Kauno technologijos universitetas

Informatikos fakultetas

Agreguoto LIME pritaikymo globaliam klasifikavimo modelių paaiškinamumui tyrimas

Baigiamasis magistro projektas

Gintarė Zokaitytė

Projekto autorė

prof. dr. Agnė Paulauskaitė-Tarasevičienė

Vadovė

Kaunas, 2025

Kauno technologijos universitetas

Informatikos fakultetas

Agreguoto LIME pritaikymo globaliam klasifikavimo modelių paaiškinamumui tyrimas

Baigiamasis magistro studijų projektas

Dirbtinio intelekto informatika (6211BX007)

Gintarė Zokaitytė

Projekto autorė

prof. dr. Agnė Paulauskaitė-Tarasevičienė

Vadovė

doc. dr. Liudas Motiejūnas

Recenzentas

Kaunas, 2025

Kauno technologijos universitetas

Informatikos fakultetas

Gintarė Zokaitytė

Agreguoto LIME pritaikymo globaliam klasifikavimo modelių paaiškinamumui 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) intelektinės 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 sąraše;

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.

Gintarė Zokaitytė

Patvirtinta elektroniniu būdu

Zokaitytė, Gintarė. Agreguoto LIME pritaikymo globaliam klasifikavimo modelių paaiškinamumui tyrimas. Magistro baigiamasis projektas / vadovė prof. dr. Agnė Paulauskaitė-Taraseviičienė; Kauno technologijos universitetas, Informatikos fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Informatikos mokslai, Informatika (B01)

Reikšminiai žodžiai: XAI, LIME, lokalus paaiškinimas, globalus paaiškinimas, atributų svarba, klasifikavimo modelių paaiškinamumas, agregavimo strategijos.

Kaunas, 2025. 52 p.

Santrauka

Kadangi mašininio mokymosi (ML) modeliai vis labiau įtakoja svarbius sprendimus sveikatos priežiūros, finansų ir politikos srityse, skaidrių ir interpretuojamų dirbtinio intelekto (DI) sistemų poreikis tampa kritinis. Lokalūs interpretuojami modelių agnostiniai paaiškinimai (LIME) yra plačiai taikomas metodas lokaliam modelių interpretavimui, tačiau jų globalus agregavimas išlieka metodologiškai ribotas. Šiame darbe nagrinėjamos ir tobulinamos globalaus paaiškinimo strategijos, agreguojant lokalius LIME paaiškinimus, sprendžiant žinomus apribojimus, tokius kaip jautrumas triukšmui ir nenuoseklus atributų aktualumas. Siekiant pagerinti globalų paaiškinimą buvo išbandytos dvi naujos strategijos: paaiškinimų svorio nustatymas pagal jų tikslumą (R^2) ir branduolio tankio įvertinimo (KDE) pagrindu sukurta agregacija. Eksperimentiniai rezultatai patvirtino, kad R^2 įtraukimas reikšmingai sumažino pasiskirstymo atstumą (JSD), pagerino stabilumą ir sušvelnino našumo pablogėjimą pašalinant atributus. Nors KDE pasiekė aukščiausią rango koreliaciją (*Spearman* = 0,91), ji atsiliko pagal kitus rodiklius, o tai rodo kompromisus. Patobulinti agregavimo metodai pranoko esamus literatūros bazinius rodiklius pagal kelis kokybės rodiklius, parodydami jų potencialą sustiprinti globalių paaiškinimų patikimumą ir aiškumą klasifikavimo modeliuose.

Zokaitytė, Gintarė. Aggregated LIME for the global explainability of classification models. Master's Final Degree Project / supervisor prof. dr. Agnė Paulauskaitė-Tarasevičienė; Faculty of Informatics, Kaunas University of Technology.

Study field and area (study field group): Computer science, Informatics (B01)

Keywords: XAI, LIME, local explanation, global explanation, feature importance, classification model interpretability, aggregation strategies.

Kaunas, 2025. 52 pages.

Summary

As machine learning (ML) models increasingly influence high-stakes decisions in healthcare, finance, and policy, the need for transparent and interpretable artificial intelligence (AI) systems becomes critical. Local Interpretable Model-Agnostic Explanations (LIME) offer a widely adopted approach for local model interpretability, yet their global aggregation remains methodologically limited. This project investigates and enhances global explanation strategies by aggregating local LIME explanations, addressing known limitations such as noise sensitivity and inconsistent feature relevance. To improve global interpretability, two novel strategies were tested: weighting explanations by their fidelity (R^2) and kernel density estimation (KDE)-based aggregation. Experimental results confirmed that incorporating R^2 significantly reduced distributional divergence (JSD), improved stability, and mitigated performance degradation under feature removal. Although KDE achieved the highest rank correlation (Spearman = 0.91), it lagged in other metrics, indicating trade-offs. The enhanced aggregation methods outperformed existing literature baselines across several quality metrics, demonstrating their potential to strengthen the reliability and clarity of global explanations in classification models.

Turinys

Lentelių sąrašas

Paveikslų sąrašas

Santrumpų sąrašas

ADI – aiškinamasis dirbtinis intelektas (angl. *Explainable artificial intelligence*).

DI – dirbtinis intelektas (angl. *Artificial intelligence*).

LIME – lokalūs interpretuojami nuo modelio nepriklausantys paaiškinimai (angl. *Local Interpretable Model-Agnostic Explanations*)

ML – mašininis mokymasis (angl. *Machine learning*)

Įvadas

Didėjanti priklausomybė nuo dirbtinio intelekto (DI) modelių įvairiose srityse, nuo sveikatos priežiūros ir finansų iki teisinių sistemų ir ne tik, padidino modelių sprendimų priėmimo procesų skaidrumo poreikį. Šie modeliai dažnai yra įpareigoti numatyti sprendimus, kurie turi didelę įtaką asmenims ir visuomenei, pvz., skaičiuoti kreditingumą ir paskolų pasiūlymus [31], asistuoti diagnozuojant ligas [39] arba nustatyti pirmenybę teisėsaugos ar sveikatos priežiūros ištekliams [15,37]. Nors mašininio mokymosi (ML) modelių, ypač gilaus mokymosi architektūrų, prognozavimo galimybės pasiekė aukštą lygį, šių „juodųjų dėžių“ (angl. *black boxes*) sudėtingumas ir neskaidrumas kelia didelių iššūkių. „Juodųjų dėžių“ pobūdis apsunkina galimybę vartotojams ir suinteresuotosioms šalims suprasti, kaip šie modeliai daro išvadas, o tai gali sukelti nepasitikėjimą ir dvejones priimant šias technologijas [5,17]. Be to, galimos klaidos, šališkumas ar nenumatytos pasekmės kritinėse srityse gali sukelti rimtų pasekmių – nuo finansinių nuostolių [7] iki pavojaus žmonių sveikatai [36]. Šie iššūkiai pabrėžia augantį skaidrių, interpretuojamų ir patikimų DI sistemų poreikį, galinčių suteikti prasmingų įžvalgų apie jų veiklą.

Kai dirbtinis intelektas vis labiau įtraukiamas į daug dėmesio reikalaujančias programas, vyriausybės ir organizacijos visame pasaulyje įveda taisykles, siekdamos užtikrinti etišką ir skaidrų jo naudojimą. Vienas iš didesnių tokios iniciatyvos pavyzdžių – Europos Sąjungos DI aktas [42], kuriuo siekiama apibrėžti DI kūrimo ir naudojimo taisykles pagal rizikos lygį. Pavyzdžiui, DI akto [42] 13 straipsnyje nurodoma, kad aukštos rizikos DI sistemos turi būti kuriamos taip, jog jų naudotojai galėtų interpretuoti jų rezultatus ir tinkamai juos naudoti, taip užtikrinant veikimo skaidrumą ir atskaitomybę. Aiškinamasis dirbtinis intelektas (ADI) atlieka lemiamą vaidmenį įgyvendinant šiuos reikalavimus, suteikdamas įrankius ir metodus paaiškinti sudėtingų DI modelių veikimo principus. Be teisinių reikalavimų laikymosi, ADI taip pat skatina pasitikėjimą, atskaitomybę ir sąžiningumą, todėl tai būtina tiek techniniam, tiek etiniam DI diegimui.

Tarp esamų ADI metodų, lokalių interpretuojamų nuo modelio nepriklausančių paaiškinimų (angl. *Local Interpretable Model-Agnostic Explanations*) (LIME) metodas [33] yra plačiai pripažintas dėl savo gebėjimo generuoti lokalius paaiškinimus, suteikiančius atributų svarbos balus atskiroms prognozėms. Nors lokalūs paaiškinimai yra vertingi, jie dažnai neatskleidžia bendro modelio elgesio įvairiuose duomenų kontekstuose [11]. Galimybės išplėsti LIME metodą globaliems modelių paaiškinimams yra pradėtos tirti [13,25,33], tačiau išlieka iššūkių parenkant tinkamus paaiškinimų agregavimo būdus tiksliems, interpretuojamiems ir stabiliems globaliems paaiškinimams. Šiuo projektu siekiama ištirti ir palyginti esamus LIME metodus globaliems paaiškinimams bei pasiūlyti naujus, inovatyvius agregavimo metodus. Šis tyrimas, analizuojant paaiškinimų ištikimumo modeliams, stabilumo ir skaičiavimo imlumo aspektus, siekia nustatyti dabartinių praktikų trūkumus ir pateikti sprendimus, prisidedančius prie ADI pažangos.

