gretinimo pasiūlymų gali būti automatiškai patvirtinami. Taigi, turint hibridinę taisyklių ir mašininio mokymosi struktūrą, galima prie pilno automatizavimo pereiti laipsniškai.

Kitas pavyzdys – grafais grįstas mašininio mokymo modelio pritaikymas operacijų gretinimui. [20] šaltinyje pateikiamas mašininio mokymosi sprendimas, kuris veikia priešingai taisyklėmis grindžiamoms sistemoms. Autoriai pastebi, jog tradiciniai automatizuoti gretinimo metodai paprastai naudoja taisykles, arba natūralios kalbos apdorojimą (angl. *natural language processing, NLP*), kad surastų sutampančius mokėjimų aprašymus, tačiau didžiąja dalimi tokie metodai „nepaiso struktūrinių sąryšių“ duomenyse [21], pavyzdžiui, ignoruojama informacija, jog keletas mažesnių mokėjimų gali priklausyti vienai didesnei sąskaitai. Šiems trūkumams spręsti, autoriai realizavo grafomis grįstą mokymą gretinimo problemai, modeliuojant operacijas kaip mazginius taškus (angl. *nodes*) grafe ir mokant modelį aptikti įrašų sutapimus tarp banko išrašų ir ERP sistemos įrašų. Toks įrašų atpažinimo būdas parodė itin gerus rezultatus, lyginant su taisyklių/NLP metodais. Būtina paminėti, kad toks modelis puikiai susitvarkė su grupių ir daug-su-daug atpažinimo užduotimis, pavyzdžiui, kai vienas didesnis mokėjimas gali mokėti keletą mažesnių sąskaitų. Tokių užduočių paprasti taisyklėmis grindžiami metodai negalėjo išspręsti.

1.3. Statistinės analizės metodai

Apibendrintas statistinės analizės metodų panaudojimo ir rezultatų aprašymas pateiktas 5 lentelėje:

5 lentelė. Statistinės analizės metodų apibendrinimas

Testas / metodas	Kada naudojame	Ką rodo rezultatas
ANOVA	Pagrindinis testas nustatyti, ar metodai statistiškai reikšmingai skiriasi tarpusavyje	F-statistika ir $p < 0.05$ rodo, kad bent vienas metodas statistiškai reikšmingai skiriasi nuo kitų

Partial η^2 (efekto dydis)	Matuoja praktinį reikšmingumą - kiek priklausomo kintamojo variacijos paaiškinama nepriklausomu kintamuoju	0.01-0.06: mažas efektas; 0.06-0.14: vidutinis efektas; >0.14: didelis efektas
Tukey HSD	Post-hoc analizė po reikšmingo ANOVA rezultato, nustatanti, kurios konkrečios metodų poros statistiškai reikšmingai skiriasi	Koreguota p reikšmė (p_{adj}) kiekvienai porai; jei $p_{adj} < 0.05$, galime atmesti nulinę hipotezę ir priimti alternatyvią (H_1)
Porinis T-testas	Tiesioginis dviejų metodų palyginimas, kai duomenys yra susiję (pvz., tie patys subjektai)	t-statistika ir p reikšmė; pateikia panašius rezultatus kaip Tukey, bet nagrinėja kiekvieną porą atskirai, nekoreguojant dėl daugkartinio testavimo
Shapiro-Wilk	Tikrina, ar duomenų skirstinys atitinka normalųjį skirstinį	Jei $p < 0.05$, duomenys reikšmingai nukrypsta nuo normalumo ir reikėtų naudoti neparametrinius testus
Friedman / Wilcoxon	Neparametriniai testai, naudojami kai duomenys neatitinka normalumo prielaidų: Friedman - ANOVA analogas, Wilcoxon - porinio t-testo analogas	Chi-kvadrato statistika (Friedman) arba W/Z statistika (Wilcoxon) ir p reikšmė; interpretuojama panašiai kaip parametrinių testų atveju

Tolimesniuose poskyriuose aprašomi kiekvienas iš šiame magistro projekte naudojamų statistinės analizės metodų detaliau.

1.3.1. ANOVA

ANOVA (angl. *Analysis of Variance*) – tai statistinis metodas, skirtas nustatyti, ar egzistuoja statistiškai reikšmingi skirtumai tarp trijų arba daugiau nepriklausomų grupių vidurkių. [22] Pagrindinė idėja yra įvertinti grupės vidurkių skirtumų ir atsitiktinės paklaidos dispersijas. Testo statistika yra F reikšmė, kuri apskaičiuojama pagal šią formulę:

$$F = \frac{MS_{\text{tarp}}}{MS_{\text{vidus}}} = \frac{\frac{SS_{\text{tarp}}}{df_{\text{tarp}}}}{\frac{SS_{\text{vidus}}}{df_{\text{vidus}}}} \quad (1)$$

čia:

- SS_{tarp} – kvadratų suma tarp grupių (grupių vidurkių skirtumai nuo bendro vidurkio);
- SS_{vidus} – kvadratų suma grupių viduje;
- $df_{\text{tarp}}, df_{\text{vidus}}$ – laisvės laipsniai (angl. *degrees of freedom*).

Pagrindinės prielaidos: duomenys turi būti normaliai pasiskirstę, dispersijos homogeniškos (panašios tarp grupių), o stebiniai – nepriklausomi.

1.3.2. Efekto dydis η^2

Efekto dydis η^2 (dalinė eta kvadratu) [23] – tai poveikio dydžio rodiklis, kuris naudojamas dispersinėje analizėje vertinant konkretaus veiksnio poveikio stiprumą. Šis rodiklis parodo, kokią variacijos dalį paaiškina konkretus nagrinėjamas veiksnys, atmetant kitų veiksnių poveikį:

$$\eta^2 = \frac{SS_{\text{veiksny}}}{SS_{\text{veiksny}} + SS_{\text{klaidos}}} \quad (2)$$

čia:

- SS_{veiksny} – kvadratų suma, susijusi su tiriamuoju veiksnium;
- SS_{klaidos} – kvadratų suma, susijusi su paklaida (neišaiškinta variacija).

Efekto dydžio η^2 interpretavimo gairės – 0,01 – mažas efektas, 0,06 – vidutinis efektas, 0,14 ir daugiau – didelis efektas. Kuo didesnis efekto dydis – tuo tyrimuose reikia mažiau dalyvių, kad statistinė analizė parodytų, jog skirtumas tarp dviejų grupių egzistuoja.

1.3.3. Tukey HSD

Tukey HSD [24] (angl. *Honest Significant Difference*) – tai daugkartinis palyginimo metodas, naudojamas po reikšmingos ANOVA analizės (*post-hoc*), skirtas nustatyti, kurios grupių poros reikšmingai skiriasi. Testo statistika (q reikšmė) apskaičiuojama pagal formulę:

$$q = \frac{|\bar{X}_i - \bar{X}_j|}{\sqrt{\frac{MS_{\text{vidus}}}{n}}} \quad (3)$$

čia:

- $|\bar{X}_i - \bar{X}_j|$ – grupių vidurkiai;
- MS_{vidus} – vidinė kvadratų suma iš ANOVA;
- n – imties dydis kiekvienoje grupėje.

Gauta q reikšmė palyginama su kritinėmis reikšmėmis, kurios priklauso nuo grupių skaičiaus ir imties dydžių grupėse.

1.3.4. Porinis T-testas

Porinis T-testas [25] – tai parametrinis testas, naudojamas nustatyti, ar duomenų porų (priklausomų imčių) vidurkiai skiriasi statistiškai reikšmingai. Dažnai taikomas pakartotinių matavimų arba prieš – po tyrimuose. Porinio t-testo statistika yra studento t kriterijus:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \quad (4)$$

čia:

- $\bar{d}$ – porinių vidurkių skirtumas;
- s_d – skirtumų standartinis nuokrypis;
- n – porų skaičius.

Pagrindinė prielaida – porinių skirtumų pasiskirstymas turi būti artimas normalumui.

1.3.5. Shapiro-Wilk

Shapiro-Wilk testas [26] – vienas efektyviausių metodų tikrinti, ar duomenys atitinka normalųjį pasiskirstymą. Testo statistika W :

$$W = \frac{(\sum_{i=1}^n a_i x_{(i)})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (5)$$

čia:

- $x_{(i)}$ – surikiuoti duomenys;
- a_i – konstantos, apskaičiuojamos remiantis imties dydžiu;
- $\bar{x}$ – imties vidurkis.

Žema W reikšmė (ir žemas p reikšmės dydis) rodo nukrypimą nuo normalumo prielaidos.

1.3.6. Friedman

Friedman'o [27] testas – neparametrinis metodas, skirtas lyginti daugiau nei dviejų priklausomų grupių (pakartotinių matavimų) rezultatus, kai nėra tenkinamos ANOVA prielaidos. Testo statistika χ^2 :

$$\chi^2 = \frac{12}{nk(k+1)} \sum_{j=1}^k R_j^2 - 3n(k+1) \quad (6)$$

čia:

- n – subjektų (respondentų) skaičius;
- k – sąlygų skaičius;
- R_j – rangų suma kiekvienai sąlygai.

Statistiškai reikšminga χ^2 reikšmė reiškia, kad bent vienas subjektas statistiškai reikšmingai skiriasi nuo kitų.

1.3.7. Wilcoxon

Wilcoxon testas [28] – neparametrinė porinio t-testo alternatyva, naudojama palyginti dvi susietų imčių grupes, kai skirtumai neatitinka normalumo prielaidos. Pirma, skirtumai tarp porų apskaičiuojami ir ranguojami pagal jų absoliučią vertę. Testo statistika (W) gaunama sumuojant rangus, atsižvelgiant į skirtumų ženklus:

$$W = \min(W^+, W^-) \quad (7)$$

čia:

- W^+ – teigiamų skirtumų rangų suma;
- W^- – neigiamų skirtumų rangų suma.

Žema reikšmė rodo statistiškai reikšmingą skirtumą tarp palygintų grupių.

1.4. Pasirinktų technologijų pagrindimas

1.4.1. Python programavimo kalba

Darbai atlikti pasirinkta Python programavimo kalba [29], nes ji yra plačiai naudojama duomenų mokslo ir dirbtinio intelekto srityse dėl savo lankstumo, paprastumo ir gausybės specializuotų bibliotekų. Python suteikia galimybę efektyviai integruoti įvairius matematinius metodus, vykdyti duomenų analizę bei apdorojimą. Taip pat ši kalba pasižymi itin gera palaikymo infrastruktūra ir bendruomene, kas yra svarbu tiek pradiniam projektui, tiek tolimesniam vystymui bei palaikymui.

Šiame projekte naudojama biblioteka scikit-learn [30], kuri suteikia plačias galimybes lengvai ir patogiai taikyti įvairius klasifikavimo algoritmus, tokius kaip atsitiktiniai miškai (angl. Random Forest), gradientinio stiprinimo metodai (angl. Gradient Boosting), logistinės regresijos modeliai ir kt. Ši biblioteka taip pat leidžia patogiai atlikti kryžminę validaciją, vykdyti modelių vertinimą bei atlikti savybių inžineriją (angl. feature engineering).

1.4.2. Azure Machine Learning Studio

Azure Machine Learning Studio [31] pasirinktas dėl jo patogumo ir funkcionalumo, leidžiančio greitai kurti, testuoti bei įgyvendinti skirtingus mašininio mokymosi modelius. Ši platforma integruojasi su Python ir leidžia naudoti Azure ML, scikit-learn modelių, bei kitas Python bibliotekas tiesiogiai per savo aplinką. Azure Machine Learning Studio pasižymi patogia naudotojo sąsaja bei galimybe stebėti modelių veikimą realiuoju laiku, analizuoti jų rezultatus, efektyviai saugoti bei valdyti mokymo duomenų rinkinius, ar modelius.

1.4.3. Azure Functions

Azure Functions [32] pasirinkta dėl savo paprastumo, lankstumo ir gero suderinamumo su kitomis Azure platformos paslaugomis. Ši technologija leidžia lengvai kurti serverių neturinčius (angl. *serverless*) architektūros sprendimus, kurie leidžia efektyviai ir greitai apdoroti gaunamus duomenis iš Business Central, ar Azure ML be papildomos infrastruktūros nuomos. Azure Functions suteikia galimybę vykdyti automatinius duomenų apdorojimo veiksmus, tokius kaip failų išskleidimas, JSON formato duomenų tvarkymas, duomenų valymas bei transformacijos.

Naudojant Azure Functions taip pat užtikrinamas lengvas ir efektyvus duomenų perdavimas į Azure Machine Learning Studio, užtikrinant greitą integraciją tarp duomenų surinkimo, jų apdorojimo bei galutinio mašininio mokymo proceso. Debesijos architektūra, neturinti konkrečių serverių, taip pat mažina infrastruktūros išlaidas, nes leidžia automatiškai pritaikyti resursus pagal poreikį, kas užtikrina ekonomišką ir efektyvų sistemos veikimą.

1.5. Egzistuojančių rinkoje sprendimų palyginimas

1.5.1. Xero

Daugiausia taisyklių pagrindu veikianti platforma, papildyta funkcionalumu, panašiu į šiame projekte realizuojamą funkcionalumą (geriausio atpažinimo metodo parinkimu) [33] bei tam tikrais statistiniais ir neišbaigto atitikmens (angl. *fuzzy*) metodais. ~70–80 % atvejų sutampa automatiškai, likusi dalis reikalauja auditoriaus peržiūros. Banko išrašų importas veikia OFX failo importo principu, todėl pilnas automatizavimas yra ribotas.

Bendros savybės:

- automatinis tikslų atitikmenų (angl. *exact match*) suradimas pagal sumas, sąskaitų numerius ir datas;
- „*Fuzzy match*“ taisyklės dėl nedidelių teksto ar skaičių neatitikimų (pvz., skaičiaus suapvalinimo klaidos);
- neatitikę atvejai atiduodami rankiniam patvirtinimui.

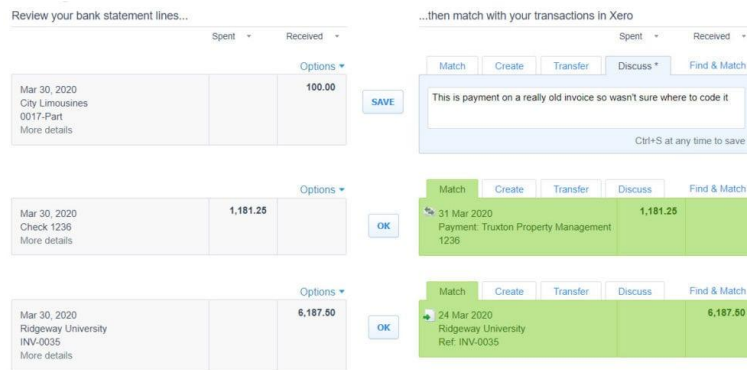
Stiprybės:

- paprasta naudoti;
- aiškiai dokumentuota;
- greita pradinė konfigūracija.

Silpnybės:

- prastai veikia daug-su-daug (angl. *many-to-many*) atveju;
- minimalus mašininio mokymosi taikymas atpažinimui;

- ribotos galimybės aptikti sudėtingesnes situacijas.



2 pav. Xero išrašų gretinimo langas

1.5.2. Intuit QuickBooks

Taisyklių ir statistinės analizės derinys, integruotas su naudotojo istorijos bei elgsenos modeliavimu. Galima automatizuoti apie 75 % darbo, kadangi pradinėje stadijoje automatinio atpažinimo neegzistuoja, reikalinga kasdieninė adaptacija per naudotojo įvestus tvirtinimus, suvedamas taisykles [34].

Bendros savybės:

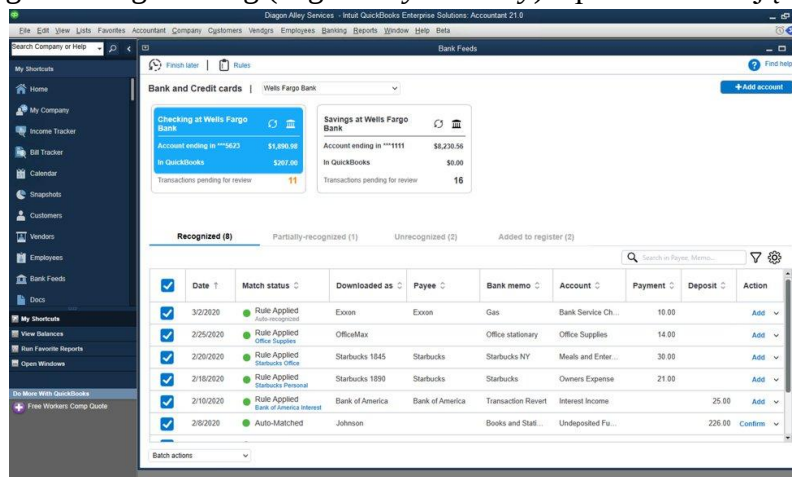
- pagrindiniai atitikmenys randami pagal taisykles, kurias susiveda naudotojas;
- veikia sprendimų priėmimo modelis, mokomas pagal naudotojo anksčiau patvirtintus ar atmestus pasiūlymus;
- pasiūlyti atitikmenys, pagal naudotojo patvirtinimus yra vertinami ir adaptuojamos papildytos taisyklės ateičiai.

Stiprybės:

- spartus mokymasis iš naudotojo elgsenos;
- geresnė adaptacija specifiniams klientų poreikiams.

Silpnybės:

- nėra pilno MM modelio – mokosi tik iš to paties kliento istorijos, todėl naujiems atvejams adaptuojasi lėtai;
- vis dar nepalengvina daug-su-daug (angl. *many-to-many*) atpažinimo atvejų.



3 pav. Quickbooks išrašų gretinimo langas

1.5.3. BlackLine

Sistemoje veikia pažangus taisyklių variklis su integruotu statistiniu ir MM komponentu, palaikantis grafų duomenų ryšius [35]. Galima automatizuoti apie 90 % banko išrašų gretinimo proceso, tačiau reikalinga jau egzistuojanti ERP sistema, su kuria BlackLine būtų integruojama. Pilnos automatizacijos vien tik su BlackLine nepavyktų įgyvendinti.

Bendros savybės:

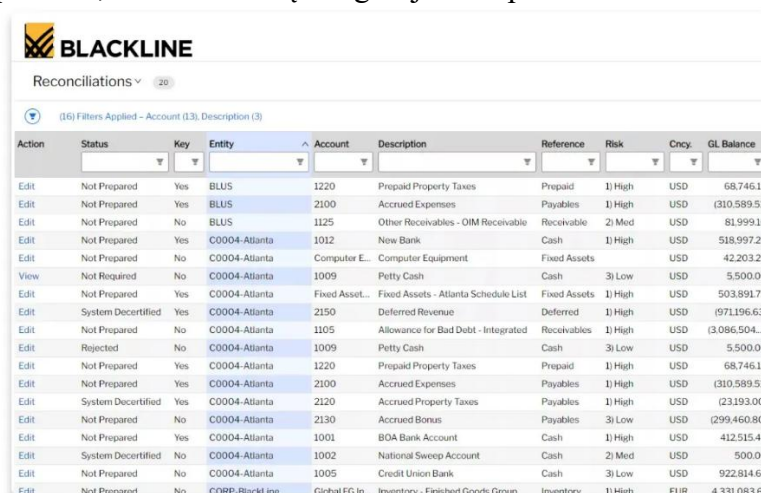
- pradinis taisyklių generavimas pagal verslo procesus;
- statistinis modelis aptinka netipinius atvejus;
- grafų duomenų bazė saugo operacijų tarpusavio ryšius, leidžia atlikti daug-su-daug atpažinimą su (angl. *graph matching*);
- interaktyvus valdymas ir auditavimo žurnalas naudotojui.

Stiprybės:

- gerai veikia sudėtingų, daug-su-daug santykių atveju;
- stipri atsekamumo ir audito integracija.

Silpnybės:

- didelės duomenų saugojimo ir skaičiavimo sąnaudos;
- brangus diegimas ir priežiūra;
- išorinio ERP poreikis, su kuriuo būtų integruojamas sprendimas.



Action	Status	Key	Entity	Account	Description	Reference	Risk	Cncl	GL Balance
Edit	Not Prepared	Yes	BLUS	1220	Prepaid Property Taxes	Prepaid	1) High	USD	68,746.14
Edit	Not Prepared	Yes	BLUS	2100	Accrued Expenses	Payables	1) High	USD	(310,589.51)
Edit	Not Prepared	No	BLUS	1125	Other Receivables - OIM Receivable	Receivable	2) Med	USD	81,999.10
Edit	Not Prepared	Yes	C0004-Atlanta	1012	New Bank	Cash	1) High	USD	518,997.24
Edit	Not Prepared	No	C0004-Atlanta	Computer E...	Computer Equipment	Fixed Assets		USD	42,203.29
View	Not Required	No	C0004-Atlanta	1009	Petty Cash	Cash	3) Low	USD	5,500.00
Edit	Not Prepared	Yes	C0004-Atlanta	Fixed Asset...	Fixed Assets - Atlanta Schedule List	Fixed Assets	1) High	USD	503,891.79
Edit	System Decertified	Yes	C0004-Atlanta	2150	Deferred Revenue	Deferred	1) High	USD	(971,196.63)
Edit	Not Prepared	No	C0004-Atlanta	1105	Allowance for Bad Debt - Integrated	Receivables	1) High	USD	(3,086,504.00)
Edit	Rejected	No	C0004-Atlanta	1009	Petty Cash	Cash	3) Low	USD	5,500.00
Edit	Not Prepared	Yes	C0004-Atlanta	1220	Prepaid Property Taxes	Prepaid	1) High	USD	68,746.14
Edit	Not Prepared	Yes	C0004-Atlanta	2100	Accrued Expenses	Payables	1) High	USD	(310,589.51)
Edit	System Decertified	Yes	C0004-Atlanta	2120	Accrued Property Taxes	Payables	1) High	USD	(23,193.00)
Edit	Not Prepared	No	C0004-Atlanta	2130	Accrued Bonus	Payables	3) Low	USD	(299,460.80)
Edit	Not Prepared	Yes	C0004-Atlanta	1001	BOA Bank Account	Cash	1) High	USD	412,515.45
Edit	System Decertified	No	C0004-Atlanta	1002	National Sweep Account	Cash	2) Med	USD	500.00
Edit	Not Prepared	No	C0004-Atlanta	1005	Credit Union Bank	Cash	3) Low	USD	922,814.62
Edit	Not Prepared	No	CORP-Blackline	Global FG In...	Inventory - Finished Goods Group	Inventory	1) High	EUR	4,331,083.61

4 pav. BlackLine išrašų gretinimo langas

1.5.4. Trintech (Cadency)

Debesijos technologijų pagrindu veikianti operacijų gretinimo platforma, apjungianti taisykles ir verslo procesų valdymą (BPM) [36]. Automatizuoti galima apie ~70 % proceso, kadangi reikia visas taisykles aprašyti pačiam, bei sistema turi būti integruojama su ERP, todėl reikalingi papildomai kaštai įdiegimui.