Projekto tikslas – pasiūlyti modifikuotą globalaus paaiškinimo metodą, pagrįstą lokaliais LIME paaiškinimais, pritaikyti jį klasifikavimo modelių paaiškinimui ir eksperimentiškai įvertinti jo kokybę, analizuojant paaiškinimų atitikimą, poveikį modeliui ir stabilumą, bei palyginti su esamais sprendimais. Šiam tikslui pasiekti iškelti šie uždaviniai:

1. atlikti globalių klasifikavimo modelių paaiškinamumo metodų literatūros analizę, išskiriant LIME pagrįstų metodų ribotumus ir galimus tobulinimo būdus, susijusius su lokalių paaiškinimų agregavimu į globalius;

2. parengti tyrimo planą, apibrėžiantį naudojamus duomenų rinkinius, klasifikavimo modelius, siūlomas naujas metodikas, vertinimo kriterijus, numatyti kiekybinio įvertinimo metrikas lokalių ir globalių paaiškinimų kokybei vertinti bei realizuoti eksperimentinę aplinką šio tyrimo vykdymui;

3. išmokyti klasifikavimo modelius, atlikti LIME parametrų paiešką ir sugeneruoti reprezentatyvų lokalių paaiškinimų rinkinį, skirtą globalių paaiškinimo metodų analizės etapui;

4. realizuoti ir eksperimentiškai ištirti modifikuotas lokalių paaiškinimų agregavimo strategijas, analizuojant skirtingų komponentų ir konfigūracijų įtaką globalių paaiškinimų kokybei;

5. atlikti modifikuotų globalių LIME paaiškinimo metodų palyginamąją analizę su esamais sprendimais, įvertinant paaiškinimų atitikimą klasifikavimo modelio atributų svarboms, poveikį modelio veikimui ir paaiškinimų stabilumą.

1. Globalių klasifikavimo modelių paaiškinamumo metodų literatūros analizė

Šiame skyriuje pateikiama paaiškinamo dirbtinio intelekto (ADI) apžvalga, daugiausia dėmesio skiriant *post-hoc* analizės metodams, kurie pagerina mašininio mokymosi modelių aiškinamumą. Nagrinėjamos vietinių ir globalių paaiškinimų sąvokos, ML modelių aiškinimo metodų klasifikacija pagal sritį, universalumą bei išvestį. Ypatingas dėmesys skiriamas LIME metodui ir jo variantams, įvertinat jų veikimo principą, privalumus ir trūkumus. Skyriuje taip pat aptariamos ištirtos strategijos, kaip lokalius paaiškinimus sujungti į globalų paaiškinimą ir vertinimo metrikos paaiškinimų kokybei įvertinti. Pabrėžiamos pagrindinės tyrimų spragos ir tobulinimo galimybės.

1.1. Aiškinamasis dirbtinis intelektas ir jo panaudojimas

Aiškinamasis dirbtinis intelektas (ADI) yra dirbtinio intelekto atšaka, orientuota į metodų ir įrankių, leidžiančių interpretuoti ir suprasti sudėtingus mašininio mokymosi (ML) modelius, kūrimą. Nors tradicinės DI sistemos optimizuoja nuspėjimo tikslumą, dėl didėjančio jų sudėtingumo ir „juodosios dėžės“ pobūdžio modelių, kuriems trūksta skaidrumo, kaip modelio sprendimas yra priimamas. ADI metodais siekiama išspręsti šią problemą, pateikiant žmonėms suprantamus paaiškinimus, paaiškinančius modelio išvesties motyvus, užtikrinant aiškinamumą nepakenkiant našumui [12].

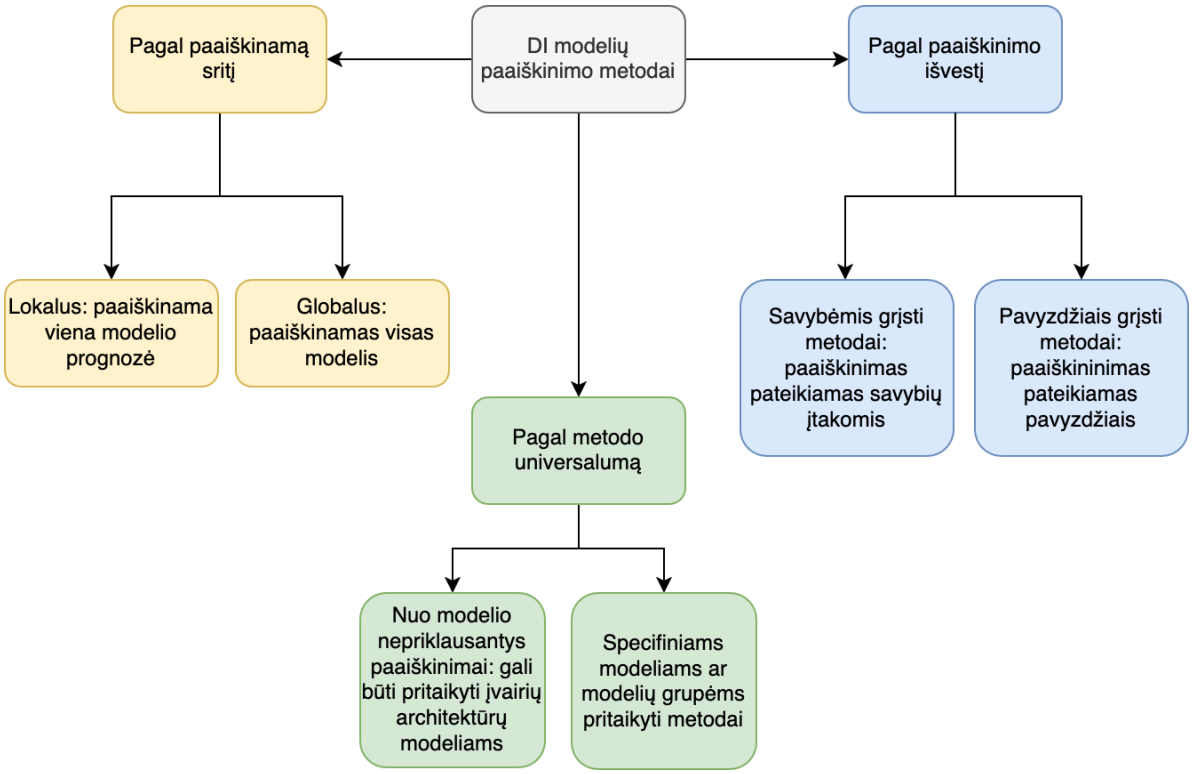
ADI naudojimas apima kelis mašininio mokymosi ciklo etapus, apimančius modelio kūrimą, diegimą ir vertinimą. Modelio mokymo ir patvirtinimo etape ADI metodai suteikia esminius įrankius, leidžiančius nustatyti ir spręsti tokias problemas kaip duomenų paklaida, modelio klaidos ir neoptimalūs atributų ryšiai. Pavyzdžiui, tokie metodai kaip *SHAP* [26] suteikia galimybę DI sistemų kūrėjams priskirti modelio prognozes atskiriems įvesties atributams, kad būtų galima išsamiai suprasti atributų indėlį. Dalinės priklausomybės diagramos (angl. *Partial Dependence Plots*) [16] papildo šias analizes, vizualizuodamos ribinį atskirų atributų poveikį išvesčiai, padedant sistemų kūrėjams aptikti linijinius ir nelinijinius ryšius tarp įvesties ir išvesčių [29]. Šios įžvalgos yra naudingos atributų pasirinkimo, hiperparametrų derinimo ir modelio mokymo stadijose.

Post-hoc analizės etape ADI metodai orientuoti į jau parengtų modelių interpretavimą, užtikrinant įdiegtų sistemų skaidrumą ir atskaitomybę. *Post-hoc* metodai leidžia suinteresuotosioms šalims pateikti paaiškinimus, kurie yra lokalios arba globalios reikšmės. Lokalūs paaiškinimai, tokie kaip LIME [33], suteikia įžvalgų apie konkrečias išvestis, apytiksliai įvertinant modelio elgesį aplink atskirus atvejus. Perturbuojant įvesties atributus ir stebint jų poveikį išvesčiai, LIME nustato, kurie atributai turi didžiausią įtaką tam tikram sprendimui. Tuo tarpu globalūs paaiškinimai analizuoja modelio elgseną visame duomenų rinkinyje ir pateikia išsamią atributų svarbos ir sąveikos apžvalgą. Naudodami tokius metodus kaip apibendrintos SHAP [26] reikšmės ir globalios atributų svarbos balai, specialistai gali suprasti modelio sprendimų modelius, aptikti paklaidas, šališkumus ir įvertinti bendrą modelio našumą [24].