Bendros savybės:

- naudotojų sudarytos taisyklės ir šablonai banko operacijų porų atitikimui;
- procesų valdymo variklis automatiškai siunčia patvirtinimus, ar atmetimus atitinkamiems naudotojams.

Stiprybės:

- geras procesų valdymas;
- integracija su ERP sistemomis;
- naudotojo vaidmenų ir darbo eigos automatizavimo parinktys.

Silpnýbès:

- reikalauja sudėtingos pradinių taisyklių konfigūracijos;
- sudėtingesnių situacijų apdorojimas ribotas, kadangi reikalingos pakankamai išsamios taisyklės tam įgyvendinti.

[illegible]

5 pav. Trintech (Cadency) išrašų gretinimo langas

6 lentelėje pateikiama apibendrinta rinkoje esančių sprendimų lyginamoji analizė. Lyginamos anksčiau aprašytos sistemos su šiame projekte realizuojama sistema:

6 lentelė. Egzistuojančių technologijų ir kuriamo sprendimo palyginimas

Sprendimas	Pagrindiniai atpažinimo metodai	MM integracija	Automatinis išrašų importas	Duomenų judėjimo procesas	Automatizacijos lygis	Pagrindinės stiprybės	Pagrindinės silpnybės
Xero	Taisyklės („exact match“, „fuzzy match“) [xero]	MM atpažinimo metodas rekomendacijų pateikimui	Nėra	ERP taisyklių sistema - sumų, sąskaitų numerių ir datų palyginimas	~75 % [alphapartners]	Greita diegti ir konfigūruoti; aiškus auditavimo kelias	Negali mokytis iš duomenų; silpnas didelio sudėtingumo situacijose;
QuickBooks (Intuit)	Taisyklės + „boosting“ iš naudotojo	Mokosi iš naudotojo patvirtinimų („light“ boosting)	Yra	ERP + naudotojo grįžtamasis ryšys - sistema siūlo atitikmenis, naudotojas patvirtina/atmeta	~75 %	Prisitaiko prie specifinio kliento istorijos	Lėtas naujų scenarijų mokymasis; netaiko ansamblių metodų
BlackLine	Taisyklės + statistiniai modeliai	Statistiniai MM komponentai + grafų duomenų bazė	Nėra (duomenys ateina iš ERP)	ERP taisyklės + statistinis filtras + grafų duomenų bazės atpažinimas + naudotojo patvirtinimas	~85 %	Gera tvarkosi su daug-su-daug situacijomis; stipri audito ir atsekamumo integracija	Didelės infrastruktūros ir duomenų saugojimo sąnaudos; brangus palaikymas
Trintech Cadency	Taisyklės	Nėra	Nėra (duomenys ateina iš ERP)	ERP taisyklės + BPM darbo eiga (užduočių nusiuntimas auditoriams)	~70 %	Integruota su procesų valdymu (BPM); galima ERP sistemų integracija	Reikalauja sudėtingos pradinės konfigūracijos
Šio darbo vystoma sistema	Taisyklės + išorinių atpažinimo metodų rezultatai (MM)	Išoriniai MM atpažinimo metodai; MM meta-klasifikatorius, parenkantis geriausią atpažinimo metodą pagal istorinius duomenis	Yra	ERP taisyklės + MM metodai + meta-klasifikatorius, parenkantis geriausią atpažinimo metodą	> 85 %	Dinamiškas geriausio atpažinimo metodo pasirinkimas kiekvienam įrašui; itin lengvas įdiegimas į Business Central ERP	Vidutiniai Azure infrastruktūros kaštai

2. Hibridinio finansinių operacijų gretinimo algoritmo projektinė dalis

2.1.1. Esama padėtis

Efektyvus finansinių operacijų gretinimas su pirkimo / pardavimo sąskaitomis yra kritiškai svarbus procesas buhalterinei apskaitai ir įmonės finansų valdymui. Klaidos šiame procese gali lemti klaidingus finansinius sprendimus, neefektyvų pinigų srautų valdymą ir neteisingas finansines ataskaitas. Esama situacija rinkoje rodo, kad dauguma įmonių naudoja taisyklėmis pagrįstus metodus arba rankinį gretinimą, kuris yra ne tik imlus laikui, bet ir gali didinti apskaitos klaidų riziką, ypač esant dideliame finansinių duomenų kiekiui.

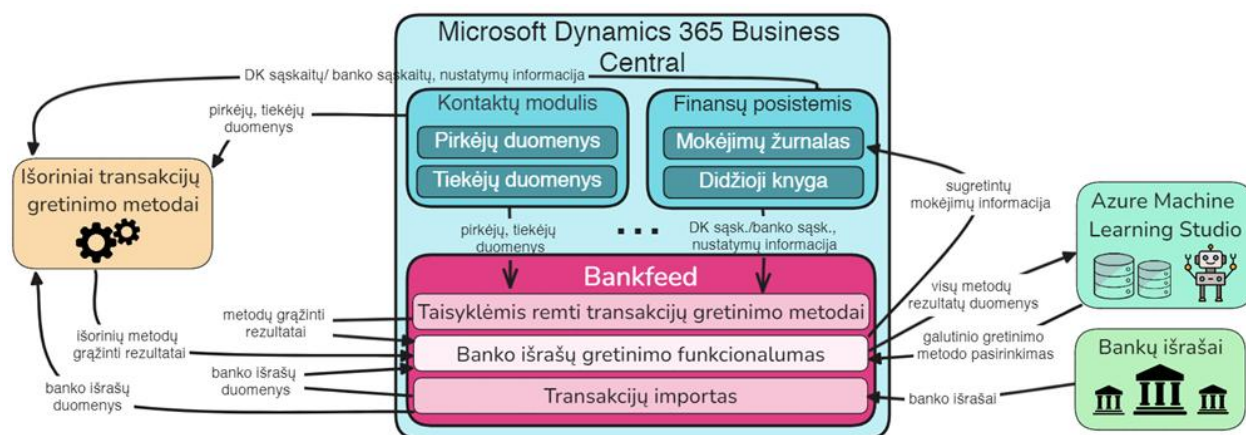
Kaip ir daugelis rinkoje esančių sistemų - „Bankfeed“ [37] šiuo metu naudoja sprendimą, kuris remiasi taisyklėmis grįstais metodais, kad automatizuotų operacijų gretinimo procesą. Dabartiniai sprendimai ne visada užtikrina aukštą tikslumą, ypač sudėtingesniais atvejais. Šiai problemai spręsti jau vystomi ir papildomi mašininio mokymosi (MM) operacijų atpažinimo metodai, kurie tam tikrose situacijose pateikia tikslesnius rezultatus. Problema, jog vienos situacijose geriau veikia MM metodai, o kitose – taisyklėmis grindžiami. Dėl šių priežasčių kilo poreikis sukurti sistemą, sujungiančią taisyklėmis grįstus ir MM metodus, kuriant hibridinį algoritmą, kuris dinamiškai parinktų geriausią gretinimo metodą konkrečiai finansinei operacijai.

2.1.2. Veiklos kontekstas

Business Central [38] sistemoje, apskaitos darbuotojai kasdieninėje veikloje registruoja pirkimo, pardavimo sąskaitas, mokesčius, susijusius su kasdiene verslo apskaita. Šie dokumentai turi atitikmenis bankinėse institucijose, tačiau kol jie nėra sudengti – nėra įrodymų sistemoje, kad už sąskaitas buvo sumokėta, ir verslo apskaita iki tol vykdoma netinkamai. Todėl yra reikalingas sudengimo žingsnis Business Central pusėje, kuomet banko išrašai būna sugretinami su anksčiau aprašytais finansiniais dokumentais.

Banko operacijų gretinimas vyksta „Bankfeed“ sprendime, kuris yra integruotas su „Microsoft Dynamics 365 Business Central“ (BC) sistema. BC aplinkoje kaupiama visa su įmonės apskaita susijusi informacija, tarp jų ir pirminių apskaitos dokumentų, pirkėjų ir tiekėjų duomenys, istorinių operacijų informacija iš mokėjimų žurnalų, ar didžiosios knygos. „Bankfeed“ produktas jau turi realizuotą bankinių operacijų importo funkcionalumą, gebantį importuoti banko išrašus iš daugelio skirtingų bankinių institucijų. Taip pat, produktas jau yra realizavęs keletą taisyklėmis grindžiamų gretinimo metodų, kurie iš esmės remiasi BC duomenų bazės informacija (pvz., pirkėjų ir tiekėjų sąskaitų ryšiai, ankstesnės operacijos, sutampantys ID, panašūs didžiosios knygos įrašai) (žr. 6 pav.).

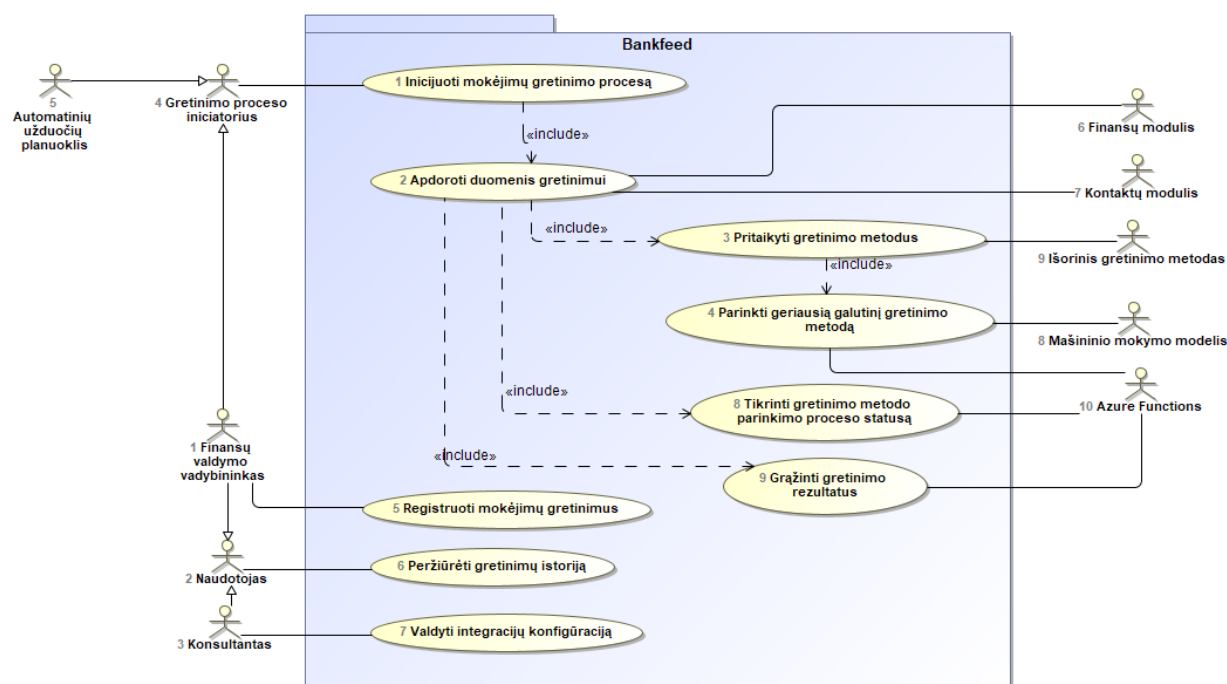
Šiame kontekste banko operacijų gretinimo funkcionalumo tikslas – nustatyti, kuriai sąskaitai (ar pirkėjui/tiekėjui) turėtų būti priskirta konkreti banko operacija. Kadangi „Bankfeed“ jau turi keletą taisyklėmis grindžiamų gretinimo metodų ir yra kuriami išoriniai MM gretinimo metodai, šiame projekte siekiama sukurti sprendimą, išrenkantį tinkamiausią esamą metodą pagal esamą kontekstą (banko išrašo savybes, istorinius duomenis ir pan.), o ne kurti dar vieną naują gretinimo sprendimą. Kitaip tariant, vykstant visam procesui, kiekvienam gretinimo metodui pateikus savo rezultatus, mašininio mokymosi pagrindu būtų priimamas galutinis sprendimas, nurodantis, kuris iš metodų yra tiksliausias konkrečiai operacijai.



6 pav. Hibridinio finansinių operacijų gretinimo algoritmo veiklos konteksto diagrama

2.1.3. Sistemos panaudojimo atvejų modelis

Hibridinio finansinių operacijų gretinimo algoritmo panaudojimo atvejai atvaizduoti diagramoje (žr. 7 pav.)



7 pav. Hibridinio finansinių operacijų gretinimo algoritmo panaudojimo atvejų diagrama

Funkcionalumas apima mokėjimų gretinimo proceso eigą nuo gretinimo inicijavimo iki rezultatų registravimo ir istorijos peržiūros. Procesuose dalyvauja šie aktoriai:

- finansų valdymo vadybininkas – pagrindinis sistemos naudotojas, kuris gali aktyvuoti mokėjimų gretinimą, registruoti gretinimus, peržiūrėti gretinimų istoriją;
- automatinis užduočių planuoklis – paveldi iniciatoriaus rolę ir gali periodiškai paleisti gretinimo procesus be tiesioginio naudotojų įsikišimo;
- naudotojas – sistemos naudotojas galintis peržiūrėti gretinimų istoriją;

- konsultantas – sistemos naudotojas, turintis galimybę peržiūrėti ankstesnių gretinimų istoriją bei valdyti integracijų konfigūraciją;
- išorinės sistemos ir moduliai – finansų modulis, kontaktų modulis, išorinis greitinimo metodas, mašininio mokymo modelis bei „Azure Functions“, kurie reikalingi duomenų gavimo bei jų apdorojimo, gretinimo sprendimų priėmimo procesų veikimui.

Hibridinio finansinių operacijų gretinimo algoritmo projekto panaudojimo atvejai detalizuojami 7-15 lentelėse:

7 lentelė. PA1 Inicijuoti mokėjimų gretinimo procesą

PA1 Inicijuoti mokėjimų gretinimo procesą	
Aktorius:	Finansų valdymo vadybininkas, automatinis užduočių planuoklis
Aprašas:	Vadybininkas arba automatinis planuoklis inicijuoja mokėjimų gretinimo procesą "Bankfeed" sistemoje
Prieš sąlyga:	Vadybininkas prisijungęs, įgalintos sistemos funkcijos, sukonfigūruotos integracijos ir paleistas naudoti produktas
Sužadinimo sąlyga:	Pasirenkama gretinimo proceso pradžios parinktis
Po-sąlyga:	Inicijuotas duomenų apdorojimo ir gretinimo procesas

8 lentelė. PA2 Apdoroti duomenis gretinimui

PA2 Apdoroti duomenis gretinimui	
Aktorius:	Sistema, kontaktų modulis, finansų modulis
Aprašas:	Reikalingų gretinimui duomenų surinkimas iš klientų, tiekėjų, ir banko sąskaitų, paruošiant duomenis tolesniam apdorojimui
Prieš sąlyga:	Inicijuotas mokėjimų gretinimo procesas
Sužadinimo sąlyga:	Duomenų surinkimo procesas pradėtas
Po-sąlyga:	Surinkti ir suformatuoti duomenys, reikalingi gretinimui, perduoti tolesniam apdorojimui

9 lentelė. PA3 Pritaikyti gretinimo metodus

PA3 Pritaikyti gretinimo metodus	
Aktorius:	Sistema, išoriniai gretinimo metodai
Aprašas:	Taisyklių pagrindu veikiantys (pvz., E2E ID atitikimo paieška), ar išoriniai gretinimo metodai (pvz. MM) pritaikomi surinktiems duomenims
Prieš sąlyga:	Surinkti duomenys gretinimui
Sužadinimo sąlyga:	Pradedamas gretinimo procesas
Po-sąlyga:	Gretinimo rezultatai iš atpažinimo metodų surinkti ir perduoti

10 lentelė. PA4 Parinkti geriausią galutinį gretinimo metodą

PA4 Parinkti geriausią galutinį gretinimo metodą	
Aktorius:	Mašininio mokymo modelis
Aprašas:	Apima procesą, kurio metu MM modelis parenka geriausią atpažinimo metodą, atsižvelgdamas į visų pritaikytų metodų rezultatus, jį grąžina į sistemą
Prieš sąlyga:	Surinkti rezultatai iš taisyklių ir išorinių metodų
Sužadinimo sąlyga:	Pradedamas galutinio metodo parinkimo procesas
Po-sąlyga:	Išrinktas geriausias gretinimo metodas, rezultatas perduotas į sistemą ir išsaugotas

11 lentelė. PA5 Registruoti mokėjimų gretinimus

PA5 Registruoti mokėjimų gretinimus	
Aktorius:	Finansų valdymo vadybininkas
Aprašas:	Apima procesą, kurio metu galutinis gretinimo rezultatas registruojamas sistemoje
Prieš sąlyga:	Parinktas galutinis gretinimo metodas
Sužadinimo sąlyga:	Pradedamas registravimo procesas
Po-sąlyga:	Gretinimo rezultatai registruoti sistemoje, informacija prieinama peržiūrai ir auditui

12 lentelė. PA6 Peržiūrėti gretinimo istoriją

PA6 Peržiūrėti gretinimo istoriją	
Aktorius:	Finansų valdymo vadybininkas, konsultantas
Aprašas:	Apima procesą, kurio metu naudotojas peržiūri ankstesnių operacijų gretinimų istoriją, gali pasirinkti peržiūrėti norimą gretinimą detaliau
Prieš sąlyga:	Yra užregistruotų gretinimų
Sužadinimo sąlyga:	Naudotojas pasirenka gretinimo istorijos peržiūros parinktį
Po-sąlyga:	Peržiūros istorija atidaryta, galima rinktis gretinimus peržiūrai ir analizei

13 lentelė. PA7 Valdyti integracijų konfigūraciją

PA7 Valdyti integracijų konfigūraciją	
Aktorius:	Konsultantas
Aprašas:	Apima procesą, kurio metu konsultantas valdo integracijų konfigūraciją, patikrina, ar funkcionalumui reikalingos integracijos yra veikiančios
Prieš sąlyga:	Konsultantas prisijungęs prie sistemos
Sužadinimo sąlyga:	Naudotojas pasirenka integracijų konfigūravimo parinktį
Po-sąlyga:	Pateikiamas integracijų konfigūravimo puslapis

14 lentelė. PA8 Tikrinti gretinimo metodo parinkimo proceso statusą

PA8 Tikrinti gretinimo metodo parinkimo proceso statusą	
Aktorius:	Azure Functions
Aprašas:	Apima procesą, kurio metu sistema siunčia užklausą į išorinę sistemą, kad gauti metodo parinkimo proceso statusą
Prieš sąlyga:	Turi būti pradėtas gretinimo procesas
Sužadinimo sąlyga:	Automatinė funkcija kas 3 minutes iškviečia statuso tikrinimą
Po-sąlyga:	Pateikiamas gretinimo proceso statusas

15 lentelė. PA9 Grąžinti gretinimo rezultatus

PA9 Grąžinti gretinimo rezultatus	
Aktorius:	Azure Functions
Aprašas:	Apima procesą, kurio metu sistema siunčia užklausą į išorinę sistemą, kad gauti galutinio atpažinimo metodo parinkimo rezultatus
Prieš sąlyga:	Turi būti pabaigtas gretinimo procesas
Sužadinimo sąlyga:	Sistema iškviečia funkciją, kai gauna statusą, jog gretinimo procesas yra baigtas
Po-sąlyga:	Grąžinami galutinio atpažinimo metodo parinkimo rezultatai

2.2. Funkciniai reikalavimai

Pateiktose 16-20 lentelėse specifikuoti pagrindiniai sistemos funkciniai reikalavimai:

16 lentelė. Funkcinis reikalavimas 1 – naudotojo sąsaja

Reikalavimo nr:	1	Reikalavimo tipas:	9	PA nr:	1
Aprašymas:	Sistema turi suteikti naudotojo sąsają finansų vadovams, skirta inicijuoti mokėjimų gretinimo užklausas, įskaitant atitinkamų banko sąskaitų pasirinkimą.				
Pagrindimas:	Finansų vadovams reikia patogios naudotojo sąsajos, kuri leistų naudoti automatinį sąskaitų gretinimo funkcionalumą.				
Šaltinis:	Finansų valdymo vadybininkas				
Tinkamumo kriterijus:	Mokėjimų gretinimo užklauso forma pasiekama iš pagrindinio sistemos meniu, su galimybe užpildyti visus reikalingus laukus.				
Užsakovo tenkinimas:	3		Užsakovo netenkinimas:	5	
Priklausomybės:	-		Konfliktai:	-	
Istorija:	Užregistruotas 2024-03-28				

17 lentelė. Funkcinis reikalavimas 2 – „Business Central“ duomenų apdorojimas

Reikalavimo nr:	2	Reikalavimo tipas:	9	PA nr:	2,3
Aprašymas:	Sistema turi apdoroti banko išrašų ir klientų/tiekėjų duomenis iš „Business Central“ modulių, juos paruošiant gretinimui su taisyklėmis grindžiamais ir išoriniais metodais.				
Pagrindimas:	Duomenų apdorojimas būtinas tiksliems mokėjimų gretinimo rezultatams generuoti.				
Šaltinis:	Sistemos analitikas				
Tinkamumo kriterijus:	Sistema sėkmingai išgauna ir apdoroja reikalingus duomenis iš Business Central be gaunamų klaidų.				
Užsakovo tenkinimas:		2	Užsakovo netenkinimas:		4
Priklausomybės:		Prieiga prie finansų, kontaktų duomenų Business Central.	Konfliktai:		-
Istorija:		Užregistruotas 2024-03-28			

18 lentelė. Funkcinis reikalavimas 3 – mašininio mokymo rezultatų apdorojimas

Reikalavimo nr:	3	Reikalavimo tipas:	9	PA nr:	4,6,9
Aprašymas:	Sistema turi apdoroti gautus mašininio mokymo modelio rezultatus ir juos atvaizduoti Business Central sąsajoje.				
Pagrindimas:	Mašininio mokymo rezultatų rodymas leidžia naudotojams suprasti gautus rezultatus ir veikti pagal juos.				
Šaltinis:	Sistemos analitikas				
Tinkamumo kriterijus:	Mašininio mokymo rezultatai yra teisingai apdorojami, intuityviai atvaizduojami Business Central sąsajoje.				
Užsakovo tenkinimas:		4	Užsakovo netenkinimas:		5
Priklausomybės:		-	Konfliktai:		-
Istorija:		Užregistruotas 2024-03-28			

19 lentelė. Funkcinis reikalavimas 4 – gretinimo istorijos peržiūra

Reikalavimo nr:	4	Reikalavimo tipas:	9	PA nr:	6
------------------------	---	---------------------------	---	---------------	---

Aprašymas:	Sistema turi suteikti finansų vadovams ir konsultantams galimybę peržiūrėti gretinimo istoriją, įskaitant praėjusių gretinimų rezultatus.		
Pagrindimas:	Prognozavimo istorijos peržiūra leidžia naudotojams sekti gretinimo modelio veiksmingumą ir įvertinti tikslumą.		
Šaltinis:	Finansų valdymo vadybininkas, konsultantas		
Tinkamumo kriterijus:	Naudotojai gali pasiekti atskirą gretinimo istorijos knygą Business Central sąsajoje, jame matoma visa ankstesnių gretinimų informacija.		
Užsakovo tenkinimas:	2	Užsakovo netenkinimas:	3
Priklausomybės:	-	Konfliktai:	-
Istorija:	Užregistruotas 2024-03-28		

20 lentelė. Funkcinis reikalavimas 5 – integracijos konfigūracijos valdymas

Reikalavimo nr:	5	Reikalavimo tipas:	9	PA nr:	7
Aprašymas:	Sistema turi leisti konsultantams valdyti integracijos konfigūracijos nustatymus, įskaitant duomenų pateikimo dažnumą, modelio mokymo intervalus ir API kredencialų konfigūracijas.				
Pagrindimas:	Galimybė keisti integracijų konfigūravimą leidžia konsultantams pritaikyti integraciją pagal konkrečius kliento reikalavimus ir pageidavimus.				
Šaltinis:	Konsultantas				
Tinkamumo kriterijus:	Konsultantai gali pasiekti ir modifikuoti integracijos konfigūracijos nustatymus per atskirą administravimo sąsają, pakeitimai yra sėkmingai pritaikomi be įtakos sistemos stabilumui.				
Užsakovo tenkinimas:	1		Užsakovo netenkinimas:	3	
Priklausomybės:	-		Konfliktai:	-	
Istorija:	Užregistruotas 2024-03-28				

2.3. Nefunkciniai reikalavimai

Pilnas nefunkcinių reikalavimų sąrašas buvo pateiktas magistro studijų metu, toliau 21-25 lentelėse išskiriami tik aktualiausi sistemos nefunkciniai reikalavimai:

21 lentelė. Nefunkcinis reikalavimas 6 – funkcionalumo kalba

Reikalavimo nr:	6	Reikalavimo tipas:	11	PA nr:	Visi
Aprašymas:	Naujasis funkcionalumas turi veikti lietuvių ir anglų kalbomis				
Pagrindimas:	„Bankfeed” produktas yra pritaikytas lietuvių ir anglų kalboms, todėl ir naujasis funkcionalumas turi palaikyti abi kalbas, siekiant išlaikyti vientisumą				
Šaltinis:	Sistemos analitikas				
Atitikimo kriterijus:	Visi sistemos elementai turi būti aprašyti lietuvių ir anglų kalbomis, patikrinama pasikeitus sistemos kalbą.				
Užsakovo tenkinimas:	5	Užsakovo netenkinimas:	5		
Prioritetas:	vidutinis	Konfliktai:	Nėra		
Istorija:	Sukurtas: 2024-04-01				

22 lentelė. Nefunkcinis reikalavimas 7 – klaidos pranešimai

Reikalavimo nr:	7	Reikalavimo tipas:	11	PA nr:	6,7
Aprašymas:	Sistema turi teikti aiškius ir lengvai suprantamus klaidos pranešimus naudotojams.				

Pagrindimas:	Suprantami klaidos pranešimai padeda naudotojams išspręsti kilusias problemas sistemos naudojimosi metu.		
Šaltinis:	Sistemos analitikas		
Atitikimo kriterijus:	Klaidos pranešimai turi būti konkretūs, suprantami naudotojams ir suteikti aiškius žingsnius klaidos šalinimui.		
Užsakovo tenkinimas:	4	Užsakovo netenkinimas:	3
Prioritetas:	vidutinis	Konfliktai:	Nėra
Istorija:	Sukurtas: 2024-04-01		

23 lentelė. Nefunkcinis reikalavimas 8 – algoritmo veikimo trukmė

Reikalavimo nr:	8	Reikalavimo tipas:	12	PA nr:	1,5,6
Aprašymas:	Bendras taisyklėmis grindžiamų metodų, išorinių modelių bei pačio meta-modelio veikimas galutiniam banko išrašų gretinimui neturėtų užtrukti ilgiau 30 minučių.				
Pagrindimas:	Greitas foninių sistemos užduočių vykdymas pagerina naudojimosi patirtį ir bendrą sistemos našumą.				
Šaltinis:	Sistemos analitikas				
Atitikimo kriterijus:	Turi būti gaunamas sugretintų operacijų paketas per ne daugiau kaip 30 minučių				
Užsakovo tenkinimas:	3	Užsakovo netenkinimas:	2		
Prioritetas:	vidutinis	Konfliktai:	Nėra		
Istorija:	Sukurtas: 2024-04-02				

24 lentelė. Nefunkcinis reikalavimas 9 – integracijos

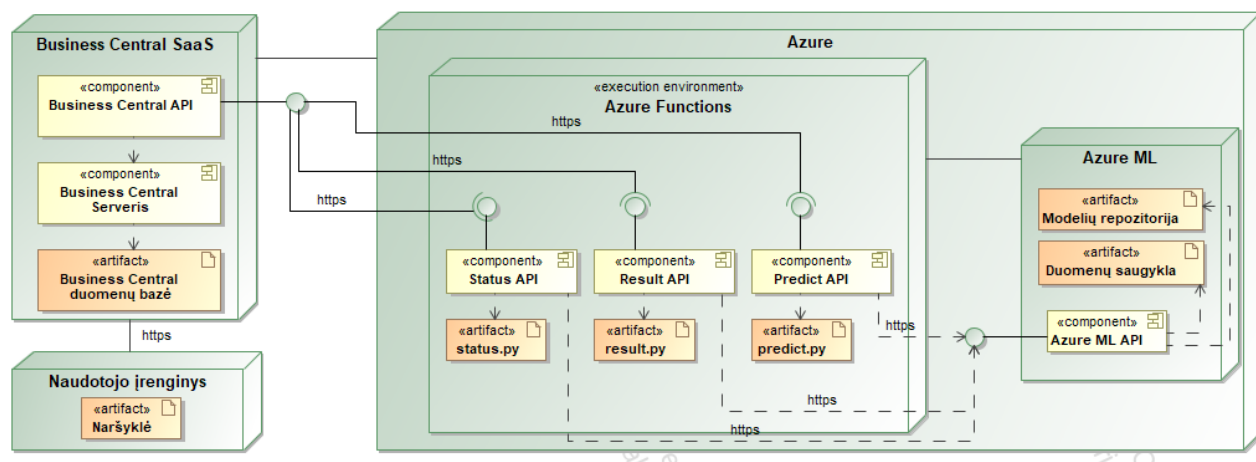
Reikalavimo nr:	9	Reikalavimo tipas:	13	PA nr:	4,5
Aprašymas:	Sistema turi būti integruota su mašininio mokymosi meta-modeliu.				
Pagrindimas:	Integracija su mašininio mokymosi modeliu užtikrins esminį kuriamo funkcionalumo veikimą.				
Šaltinis:	Sistemos analitikas				
Atitikimo kriterijus:	Sistema turi palaikyti integraciją su Azure Machine Learning Studio				
Užsakovo tenkinimas:	3	Užsakovo netenkinimas:	5		
Prioritetas:	vidutinis	Konfliktai:	Nėra		
Istorija:	Sukurtas: 2024-04-02				

25 lentelė. Nefunkcinis reikalavimas 10 – integracijos konfigūracijos valdymo teisės

Reikalavimo nr:	10	Reikalavimo tipas:	15	PA nr:	7
Aprašymas:	Tik konsultantai ir paskirti kliento įmonės naudotojai turi turėti galimybes keisti integracijos konfigūraciją.				
Pagrindimas:	Integracijos konfigūracijos redagavimą turi galėti atlikti tik konsultantai ir paskirti kliento įmonės naudotojai, kad paprasti naudotojai netrikdytų funkcionalumo pakeičiant jiems nežinomus kintamuosius.				
Šaltinis:	Sistemos analitikas				
Atitikimo kriterijus:	Turi būti sukurtas atskiras, konfigūravimui skirtas teisių rinkinys jis priskirtas tik tam tikriems sistemos naudotojams.				
Užsakovo tenkinimas:	2	Užsakovo netenkinimas:	5		
Prioritetas:	vidutinis	Konfliktai:	Nėra		
Istorija:	Sukurtas: 2024-04-02				

2.4. Sistemos išdėstymo vaizdas

Toliau pavaizduotas sistemos išdėstymas diegimo diagramoje (žr. 8 pav.). Matomi ryšiai tarp projekto komponentų, jų išdėstymas.



8 pav. Hibridinio finansinių operacijų gretinimo algoritmo diegimo diagrama

Naudotojas inicijuoja gretinimo procesą per interneto naršyklę, o naršyklė jungiasi prie Business Central SaaS debesijos paslaugų aplikacijos, naudodant HTTPS protokolą. Business Central mazge galimi matyti aktualūs komponentai – Business Central serveris, vykdamas verslo logiką, ir Business Central API, naudojamas integracijai su Azure Functions, išsiunčiant mokymui bei spėjimų gavimui reikalingus duomenis, gaunant atsakymus dėl gretinimo statuso, ar spėjimo rezultatų.

Azure Functions mazge išskirti trys pagrindiniai komponentai, kurie aprašyti 26 lentelėje:

26 lentelė. Azure Functions komponentų aprašymas

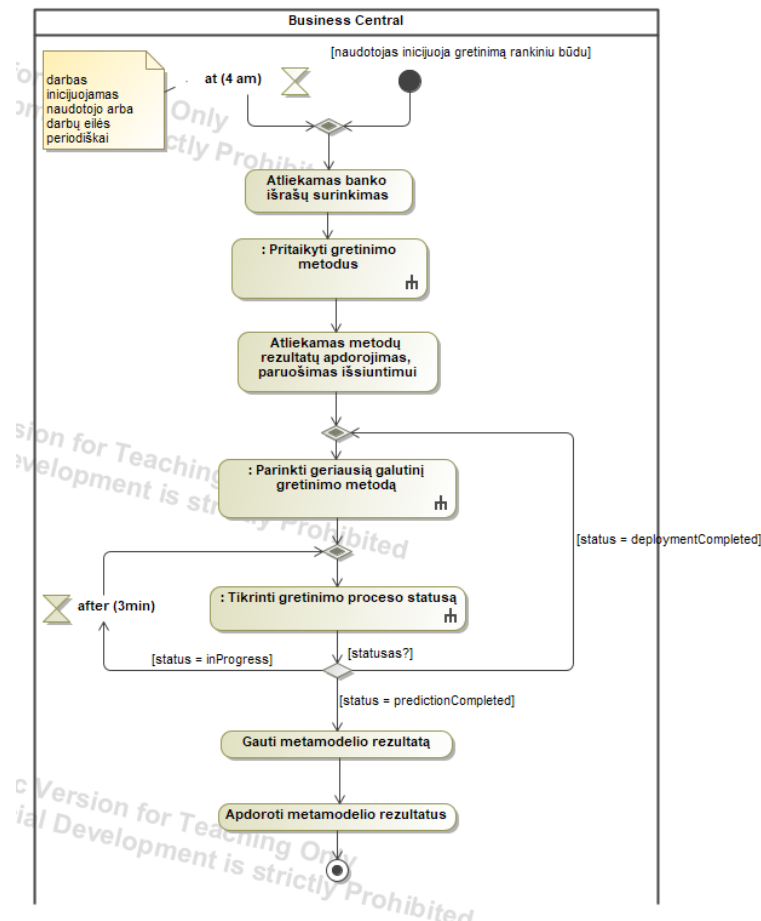
Komponentas	Failas	Paskirtis
Result	result.py	Kreipiasi į Azure ML, kad gautų gretinimo rezultatų duomenis, juos tinkamai apdoroja grąžinimui į Business Central.
Status	status.py	Gauna ir grąžina informaciją apie mokinamo modelio, ar duomenų spėjimo procesą būsenas.
Predict	predict.py	Siunčia užklausą į Azure ML naujiems modeliams mokytis iš istorinių duomenų, arba, jeigu modelis jau mokytas – spėjimams gauti.

Azure Machine Learning (Azure ML) mazge išskirti modelių repozitorijos bei kliento duomenų saugyklos artefaktai. Azure ML naudojimas leidžia patogiai versijuoti tam tikros įmonės mokytus modelius, esant poreikiui grįžti prie anksčiau mokyto modelio. Taip pat, čia saugomi visų spėjimų metaduomenys, rezultatai. Specialistas, norėdamas dirbti tik prie mašininio mokymo modelio funkcionalumo, gali tą atlikti be prisijungimo prie Business Central. Azure ML API komponentas leidžia išorinėms paslaugoms užregistruoti naujus modelius, paleisti mokymą, iškviesti modelio spėjimus.

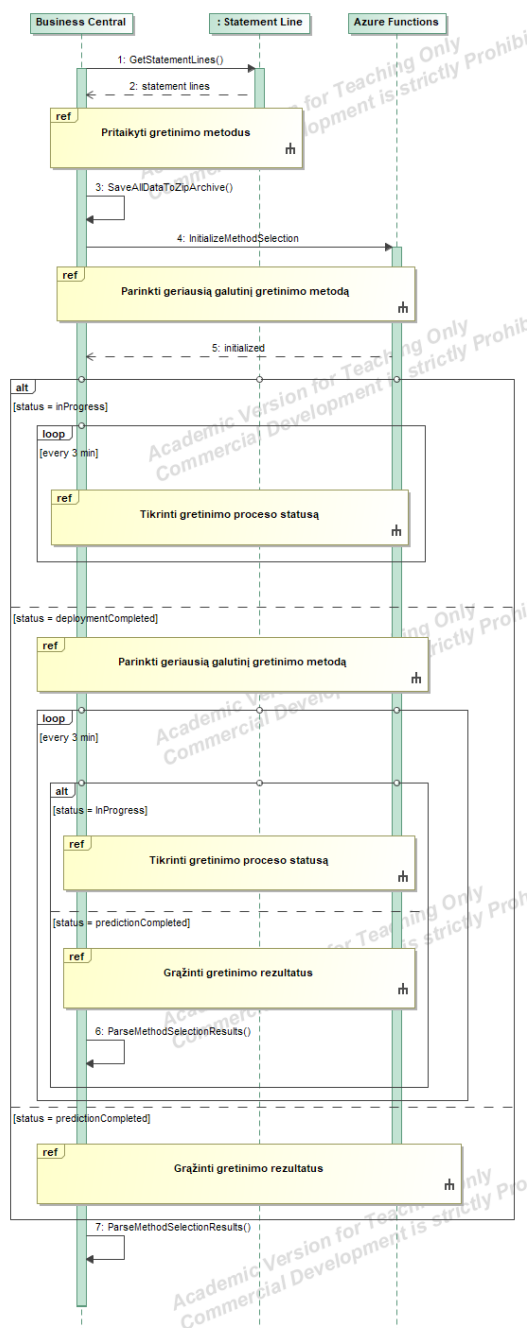
2.5. Sistemos architektūra

Duomenų apdorojimo gretinimui (žr. 9 - 10 pav.) procesas prasideda nuo visų reikalingų apdoroti banko išrašo eilučių surinkimo ir suspaudimo į vieną ZIP archyvą. Toliau, minėtas ZIP failas yra siunčiamas į išorinę „Azure Functions“ funkciją, gaudamas atgal proceso numerį *JobID*, pagal kurį

vėliau bus galima tikrinti proceso būseną, ar parsisiųsti skaičiavimo rezultatus. Toliau, reguliariai siunčiamos statuso užklausos, tikrinančios, ar atpažinimo užduotis jau baigta (būsena „*PredictionCompleted*“). Jei procesas dar nebaigtas – yra laukiama ir tikrinama iš naujo, kol pasiekiamas galutinis rezultatas. Kai gaunama informacija, kad procesas baigtas – modulis siunčia užklausą gauti atpažinimo duomenis – masyvą su kiekvienos eilutės identifikatoriumi ir skirtingų atpažinimo metodų įvertinimais. Gauti atsakymai yra apdorojami ir išsaugojami reikiamoje lentelėje, kad rezultatai ateityje būtų matomi naudotojui. Taip visas ciklas – nuo duomenų surinkimo iki atpažinimo rezultatų įrašymo – vyksta automatiškai be tiesioginio žmogaus indėlio.



9 pav. Panaudojimo atvejo „Apdoroti duomenis gretinimui“ veiklos diagrama

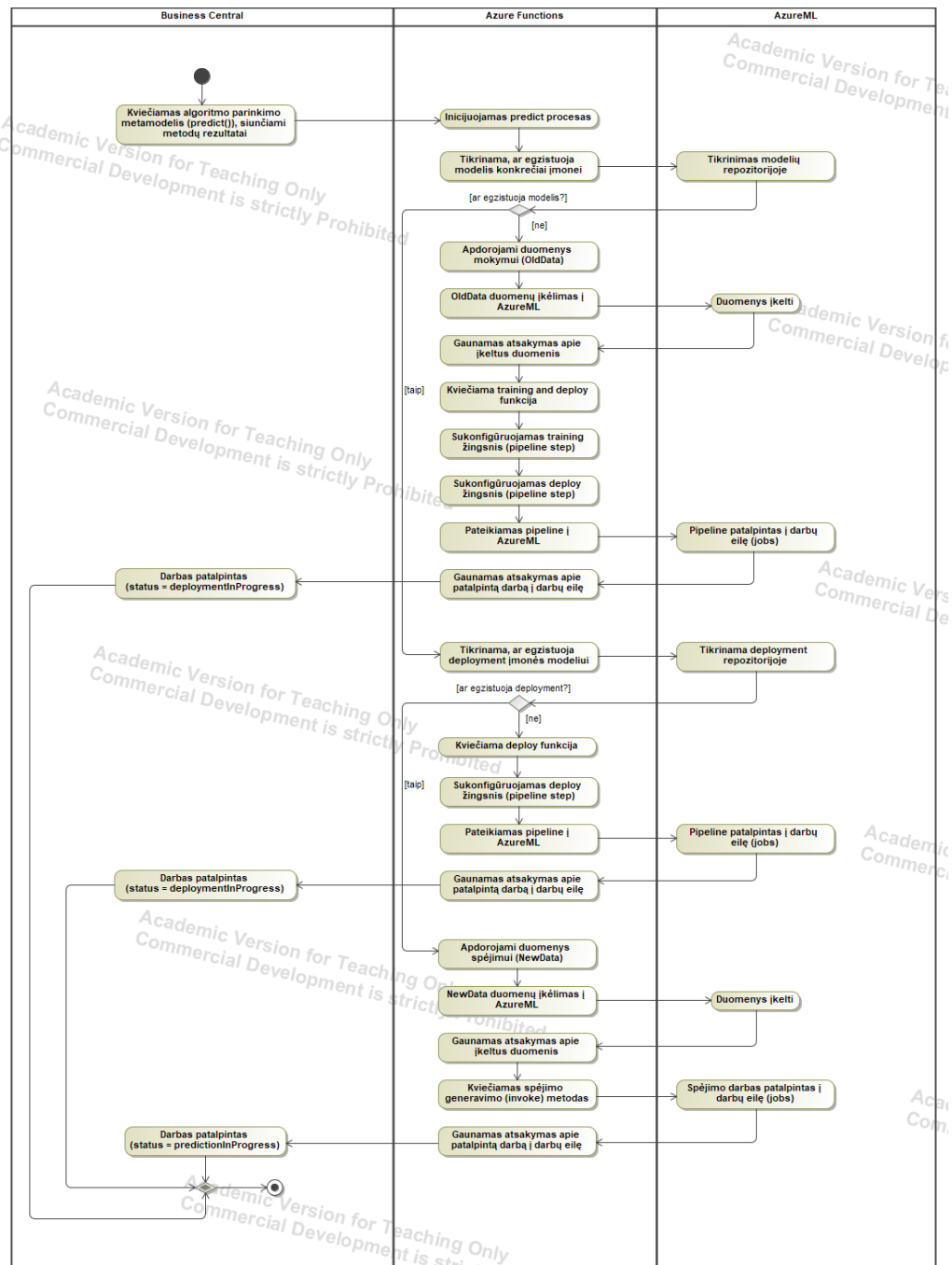


10 pav. Panaudojimo atvejo „Apdoroti duomenis gretinimui“ sekos diagrama

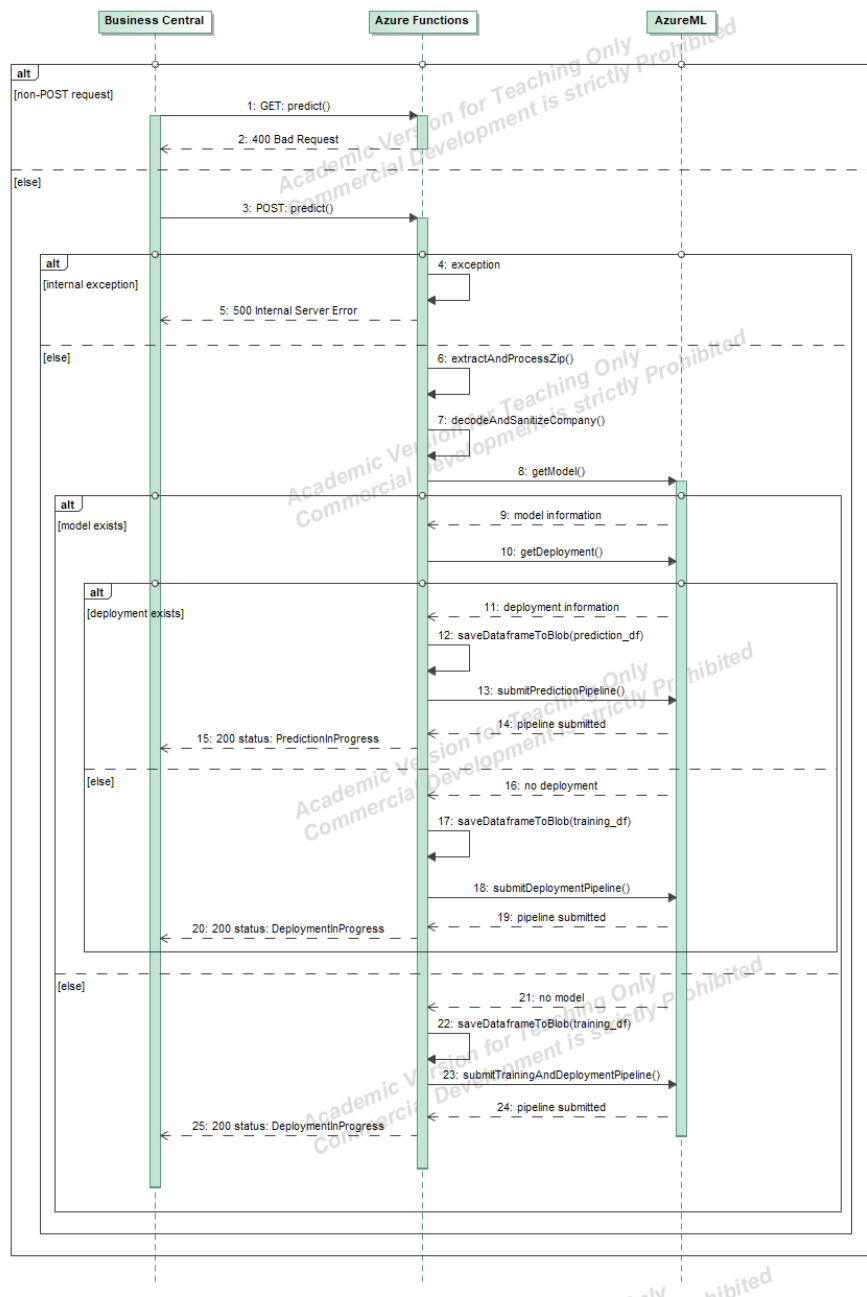
Geriausio galutinio gretinimo metodo parinkimo procese (žr. 11 – 12 pav.), „Azure Functions“ pirmiausia išskiria duomenis, kurie buvo perduoti modelio mokymui, o kurie prognozavimui, vėliau tikrina, ar tam tikrai įmonei modelis jau yra mokytytas Azure ML platformoje:

- jei modelio dar nėra, sukuriami nauji užduoties Azure ML sistemoje, duomenys nusiųsti mašininio mokymo modelio mokymui bei diegimui;
- jeigu tam tikrai įmonei modelis jau yra mokytytas, bet jis dar neįdiegtas – tada atliekamas tik diegimo žingsnis;
- jeigu tam tikrai įmonei modelis ir diegiamas jau egzistuoja – perduodami nauji duomenys prognozėms atlikti.