ADI integravimas visuose DI sistemos kūrimo etapuose užtikrina, kad mašininio mokymosi modeliai būtų interpretuojami ir patikimi, padedant modelių derinimui, validavimui ir diegimui. Atsižvelgdami į lokalias ir globalias perspektyvas, ADI metodai padidina modelio skaidrumą, todėl DI diegėjai ir naudotojai gali priimti pagrįstus sprendimus ir kartu užtikrinti atitiktį reguliavimo bei etikos standartams.

1.2. Aiškinamojo dirbtinio intelekto metodų klasifikavimas

ADI metodus galima suskirstyti į kategorijas pagal jų taikymo sritį, pritaikomumą ir požiūrį į paaiškinamumą. Šiame skyriuje nagrinėjami pagrindiniai ADI metodų klasifikavimo aspektai (1 pav.), iliustruojantis ADI metodų klasifikavimą pagal apimtį (lokalus ir globalus), pagal išvestį (pavyzdžiu pagrįstas ir atributais pagrįstas) bei pagal modelio universalumą (nuo modelio nepriklausomas ar specifiniams modeliams pritaikytas).

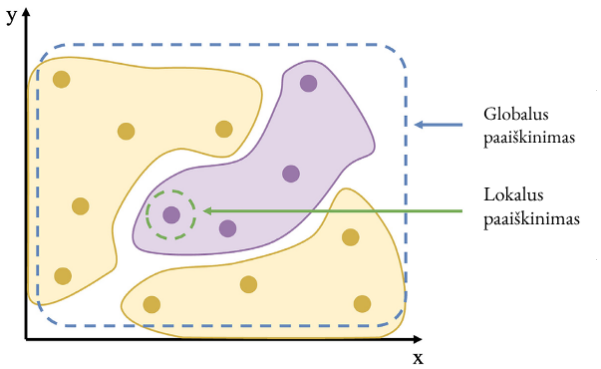


1 pav. DI modelių paaiškinimo metodų klasifikavimas

1.2.1. Lokalus ir globalus paaiškinimas

Pagrindinis skirtumas tarp lokalaus ir globalaus paaiškinimo yra paaiškinama modelio sritis ir detalumas (2 pav.). Lokalūs paaiškinimai analizuoja, kaip daromos atskiros prognozės, sutelkiant dėmesį į vieną įvesties atvejį. Šie metodai veikia mažoje, tiksliai apibrėžtoje įvesties erdvės srityje, nustatydami konkrečių atributų indėlį į konkretų modelio sprendimą. Išskirdami veiksnius, turinčius įtakos atskiriems rezultatams, lokalūs paaiškinimai suteikia tikslinių įžvalgų, kurios ypač naudingos DI sistemoms, kurioms reikalingas skaidrumas sprendimų lygiu, pvz., klinikinės diagnozės, sukčiavimo aptikimas ar kredito patvirtinimai.

Tuo tarpu globalūs paaiškinimai suteikia išsamų supratimą apie bendrą modelio elgesį visame duomenų rinkinyje. Skirtingai nuo vietinių metodų, kuriuose pagrindinis dėmesys skiriamas individualiems sprendimams, globalūs ADI metodai analizuoja atributų svarbą, įvesties ir išvesties priklausomybes ir sistemingus modelio numatymo modelius. Šie paaiškinimai apibendrina įžvalgas apie visus duomenų atvejus, leidžia plačiau įvertinti modelio logiką, nustatyti šališkumą ir užtikrinti, kad būtų laikomasi etikos standartų. Globalūs paaiškinimai yra ypač vertingi vertinant modelių patikimumą, optimizuojant našumą ir patvirtinant jų pritaikymą įvairiems duomenų kontekstams.



2 pav. Globalaus ir lokalaus paaiškinimo schema [9]

Vietiniai ir globalūs paaiškinimai papildo vienas kitą, suteikdami tiek mikro, tiek makro lygmens modelio interpretavimo perspektyvas. Derinant šiuos metodus, DI sistemų kūrėjai ir naudotojai gali pasiekti išsamesnį ir visapusiškesnį modelio sprendimų supratimą.

1.2.2. Aiškinamieji dirbtinio intelekto metodai pagal pritaikymo universalumą

Kitas svarbus ADI metodų požymis yra jų pritaikymas įvairių tipų ML modeliams. Nuo modelio nepriklausantys metodai yra bendrosios paskirties metodai, kurie gali būti taikomi bet kokiam „juodos dėžės“ modeliui. Jie analizuoja ryšį tarp įvesčių ir išvesčių, nereikalaujant prieigos prie modelio vidinės architektūros. Bendra šių metodų savybė yra tai, kad jie priklauso nuo perturbacijomis pagrįstų metodų, kai įvesties atributai yra sistemingai modifikuojami, kad būtų galima stebėti išvesties pokyčius. Įvertinant, kaip modelis reaguoja į atskirų ypatybių svyravimus, šie metodai apytiksliai nustato atributų svarbą ir sprendimų ribas, sukuriant įžvalgas, kurias galima interpretuoti ir taikyti įvairiems modelių tipams. Tokių metodų pavyzdžiai metodai – LIME [33], *SHAP* [26], *Anchors* [39].

Kita vertus, modeliams būdingi metodai panaudoja tam tikrų ML modelių vidinę struktūrą, kad pateiktų paaiškinimus. Šie metodai paprastai yra efektyvesni, tačiau taikomi tik konkretiems ML modelių tipams. Pavyzdžiui, *Layer-wise Relevance Propagation* [4] yra neuroniniams tinklams pritaikyta technika, perskirstanti prognozavimo aktualumą sluoksnis po sluoksnio, siekiant nustatyti įtakingiausias savybes. Panašiai *Tree SHAP* [27] yra sukurtas specialiai medžiu pagrįstiems modeliams, efektyviai apskaičiuojant atributų indėlį, panaudojant sprendimų medžių struktūrą. Šie modeliui būdingi metodai yra naudingi dėl jų suderinimo su pagrindine architektūra, todėl pateikiami tikslūs ir skaičiavimo požiūriu efektyvūs paaiškinimai.

1.2.3. Aiškinamieji dirbtinio intelekto metodai pagal išvestį

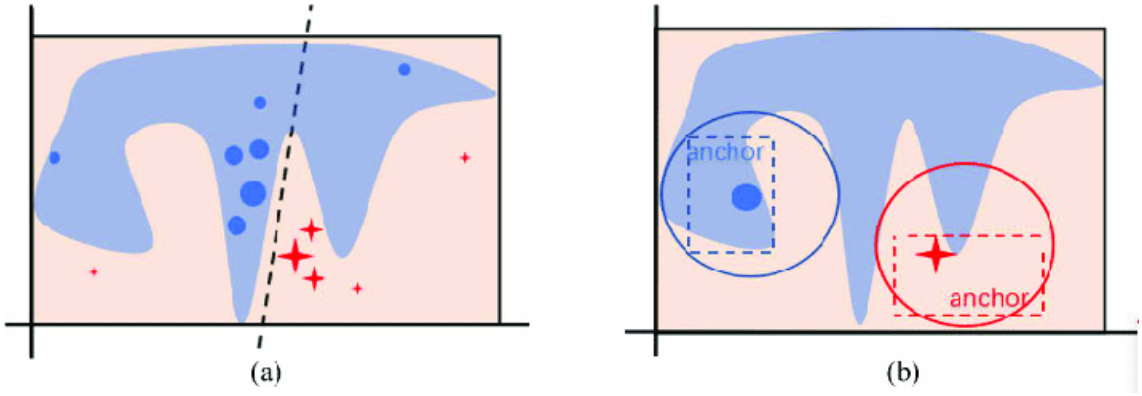
ADI metodai gali būti toliau skirstomi į kategorijas pagal jų pateiktų paaiškinimų išvesties formą. Pavyzdžiais pagrįsti metodai remiasi konkrečiais atvejais, kad iliustruotų modelio elgesį, o atributais pagrįsti metodai yra skirti kiekybiškai įvertinti įvesties atributų įtaką modelio išvestims.

1.2.3.1. Pavyzdžiais grįsti metodai

Pavyzdžiais pagrįsti ADI metodai suteikia įžvalgų apie ML modelio elgesį, sutelkiant dėmesį į konkrečius duomenų atvejus. Šie metodai iš esmės yra lokalaus pobūdžio ir suteikia paaiškinimus, susietus su individualiomis išvestimis arba konkrečiais įvesties erdvės regionais. Jie naudoja perturbacijas, optimizavimą ar atrankos kriterijus, kad nustatytų atvejus, kurie atspindi tipinius

dėsningumus, kritines sprendimų ribas arba minimalius pakeitimus, reikalingus modelio išvesčiai pakeisti. Šiuos metodus sieja dėmesys lokaliam aiškinamumui, pasitikėjimas duomenimis pagrįstais (angl. *data-driven*) scenarijais ir gebėjimas paaiškinti ML modelius nereikalaujant prieigos prie jų architektūros. Populiarūs tokių metodų pavyzdžiai yra *Anchors* [32], priešininigi paaiškinimai (angl. *Counterfactual Explanations*) (CE) [38] ir kontrastingi paaiškinimai (angl. *Contrastive Explanations Method*) (CEM) [10].

Anchors metodas generuoja „jei-tuomet“ (angl. *„if-then“*) taisykles, kurios atspindi stabilius regionus įvesties erdvėje, kur modelis nuosekliai pateikia tą pačią išvestį. Šie paaiškinimai gaunami perturbacijų ir sąlyginių paskirstymų atrankos būdu, siekiant nustatyti atributų ir reikšmių derinius (3 pav. (b)), kurie įtvirtina modelio prognozę. Metodas pateikia tikslius lokalius paaiškinimus, kuriuos galima interpretuoti ir taikyti įvairiems duomenų tipams. Tačiau didelių dimensijų duomenų erdvėse jie reikalauja daug skaičiavimo resursų ir yra apriboti specifiniais regionais, kuriuos paaiškina [32].



3 pav. Paaiškinamos srities palyginimo schema: LIME metodas (a) ir Anchors metodas (b). [28]

CE metodas nustato minimalius įvesties pakeitimus, kurie pakeistų modelio išvestį. Šie paaiškinimai remiasi optimizavimo problemų sprendimu, kurie sumažina atstumą tarp pradinio egzemplioriaus ir priešingos padėties, tuo pačiu užtikrinant patikimumą. CE suteikia veiksmingų įžvalgų ir yra veiksmingi norint suprasti sprendimų ribas, tačiau šis metodas gali reikalauti didelių skaičiavimo resursų ir pateikti neunikalius paaiškinimus, todėl gali atsirasti dviprasmybių rekomendacijose [38].

CEM metodas išplečia priešingus paaiškinimus nustatydamas atitinkamus teigiamus aspektus (atributus, būtinas prognozei) ir susijusias neigiamas (atributus, kuriuos pašalinus pasikeis modelio prognozė). Šis metodas naudoja optimizavimą numatyti minimalius pakeitimus ir auto-enkoderius užtikrinti pakeistų duomenų tikroviškumą. CEM siūlo išsamius dvipusius paaiškinimus, tačiau tam reikalingi dideli skaičiavimo ištekliai ir, norint gauti tikslius rezultatus, jis priklauso nuo auto-enkoderio kokybės [10].

Pavyzdžiais pagrįsti ADI metodai suteikia vertingus lokalius paaiškinimus, sutelkiant dėmesį į konkrečius duomenų atvejus. *Anchors* suteikia interpretuojamomis taisyklėmis pagrįstas įžvalgas, CE išryškina veiksmingus kelius modelio prognozei pakeisti, o CEM pateikia bendrą palaikančių ir priešingų veiksnių vaizdą. Nors kiekvienas metodas turi skirtingas stipriąsias puses, jų apribojimai pabrėžia hibridinių metodų, kurie derina paaiškinimų tikslumą, metodo plečiamumą (dideliems duomenų kiekiams bei sudėtingiems aiškinamiems modeliams) ir globalią paaiškinimo aprėptį, poreikį.

1.2.3.2. Atributais grįsti metodai

Atributais pagrįsti ADI metodai skirti kiekybiškai įvertinti įvesties atributų indėlį į mašininio mokymosi modelių prognozes. Skirtingai nuo pavyzdžiais pagrįstų metodų, kuriuose dėmesys sutelkiamas į konkrečius atvejus, šie metodai analizuoja duomenų rinkinio atributų įtaką, pateikdami tiek lokalias (konkrečių atvejų), tiek globalias (viso duomenų rinkinio) įžvalgas. Išaiškinant atskirų atributų vaidmenis, šie metodai padeda interpretuoti sudėtingus modelius, didinti skaidrumą ir padeda atributų parinkime. Pagrindiniai metodai tokio pobūdžio metodai sutelkti į:

1. atributų ryšius (dalinės priklausomybės diagramos (PDP) [16] individualių sąlyginių lūkesčių (ICE) [18] grafikai, sukauptų lokalių efektų (ALE) [3] grafikai);
2. atributų įtaką (atributų svarbos matai [14] , *Shapley* paaiškinimai (SHAP) [26];
3. modelio aproksimaciją (globalūs surogatiniai modeliai ir LIME [33]).

Atributų ryšiai

Dalinės priklausomybės diagramos (PDP) vizualizuoja vienos ar dviejų savybių ribinį poveikį modelio prognozėms, apskaičiuojant visų kitų savybių įtaką. PDP apskaičiuoja numatomą modelio išvestį, keisdami tikslinę funkciją, o kitus atributus išsaugant fiksuotus. Tai atliekama integruojant per bendrą netikslinių funkcijų paskirstymą. Gautoje diagramoje pavaizduotas ryšys tarp pasirinktos funkcijos ir modelio išvesties. PDP yra veiksmingos siekiant atskleisti globalias tendencijas ir nelinijinius ryšius. Tačiau priimama prielaida, kad atributai yra nepriklausomi, todėl duomenų rinkiniuose su koreliuotais atributais gali atsirasti klaidinančių rezultatų. Be to, PDP gali pateikti nepatikimas interpretacijas dėl ekstrapoliacijos, kai ribinis pasiskirstymas viršija mokymo duomenis [16].

ICE diagramos suteikia konkrečiam atvejui įžvalgų, vizualizuodami funkcijos poveikį individualioms prognozėms, papildydami PDP siūlomą pasaulinę perspektyvą. ICE diagramos sukuria kreivę kiekvienam duomenų rinkinio egzemplioriui, keičiant tikslinę funkciją, o visas kitas išlaikant pastovias. Tai atskleidžia nevienalytį funkcijos poveikį įvairiais atvejais, todėl galima identifikuoti pogrupius su skirtingais modeliais ICE diagramos puikiai išryškina kintamumą ir sąveiką, kurią PDP gali užgožti. Tačiau jie gali būti netvarkingi dideliuose duomenų rinkiniuose, o tai apsunkina interpretavimą. Norint pagerinti skaitomumą, dažnai reikalingi agregavimo arba grupavimo metodai. [18].

ALE diagramose atsižvelgiama į PDP apribojimus, atsižvelgiama į funkcijų koreliacijas ir išvengiama ekstrapoliacijos, o tai užtikrina patikimesnį funkcijų efektų atvaizdavimą. ALE diagramose apskaičiuojamas lokalus poveikis, apskaičiuojant modelio gradiento vidurkį per mažus intervalus, atsižvelgiant į faktinį atributo pasiskirstymą. Tada šie vietiniai efektai kaupiami, kad būtų sukurta globali diagrama. Atsižvelgdamos į atributų priklausomybes, ALE diagramos yra patikima PDP alternatyva koreliuojantiems duomenims. Jie yra efektyvūs skaičiavimo požiūriu ir yra mažiau linkę klaidingai interpretuoti. Tačiau jų pasikliovimas intervalais pagrįstais skaičiavimais gali pernelyg supaprastinti labai netiesinius ryšius [3].

Atributų įtaka

Atributų svarbos metodai [14] įvertina santykinį kiekvieno atributo indėlį į modelio prognozes. Jie plačiai naudojami norint nustatyti svarbiausius atributus ir supaprastinti modelio kūrimą. Permutacijos svarba, populiarus metodas, įvertina atributų svarbą atsitiktinai sumaišydamas ypatybės

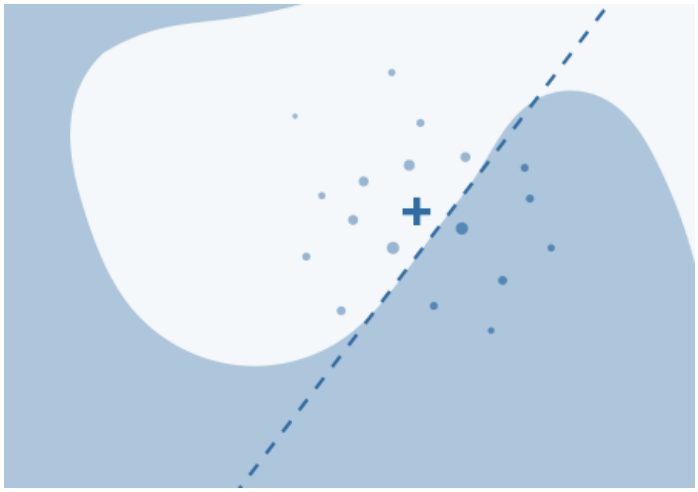
reikšmes ir išmatuodamas sumažėjusį modelio našumą. Šios permutacijos išskiria atributo poveikį, stebint numatymo tikslumo pokytį. Skirtingai nuo vidinių metodų (pvz., medžių savybių svarba atsitiktiniuose miškuose), permutacijos svarba veikia nepriklausomai nuo modelio struktūros, todėl ji taikoma bet kuriam nuspėjamajam modeliui. Atributų svarba suteikia intuityvių įžvalgų apie atributų tinkamumą ir padeda sumažinti duomenų dimensijas. Tačiau tai yra skaičiavimo prasme intensyvus procesas, ypač dideliems duomenų rinkiniams ar sudėtingiems modeliams, o rezultatai gali skirtis dėl permutacijų stochastinio pobūdžio. Norint padaryti tvirtas išvadas, reikia pakartotinai vertinti.