Galutinis proceso rezultatas – grąžinama užduoties būsena, priklausomai nuo atliekamo proceso (mokymas/diegimas/prognozė).

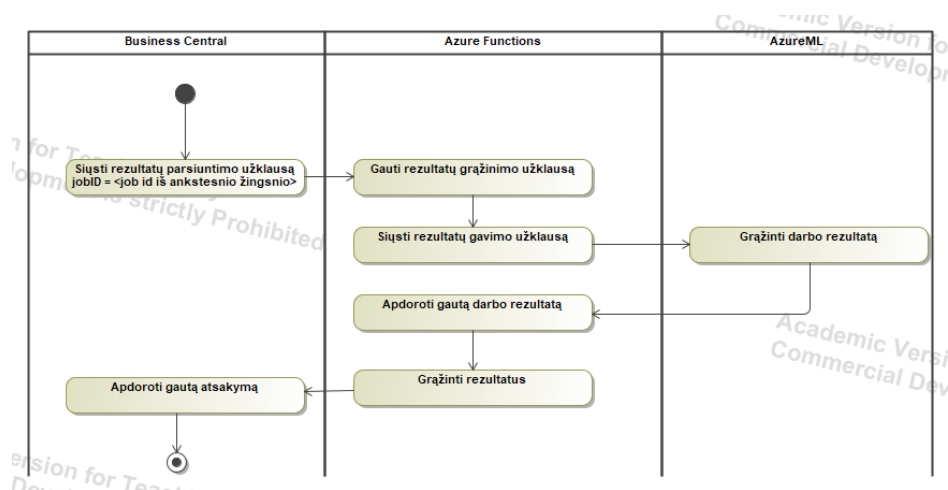


11 pav. Panaudojimo atvejo „Parinkti geriausią galutinio gretinimo metodą“ veiklos diagrama

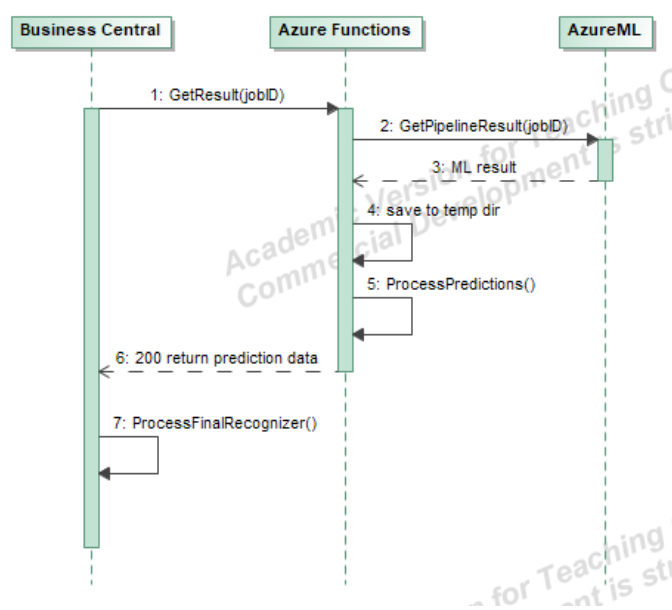


12 pav. Panaudojimo atvejo „Parinkti geriausią galutinio gretinimo metodą“ sekos diagrama

Gretinimo rezultatų grąžinimo procesas (žr. 13 – 14 pav.), gavęs užklausą su užduoties identifikatoriumi *JobID*, prisijungia prie Azure ML platformos ir bando parsisiųsti anksčiau paleistos analizės rezultatą. Sėkmingai parsisiuntus CSV, eilutė po eilutės nuskaityti duomenys apie skirtingų atpažinimo metodų (*Default* ir *External*) tikslumą ir tikimybę. Kiekvienam įrašui pasirenkama, kuris metodas pasirodė geriausias, ir surenkamas galutinis atsakymas. Gauti rezultatai su kiekvieno eilutės sprendimu bei papildoma informacija grąžinami JSON formatu į Business Central, kur yra toliau apdorojami.

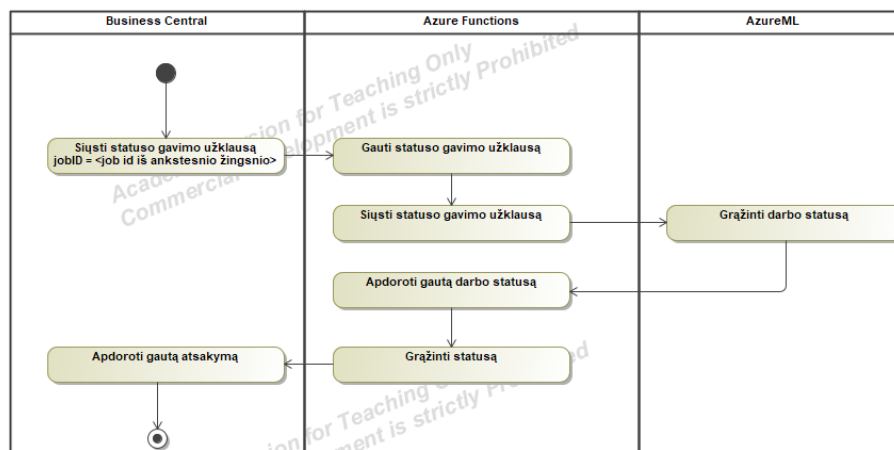


13 pav. Panaudojimo atvejo „Grąžinti gretinimo rezultatus“ veiklos diagrama

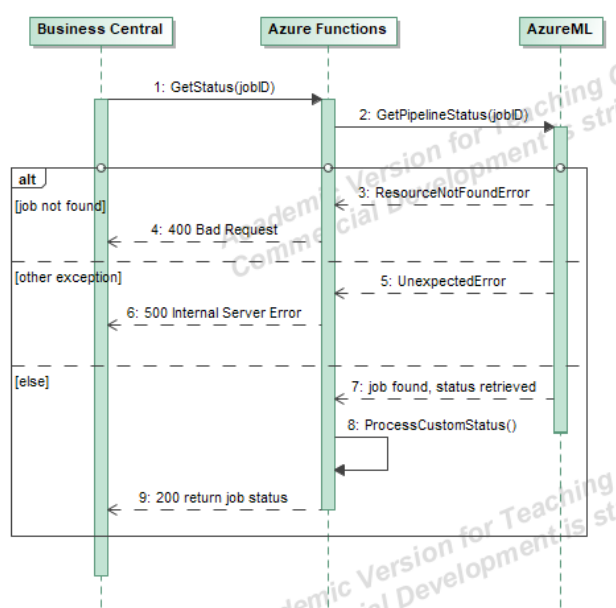


14 pav. Panaudojimo atvejo „Grąžinti gretinimo rezultatus“ sekos diagrama

Gretinimo statuso tikrinimo procese (žr. 15 – 16 pav.), Business Central sistema nusiunčia užklausą su darbo identifikatoriumi *JobID* į Azure Functions funkciją, kuri autentifikuoja „Azure ML“ platformoje iš kreipiasi dėl darbo būsenos grąžinimo pagal gautą *JobID*. Gautas statusas iš Azure ML yra papildomai apdorojamas – priskiriamos būsenos (pvz., „*InProgress*“, „*PredictionCompleted*“, „*Error*“ ir kt.). Galiausiai į Business Central grąžinamas atsakymas su apdorota darbo būsena arba klaidos pranešimu (jei darbas nerastas pagal *JobID* arba įvyko nenumatyta klaida).



15 pav. Panaudojimo atvejo „Tikrinti gretinimo proceso statusą“ veiklos diagrama



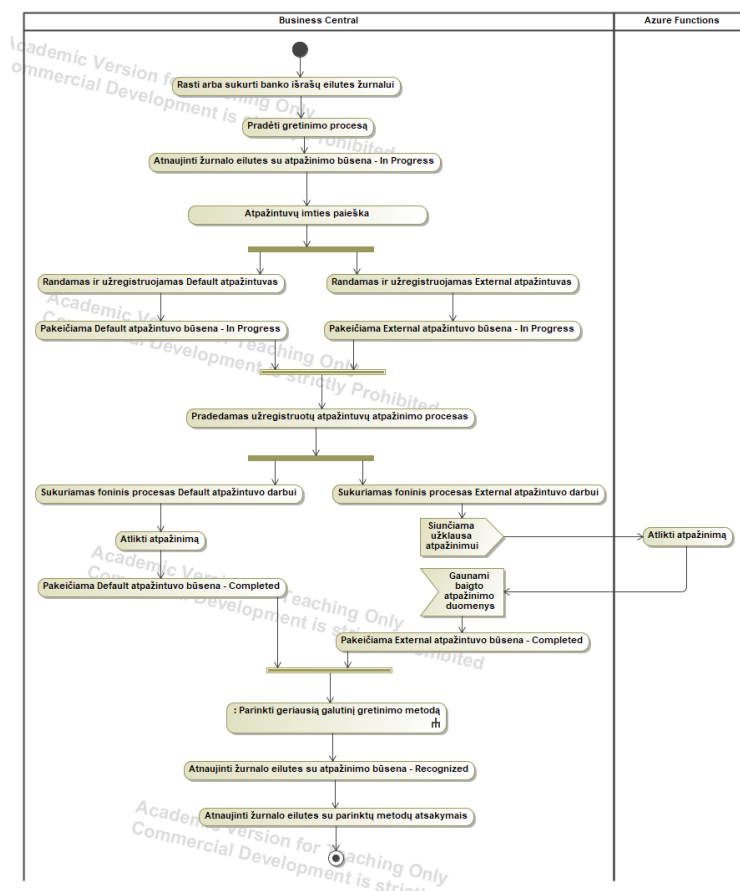
16 pav. Panaudojimo atvejo „Tikrinti gretinimo proceso statusą“ sekos diagrama

Gretinimo metodų pritaikymo procese (žr. 17 – 18 pav.), sistema pasiima banko išrašų eilutes (atsižvelgdama į naudotojo nurodytus filtrus arba sąlygas), taip pat atpažinimo žurnalo įrašuose pažymima būseną „In Progress“, kad būtų matyti, jog vyksta atpažinimo veiksmai. Toliau sistema patikrina, kokie atpažinimo metodai yra įjungti – šiuo metu tai yra vidinis (*Default*) ir išorinis (*External*) mechanizmai:

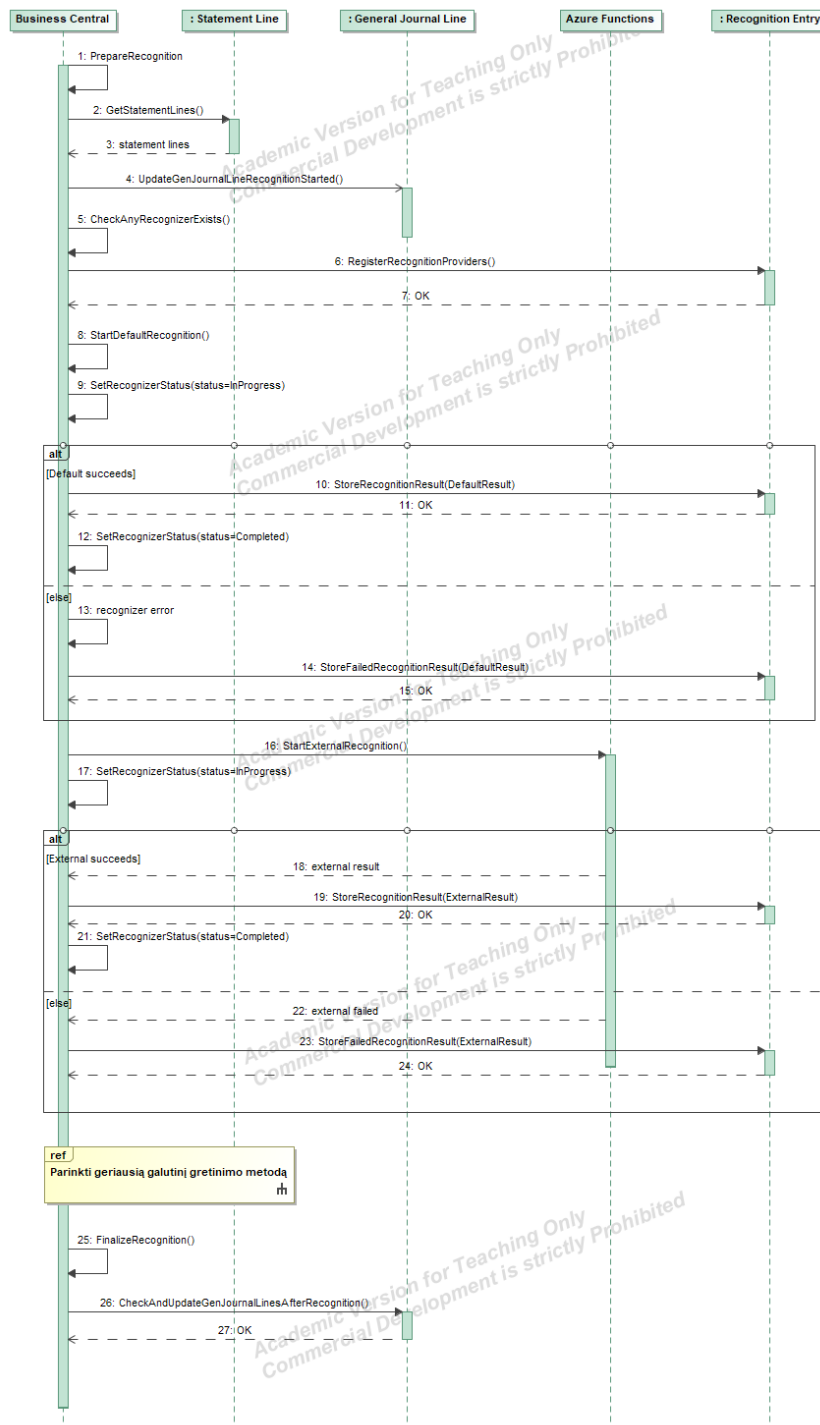
- vidinis metodas (*Default*) – pritaiko taisyklėmis grįstus metodus ir grąžina atpažinimo rezultatus;
- išorinis metodas (*External*) – išorinėje sistemoje atliekamas banko išrašų duomenų atpažinimas ir savų atsakymų grąžinimas.

Kai abu mechanizmai pateikia savo sprendimus, atnaujinami laukai su sąskaitos numeriu, operacijos tipu, suma. Visi atpažinimo įrašai perkeliama iš laikinos saugyklos į nuolatinę duomenų bazę, ir žurnalo eilutės pažymimos kaip „Completed“. Toliau atliekamas galutinio gretinimo metodo parinkimas, parenkantis tinkamiausią metodą pagal istorinius naudotojų pasirinkimus, kuris ir yra

realizuojamas šiame projekte. Atlikus geriausio gretinimo parinkimą, žurnalo eilutės pažymimos, kaip „Recognized“ ir procesas užbaigiamas.



17 pav. Panaudojimo atvejo „Pritaikyti gretinimo metodus“ veiklos diagrama



18 pav. Panaudojimo atvejo „Pritaikyti gretinimo metodus“ sekos diagrama

3. Tyrimo dalis

3.1. Tyrimo tikslas

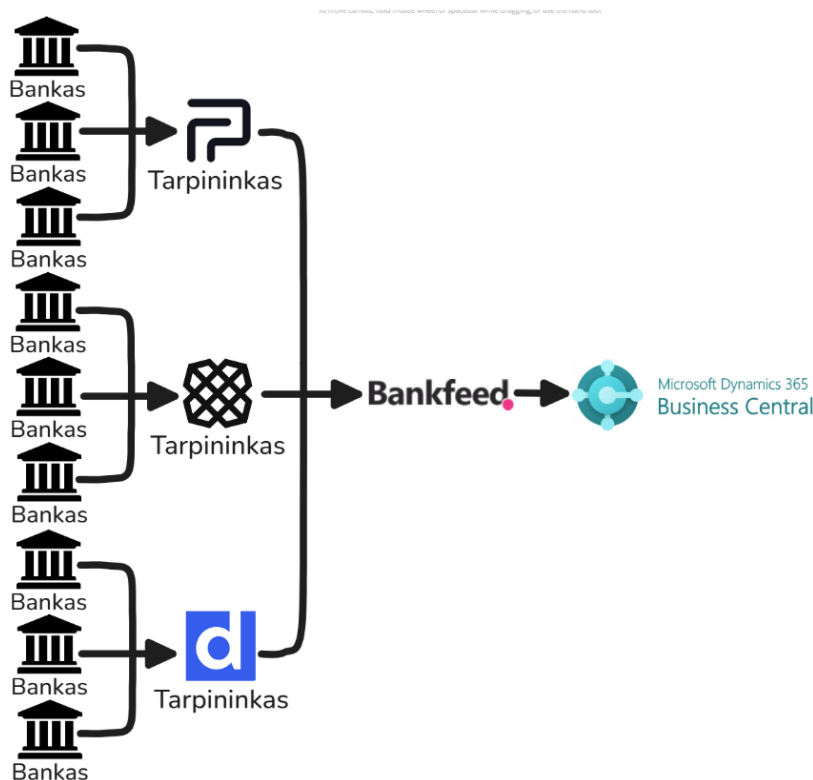
Tyrimo tikslas yra surinkti ir tinkamai apdoroti banko išrašų ir sąskaitų atpažinimo metodų rezultatų duomenis, sukurti ir mokyti mašininio mokymosi algoritmus, leidžiančius atrinkti optimalų atpažinimo metodą kiekvienai banko išrašo ir sąskaitos porai. Tyrimo metu dėmesys skiriamas duomenų valymo, požymių inžinerijos procesams, kadangi banko išrašų duomenys pasižymi dideliu struktūriniu ir semantiniu sudėtingumu, ypač lyginant banko išrašus, importuojamus iš skirtingų šalių. Netinkamai apdoroti duomenys, ar neišskirti tam tikri požymiai gali ženkliai sumažinti modelių tikslumą, todėl svarbu užtikrinti mokymo duomenų kokybę. Taip pat, tyrimo metu palyginami keli modeliai ir pasirenkamas geriausias sprendimas, atitinkantis tiek tikslumo, tiek greitaveikos reikalavimus, bei įvertinamas jo pritaikomumas realiam naudojimui.

3.2. Tyrimo eiga

3.2.1. Tinkamų mokymui duomenų paruošimas ir apdorojimas

Banko duomenų surinkimas

Business Central pusėje surenkami banko išrašų duomenys iš skirtingų prijungtų banko institucijų sąskaitų. Bankfeed banko išrašų importui yra naudojami Open Banking API palaikantys tarpininkai, sujungiantys daugelį skirtingų bankų į suvienodintą formatą. Tokių tarpininkų yra keletas, ir jie integruoti su įvairiausių šalių bankais. Šie tarpininkai suvienodina informaciją, kad ji atitiktų Open Banking standartus, o Bankfeed importuoja visų banko išrašų duomenis į Business Central, taip dar labiau pritaikant gaunamus duomenis naudojimui ERP sistemoje.



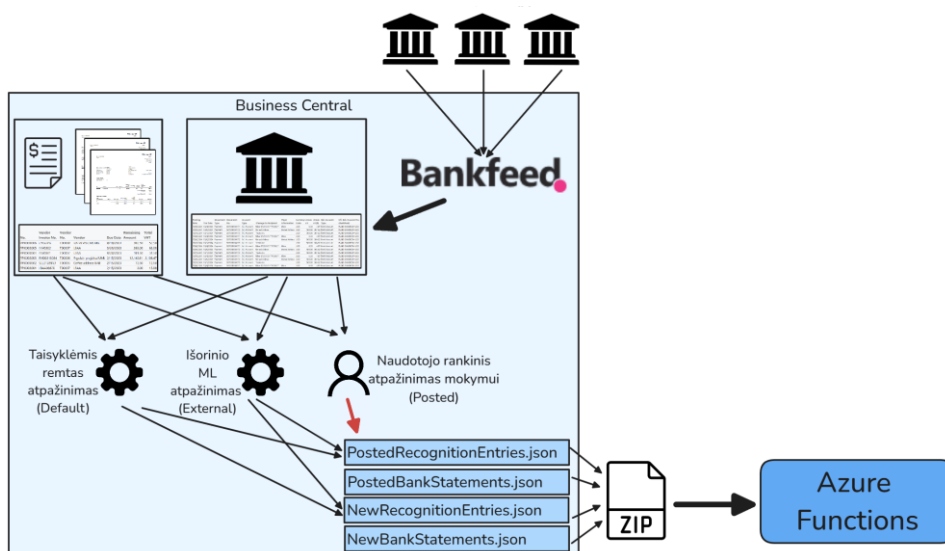
19 pav. Apibendrinta banko išrašų importo į Business Central schema

Business Central duomenų surinkimas

Bankfeed pagalba surinkti banko išrašų duomenys lygiagrečiai pereina per atpažinimo metodus Business Central aplinkoje, kuomet yra atliekami visų atvirų finansinių įrašų gretinimai su banko išrašais. Tolimesniam galutinio atpažinimo metodo parinkimui, aktualu gauti visus gretinimo bandymus iš Business Central. Į šią informaciją įeina atpažinimo metodo pavadinimai, banko sąskaitų numeriai, dokumentų, kurie gretinami, numeriai, dokumentų aprašymai, pasirinktų gretinimo būdų pavadinimai (vidiniai, grąžinami iš atpažinimo metodų), atspėtos reikšmės, vidiniai pasitikėjimo įverčiai ir t.t.

Šie iš Business Central ateinantys duomenys yra išskiriami į dar dvi dalis – į senus (istorinius) įrašus ir į naujus įrašus. Pagrindinis skirtumas tarp naujų ir istorinių įrašų, ateinančių iš BC, yra papildoma „Posted“ eilutė – nusakanti koks buvo galutinis naudotojo pasirinkimas, kai buvo gretinami įrašai. Turint šią „Posted“ eilutę būtų galima vykdyti prižiūrimą (angl. *supervised*) mokymą, kadangi iš šios eilutės galima spręsti, ar gretinimo metodai buvo teisūs istoriniuose duomenyse.

Prieš atliekant duomenų atidavimą į Azure Functions, yra atliekamas minėtų keturių duomenų struktūrų formatavimas į JSON formatą, bei archyavimas į ZIP.



20 pav. Apibendrinta Business Central duomenų surinkimo ir apdorojimo schema

Duomenų apdorojimas Azure Functions sistemoje

Perdavus duomenis iš Business Central į Azure Functions, pradinis etapas Azure Functions pusėje apima neapdorotų duomenų, saugomų suglaudintame ZIP archyve, programinį išskleidimą į numatytą katalogą. Šiame kataloge identifikuojami JSON formato failai, kurie tada sukeliami į duomenų struktūras (angl. *data frames*). Pagrindinis dėmesys skiriamas dviem konkrečioms failams, atstovaujantiems skirtingus duomenų rinkinius – viename iš jų saugomi pagrindiniai banko išrašų įrašai (čia vadinami „senaisiais duomenimis“), antrame saugoma kontekstinė informacija apie šiuos įrašus (spėjimų duomenys iš Business Central pusės).

Įkėlus duomenis, kitas žingsnis yra duomenų rinkinių atskyrimas ir paruošimas analizei. Įkelto duomenų rinkinio poaibis, identifikuotas pagal konkrečią žymą („Posted“), yra išskiriamas kaip

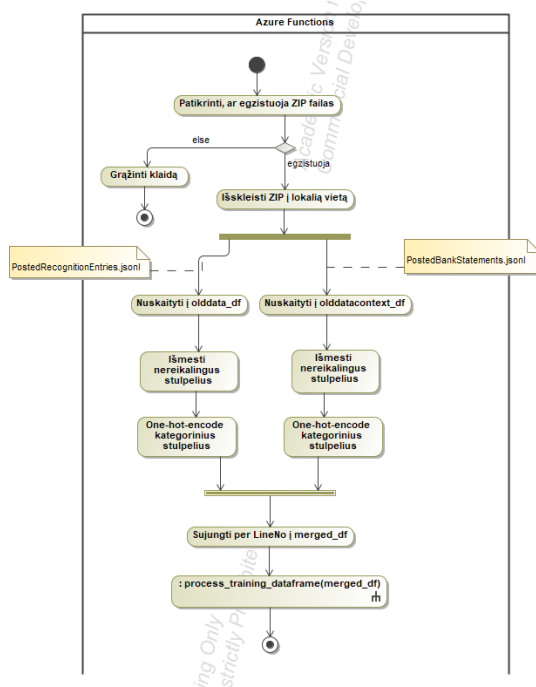
etalonas (angl. *ground truth*). Šis etalonas atspindi naudotojų rankiniu būdu patikrintus ir patvirtintus įrašus, pagal kuriuos bus vertinami kiti atpažinimo metodai.

Tolimesniame etape dėmesys skiriamas struktūrinių skirtumų ir galimų duomenų kokybės problemų sprendimui:

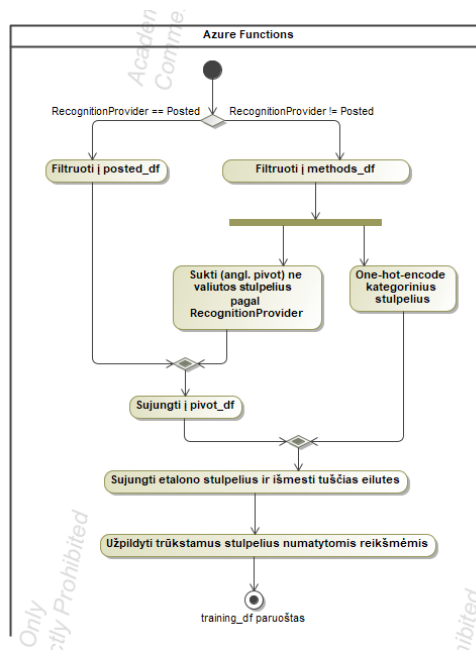
- dublikatų tvarkymas - agreguojant duomenis naudojant ašies operacijas, pasirenkamas pirmas ne tuščias pasikartojimas dublikatų atveju;
- tuščių reikšmių valymas - papildomas valymas apima eilučių, kuriose trūksta esminių etalono dalių, pašalinimą. Trūkstamos vertės, konkrečių atpažinimo metodų rezultatams, pakeičiamos numatyta reikšme pvz., „nežinomas“ (angl. *unknown*), siekiant užtikrinti nuoseklumą ir tuščių stulpelių nebuvimą prieš mokinant modelį.

Savybių inžinerija (angl. *feature engineering*) buvo atliekama siekiant transformuoti apdorotus duomenis į nuoseklų mokymo rinkinį. Metodologija apėmė kelias pagrindines transformacijas:

- Buvo konvertuojamos atpažinimo metodų rezultatų eilutės vienai bankinei operacijai į plačią lentelę, kurioje kiekviena operacija būtų pateikiama viena eilute su visų atpažinimo metodų rezultatais;
- transformuoti duomenys buvo sujungti su anksčiau išskirtu etalonu. Šis sujungimas užtikrino, kad mokymo rinkinyje būtų tiek automatizuotų atpažinimo metodų rezultatai, tiek patvirtinti spėjimų duomenys, pagal kuriuos modeliai galėtų būti mokomi;
- panaudojant *One-hot-encoding* buvo transformuojamos kategorinės vertės į dvejetainės matricos atvaizdavimą. Šis žingsnis leido mašininio mokymosi algoritmams apdoroti ir išnaudoti kategorinę informaciją.



21 pav. Apibendrinta duomenų apdorojimo Azure Functions sistemoje schema



22 pav. Apibendrinta funkcijos *process_training_dataframe()* schema

Galutinis mokymui naudotų duomenų laukų sąrašas su aprašymais pateiktas 1 priede.

3.2.2. Geriausio mašininio mokymo modelio parinkimo procesas

Siekiant sistemingai identifikuoti efektyviausią atpažinimo metodo pasirinkimo strategiją, buvo sukurta suvienodinta modelių mokymo sistema, gebanti apdoroti rezultatus, sukurti maišos matricas ir įvertinti kiekvieno modelio efektyvumą pagal kelias metodikas.

Į neapdorojamus atpažinimo duomenis įėjo pačių atpažinimo metodų išvesties duomenys (atpažinti sąskaitų numeriai, jų tipai) kartu su etalono klasėmis ir metaduomenų laukais (pvz., eilučių numeriais). Nesvarbūs ir pertekliniai stulpeliai – tokie kaip etalono laukai, eilučių indeksai ir bet kokie anksčiau apskaičiuoti teisingumo rodikliai buvo pašalinami, siekiant sumažinti triukšmą ir išvengti duomenų nutekėjimo.

Siekiant užtikrinti, kad kiekvienas duomenų rinkinys ir kiekvienas kandidatinis modelis matytų identišką požymių rinkinį, buvo fiksuojamas galutinis stulpelių sąrašas mokymo metu ir užtikrinama ta pati schema transformavimo metu – automatiškai įterpiant trūkstamus stulpelius kaip nulius, kai reikia.

Suvokiant, kad kiekviena bankinės operacijos eilutė gali būti teisingai atpažinta vieno ar kelių atpažinimo metodų, buvo performuluotos pradinės atpažinimo žymės (angl. *labels*) į vieną bendrą klasifikatorių. Konkrečiai, kiekvienam pavyzdžiui priskiriama viena iš keturių klasių:

- „None“, jei nei vienas atpažinimo metodas neatitinka etalono;
- „Default“, jei taisyklėmis grindžjiamas atpažinimo metodas yra teisingas;
- „External“, jei alternatyvus išorinis atpažinimo metodas yra teisingas;
- "Both", jei abiejų išvesties stulpelių duomenys sutampa su etalonu.

27 lentelė. Klasifikatoriaus sudarymo taisyklės

Default (teisingas = 1, ne = 0)	External (teisingas = 1, ne = 0)	Galutinė klasė
---------------------------------	----------------------------------	----------------

0	0	None
1	0	Default
0	1	External
1	1	Both

Galutinės klasės buvo konvertuojamos į sveikųjų skaičių kodus, kad jie būtų suderinami su *scikit-learn* vertintojo (angl. *evaluator*) API.

Buvo įvertintas platus prižiūrimo (angl. *supervised*) mokymosi algoritmų spektras, įskaitant medžiais grįstus ansamblius (atsitiktinis miškas, gradientinis stiprinimas, histograma grįstas gradientinis stiprinimas, papildomi medžiai), linijinius modelius (logistinė regresija), branduolio metodus (atraminių vektorių mašinos), atstumu grįstus klasifikatorius (k artimiausių kaimynų), sprendimų medžius, stiprinimo (angl. *boosting*) metodus (*AdaBoost*, *XGBoost*, *LightGBM*, *CatBoost*), naivųjį Bayes'ą ir daugiasluoksnius perceptronus. Siekiant užtikrinti atkuriamumą ir sąžiningą palyginimą, kiekvienas algoritmas buvo inicijuotas su fiksuotais, dažnai naudojamais hiperparametrais (pvz., 100 vertintojų medžių ansambliams, numatytasis mokymosi greitis = 0,1, kai taikoma, maksimalus medžio gylis nustatytas pagal bibliotekų numatytąsias reikšmes), ir buvo taikoma pastovi atsitiktinių skaičių seka.

28 lentelė. Tyrime išbandytų mašininio mokymo modelių sąrašas

Classifier	n_estimators	random_state	max_iter	probability	use_label_encoder	eval_metric	verbose
Gradient Boosting	100	42					
Random Forest	100	42					
HistGradient Boosting		42					
Logistic Regression		42	1000				
SVC		42		True			
KNeighbors							
DecisionTree		42					
AdaBoost		42					
ExtraTrees	100	42					
MLP		42	1000				
GaussianNB							
XGBoost		42			False	mlogloss	
LightGBM		42					
CatBoost		42					0

Analizei buvo išskirtos papildomos metrikos, siekiant kiekybiškai įvertinti meta-klasifikatoriaus realią naudą, lyginant su bazinėmis strategijomis (visada pasirenkant numatytąjį atpažinimo metodą arba visada pasirenkant išorinį atpažinimo metodą). Buvo matuojama:

- dažnis, kuriuo meta-klasifikatorius susilaiko („None“);
- atvejus, kai jis padaro „kritinį“ klaidingą spėjimą (prognozuodamas atpažinimo metodą, kai nė vienas nėra teisingas, arba pasiūlydamas vieną metodą, kai kitas yra teisingas);
- papildomus teisingus išgavimus, palyginti su kiekvienu baziniu metodu;

Naudojant papildomų metrikų rinkinį įvertinamas ne tik grynas tikslumas, bet ir praktinė nauda bei rizika, susijusi su modelio įdiegimu realiuose darbinuose procesuose.

Kiekvienam algoritmui buvo registruojamas mokymo laikas, kurį užtrukdavo algoritmas visam mokymo rinkiniui. Šie laiko matavimai, kartu su tikslumo rezultatais, leido subalansuotai įvertinti modelius tarp prognozavimo galios ir skaičiavimo sąnaudų – kas yra kritiškai svarbu realaus laiko ar didelio masto diegimui.

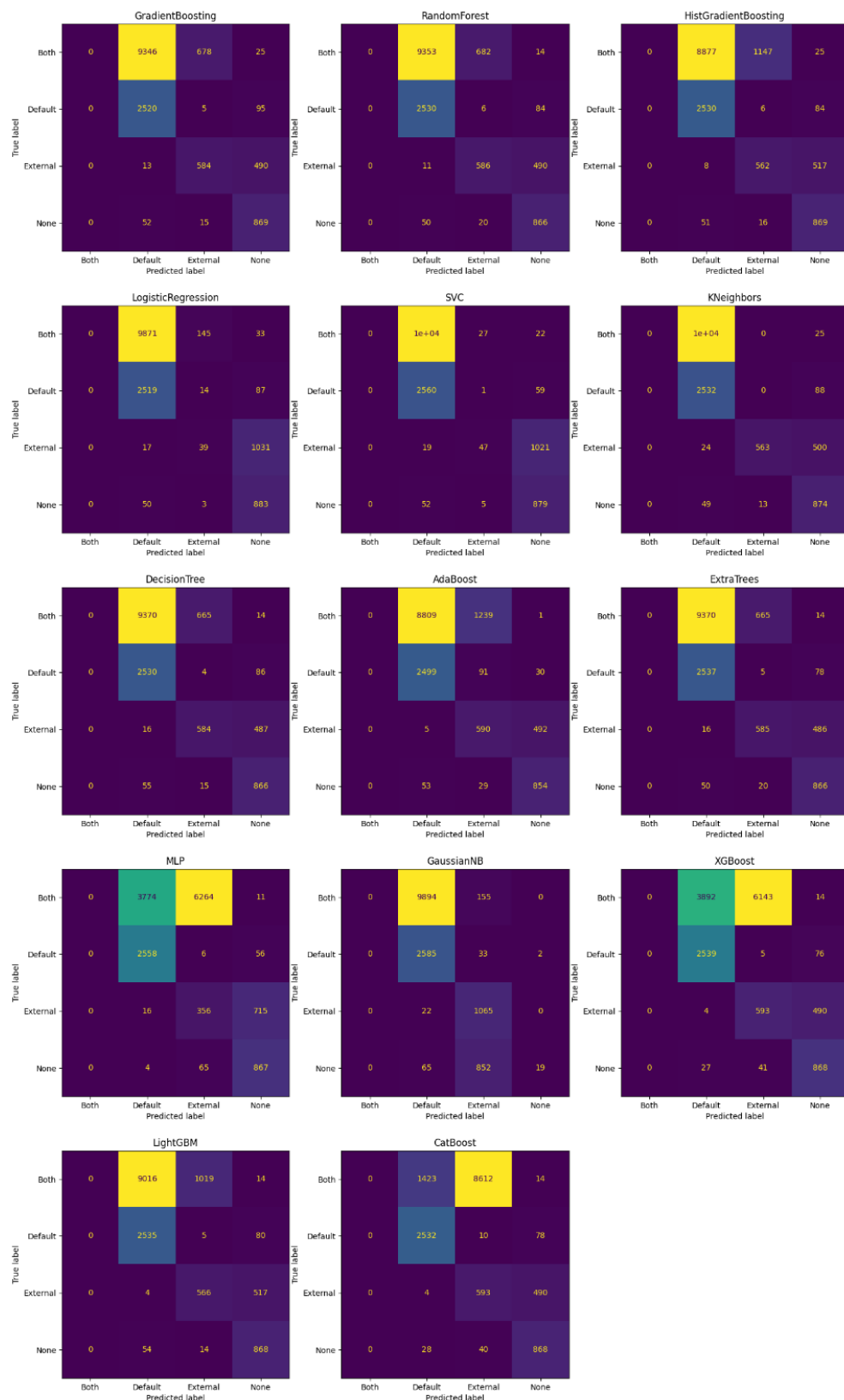
Kiekvienam kandidatui buvo apskaičiuotos:

- metrikos kiekvienam atpažinimo metodui (tikslumai, jautrumas, F₁-balas), lyginant meta-klasifikatoriaus dvejetainį rodiklį kiekvienam atpažinimo metodui su etalonu;
- bendras daugiaklasis tikslumas, atspindintis pavyzdžių dalį, kuriai meta-klasifikatorius teisingai identifikavo teisingų atpažinimo metodų rinkinį;
- vidutines metrikas, gautas apskaičiuojant kiekvieno atpažinimo metodo rezultatus, siekiant užfiksuoti subalansuotą efektyvumą visuose atpažinimo metoduose.

29 lentelė. Apibendrintas modelių rezultatų, tikslumo, rizikos įvertinimo, mokymo laiko sąrašas

Classifier	Overall Accuracy	Avg F1 Score	Avg Precision	Avg Recall	Default Penalty	External Penalty	ML penalty	Training Time	Thres-hold	Cover-age %
XGBoost	0.2723	0.7121	0.9942	0.5563	0.2322	0.6735	0.1510	3.91 sec	0.41	90.14
Light GBM	0.2701	0.6002	0.9916	0.5270	0.2322	0.6735	0.1531	3.00 sec	0.41	89.93
CatBoost	0.2718	0.6889	0.9933	0.5694	0.2322	0.6735	0.1545	7.99 sec	0.39	90.13
Hist Gradient Boosting	0.2696	0.6055	0.9911	0.5269	0.2322	0.6735	0.1569	10.40 sec	0.41	89.82
Random Forest	0.2710	0.5848	0.9874	0.5259	0.2322	0.6735	0.1582	4.02 sec	0.59	90.10
Gradient Boosting	0.2704	0.5840	0.9895	0.5250	0.2322	0.6735	0.1585	5.10 sec	0.49	89.93
K Neighbors	0.2701	0.5444	0.9858	0.5208	0.2322	0.6735	0.1597	6.10 sec	0.80	89.88
Decision Tree	0.2709	0.5836	0.9895	0.5257	0.2322	0.6735	0.1602	4.35 sec	0.38	90.11
Extra Trees	0.2714	0.5839	0.9874	0.5261	0.2322	0.6735	0.1602	2.94 sec	0.38	90.17
MLP	0.2574	0.7042	0.9931	0.5471	0.2322	0.6735	0.1742	7.73 sec	0.79	88.78
SVC	0.2373	0.5030	0.9597	0.4990	0.2322	0.6735	0.1872	5.00 sec	0.55	86.52
Logistic Regress.	0.2342	0.5093	0.9550	0.4973	0.2322	0.6735	0.1956	3.04 sec	0.71	86.16
Ada Boost	0.2684	0.6103	0.9667	0.5284	0.2322	0.6735	0.2149	5.58 sec	0.50	90.63
Gaussian NB	0.2497	0.5866	0.7863	0.5473	0.2322	0.6735	0.6630	2.46 sec	0.00	99.86

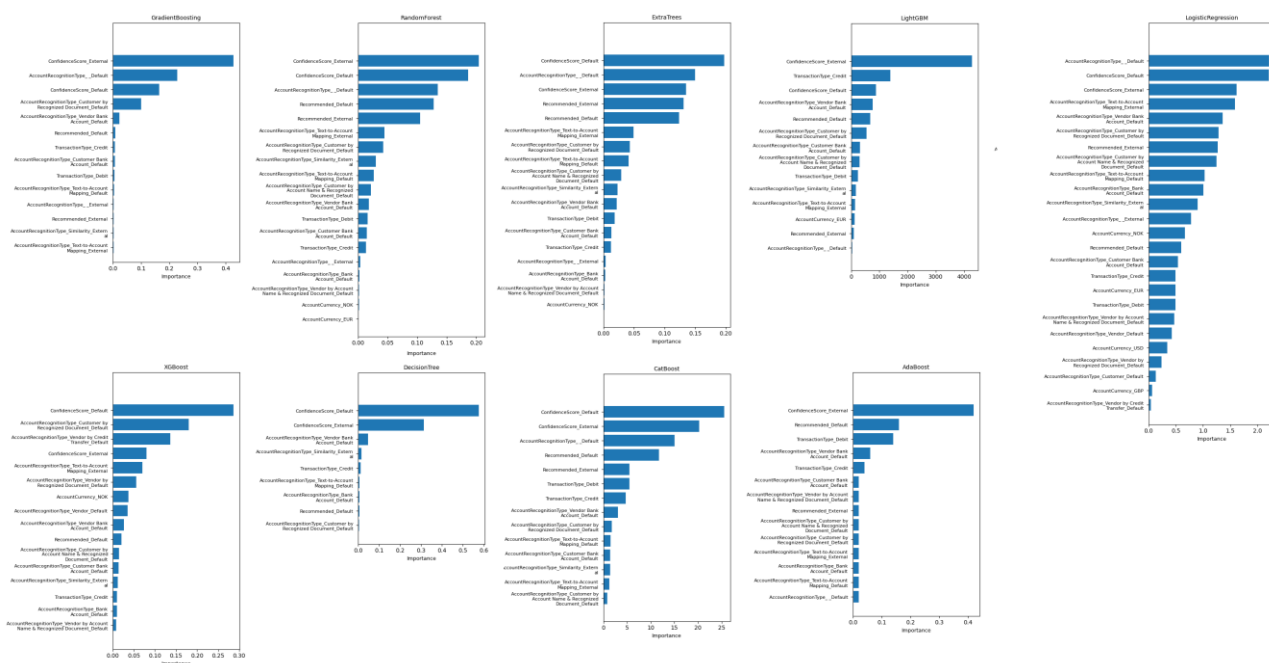
Siekiant vizualizuoti klaidų schemas, buvo generuojamos maišos matricos (žr. 23 pav.) kiekvienam modeliui, iliustruojančios klaidingai teigiamus ir klaidingai neigiamus rezultatus:



23 pav. Tyrimė išbandytų prižiūrimo mokymo algoritmų rezultatų maišos matricos

Modeliams su integruotomis požymių svarbos nustatymo priemonėmis (medžių ansambliai) arba koeficientais pagrįsta svarba (linijiniai modeliai), buvo išgauti požymiai pagal jų prognozuojamąją

galia. Šie įvertinimai buvo agreguoti į grafikus, atskleidžiančius įtakingiausius atributus, kurie ir buvo panaudoti galutiniame mokyme.