SHAP [26] pasitelkia bendradarbiavimo žaidimų teoriją, kad pateiktų nuoseklius ir teoriškai pagrįstus funkcijų priskyrimus. Metodas paaiškina tiek individualias prognozes, tiek globalią modelio elgseną, priskirdamas svarbos išvesčiai reikšmes atributams. SHAP reikšmės gaunamos lyginant modelio prognozę su kiekvienu atributu ir be jos visuose galimuose atributų pogrupiuose. Šis išsamus įvertinimas užtikrina teisingumą ir nuoseklumą, o visų savybių indėlis į prognozavimo skirtumą sumuoja globalų vaizdą. SHAP suteikia tvirtą pagrindą tiek lokaliems, tiek globaliems paaiškinimams, derindamas aiškinamumą su teoriniu griežtumu. Jis yra universalus ir patikimas. Tačiau sudėtingų modelių SHAP verčių skaičiavimas yra brangus, todėl praktikoje reikia naudoti apytikslius duomenis.

Modelių aproksimacija

Globalūs surogatiniai modeliai apytiksliai suderina sudėtingų „juodųjų dėžių“ modelių prognozes, naudodami interpretuojamus modelius, tokius kaip tiesinė regresija, sprendimų medžiai arba taisyklėmis pagrįstos sistemos. Jie suteikia globalų modelio elgsenos supratimą mokantis supaprastinto „juodosios dėžės“ modelio vaizdavimo. Pakaitinis modelis mokomas naudojant tuos pačius įvesties duomenis kaip ir juodosios dėžės modelis, o juodosios dėžės prognozės yra tikslinė išvestis. Šis trijų etapų procesas apima prognozių generavimą iš „juodosios dėžės“ modelio, interpretuojamo surogatinio modelio tipo parinkimą ir pakaitinio modelio mokymą įvesties ir išvesties porose. Surogatinio modelio patikimumas įvertinamas lyginant jo prognozes su juodosios dėžės modelio prognozėmis, o mažesnė paklaida rodo geresnį aproksimavimą. Globalūs surogatiniai modeliai yra labai lankstūs, nes leidžia pakeisti interpretuojamus ir „juodosios dėžės“ modelius. Tam pačiam „juodosios dėžės“ modeliui galima paruošti kelis pakaitinius modelius, kad būtų galima matyti įvairias perspektyvas. Tačiau surogatiniai modeliai tik apytiksliai įvertina „juodosios dėžės“ modelio elgesį ir gali nesugebėti užfiksuoti tam tikrų duomenų pogrupių niuansų. Jie taip pat yra mažiau veiksmingi paaiškinant individualias prognozes.

LIME [33], skirtingai nuo globalių surogatų, LIME sutelkia dėmesį į vieną atvejį ir sukuria vietinį paaiškinimą. LIME perturbuoja įvesties duomenis keičiant atributų reikšmes ir stebi, kaip keičiasi „juodosios dėžės“ modelio prognozės. Šios perturbacijos naudojamos norint išmokyti paprastą interpretuojamą modelį (pvz., tiesinę regresiją), kuris lokaliai atitinka „juodosios dėžės“ modelį. Modelis priskiria atributų svorius pagal jų indėlį į išvestį, leidžiantį interpretuoti atvejį. LIME pateikia tikslius, konkrečiam atvejui būdingus paaiškinimus, todėl jis ypač naudingas norint suprasti lokalizuoto modelio elgseną. Metodas nepriklauso nuo aiškinamo modelio ir gali būti taikomas įvairiems duomenų tipams, įskaitant tekstą ir vaizdus. Tačiau jo priklausomybė nuo perturbacijų gali sukelti nestabilumą, o jo paaiškinimai gali būti netinkami kitiems atvejams.

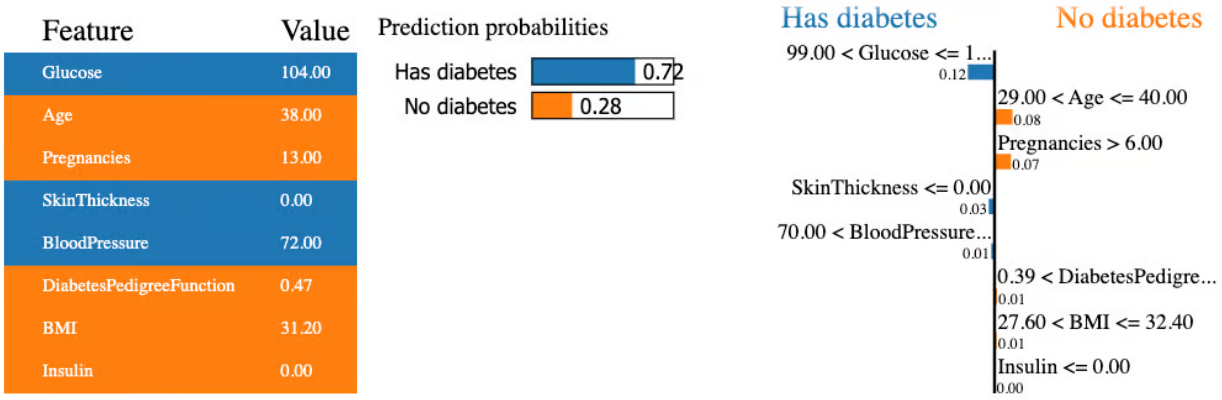


4 pav. LIME metodo iliustracinis pavyzdys

Atributais pagrįsti ADI metodai suteikia esminius įrankius mašininio mokymosi modeliams interpretuoti, išaiškinant įvesties atributų indėlį. Tokie metodai kaip atributų svarba, PDP, 4 pav. diagramos, ALE diagramos, SHAP, pasauliniai surogatiniai modeliai ir LIME patenkina įvairius interpretavimo poreikius, nuo pasaulinių tendencijų iki konkrečių atvejų įžvalgų. Nors kiekvienas metodas turi skirtingas stipriąsias puses, jų apribojimai, tokie kaip skaičiavimo sąnaudos, prielaidos dėl savybių nepriklausomumo arba pasikliovimas apytiksliais skaičiavimais, pabrėžia kruopštaus taikymo svarbą. Integruodami papildomus metodus, specialistai gali padidinti AI sistemų skaidrumą ir patikimumą.

1.3. Lokalus interpretuojamas nuo modelio nepriklausantis paaiškinimas (LIME)

LIME [33] yra lokalus, nuo modelio nepriklausantis, atributais grįstas ADI metodas. LIME paaiškina individualias prognozes, apytiksliai nustatant vietinį „juodosios dėžės“ modelio elgesį, naudojant interpretuojamą surogatinį modelį. Šis metodas grąžina paaiškinimą atributų įtakos forma, surogatinio modelio koeficientai pateikiami kaip atributų įtakos reikšmės. Paveikslėlyje (5 pav.) galime matyti tokio paaiškinimo pavyzdį.



a) b) c)

5 pav. LIME paaiškinimo pavyzdys: a) įvesties vienas duomenų atvejis; b) aiškinamo modelio išvesties klasių tikimybės; c) suskaičiuotas LIME paaiškinimas šiam duomenų atvejui, pateiktas atributų įtakos forma.

1.3.1. LIME metodika

LIME metodika apima aiškinamo modelio traktavimą kaip orakulą, galintį generuoti bet kurios įvesties prognozes. Norint išmokyti lokalius surogatinius modelius, procesas prasideda pasirinkus dominantį duomenų atvejį, kurio „juodosios dėžės“ prognozei reikia paaiškinimo. Tada duomenų rinkinys perturbuojamas, kad aplink atvejį būtų generuojami nauji pavyzdžiai, ir gaunamos aiškinamo modelio prognozės šiems perturbuotiems taškams. Kiekvienam atvejui priskiriamas svoris, atsižvelgiant į jo atstumą nuo aiškinamo atvejo, užtikrinant, kad arčiau esantys taškai turėtų didesnę įtaką. Tada mokomas svertinis interpretuojamas modelis, naudojant perturbuotą duomenų rinkinį, aproksimuojant aiškinamo modelio elgesį vietiniame regione. Galiausiai, prognozė paaiškinama analizuojant išmokytą surogatinį modelį, suteikiant įžvalgų apie veiksnius, lemiančius aiškinamo modelio sprendimą. LIME paaiškinimas išreikštas lygtimi:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (1)$$

kur:
x – vienas duomenų atvejis
ξ – paaiškinimas
f – aiškinamas modelis
g – interpretuojamas surogatinis modelis (pvz.: tiesinė *Ridge* regresija)
π_x– artumo funkcija svoriams skaičiuoti, įvertinanti kiek kiekvienas perturbuotas taškas panašus į x (parametras - kaimynystės dydis aplink atvejį x, kurią įtraukiame į paaiškinimo skaičiavimus)
L – nuostolio funkcija (pvz.: vidutinė kvadratinė paklaida)
Ω – surogatinio modelio sudėtingumas (maksimalus atributų skaičius)

1.3.2. LIME paaiškinimų patikimumo vertinimas

Šalia kiekvieno paaiškinimo taip pat grąžinamas patikimumo įvertinimas, išreikštas determinacijos koeficientu. Įvertinamas vietinių paaiškinimų tikslumas, siekiant išsiaiškinti, ar pagal spėjimo tikslumą, surogatinis modelis atitinka pirminį sudėtingą modelį, kurį mes paaiškiname. Lygtimi tai išreiškiama:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2)$$

R^2 - determinacijos koeficientas
n – perturbuotų taškų skaičius
y_i – aiškinamo modelio prognozė duomenų taškui *i* (klasės tikimybė)
$\hat{y}_i$ - surogatinio modelio prognozė duomenų taškui *i* (klasei pritaikytos regresijos išvesties tikimybė)
$\overline{y}$ – aiškinamo modelio prognozių klasei vidurkis

1.4. Globalūs paaiškinimai iš lokalių LIME paaiškinimų

Nors pavienio sprendimo paaiškinimas suteikia vertingų lokalios modelio elgsenos įžvalgų, jis nėra pakankamas norint įvertinti bendrą pasitikėjimą modeliu. Vis dėlto, kadangi lokalūs paaiškinimai dažnai atskleidžia reikšmingą atributų įtaką konkrečiuose kontekstuose, jie turi pakankamai informacinės vertės, kad galėtų būti agreguojami į globalius apibendrinimus. Keletas tyrimų [1,13,23,25,33] jau pasiūlė tokius agregavimo metodus, kurie leidžia globaliai paaiškinti modelį,

laikant jį „juodąja dėže“. Metodai pagrinde skiriasi pavyzdžių atrinkimo bei svorių skaičiavimo strategijomis.

1.4.1. Globalūs metodai pagal pavyzdžių atrinkimą

1.4.1.1. SP-LIME metodas

Kartu su originaliu LIME metodu, Ribeiro et al. [33] pristatė pirmąjį globalaus modelių paaiškinimo metodą, susidedantį iš lokalių paaiškinimų. Metodas susideda iš reprezentatyvių pavyzdžių atrankos, kurių paaiškinimai kartu atspindi bendrą modelio elgseną. Šiam tikslui naudojamas submodulinės optimizacijos algoritmas, vadinamas SP-LIME, kuris iš visų galimų lokalių paaiškinimų atrenka nedidelį, informacijos pertekliaus neturintį pavyzdžių rinkinį. Atranka SP-LIME metode grindžiama atributų padengimo optimizavimu. Iš pradžių kiekvienam duomenų taškui $x_i \in X$ sugeneruojamas lokalus paaiškinimas, sudarant paaiškinimų matricą $W \in \mathbb{R}^{nxd'}$. Tuomet sprendžiama optimizavimo užduotis, kuri siekia atrinkti B pavyzdžių rinkinį $V \subseteq X$, maksimaliai padengiantį svarbiausius atributus pagal funkciją:

$$SPPick(W, I) = arg \max_{\substack{V \subset X \\ |V| \leq B}} \sum_{j=1}^{d'} 1\left[\exists i \in V : W_{ij} > 0\right] \cdot I_j \quad (3)$$

Po atrankos, globali kiekvieno atributo j svarba apskaičiuojama tik pagal atrinktus pavyzdžius V, taikant šią formulę:

$$I_{SPLIME\_j} = \sqrt{\sum_{i \in V} |W_{ij}|} \quad (4)$$

čia V – atrinktų lokalių paaiškinimų aibė; W_{ij} – atributo j svarba i-ojo atvejo paaiškinime.

Tokiu būdu SP-LIME leidžia suformuoti kompaktišką ir reprezentatyvų lokalių paaiškinimų rinkinį, kuris atskleidžia svarbiausius atributus visame modelyje. Optimizavimo funkcija yra submodulinė, todėl ji sprendžiama godžiu algoritmu.

1.4.1.2. ILIME metodas

ILIME (*Interpretable LIME*) [11] praplečia LIME metodą, siekiant pateikti tiek lokalius, tiek globalius paaiškinimus. Norint suformuoti globalų paaiškinimą, ILIME pirmiausia išrenka atstovaujančių atvejų aibę naudojant klasterizavimą. Atvejai grupuojami į k klasterių (naudojant k-vidurkių algoritmą), o iš kiekvieno klasterio pasirenkamas centrinis taškas kaip reprezentatyvus pavyzdys. Kiekvienam šiam taškui sugeneruojamas lokalus paaiškinimas naudojant LIME. Globalus atributo j svarbos įvertis gaunamas skaičiuojant šių pasirinktų atvejų paaiškinimų atributo svarbos vidurkį:

$$I_{ILIME\_j} = \frac{1}{|S|} \sum_{i \in S} |w_{ij}| \quad (5)$$

čia S – atrinktų atvejų aibė; w_{ij} – atributo j svarba atvejui i priskirtame lokaliame paaiškinime.

1.4.1.3. OLEA metodas

OLEA (*Optimal Local Explainer Aggregation*) [23] metodas suformuluotas kaip kombinatorinio optimizavimo užduotis, kurios tikslas – iš lokalių paaiškinimų parinkti optimalią aibę, balansuojant padengimą ir tikslumą (angl. fidelity). Skirtingai nei heuristiniai metodai (pvz., SP-LIME), OLEA taiko sveikaskaitinį programavimą, siekiant išrinkti fiksuoto dydžio *K* lokalių paaiškinimų rinkinį taip, kad būtų maksimaliai padengti duomenų taškai ir kartu išlaikytas minimalus kiekvieno paaiškinimo tikslumo lygis. Optimizavimo užduotis formuluojama taip:

$$\max_{\gamma}\{\, Cov(\gamma, D) : Fid(\gamma, D) \geq \varphi, |\gamma| \leq K\} \tag{6}$$

čia γ – pasirinktų lokalių paaiškinimų aibė; $Cov(\gamma, D)$ – duomenų taškų, padengtų bent vienu paaiškinimu iš γ, skaičius; $Fid(\gamma, D)$ – mažiausias lokalaus paaiškinimo tikslumas aibėje γ; φ – nustatytas minimalus tikslumo slenkstis; K – didžiausias leidžiamas paaiškinimų skaičius.

Kiekvienas lokalus paaiškinimas taikomas regione (rutulyje su spinduliu *r* aplink centro tašką), o tikslumas matuojamas kaip proporcija atvejų, kurių klasifikacija pagal paaiškinimą sutampa su „juodosios dėžės“ modelio prognozėmis. Po atrankos atributų svarbos gali būti agreguojamos, tačiau pats OLEA metodas konkretaus agregavimo būdo nenurodo – tai paliekama interpretuoti taikymo kontekste, tačiau straipsnyje palyginimams naudotas absoliučių verčių vidurkis.

1.4.2. Globalūs metodai naudojantys svorį vidurkiui

Linden et al. [25] savo tyrime palygina kelis lokalių LIME paaiškinimų agregavimo būdus. Pirmiausia, GALE-AVG metodas apskaičiuoja vidutinį absoliutų reikšmės priskyrimą kiekvienam atributui visose situacijose, kuriose jis pasirodo:

$$I_{GALE-AVG,j} = \frac{\sum_{i=1}^{N} |W_{ij}|}{\sum_{i:W_{ij}\neq 0} 1} \tag{7}$$

čia W_{ij} – atributo *j* lokalus paaiškinimas atvejui *i*; skaitiklis sumuoja absoliučias atributų reikšmes per visus atvejus; vardiklis skaičiuoja, kiek ne 0 reiškmės atvejų turi atributas *j*; poveikis – neleidžia dažnai pasitaikantiems atributams dominuoti vien dėl to, kad jie dažnesni.

GALE-AVG tyrime lyginamas naujai siūlomu metodu – GALE-H-WEIGHTED, homogeniškai pasverta agregavimo strategija, skirta išryškinti nuoseklumą tarp skirtingų klasių.

Šio metodo skaičiavimai susideda iš 3 žingsnių:

1. Normalizuoti LIME svarbą pagal klasę (užtikrina, kad atributo svarba būtų vertinama atsižvelgiant į klasių pasiskirstymą).

$$p_j^c = \frac{\sqrt{\sum_{i\in S_c} |W_{ij}|}}{\sum_{c'\in L} \sqrt{\sum_{i\in S_{c'}} |W_{ij}|}} \tag{8}$$

čia W_{ij} – atributo *j* lokalus paaiškinimas atvejui *i;* S_c – visų atvejų, priskirtų klasei *c*, aibė; skaitiklis apskaičiuoja atributo *j* svarbumą klasėje *c*; vardiklis normalizuoja visose klasėse *L*.

2. Apskaičiuoti Šanono (angl. *Shannon*) entropiją. Ji matuoja, kaip plačiai atributo svarba pasiskirsčiusi tarp skirtingų klasių.

$$H_j = -\sum_{c \in L} p_j^c \log p_j^c \tag{9}$$

3. Apskaičiuoti homogeniškai pasvertą svarbą (pasveriama LIME svarba pagal entropiją, suteikiant daugiau svorio nuosekliems atributams). Žemiau pateiktoje formulėje entropijos reikšmės H_j normalizuojamos tarp 0 ir 1. Atimama iš 1 tam, kad mažos entropijos atributai (nuoseklūs) išlaikytų savo svarbą, didelės entropijos atributai (nenuoseklūs) būtų sumažinto svorio. Tai užtikrina, kad globaliai svarbūs atributai būtų būdingi konkrečioms klasėms.

$$I_{H,j} = \left(1 - \frac{H_j - H_{min}}{H_{max} - H_{min}}\right) I_{LIME,j}, \qquad kur\ I_{LIME,j} = \sqrt{\sum_{i=1}^{N} |W_{ij}|} \tag{10}$$

čia $I_{LIME,j}$ – naudoja kvadratinę šaknį iš sumos, kad pabrėžtų atributus, kurie pasirodo dažnai ir yra svarbūs.