24 pav. Tyrimo išbandytų prižiūrimo mokymo algoritmų svarbiausių požymių grafikai

3.2.3. Mokymo pritaikymas teikti prioritetą aukštam tikslumui

Kadangi projekto metu yra bandoma atpažinti kritiško jautrumo duomenis, norint atsižvelgti į skirtingų klaidų pasekmes, buvo konsultuotasi su srities ekspertais, dirbančiais su šiuo funkcionalumu, siekiant išsiaiškinti, kokie svoriai turi būti pritaikomi grąžinamiems spėjimams. Konkrečiau, buvo įvardinta, jog klaidingas spėjimas C_{FP} (angl. „false positive“) apie netinkamą atpažinimo metodą gali kainuoti iki 10-ies kartų daugiau, nei jokio pasiūlymo nepateikimas C_A .

Užregistruojant klaidingą gretinimą, sukeliamos itin didelės problemos, nes neatitikimai gali būti nepastebėti iš karto, o po ilgo laiko, klaidas gali tekti tvarkyti einant per istoriją ir operacijas gretinant iš naujo.

Šis svoris (10:1) buvo įtrauktas į modelių mokymą, siekiant juos padaryti „atsargesniais“, taip pat, buvo reikalinga optimizuoti sprendimų priėmimo slenksčius, kad sprendimai būtų priimami tik tada, kai jie yra užtikrinti. Sprendimas įtraukti svorius ir optimizuoti sprendimų priėmimo slenksčius itin sustiprino modelių grąžinamų rezultatų tikslumą, pernelyg nesumažinant bendro atliekamų spėjimų skaičiaus. Modelių spėjimų kiekio ir tikslumo metrikos palygintos 29 lentelėje.

Slenksčio apskaičiavimui buvo išskirti svoriai šiems atvejams:

$$w_i = \begin{cases} C_A, & \text{jei bent vienas Correct_Recognizer}_X = 1 \\ C_{FP}, & \text{kitais atvejais} \end{cases} \quad (8)$$

čia:

- w_i – svoris, priskiriamas konkrečiam atvejui/modelio spėjimui;
- C_A – kaina, kai modelis susilaiko nuo sprendimo;

- C_{FP} – kaina, kai modelis klaidingai mano, kad atpažinimas teisingas, bet iš tikrųjų jis neteisingas.

Taikyti svoriai w_i tiesiogiai patikslino klasifikatoriaus praradimo (angl. *loss*) funkciją, verčiant modelį rečiau spėti, jei rizika didelė.

Kad galutinis sprendimas susilaikyti, ar spėti būtų pagrįstas statistiškai, o ne intuicija, ar iš anksto numatytais reikšmėmis – buvo reikalinga gauti optimaliausio slenksčio įvertį, jį skaičiuojant iš mokomųjų duomenų rinkinio. Nuo visos mokymo imties buvo atskirta ~30 % tikrinimo rinkiniui, užtikrinant, kad teigiamų ir neigiamų atvejų proporcija būtų panaši. Iš tikrinimo rinkinio kiekvienos eilutės i , buvo renkamos kiekvieno atpažinto metodo atpažinimo tikimybės $\hat{P}$, o iš jų – maksimalios eilutėje:

$$p_i = \max_r \hat{P}(r | x_i) \quad (9)$$

čia:

- p_i – aukščiausia atpažinimo tikimybė tarp visų metodų konkrečiai eilutei;
- $\hat{P}$ – modelio numatyta tikimybė, kad atitinkamas atpažinimo metodas yra tinkamas konkrečiam įrašui;
- r – konkretus atpažinimo metodas;
- x_i – įrašo i charakteristikos

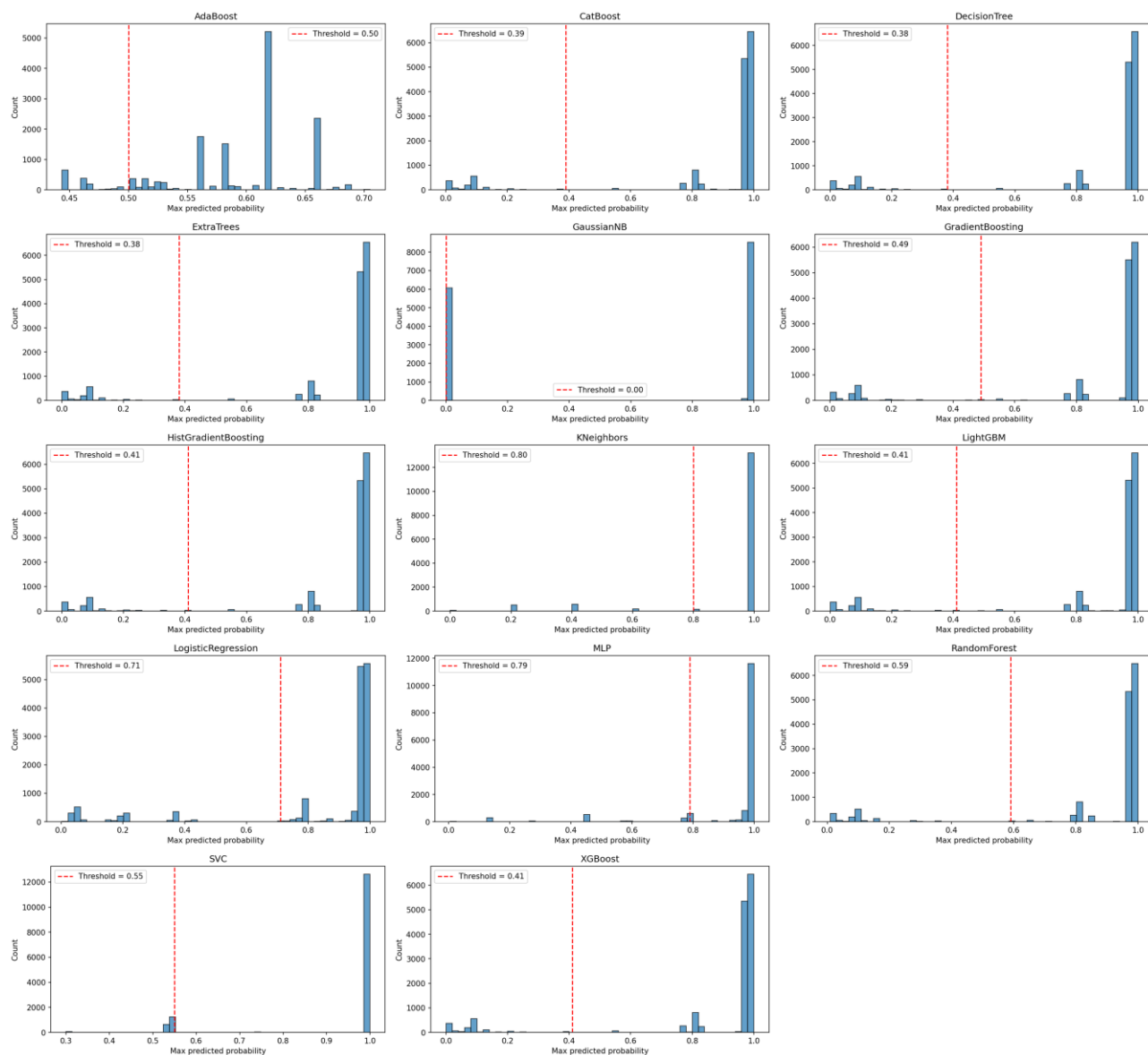
Toliau, iteratyviai didinant slenkstį $t \in [0,1]$, kas 0.01, buvo ieškomos mažiausia reikšmė $kaina(t)$, kuri ir nusako žemiausią tinkantį slenkstį sprendimo priėmimui, pagal formules:

$$\begin{aligned} kaina(t) &= \hat{P}\{FP|P_i > t\} \cdot C_{FP} + \hat{P}\{susilaikyti|p_i \leq t\} \cdot C_A. \\ t^* &= \arg \min_t kaina(t) \end{aligned} \quad (10)$$

čia:

- t – sprendimo slenkstis, nuo kurio sprendžiama, ar verta spėti, ar susilaikyti;
- $\hat{P}\{FP|P_i > t\}$ – sąlyginė tikimybė, kad bus klaidingas spėjimas, jei P_i viršija slenkstį t ;
- $\hat{P}\{susilaikyti|p_i \leq t\}$ – tikimybė, kad susilaikoma nuo spėjimo, jei P_i nepakankamai didelė;
- t^* – optimalus slenkstis, kuris minimizuoja bendrą kainą

Gautas t^* buvo pritaikytas mokymo rinkinyje – jei naujai spėjamos eilutės $p_i > t^*$, tada sprendimas priimamas, kitu atveju – susilaikoma. Slenksčių įtaką skirtingų modelių sprendimo priėmimui vizualizuoja histogramos su t^* linija, kuriose matoma, nuo kurios vietos sprendimai būna priimami.



25 pav. Tyrimė išbandytų prižiūrimo mokymo algoritmų slenksčių vizualizacijos

3.3. Tyrimo rezultatai

3.3.1. Hibridinis gretinimo algoritmas veikia tiksliau už vien taisyklėmis grįstą metodą

H_0 : Sukurta algoritmo parinkimo sistema veikia tokiu pačiu tikslumu kaip taisyklėmis grįsti, ar išoriniai metodai

H_1 : Naudojant algoritmo parinkimo sistemą pagerinamas pasiūlymų tikslumas, lyginant su taisyklėmis grįsto, ar išorinio metodo tikslumu.

Panaudojus 3.2 skyriuje aprašytą metodiką, mokymo duomenims buvo atliktas aprašytas mokymo procesas, ir gauti rezultatai, aprašyti 30 bei 32 lentelėse:

30 lentelė. Mašininio mokymo modelių rezultatai, lyginant su atpažinimo metodais

Classifier	Test Size	Def Rec (count)	Def Rec (%)	Def Corr. if Rec (count)	Def Corr. if Rec (%)	Ext Rec (count)	Ext Rec (%)	Ext Correct if Rec (count)	Ext Corr. if Rec (%)	ML None (count)	ML None (%)	ML Guess (count)	ML Guess (%)	ML Corr. when Guess (%)	ML Crit. Bad (count)	ML Crit. Bad (%)	Default Penalty	Ext. Penalty	ML Penalty
XGBoost	14692	12061	82,09	11983	81,56	11407	77,64	10746	73,14	1448	9,86	13244	90,14	99,42	77	0,581	0,2322	0,6735	0,1510
Light GBM	14692	12061	82,09	11983	81,56	11407	77,64	10746	73,14	1479	10,07	13213	89,93	99,42	77	0,583	0,2322	0,6735	0,1531
Cat Boost	14692	12061	82,09	11983	81,56	11407	77,64	10746	73,14	1450	9,87	13242	90,13	99,38	82	0,619	0,2322	0,6735	0,1545
Hist Gradient Boosting	14692	12061	82,09	11983	81,56	11407	77,64	10746	73,14	1495	10,18	13197	89,82	99,39	81	0,614	0,2322	0,6735	0,1569
Random Forest	14692	12061	82,09	11983	81,56	11407	77,64	10746	73,14	1454	9,9	13238	90,1	99,34	87	0,657	0,2322	0,6735	0,1582
Gradient Boosting	14692	12061	82,09	11983	81,56	11407	77,64	10746	73,14	1479	10,07	13213	89,93	99,36	85	0,643	0,2322	0,6735	0,1585
K Neighbors	14692	12061	82,09	11983	81,56	11407	77,64	10746	73,14	1487	10,12	13205	89,88	99,35	86	0,651	0,2322	0,6735	0,1597
Decision Tree	14692	12061	82,09	11983	81,56	11407	77,64	10746	73,14	1453	9,89	13239	90,11	99,32	90	0,680	0,2322	0,6735	0,1602
Extra Trees	14692	12061	82,09	11983	81,56	11407	77,64	10746	73,14	1444	9,83	13248	90,17	99,31	91	0,687	0,2322	0,6735	0,1602
MLP	14692	12061	82,09	11983	81,56	11407	77,64	10746	73,14	1649	11,22	13043	88,78	99,3	91	0,698	0,2322	0,6735	0,1742
SVC	14692	12061	82,09	11983	81,56	11407	77,64	10746	73,14	1981	13,48	12711	86,52	99,39	77	0,606	0,2322	0,6735	0,1872
Logistic Regress.	14692	12061	82,09	11983	81,56	11407	77,64	10746	73,14	2034	13,84	12658	86,16	99,34	84	0,664	0,2322	0,6735	0,1956
AdaBoost	14692	12061	82,09	11983	81,56	11407	77,64	10746	73,14	1377	9,37	13315	90,63	98,66	178	1,337	0,2322	0,6735	0,2149
GaussianNB	14692	12061	82,09	11983	81,56	11407	77,64	10746	73,14	21	0,14	14671	99,86	93,37	972	6,625	0,2322	0,6735	0,6630

Siekiant sumažinti variaciją, kylančią iš vieno mokymo ir testavimo rinkinio padalijimo, buvo atliktas dešimties k kryžminis validavimas. Kiekviename k kartojime visas mokymo rinkinys buvo dalijamas į sub-mokymo ir sub-validavimo dalis, tada modelis buvo iš naujo mokomas, o k -tojo karto metrikos užfiksuojamos. Dešimties k rezultatų vidurkis suteikė patikimą, mažos variacijos įvertį apie efektyvumą su nematytais duomenimis, taip pat leido apskaičiuoti pasikliautinus intervalus, standartinę deviaciją. Tyrimo rezultatai pateikti 31 lentelėje.

31 lentelė. Tyrimo metu atliktos 10 k kryžminės patikros rezultatai

Classifier	Mean Accuracy	CI Lower - CI Upper	St. Dev. Accuracy	Mean F1	CI Lower - CI Upper	Std F1	Mean Precision	CI Lower - CI Upper	St. Dev. Precision	Mean Recall	CI Lower - CI Upper	St. Dev. Recall
Gradient Boosting	0.9423	0.937 - 0.947	0.0074	0.8379	0.8248 - 0.8510	0.0183	0.8755	0.8678 - 0.8832	0.0108	0.8534	0.8434 - 0.8634	0.0140
XGBoost	0.9412	0.9369 - 0.9455	0.0060	0.8360	0.8251 - 0.8468	0.0151	0.8748	0.8675 - 0.8822	0.0102	0.8504	0.8419 - 0.8589	0.0119
CatBoost	0.9412	0.9369 - 0.9456	0.0061	0.8359	0.8240 - 0.8479	0.0167	0.8746	0.8673 - 0.8820	0.0103	0.8504	0.8399 - 0.8608	0.0146
HistGrad Boosting	0.9404	0.9368 - 0.9441	0.0051	0.8343	0.8231 - 0.8456	0.0157	0.8746	0.8687 - 0.8805	0.0083	0.8487	0.8383 - 0.8590	0.0145
Decision Tree	0.9406	0.9359 - 0.9453	0.0065	0.8350	0.8239 - 0.8461	0.0156	0.8742	0.8668 - 0.8816	0.0104	0.8493	0.8401 - 0.8585	0.0128
Light GBM	0.9406	0.9370 - 0.9442	0.0050	0.8342	0.8225 - 0.8459	0.0163	0.8738	0.8674 - 0.8803	0.0090	0.8488	0.8378 - 0.8598	0.0154
Extra Trees	0.9409	0.9363 - 0.9455	0.0064	0.8350	0.8250 - 0.8450	0.0140	0.8735	0.8653 - 0.8817	0.0115	0.8493	0.8412 - 0.8574	0.0113
Random Forest	0.9416	0.9375 - 0.9456	0.0056	0.8357	0.8253 - 0.8462	0.0146	0.8729	0.8652 - 0.8806	0.0107	0.8480	0.8376 - 0.8584	0.0145
K Neighbors	0.9425	0.9379 - 0.9471	0.0064	0.8305	0.8110 - 0.8500	0.0272	0.8542	0.8411 - 0.8673	0.0183	0.8284	0.8118 - 0.8450	0.0232
MLP	0.9366	0.9313 - 0.9420	0.0075	0.8159	0.7931 - 0.8388	0.0320	0.8475	0.8285 - 0.8665	0.0266	0.8092	0.7878 - 0.8305	0.0299
Ada Boost	0.9039	0.8856 - 0.9222	0.0256	0.7879	0.7660 - 0.8098	0.0306	0.7998	0.7650 - 0.8347	0.0487	0.8188	0.7984 - 0.8392	0.0285
SVC	0.8964	0.8881 - 0.9048	0.0116	0.6338	0.6190 - 0.6486	0.0206	0.7453	0.6484 - 0.8423	0.1355	0.7100	0.7029 - 0.7172	0.0100
Logistic Regression	0.8976	0.8894 - 0.9057	0.0114	0.6351	0.6203 - 0.6498	0.0206	0.7303	0.6356 - 0.8250	0.1324	0.7105	0.7035 - 0.7174	0.0097
Gaussian NB	0.3394	0.2119 - 0.4670	0.1783	0.3668	0.3029 - 0.4308	0.0894	0.4868	0.4564 - 0.5171	0.0424	0.5370	0.4914 - 0.5825	0.0637

Iš gautų rezultatų 30 lentelėje galime teigti, jog taisyklėmis grindžiamas atpažinimo metodas (*Default*) bando atpažinti apie 82% visų eilučių, iš kurių teisingai atpažįsta apie 81.6% įrašų, išorinis gretinimo metodas (*External*) spėja mažiau, apie 77.6% visų eilučių ir spėjimo tikslumas yra dar mažesnis – 73.1%. Tačiau, būtina paminėti, jog išorinis gretinimo metodas yra visiškai automatinis - jam nereikalingas joks darbuotojų įsikišimas, todėl ir šis rezultatas yra ganėtinai geras. Galima daryti išvadą, jog nei vienas atpažinimo metodas neapima pageidaujamo eilučių skaičiaus, taip pat, spėjimų tikslumas nėra pageidautinas, žiūrint į kontekstą, kai yra dirbama su finansinėmis operacijomis, kai tikslumas yra itin svarbus.

K-kartotinių kryžminio tikrinimo rezultatai 31 lentelėje atspindi aprašytą „atsargų“ mokymą, o galutinių rezultatų lentelėje (33 lentelė) matoma, kaip „atsargumo“ įvedimas su perskaičiuotų slenksčių panaudojimu priveda tik prie ~9 % spėjimų atsisakymų, kas leidžia išlaikyti aukštą spėjimų skaičiaus ir tikslumo balansą.

32 lentelė. Atpažinimo metodų tikslumo santrauka

Atpažinimo metodas	Bandyta spėti eilučių %	Teisingai atpažinta iš bandymų spėti %
Default	82.09 %	81.56 %
External	77.64 %	73.14 %
MM metodų parinkėjas	90.14 %	99.42 %

Lyginant bendrus klasifikatorių rezultatus, matome, jog mašininio mokymo algoritmai, renkantys tinkamiausius atpažinimo metodus, didžiąja dalimi grąžina panašius rezultatus, tačiau kai kalbama apie susilaikymo / spėjimo procentus, ar modelių tikslumą - galime matyti nemažą skirtumą ties kai kuriais modeliais, tokiais kaip AdaBoost, ar GaussianNB).

33 lentelė. Mašininio mokymo modelių rezultatų apibendrinimas

Rodikliai	XGBoost (geriausias modelis)	Rezultatų aibė tarp modelių
Susilaikymo (ML None %)	9.86 %	0.14 % (GNB) → 13.84 % (GNB)
Pasiūlymo (ML Rec %)	90.16 %	86.16 % (LogReg) → 96.10 % (GNB)
Teisingai atpažinta, kai buvo bandoma spėti (ML Corr. if Rec. %)	99.42 %	93,37 % (GNB) → 99.42 % (XGB)
Kritiškai blogų spėjimų (kai parinktas visiškai neteisingas atpažinimo metodas) (kiekis viso / %)	77 / 0.59 %	77 / 0.59 % (CatB) → 972 / 6.625 % (GNB)
Nuobauda už neteisingą spėjimą (ML Penalty), (kuo mažesnis skaičius tuo geriau)	0.1510	0.1510 (XGB) → 0.6630 (GNB)

Apibendrinant gaunamus rezultatus, galime interpretuoti tokias išvadas:

- kai mašininio mokymosi (MM) modelis pateikia rekomendaciją, ji beveik visada būna teisinga: XGBoost pirmauja su 99,42% tikslumu, LightGBM, HistGradientBoosting, SVC ~99,39%. Tai žymus pagerėjimas, lyginant su taisyklėmis grindžiamo metodo tikslumu (81,6%), ar išorinio atpažinimo metodo tikslumu (73,1%);
- naudojant XGBoost modelį, tik 0,58% atvejų, kai MM rekomenduoja atpažinimo metodą, šis metodas būna neteisingas;
- kai pasveriname ir apskaičiuojame kritiškai blogo atpažinimo, ar susilaikymo nuobaudą – gauta metrika sumažėja nuo 0,2322 (taisyklėmis grindžiamo metodo) iki 0,151 (naudojant

XGBoost), t.y. ~30% pagerėjimas. Realioje sistemoje tai tiesiogiai lemtų mažesnę tolimesnių klaidų skaičių ir sumažintų pakartotinio rankinio tikrinimo poreikį;

- MM susilaiko nuo sprendimo ~9,8% atvejų, palyginti su 18% – 23% atvejų baziniuose metoduose. Modelis atsisako priimti sprendimą tik tada, kai tikrai abejoja, taip išlaikydamas aukštą bandomų atpažinti eilučių skaičių.