1.4.2.1. NormLIME metodas

NormLIME [1] yra LIME metodo išplėtimas, skirtas gauti globalius arba klasių lygmens paaiškinimus agreguojant kelis lokalios svarbos vektorius. Vietoje tiesioginio vidurkinimo, NormLIME normalizuoja kiekvieno lokalaus paaiškinimo svorius pagal jų *L1* normą. Konkretaus atributo j globali svarba skaičiuojama tik iš tų lokalių paaiškinimų, kuriuose jis pasireiškia, ir apskaičiuojama pagal formulę:

$$I_j^{(NormLIME)} = \frac{1}{|E(j)|} \sum_{w \in E(j)} \frac{|w_j|}{\|w\|_1} \cdot |w_j| \tag{11}$$

čia *E(j)* – lokalių paaiškinimų aibė, kuriose atributas *j* turi nenulinę reikšmę; *w* – vieno paaiškinimo atributų svarbos vektorius; $\|w\|_1$ – *w* vektoriaus L1 norma.

Šis metodas išskiria atributus, kurie yra svarbūs santykinai savo kontekste, todėl dažnai pasikartojantys, bet ne dominuojantys atributai gali būti nuvertinti, o pastoviai svarbūs – išryškinti. NormLIME taip pat leidžia skaičiuoti klasės lygmens svarbą, kai *E(j)* apribojama tik atvejais, kurie priklauso konkrečiai prognozuotai klasei.

1 lentelė. Globalių paaiškinimų iš lokalių LIME paaiškinimų metodų palyginimas

Tyrimas	Metodo trumpinys	Agregavimo metodo sudėtinės dalys		
		Pavyzdžių atrinkimo algoritmas	**Svoris**	**Agregacijos matas**
Ribeiro et al. (2016) [33]	SP-LIME	Optimizuojamas atributų padengimas	-	Sumos kvadratinė šaknis
Linden et al. (2019) [25]	GALE-AVG	-	-	Absoliučių verčių vidurkis
	GALE-H-WEIGHTED	-	Atributų entropija per klases	Sumos kvadratinė šaknis
Ahern et al. (2019) [1]	NORM-LIME	-	L1 normalizuotos paaiškinimų vertės	Absoliučių verčių vidurkis
ElShawi et al. (2019) [13]	ILIME	Hierarchinis klasterizavimas	-	Absoliučių verčių vidurkis
Li et al. (2022) [23]	OLEA	Sveikaskaitis programavimas	-	Absoliučių verčių vidurkis

1 lentelėje apibendrintos žinomiausios globalių LIME paaiškinimų strategijos, išskaidytos pagal jų sudėtinius komponentus: pavyzdžių atrinkimo algoritmą, svėrimo mechanizmą ir agregavimo matą. Dauguma metodų remiasi absoliučių reikšmių vidurkiu, tačiau kai kurie – kaip SP-LIME ar GALE-H – taiko kvadratinės šaknies sumas. Tik keli metodai integruoja papildomus svorius ar atrankos strategijas, o tai riboja jų pritaikomumą įvairiose situacijose.

1.5. Literatūros analizės išvados

Aiškinamojo dirbtinio intelekto (ADI) metodų įvairovė leidžia interpretuoti tiek atskiras modelio prognozes (lokalūs metodai), tiek bendrą modelio elgseną (globalūs metodai). Nors šie metodai papildo vienas kitą, literatūroje išryškėja keli esminiai trūkumai, kurie riboja esamų sprendimų taikymą praktikoje.

Lokalūs metodai, tokie kaip LIME, pasižymi lankstumu ir modelio nepriklausomumu, tačiau jų patikimumas priklauso nuo lokalaus surogatinio modelio tikslumo, kuris dažnai lieka neįvertintas. Tuo tarpu populiarūs globalūs metodai, kaip SHAP, nors teoriškai pagrįsti, yra brangūs skaičiavimo požiūriu ir reikalauja supaprastinimų, kurie gali iškreipti paaiškinimų kokybę. Dėl to LIME pagrindu sukurti metodai vis dar laikomi viena praktiškiausių alternatyvų globaliam paaiškinimui.

Literatūroje išanalizuoti globalūs paaiškinimo metodai dažniausiai naudoja paprastas agregavimo strategijas – absoliučių verčių vidurkį ar kvadratinės šaknies sumą. Daug dėmesio skiriama pavyzdžių atrankos algoritmams, tačiau svėrimo komponentai, ypač paremti lokalaus paaiškinimo patikimumu (pvz., determinacijos koeficientu R^2), praktiškai nenaudojami. Be to, esami metodai retai lyginami tiesiogiai vienas su kitu, o jų taikymas apsiriboja dvejetainiais modeliais, ignoruojant daugiaklasių atvejų specifiką.

2. LIME agregacijų globaliam paaiškinimui tyrimo projektas

Šiame skyriuje pristatomas globalių LIME paaiškinimų metodų tyrimo projektas. Pirmiausia aprašomi parinkti duomenų rinkiniai. Toliau pateikiami tiriami globalių paaiškinimo metodų patobulinimai: paaiškinimų svėrimas pagal patikimumą (R^2) ir agregavimas branduolio tankio įverčiu (KDE). Pristatytos vertinimo metrikos apima klasifikavimo modelių našumą, paaiškinimų kokybę ir stabilumą. Galiausiai aprašoma eksperimentų eiga ir naudota infrastruktūra – naudojamos bibliotekos, duomenų bazės schema ir rezultatų apdorojimo ir valdymo priemonės.

2.1. Duomenų rinkiniai

Eksperimentuose naudojami trys tabuliuoti duomenų rinkiniai, atrinkti atsižvelgiant į jų aktualumą dirbtinio intelekto skaidrumo reikalavimams bei paaiškinamumo svarbą realiose taikymo srityse. DIAB ir CRED rinkiniai atspindi sveikatos priežiūros ir finansų sektorius – tai sritys, kuriose DI sprendimų skaidrumas ir pagrįstumas yra ypač svarbūs tiek dėl etinių, tiek dėl teisinių priežasčių. Šie rinkiniai dažnai pasitelkiami sprendimams, turintiems tiesioginę įtaką žmonių gerovei ar finansinei padėčiai, todėl jie tinka vertinti globalių paaiškinimų gebėjimą suteikti suprantamą modelio elgesio apžvalgą. Tuo tarpu FOREST daugiasluoksnis klasifikavimo uždavinys įtrauktas siekiant išbandyti paaiškinamumo metodų veikimą sudėtingesnėje, didesnės apimties klasifikavimo aplinkoje, turinčioje daugiau klasių ir požymių. Tai leidžia įvertinti, kaip globalūs paaiškinimai veikia platesniame, labiau techniniame kontekste, kuriame etinis aspektas gali būti mažiau ryškus, tačiau techninis patikimumas itin svarbus.

Diabeto rodiklių rinkinys (DIAB)

Pirmasis duomenų rinkinys yra CDC diabeto sveikatos rodiklių [8] pogrupis (*Diabetes_binary*), skirtas nustatyti diabeto riziką. Jį sudaro 70 692 atvejai ir 21 atributas, iš kurių 3 yra skaitiniai (pvz., kūno masės indeksas – BMI, fizinė ir psichinė sveikata), o likusieji – binariniai ar ranginiai kategoriniai kintamieji. Binariniai kategoriniai atributai atspindi sveikatos būklę (aukštą kraujospūdį, aukštą cholesterolio, didelį cholesterolio kiekį), elgesio veiksnius (rūkymą, fizinį aktyvumą, vaisių vartojimą, daržovių vartojimą), ir prieigą prie sveikatos priežiūros. Ranginiai atributai: amžiaus grupė (sugrupuota į 5 metų intervalus), bendra sveikata (savarankiškai pranešta sveikatos būklė pagal skalę nuo 1 iki 5), išsilavinimas ir pajamų grupė. Tikslinis kintamasis turi dvi reikšmes: 0 – nėra diabeto, 1 – prediabetas arba diabetas. Duomenų rinkinys yra visiškai subalansuotas, atvejų skaičius kiekvienai klasei lygus.

Kreditingumo rizikos rinkinys (CRED)

Antrasis rinkinys – vokiečių kredito rizikos duomenys (CRED) [21], apimantys 1000 atvejų ir 35 atributus, iš kurių 28 yra kategoriniai, 7 – skaitiniai. Tikslinis kintamasis – kredito rizika – turi dvi klases: 1 (maža rizika) ir 0 (didelė rizika). Šis rinkinys yra nesubalansuotas: 700 mažos rizikos atvejų ir 300 didelės rizikos.