4. Eksperimentinė dalis

4.1. Eksperimento tikslas

Eksperimento tikslas yra įvertinti, kaip tiksliai tyrimo metu pasirinktas, geriausiai pasirodęs MM modelis, gali atlikti spėjimus naudojant realių įmonių duomenis. Taip pat, ar sukurtas algoritmas, optimizuotas minimalizuoti 10:1 baudos už klaidą / neatliktą spėjimą, statistiškai reikšmingai sumažina vidutinę baudos vertę (angl. *penalty*) lyginant su:

- vien taisyklėmis grįstais metodais („*Default*“), kurie turi būti aprašomi darbuotojų rankiniu būdu;
- išoriniu („*External*“) MM atpažinimo įrankiu.

H₀: Visų trijų tirtų metodų („*Default*“, „*External*“, „*ML*“) vidutinės baudos vertės ir tikslumai statistiškai nesiskiria

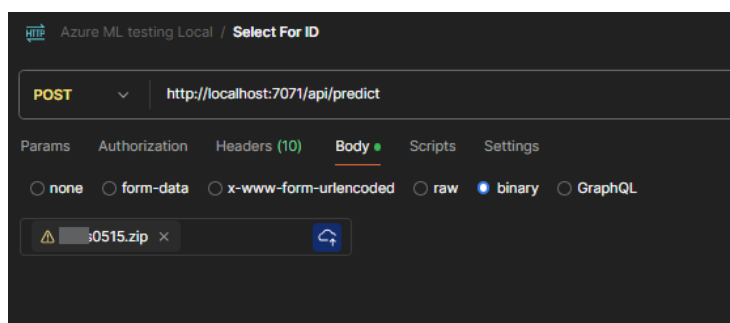
H₁: MM modelis sumažina vidutinę baudos vertę (angl. *penalty*) reikšmingai daugiau nei „*Default*“, ar „*External*“ metodai.

H₂: MM modelio tikslumas (angl. *precision*) yra reikšmingai aukštesnis nei „*Default*“, ar „*External*“ metodų.

4.2. Eksperimento eiga

4.2.1. Eksperimento atlikimo procesas

Iš įmonių, sutikusių duoti duomenis eksperimentui bei turinčių tinkamą Bankfeed produkto versiją, buvo eksportuojami ZIP archyvai, turintys tiek istorinius, tiek naujus duomenis, kuriuos būtų galima panaudoti mokymui. ZIP paruošimo ir išsiuntimo procesas jau yra realizuotas Bankfeed produkte, tačiau šio eksperimento metu duomenų užkėlimas atliekamas ranka, naudojant testavimo programą „Postman“. ZIP archyvo išsiuntimas pradeda duomenų apdorojimo ir mokymo procesus Azure ML aplinkoje:

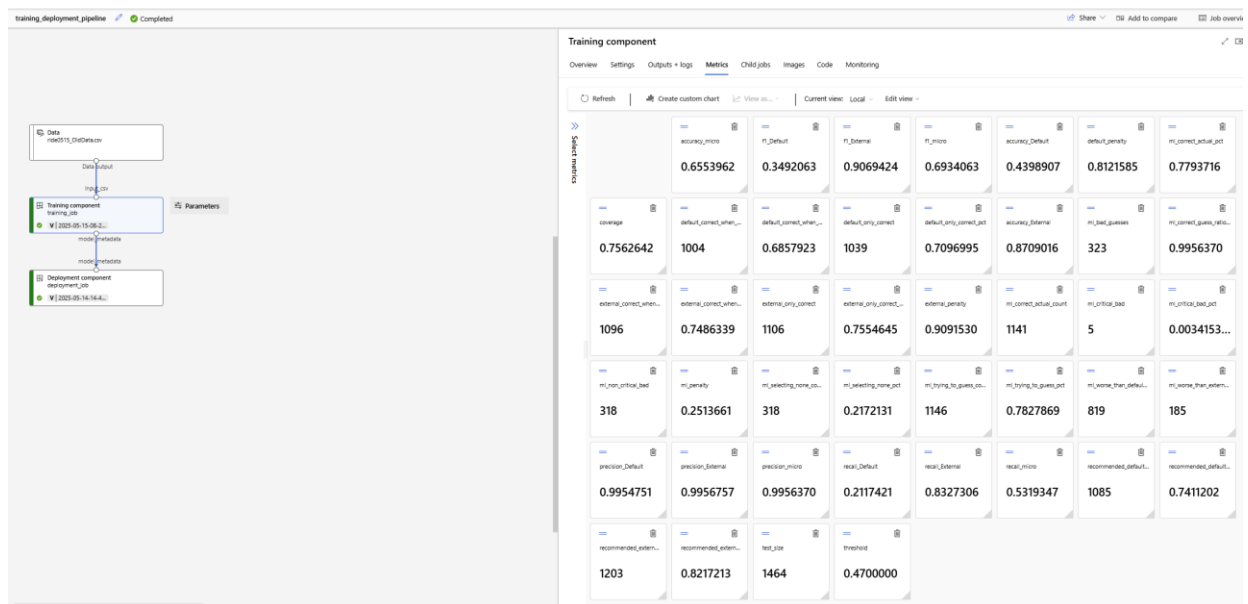


26 pav. POST užklausa, siunčianti įmonės ZIP archyvą modelio mokymui

Display name (25 visualized)	Parent job name	Experiment	Status
▼ training_deployment_pipeline (2)		ride0515_deployment_pipeline	✓ Completed
deployment_job	red_curtain_nybm1860d	ride0515_deployment_pipeline	✓ Completed
training_job	red_curtain_nybm1860d	ride0515_deployment_pipeline	✓ Completed

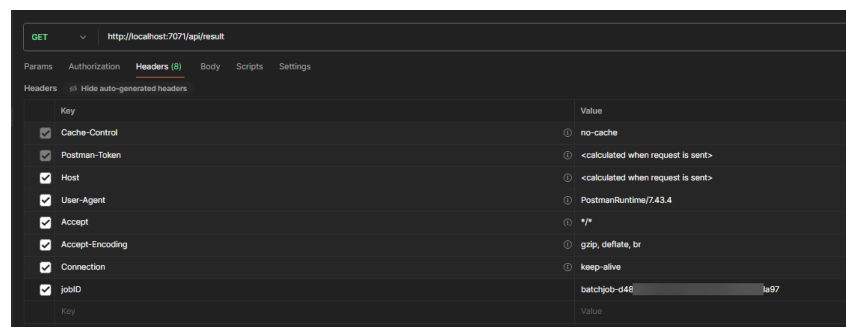
27 pav. Mokymo ir patalpavimo procesas Azure ML aplinkoje

Taip pat, kiekvieno mokymo metu užfiksuojama aibė metrikų, skirtų analizei:

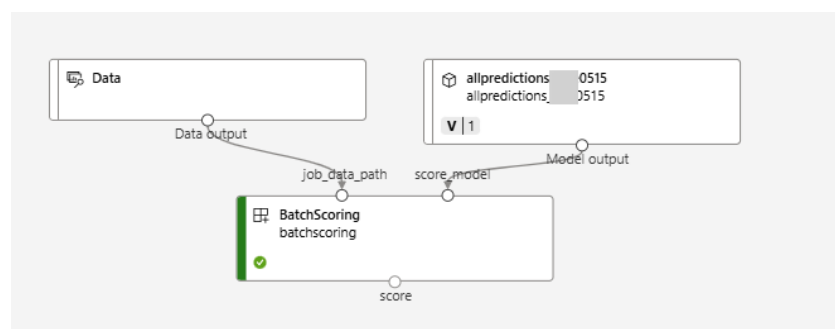


28 pav. Detalesnis Azure ML mokymo procesas su metrikomis

Atlikus mokymą, toliau kreipiamasi į rezultatų gavimo funkciją, kad išgauti spėjimus naujiems duomenims, siunčiama GET užklausa:



29 pav. GET užklausa, skirta rezultatų gavimui iš Azure ML



30 pav. Naujų duomenų spėjimo procesas Azure ML platformoje

Gauti rezultatai jau būna apdoroti JSON formatu, tinkami naudojimui Business Central aplinkoje:

4	105423	976,14	2025-01-20 – 2025-05-14	00:02:11	00:02:03	00:02:10
5	1466	48,86	2025-04-15 – 2025-05-14	00:01:30	00:01:50	00:01:03
6	1993	66,43	2025-04-14 – 2025-05-14	00:01:18	00:01:48	00:01:24
7	2140	62,94	2025-04-10 – 2025-05-14	00:00:30	00:02:01	00:02:00
8	1464	73,2	2025-04-25 – 2025-05-14	00:01:43	00:01:48	00:01:01

Gautos duomenų imtys itin svyruoja su turimu įrašų skaičiumi (nuo 1400 – 105000 įrašų), tačiau mokymo, patalpinimo ir rezultatų gavimo trukmė iš esmės nesiskiria. Didžiąją dalį procesų trukmės užima infrastruktūros startavimas ir sustabdymas. Procesai trunka pakankamai trumpai, atitinka greitaveikos reikalavimus.

4.3. Modelio tikslumo vertinimas naudojant realių įmonių duomenis

Pagrindinės metrikos, kurios lyginamos eksperimento metu, yra baudos vertė (angl. *penalty*), tikslumas (angl. *precision*) ir apimtis (angl. *coverage*):

- Baudos vertė (angl. *penalty*) – metrika, pritaikanti svorius atliktų klaidų ir atsisakymų spėti skaičiui:

$$\text{Baudos vertė} = \frac{10 \times N_{\text{klaidų}} + 1 \times N_{\text{atsisakymų spėti}}}{N_{\text{atliktų spėjimų}}} \quad (11)$$

- Tikslumas (angl. *precision*) – metrika, skaičiuojanti teisingų spėjimų dalį iš atliktų spėjimų:

$$\text{Tikslumas} = \frac{N_{\text{teisingų spėjimų}}}{N_{\text{atliktų spėjimų}}} \quad (12)$$

- Apimtis (angl. *coverage*) – metrika, nurodanti kiek bendrai buvo atlikta spėjimų

$$\text{Apimtis} = \frac{N_{\text{atliktų spėjimų}}}{N_{\text{visų įrašų}}} \quad (13)$$

Geriausiai pasirodęs modelis buvo išbandytas ant 8 įmonių realių duomenų ir gauti rezultatai, aprašyti 35 ir 36 lentelėse:

35 lentelė. Eksperimento, atlikto ant 8 įmonių realių duomenų, rezultatai

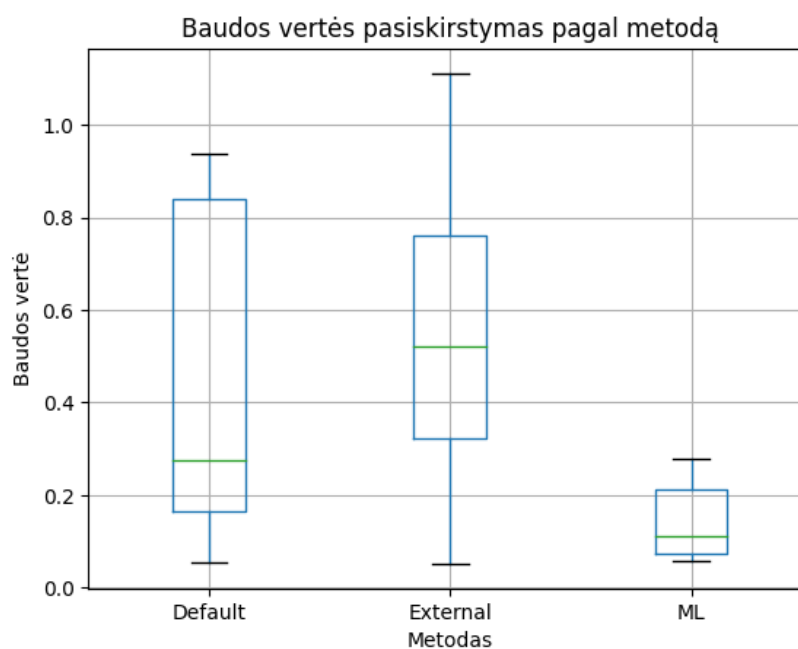
Įm. nr.	Test. imties dydis	Default apimtis % (baz.)	Default tikslumas % (baz.)	External apimtis % (baz.)	External tikslumas % (baz.)	MM apimtis %	MM tikslumas %	Default baudos vertė	External baudos vertė	MM baudos vertė
1	3108	90,315	98,896	91,023	96,854	94,208	99,556	0,197	0,376	0,100
2	5086	80,220	100	78,097	99,924	82,344	99,713	0,198	0,225	0,200
3	6305	77,335	98,339	94,750	93,488	88,406	98,152	0,355	0,669	0,279
4	105423	96,186	99,670	35,933	86,907	96,599	99,739	0,070	1,111	0,059
5	1466	97,954	99,652	97,544	99,720	99,318	99,313	0,055	0,052	0,075
6	1993	8,329	97,590	75,213	93,863	94,431	99,256	0,937	0,709	0,126
7	2140	8,178	100	78,458	98,213	99,112	99,340	0,918	0,356	0,074
8	1464	74,112	92,535	82,172	91,106	78,279	99,564	0,812	0,909	0,251

Iš rezultatų galima teigti, kad MM metodų parinkimo funkcionalumas visais atvejais grąžina geresnius rezultatus, nei naudojant kažkurį vieną iš metodų. Spėjimų apimtis yra didesnė bei atliekami spėjimai yra tikslesni, šias metrikas apibendrina 36 lentelė.

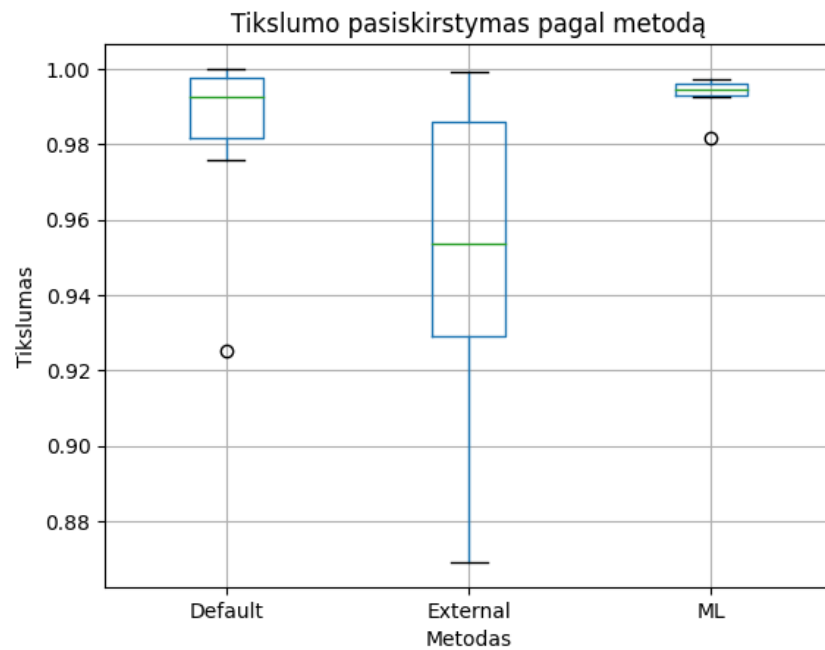
36 lentelė. Eksperimento rezultatų vidurkių apibendrinimas

Metodas	Baudos vertė (vid. $\pm$ SD)	Tikslumas (vid. $\pm$ SD)	Apimtis (vid. $\pm$ SD)
Default	0.4426 ± 0.3826	$98.34 \% \pm 2.49 \%$	$66.58 \% \pm 37.01 \%$
External	0.5510 ± 0.3596	$95.01 \% \pm 4.54 \%$	$79.15 \% \pm 19.33 \%$
MM	0.1456 ± 0.0862	$99.33 \% \pm 0.51 \%$	$91.59 \% \pm 7.83 \%$

Metodų rezultatuose galima matyti, kad tiek baudos vertės, tiek tikslumo pasiskirstymas per metodus yra glaudesnis, kai naudojamas MM metodų parinkimas (žr. 33, 34 pav.)

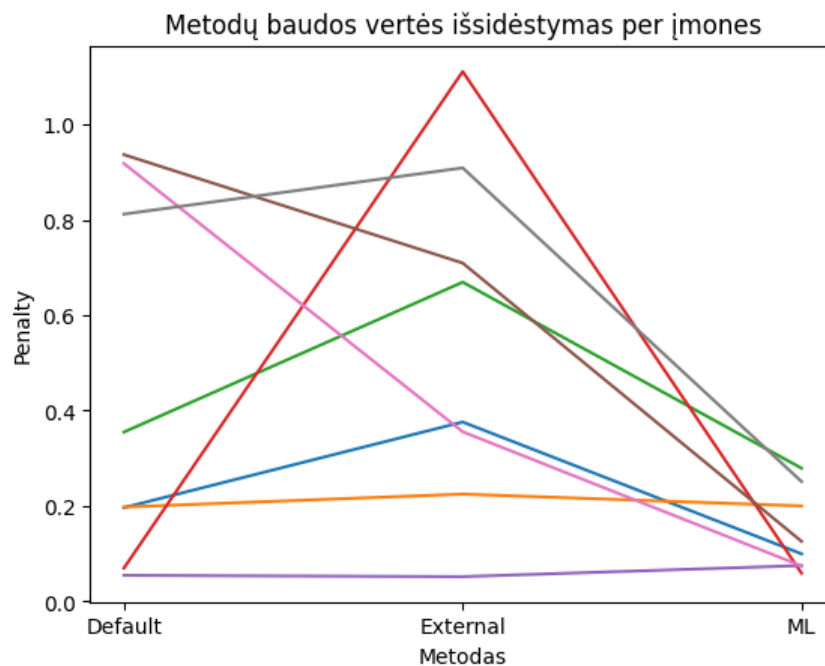


33 pav. Baudos vertės pasiskirstymas pagal metodą



34 pav. Tikslumo pasiskirstymas pagal metodą

Linijinėje diagramoje (žr. 35 pav.) matoma, kad didžiojoje dalyje analizuojamų įmonių yra baudos vertės mažėjimo tendencija, kai einama link MM metodo:



35 pav. Metodų baudos vertės išsidėstymas per įmones

4.4. Eksperimento rezultatų statistinė analizė

Rezultatų statistinio reikšmingumo analizei buvo panaudoti metodai, aprašyti 1.3 skyriuje. Tolimesniuose poskyriuose pateikiami kiekvieno iš metodų rezultatai ir gaunamos išvados.

4.4.1. ANOVA rezultatai ir efektų dydžiai

Baudos vertės (angl. penalty) reikšmės rezultatai $F(2,14)=4.394$, $p=0.033$ rodo statistiškai reikšmingą skirtumą tarp metodų, baudos vertės atžvilgiu.

Tikslumo (angl. precision) reikšmės rezultatas $F(2,14)=5.491$, $p=0.017$ rodo statistiškai reikšmingą skirtumą tarp metodų, tikslumo atžvilgiu.

Epektų dydžiai $\eta^2_p \approx 0.39$ (baudos vertės) ir 0.44 (tikslumo) viršija 0.14 ribą, kas rodo itin didelius efektus.

37 lentelė. ANOVA rezultatų apibendrinimas

Metrika	F	df (num, den)	p-reikšmė	η^2_p
Penalty	4.394	(2, 14)	0.0330	0.3856
Precision	5.491	(2, 14)	0.0174	0.4396

4.4.2. Tukey HSD post-hoc palyginimas

MM metodas statistiškai reikšmingai sumažina baudos vertę ir padidina tikslumą, lyginant su „External“ metodu:

- baudos verčių vidurkių skirtumas (Δ) -0.4053 rodo, kad MM metodas turi mažesnę baudos vertę, o koreguota p-reikšmė – $0.0391 < 0.05$;
- tikslumo vidurkių skirtumas (Δ) $+0.0432$ rodo, kad MM metodas turi didesnę tikslumą, o koreguota p-reikšmė – $0.0237 < 0.05$.

MM metodas nepasiekė statistinio reikšmingumo lyginant su „Default“ metodu:

- „Default“ vs ML palyginimai nerodė statistiškai reikšmingų skirtumų (koreguotos p-reikšmės > 0.05)

38 lentelė. Tukey HSD rezultatų apibendrinimas

Metrika	Palyginimas	Vid. Skirtumas Δ	Kor. p-reikšmė	Atmeta H_0
Baudos vertė	Default vs. External	+0.1083	0.7631	Ne
Baudos vertė	Default vs. ML	-0.2970	0.1541	Ne
Baudos vertė	External vs. ML	-0.4053	0.0391	Taip
Tikslumas	Default vs. External	-0.0333	0.0924	Ne
Tikslumas	Default vs. ML	+0.0099	0.7882	Ne
Tikslumas	External vs. ML	+0.0432	0.0237	Taip

4.4.3. T-testų rezultatai

Iš porinių t-testų rezultatų, matomų 39 lentelėje, prieinama prie šių išvadų:

- palyginus „External“ vs. ML matoma, kad MM metodas statistiškai reikšmingai sumažina baudos vertę tarp metodų ir reikšmingai pagerina tikslumą, o tai patvirtina Tukey HSD rezultatus;
- „Default“ vs. ML palyginime pastebima tendencija, kad MM metodas gali turėti mažesnę baudos vertę nei „Default“, tačiau tai nėra statistiškai patvirtinta 0.05 reikšmingumo lygmeniu (rezultatas yra labai arti reikšmingumo ribos ($p \approx 0.06$), bet jos nepasiekia).

39 lentelė. T-testų rezultatų apibendrinimas

Metrika	Palyginimas	t	p
Baudos vertė	Default vs. External	-0.6620	0.5292
Baudos vertė	Default vs. ML	2.2292	0.0610
Baudos vertė	External vs. ML	3.2441	0.0142
Tikslumas	Default vs. External	2.2602	0.0583
Tikslumas	Default vs. ML	-1.1042	0.3060
Tikslumas	External vs. ML	-2.6704	0.032

4.4.4. Shapiro-Wilk testo rezultatai

Gautas rezultatas ties „Default“ vs. ML palyginimu ($W = 0.7750$, $p = 0.0154 < 0.05$) rodo, kad normalumo prielaida yra pažeista – duomenų skirstinys statistiškai reikšmingai skiriasi nuo normalaus skirstinio, dėl šios priežasties parametriniai testai (t-testas) nėra tinkami šiai porai ir būtinas neparametrinis Wilcoxon testas:

40 lentelė. Shapiro-Wilk rezultatų apibendrinimas

Lyginimas	Shapiro–Wilk W	p-reikšmė
Default vs External	0.9247	0.4692
Default vs ML	0.7750	0.0154
External vs ML	0.9486	0.6970

4.4.5. Neparametriniai Friedman / Wilcoxon testai

Baudos vertės (angl. *penalty*) reikšmės rezultatai $\chi^2 = 5.250$, $p = 0.0724$ rodo nepakankamą statistinį reikšmingumą tarp metodų.

Tikslumo (angl. *precision*) reikšmės rezultatas $\chi^2 = 4.750$, $p = 0.0930$ taip pat rodo nepakankamą statistinį reikšmingumą tarp metodų, tačiau abu testai rodo artėjimą prie reikšmingumo ribos.

Porinis Wilcoxon testas su Bonferroni korekcija patvirtino statistinį „External“ vs. ML skirtumų reikšmingumą (koreguota p-reikšmė < 0.05 abiem metrikoms). Nors Shapiro–Wilk testas parodė normalumo prielaidos pažeidimą „Default“ vs. ML skirtumams, Wilcoxon testo rezultatai patvirtino, kad statistinio reikšmingumo nebuvimas „Default“ vs. ML atveju nepriklausė nuo duomenų skirstinio skirtumų.

41 lentelė. Friedman rezultatų apibendrinimas

Metrika	χ^2	p-value
Baudos vertė	5.250	0.0724
Tikslumas	4.750	0.0930

42 lentelė. Wilcoxon rezultatų apibendrinimas

Metrika	Lyginimas	T-statistika	Kor. p-reikšmė
Baudos vertė	Default vs External	13.00	1.0000
Baudos vertė	Default vs ML	4.00	0.1641
Baudos vertė	External vs ML	1.00	0.0469
Tikslumas	Default vs External	1.00	0.0469
Tikslumas	Default vs ML	21.00	1.0000
Tikslumas	External vs ML	0.00	0.0000

Apibendrinant eksperimento rezultatus, galime daryti tokias išvadas:

- MM modelis, naudojant realių įmonių duomenis, atitinka greitaveikos reikalavimus;
- išrinktas XGBoost MM modelis, optimizuotas 10:1 baudos įverčio svoriui, reikšmingai sumažina vidutinę baudos vertę ir padidina tikslumą, kai yra lyginamas su išoriniu „*External*“ metodu:
 - ANOVA rodo reikšmingą skirtumą tarp metodų baudos įverčio lyginyje ($F(2,14)=4,394$; $p=0,033$) ir tikslume ($F(2,14)=5,491$; $p=0,017$);
 - Tukey HSD post-hoc testas patvirtina, kad MM metodas, lyginant su „*External*“, reikšmingai sumažina baudos vertę ($\Delta=-0,4053$; $p=0,0391$) ir padidina tikslumą ($\Delta=+0,0432$; $p=0,0237$);
 - neparametriniai testai taip pat patvirtina MM pranašumą prieš „*External*“ metodą, ($T\text{-stat}=1.00$; $p=0,0469$);
- vidutinė MM baudos vertė ($0,1456 \pm 0,0862$) yra ženkliai mažesnė nei „*Default*“ ($0,4426 \pm 0,3826$) ir „*External*“ ($0,5510 \pm 0,3596$), o tikslumas – aukštesnis ($99,33 \% \pm 0,51 \%$ vs. $98,34 \% \pm 2,49 \%$ ir $95,01 \% \pm 4,54 \%$), kartu užtikrinant aukštesnę spėjimų apimtį ($91,59 \% \pm 7,83 \%$ vs. $66,58 \% \pm 37,01 \%$ ir $79,15 \% \pm 19,33 \%$);
- eksperimento metu lyginant MM metodą su „*Default*“ metodu, niekur nebuvo pasiekta 0,05 reikšmingumo riba. Tai rodo, kad H_0 hipotezė būna paneigiama, ir H_1 ir H_2 hipotezės būna pagrįstos tik dalinai – tik tada, kai lyginimas yra atliekamas su „*External*“ metodu.

Išvados

1. Atlikta literatūros analizė parodė, kad taisyklėmis grįsti metodai užtikrina didelį aiškumą ir stabilumą paprastose, struktūruotose atitikimo užduotyse, tačiau prastai prisitaiko prie sudėtingų ir besikeičiančių scenarijų. Mašininio mokymosi algoritmai (įskaitant ansamblius, grafų modelius ir meta-klasifikatorius) geriau valdo daug-su-daug atvejus ir adaptuojasi pagal istorinius duomenis, tačiau reikalauja daugiau skaičiavimo išteklių ir pradinio žymėjimo. Išanalizavus jau sukurtus metodus prieita prie išvados, kad optimalus sprendimas – hibridinė architektūra, jungianti taisyklių sistemą su MM komponentais, leidžianti palaipsniui didinti automatizacijos lygį ir pasiekti aukštą atpažinimo tikslumą bei patikimumą.
2. Atlikta esamų konkurentinių sprendimų lyginamoji analizė parodė, kad rinkoje esą platformų, siūlančių gretinimo funkcionalumą, tačiau dauguma platformų naudoja tik taisyklių pagrindu veikiančius sprendimus, reikalaujančius rankinio paruošimo, arba siūlančius ribotą mašininio mokymosi modelių integraciją.
3. Sudarytas kuriamo algoritmo parinkimo meta-modelio projektas, aprašantis duomenų manipuliavimo logiką tarp „Business Central“, „Azure Functions“ ir „Azure ML“ modulių, rodantis, kaip bus atliekamas taisyklių bei išorinių MM metodų taikymas, optimaliausio atpažinimo būdo parinkimas kiekvienai banko operacijai. Apibrėžti funkciniai ir nefunkciniai reikalavimai, panaudojimo atvejų modelis ir diegimo diagrama, nurodantys, kaip algoritmų parinkimo meta-modelis turi veikti.
4. Realizuotas siūlomas hibridinis finansinių operacijų gretinimo algoritmas pagal iškeltus funkcinius bei nefunkcinius reikalavimus, panaudojus analizės metu nustatytas projekto realizavimo technologijas. Panaudotos tokios technologijos, kaip „Azure Functions“ ir „Azure ML“, algoritmas sukurtas naudojant „Python“ programavimo kalbą. Duomenų rinkiniui panaudoti demonstraciniai duomenys, gauti iš užsakovo, buvo dokumentuota jų apdorojimo logika, atliktas duomenų valymas, dublikatų ir trūkstamų reikšmių apdorojimas bei *one-hot* kodavimas.
5. Tyrimo metu ištirta plati prižiūravimo mokymosi algoritmų aibė, įskaitant ansamblių metodus, linijinius modelius, kaimynų metodus, sprendimų medžius, naivųjį Bayes'ą, daugiasluoksnius perceptronus, o kiekvieno modelio veikimas buvo iliustruotas maišos matricomis, požymių svarbos grafikais. Taikyta 10-ies kartotinių kryžminė patikra, o įtraukus klaidingų spėjimų ir susilaikymo kaštus (10:1) į mokymą bei optimizavus sprendimo priėmimo slenkstį, pasirinktas geriausiai pasirodęs „XGBoost“ meta-klasifikatorius, užtikrinantis >90 % padengimą, ~99,4 % tikslumą ir rodantis žymiai geresnius rezultatus nei baziniai taisyklėmis grįsti (tikslumas 81,6%), ar išoriniai (tikslumas 73,1%) metodai.
6. Atliktas eksperimentas ant realių aštuonių įmonių duomenų, parodė, kad realizuotas MM modelis (XGBoost su 10:1 svoriais) visais atvejais pasiekė didesnę spėjimų apimtį (~91,6 %) ir tikslumą (~99,3 %) bei žymiai mažesnę vidutinę baudos vertę (~0,146) nei taisyklėmis grįsti (~0,443) ir išoriniai (~0,551) metodai. Statistinis reikšmingumas tikrintas ANOVA (baudos $F(2,14)=4,394$, $p=0,033$; tikslumas $F(2,14)=5,491$, $p=0,017$) ir Tukey HSD bei poriniais Wilcoxon testais. Gauti rezultatai parodė, kad MM metodas reikšmingai tiksliau spėja banko eilutes prieš išorinį metodą, tačiau palyginimas su taisyklėmis grindžiamais metodais nepasiekė 0,05 ribos ($p \approx 0,06$), kas rodo, kad skirtumas linkęs į reikšmingumą, bet duomenų imtis per maža jį tvirtai patvirtinti. Nepaisant aiškaus praktinio pagerėjimo (baudos sumažėjimas ~30 %, kritiškai blogų spėjimų skaičiaus mažėjimas), ribotas realių klientų duomenų kiekis sumažino statistinio testo galią, todėl reikalinga platesnė imtis pilnam algoritmo veiksmingumo įrodymui.

Literatūros sąrašas

1. „Foundations and Innovations in Data Fusion and Ensemble Learning for Effective Consensus“. Žiūrėta: 2025 m. gegužės 19 d. [Interaktyvus]. Adresas: <https://www.mdpi.com/2227-7390/13/4/587>
2. K. Tumer ir J. Ghosh, „Analysis of decision boundaries in linearly combined neural classifiers“, *Pattern Recognit.*, t. 29, nr. 2, p. 341–348, vas. 1996, doi: 10.1016/0031-3203(95)00085-2.
3. D. Von Winterfeldt ir W. Edwards, *Decision analysis and behavioral research*. Cambridge [Cambridgeshire]; New York: Cambridge University Press, 1986.
4. B. Kamiński, M. Jakubczyk, ir P. Szufel, „A framework for sensitivity analysis of decision trees“, *Cent. Eur. J. Oper. Res.*, t. 26, nr. 1, p. 135–159, 2018, doi: 10.1007/s10100-017-0479-6.
5. R. E. Schapire, „The strength of weak learnability“, *Mach. Learn.*, t. 5, nr. 2, p. 197–227, birž. 1990, doi: 10.1007/BF00116037.
6. L. Breiman, „Bagging predictors“, *Mach. Learn.*, t. 24, nr. 2, p. 123–140, rugpj. 1996, doi: 10.1007/BF00058655.
7. D. Wood, T. Mu, A. Webb, H. Reeve, M. Luján, ir G. Brown, „A Unified Theory of Diversity in Ensemble Learning“, 2024 m. vasario 7 d., *arXiv*: arXiv:2301.03962. doi: 10.48550/arXiv.2301.03962.
8. L. Breiman, „BIAS, VARIANCE, AND ARCING CLASSIFIERS“.
9. Y. Freund ir R. E. Schapire, „A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting“, *J. Comput. Syst. Sci.*, t. 55, nr. 1, p. 119–139, rugpj. 1997, doi: 10.1006/jcss.1997.1504.
10. T. Chen ir C. Guestrin, „XGBoost: A Scalable Tree Boosting System“, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, rugpj. 2016, p. 785–794. doi: 10.1145/2939672.2939785.
11. L. Breiman, „Random Forests“, *Mach. Learn.*, t. 45, nr. 1, p. 5–32, spal. 2001, doi: 10.1023/A:1010933404324.
12. S. V. Stehman, „Selecting and interpreting measures of thematic classification accuracy“, *Remote Sens. Environ.*, t. 62, nr. 1, p. 77–89, spal. 1997, doi: 10.1016/S0034-4257(97)00083-7.
13. G. James, D. Witten, T. Hastie, ir R. Tibshirani, *An Introduction to Statistical Learning*, t. 103. Springer Texts in Statistics, vol. 103. New York, NY: Springer New York, 2013. doi: 10.1007/978-1-4614-7138-7.
14. K. Gopalakrishnan, „Toward Transparent and Interpretable AI Systems in Banking: Challenges and Perspectives“, 2023.
15. „Matching Rules in Financial Reconciliation and Reporting“, *Accounting Insights*. Žiūrėta: 2025 m. gegužės 19 d. [Interaktyvus]. Adresas: <https://accountinginsights.org/matching-rules-in-financial-reconciliation-and-reporting/>
16. O. O. Sarumi, B. A. Ojokoh, O. A. Sarumi, ir O. S. Adewale, „Financial accounts reconciliation system using enhanced mapping algorithm“, *Int. J. Bus. Intell. Data Min.*, t. 21, nr. 2, p. 171, 2022, doi: 10.1504/IJBIDM.2022.124843.
17. Y. Rhee, S. Yoon, ir H. Park, „Exploring Knowledge Trajectories of Accounting Information Systems Using Business Method Patents and Knowledge Persistence-Based Main Path Analysis“, *Mathematics*, t. 10, nr. 18, p. 3349, rugs. 2022, doi: 10.3390/math10183349.

18. R. services Inc, „Record-to-Report Intelligent Automation: The Top 4 Use Cases“, Rely services Inc. Žiūrėta: 2025 m. gegužės 19 d. [Interaktyvus]. Adresas: <https://www.relyservices.com/blog/record-to-report-intelligent-automation-use-cases>
19. S. L. Lúcio, „Supervised Record Linkage For Banking Reconciliations“, vas. 2020, Žiūrėta: 2025 m. balandžio 3 d. [Interaktyvus]. Adresas: <https://repositorio-aberto.up.pt/handle/10216/133580>
20. J. Munoz, M. Jalili, ir L. Tafakori, „Automating Bank Reconciliation Through Graph Representation Learning and Downstream Record Linkage“, 2024 m. sausio 17 d., *Social Science Research Network, Rochester, NY*: 4698441. doi: 10.2139/ssrn.4698441.
21. P. A. Chew ir D. G. Robinson, „Automated account reconciliation using probabilistic and statistical techniques“, *Int. J. Account. Inf. Manag.*, t. 20, nr. 4, p. 322–334, saus. 2012, doi: 10.1108/18347641211272722.
22. „ANOVA Test: An In-Depth Guide with Examples“. Žiūrėta: 2025 m. gegužės 19 d. [Interaktyvus]. Adresas: <https://www.datacamp.com/tutorial/anova-test>
23. J. T. E. Richardson, „Eta squared and partial eta squared as measures of effect size in educational research“, *Educ. Res. Rev.*, t. 6, nr. 2, p. 135–147, saus. 2011, doi: 10.1016/j.edurev.2010.12.001.
24. J. W. Tukey, „Comparing Individual Means in the Analysis of Variance“, *Biometrics*, t. 5, nr. 2, p. 99–114, 1949, doi: 10.2307/3001913.
25. „Student’s *t*-test“, *Wikipedia*. 2025 m. balandžio 9 d. Žiūrėta: 2025 m. gegužės 19 d. [Interaktyvus]. Adresas: https://en.wikipedia.org/w/index.php?title=Student%27s_t-test&oldid=1284704142
26. „analysis of variance test for normality (complete samples)† | Biometrika | Oxford Academic“. Žiūrėta: 2025 m. gegužės 19 d. [Interaktyvus]. Adresas: <https://academic.oup.com/biomet/article-abstract/52/3-4/591/336553?redirectedFrom=fulltext&login=false>
27. W. W. Daniel, *Applied Nonparametric Statistics*. PWS-KENT Pub., 1990.
28. F. Wilcoxon, „Individual Comparisons by Ranking Methods“, *Biom. Bull.*, t. 1, nr. 6, p. 80–83, 1945, doi: 10.2307/3001968.
29. „General Python FAQ“, Python documentation. Žiūrėta: 2025 m. gegužės 19 d. [Interaktyvus]. Adresas: <https://docs.python.org/3/faq/general.html>
30. „Documentation scikit-learn: machine learning in Python — scikit-learn 0.21.3 documentation“. Žiūrėta: 2025 m. gegužės 19 d. [Interaktyvus]. Adresas: <https://scikit-learn.org/0.21/documentation.html>
31. sdgilley, „What is Azure Machine Learning? - Azure Machine Learning“. Žiūrėta: 2025 m. gegužės 19 d. [Interaktyvus]. Adresas: <https://learn.microsoft.com/en-us/azure/machine-learning/overview-what-is-azure-machine-learning?view=azureml-api-2>
32. ggailey777, „Azure Functions overview“. Žiūrėta: 2025 m. gegužės 19 d. [Interaktyvus]. Adresas: <https://learn.microsoft.com/en-us/azure/azure-functions/functions-overview>
33. K. Gorman ir C. Walker, „Data reconciliation based on computer analysis of data“, US20220405836A1, 2022 m. gruodžio 22 d. Žiūrėta: 2025 m. gegužės 10 d. [Interaktyvus]. Adresas: <https://patents.google.com/patent/US20220405836A1/en>
34. „Connect bank and credit card accounts to QuickBooks Online“. Žiūrėta: 2025 m. gegužės 10 d. [Interaktyvus]. Adresas: https://quickbooks.intuit.com/learn-support/en-us/help-article/banking/connect-bank-credit-card-accounts-quickbooks/L4yDAHMH_US_en_US

35. „Blackline Reconciliation: What It Is and How it Works | Numeric“. Žiūrėta: 2025 m. gegužės 19 d. [Interaktyvus]. Adresas: <https://www.numeric.io/blog/blackline-reconciliation>
36. „Cadency Reconciliation Match Product Sheet“, Trintech. Žiūrėta: 2025 m. gegužės 19 d. [Interaktyvus]. Adresas: <https://www.trintech.com/brochure/cadency-reconciliation-match-product-sheet/>
37. „Bankfeed“. Žiūrėta: 2025 m. gegužės 19 d. [Interaktyvus]. Adresas: <https://bankfeed.com/features/>
38. jswymer, „Welcome to Microsoft Dynamics 365 Business Central - Business Central“. Žiūrėta: 2025 m. gegužės 19 d. [Interaktyvus]. Adresas: <https://learn.microsoft.com/en-us/dynamics365/business-central/welcome>

Priedai

1 priedas. Sujungto galutinio mokymo duomenų rinkinio laukų paaiškinimai

Požymis	Įrašų sk.	Duom. tipas	Aprašymas
LineNo	72189	string	Įrašo nr.
RecognitionProvider	72189	string	Atpažinimo metodo pavadinimas
Description	72189	string	Banko operacijos aprašymas
RecognizedAccountType	72189	string	Atpažintos sąskaitos tipas
RecognizedAccountNo	72189	string	Atpažintos sąskaitos numeris
ConfidenceScore	72189	double	Vidinis pasitikėjimo įvertis
Recommended	72189	bool	Ar atp. metodas rekomenduoja atpažintą įrašą
AccountRecognitionType_	72189	integer	Atpažinimo tipas - tuščias
AccountRecognitionType_Bank Account	72189	integer	Atpažinimo tipas – filtrai banko sąskaitoms
AccountRecognitionType_Customer	72189	integer	Atpažinimo tipas – filtrai pirkėjams
AccountRecognitionType_Customer Bank Account	72189	integer	Atpažinimo tipas – filtrai pirkėjų banko sąskaitoms
AccountRecognitionType_Customer by Account Name & Recognized Document	72189	integer	Atpažinimo tipas - filtrai pagal pirkėjų dokumentus
AccountRecognitionType_Customer by E2E ID	72189	integer	Atpažinimo tipas - filtrai end2end ID
AccountRecognitionType_Customer by Recognized Document	72189	integer	Atpažinimo tipas - filtrai pirkėjo dokumentams
AccountRecognitionType_Similarity	72189	integer	Atpažinimo tipas – panašumas
AccountRecognitionType_Text-to-Account Mapping	72189	integer	Atpažinimo tipas – aprašymo su sąsk. lyginimai
AccountRecognitionType_Vendor	72189	integer	Atpažinimo tipas - filtrai tiekėjams
AccountRecognitionType_Vendor Bank Account	72189	integer	Atpažinimo tipas - filtrai tiekėjų banko sąskaitoms
AccountRecognitionType_Vendor by Account Name & Recognized Document	72189	integer	Atpažinimo tipas - filtrai pagal tiekėjų dokumentus
AccountRecognitionType_Vendor by Credit Transfer	72189	integer	Atpažinimo tipas - filtrai tiekėjo pagal kredito pervedimus
AccountRecognitionType_Vendor by Recognized Document	72189	integer	Atpažinimo tipas - filtrai tiekėjo dokumentams
BankAccountNo	72189	string	Banko sąskaitos numeris
BankAccountIBAN	72189	string	Banko sąskaitos IBAN
SignedStatementAmount	72189	double	Pasirašytas kiekis
StatementAmount	72189	double	Banko išrašo įvestas kiekis
StatementChargesAmount	72189	double	Banko išrašo mokesčių kiekis
DebtorName	72189	string	Skolininko vardas
DebtorIBAN	72189	string	Skolininko IBAN
DebtorRegistrationNo	72189	string	Skolininko reg. nr
CreditorName	72189	string	Kreditoriaus vardas
CreditorIBAN	72189	string	Kreditoriaus IBAN
CreditorRegistrationNo	72189	string	Kreditoriaus reg. nr
TransactionText	72189	string	Banko išrašo tekstas
AdditionalTransactionInfo	72189	string	Pap. banko išrašo tekstas
CreditorReference	72189	string	Kreditoriaus kodas
EndtoEndID	72189	string	End-to-end identification code

TransactionType_Credit	72189	bool	Pervedimo tipas - kreditinis
TransactionType_Debit	72189	bool	Pervedimo tipas - debetinis
AccountCurrency_AUD	72189	bool	Valiuta AUD
AccountCurrency_CAD	72189	bool	Valiuta CAD
AccountCurrency_DKK	72189	bool	Valiuta DKK
AccountCurrency_EUR	72189	bool	Valiuta EUR
AccountCurrency_GBP	72189	bool	Valiuta GBP
AccountCurrency_NOK	72189	bool	Valiuta NOK
AccountCurrency_PLN	72189	bool	Valiuta PLN
AccountCurrency RON	72189	bool	Valiuta RON
AccountCurrency_SEK	72189	bool	Valiuta SEK
AccountCurrency_THB	72189	bool	Valiuta THB
AccountCurrency_USD	72189	bool	Valiuta USD