Miško dangos tipų rinkinys (FOREST)

Trečiasis duomenų rinkinys – „*Forest Covertype*“ [6] sudarytas iš 581 012 įrašų su 54 atributais. Tai daugiaklasis klasifikavimo uždavinys su 7 skirtingais miško dangos tipais (pvz., *„Lodgepole Pine“, „Aspen“, „Douglas-fir“*). Atributai apima 10 skaitinių topografinių kintamųjų (pvz., aukštis, nuolydis, atstumas iki kelių, vandens telkinių) ir 44 binarinius kintamuosius, atspindinčius dirvožemio tipą bei buveinės tipą. Klasės pasiskirsčiusios labai netolygiai – nuo daugiau kaip 280 tūkst. iki mažiau nei 3 tūkst. atvejų.

2.2. Tiriami LIME pagrindo globalių paaiškinimų metodų patobulinimai

Šiame darbe siekiama pagerinti globalių LIME paaiškinimų metodų tikslumą, stabilumą ir atitikimą modelio elgsenai, pritaikant du esminius naujumus: (1) paaiškinimų patikimumo (R^2) įtraukimą į agregavimo žingsnį ir (2) alternatyvų agregavimo metodą – branduolio tankio įverčio (KDE) funkcijos piką**.**

Visi analizuoti metodai grindžiami LIME algoritmu, kuris generuoja lokalios atributų svarbos reikšmes kiekvienam duomenų atvejui. Skirtumai tarp globalių metodų atsiranda agregavimo etape – t. y., kaip iš šių lokalių reikšmių formuojamas globalus svarbos pasiskirstymas. Tradiciškai taikomas vidurkis gali būti jautrus triukšmui ar nepatikimiems paaiškinimams, todėl tikslumo ir interpretacijos tikslais buvo taikomi alternatyvūs metodai.

Remiantis vertinimo metrikomis – atributų svarbos atitikimu**,** modelio jautrumu svarbiausiems atributams bei paaiškinimų stabilumu tarp eksperimentų – suformuluotos šios tyrimo hipotezės:

- Paaiškinimų patikimumo (R^2) įtraukimas pagerina globalių paaiškinimų kokybę, nes leidžia sumažinti triukšmingų ar mažai patikimų lokalių vektorių įtaką.
- KDE pagrindu atliekamas agregavimas leidžia geriau atspindėti dažniausiai pasikartojančią atributų svarbą, todėl turėtų pagerinti interpretacinę vertę.

2.2.1. Patikimumo (R^2) įtraukimas į esamus globalių paaiškinimų metodus

Siekiant įvertinti, ar paaiškinimų patikimumas turi teigiamą poveikį globalių atributų svarbos įvertinimui, literatūroje aprašyti metodai papildomi determinacijos koeficientu (R^2) agregacijos žingsnyje. Agregacijos strategijos skirstomos į dvi pagrindines grupes:

- Kvadratinės šaknies iš sumos metodai, kurie agreguoja absoliučias atributų reikšmes: SP-LIME, SP-LIME-ALL, GALE-H-WEIGHTED,
- Vidurkio metodai, kurie skaičiuoja lokalios svarbos reikšmių vidurkį: GALE-AVG, ILIME, OLEA, NORM-LIME

Į šias grupes integruoti R^2 svertiniai variantai, kuriuose kiekvienam paaiškinimui priskiriamas svoris pagal jo determinacijos koeficientą d_i. Abiem atvejais naudojamos absoliučios lokalios svarbos reikšmės. R^2 svertinis vidurkis, taikytas naujuose metoduose: GALE-AVG-R2, ILIME-R2, OLEA-R2, NORM-LIME-R2:

$$Svarba_j = \frac{\sum_{i=1}^{n} d_i \cdot |e_{ij}|}{\sum_{i=1}^{n} d_i} \tag{12}$$

- R^2 svertinė kvadratinės šaknis iš sumos, taikyta naujuose metoduose SP-LIME-ALL-R2, GALE-H-R2:

$$Svarba_j = \sqrt{\sum_{i=1}^{n} d_i \cdot |e_{ij}|} \tag{13}$$

čia:

- d_i – i-ojo duomenų atvejo lokalaus paaiškinimo determinacijos koeficientas (R²)
- e_{ij} – atributo j lokalus paaiškinimas i-ajam duomenų atvejui
- n – duomenų atvejų skaičius

2.2.2. KDE pagrįstas agregavimas

Analizuojant lokalios svarbos paaiškinimus, vienas iš pagrindinių iššūkių yra tai, kad šios reikšmės dažnai pasižymi dideliu kintamumu tiek tarp duomenų atvejų, tiek tarp atributų. Paprastas vidurkis (angl. *mean aggregation*) gali būti jautrus išskirtinėms reikšmėms ir neišryškina dažniausiai pasitaikančių svarbos tendencijų. Todėl šiame darbe taikomas branduolio tankio įverčio metodas (KDE), kuris leidžia identifikuoti kiekvieno atributo svarbos reikšmių modalinę (dažniausiai pasitaikančią) reikšmę. Šis metodas laikomas atsparesniu triukšmui ir labiau atspindinčiu stabilias svarbos tendencijas globaliu mastu.

2 lentelė. KDE agregacija grįstų globalių paaiškinimų konfigūracijos

Konfigūracija	**Normalizacija**	**Svėrimas**
KDE-N0-R0	Nėra	Nėra
KDE-N0-R2	Nėra	Pagal R²
KDE-NR-R0	Eilučių L1 normalizacija	Nėra
KDE-NR-R2	Eilučių L1 normalizacija	Pagal R²
KDE-NRMM-R0	L1 normalizacija + stulpelių min-max	Nėra
KDE-NRMM-R2	L1 normalizacija + stulpelių min-max	Pagal R²

Siekiant įvertinti, kaip skirtingos duomenų apdorojimo strategijos veikia KDE rezultatus, taikomos šešios eksperimentinės konfigūracijos, pateiktos 2 lentelėje. Jos sudarytos iš dviejų nepriklausomų komponentų:

- **Normalizacijos strategija**: nulemia, kaip tvarkomos lokalių paaiškinimų reikšmės prieš taikant KDE. Tai svarbu norint užtikrinti jų palyginamumą tarp duomenų atvejų ir išvengti mastelio iškraipymų.
- **Svoris**: apibrėžia, ar kiekvienam paaiškinimui suteikiamas svoris pagal jo patikimumą (R²). Tokiu būdu aukštesnio tikslumo paaiškinimai turi didesnę įtaką globaliam KDE rezultatui.

2.2.3. KDE realizacija

KDE (branduolio tankio įverčio) metodas leidžia nustatyti dažniausiai pasitaikančią atributo svarbos reikšmę per visų duomenų atvejų lokalius paaiškinimus. Tegul:

- $e_{ij}^1, e_{ij}^2, \ldots, e_{ij}^n$ – atributo j lokalios svarbos reikšmės n skirtingų duomenų atvejų;
- $f_j(x)$ – KDE pagrindu įvertinta atributo j svarbos reikšmių pasiskirstymo tankio funkcija.

Tuomet atributo j globali svarba skaičiuojama kaip KDE funkcijos maksimumas:

$$Svarba_j = arg\max_x f_j(x) = arg\max_x \left(\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{\left(x - e_{ij}\right)}{h} \right) \right) \qquad (14)$$

čia:

- $K(\cdot)$ – branduolio funkcija (Gauso): $K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$
- h – glotninimo (angl. bandwidth) parametras, kontroliuojantis KDE jautrumą.

Šiame darbe h parenkamas pagal *Silverman[35]* taisyklę: $h = 1.06 \cdot \sigma \cdot n^{-\frac{1}{5}}$, kur σ – atributo j svarbos reikšmių standartinis nuokrypis, o n – reikšmių skaičius.

2.2.4. Normalizacijos strategijos

Siekiant užtikrinti lokalių paaiškinimų palyginamumą tarp skirtingų duomenų atvejų ir sumažinti mastelio įtaką KDE skaičiavimui, taikomos dvi skirtingos normalizacijos strategijos.

2.2.4.1. NR - L1 eilutės normalizacija

Kiekvieno duomenų atvejo i paaiškinimo vektorius normalizuojamas taip, kad visų jo komponentų absoliučių reikšmių suma būtų lygi 1:

$$\tilde{e}_{ij} = \frac{| e_{ij} |}{\sum_{k=1}^{d} | e_{ij} |} \tag{15}$$

čia:

- d – atributų skaičius,
- $\tilde{e}_{ij}$ – normalizuota atributo j svarba duomenų atvejui i.

2.2.4.2. NRMM – L1 eilutės normalizacija + stulpelio mastelio keitimas (min-max)

Kitoje konfigūracijoje, po L1 normalizacijos (15) kiekvienas atributas j per visus duomenų atvejus papildomai normalizuojamas (*min-max*) atributo lygmeniu:

$$\hat{e}_{ij} = \frac{\bar{e}_{ij} - min_i(\bar{e}_{ij})}{max_i(\bar{e}_{ij}) - min_i(\bar{e}_{ij})}$$

čia:

- $\hat{e}_{ij}$ – galutinė NRMM normalizuota reikšmė,
- Normalizavimas vykdomas stulpelio (atributo) lygmenyje.

2.3. Vertinimo metrikos

Vertinant paaiškinamumo metodų veikimą, atliekamas nuoseklus klasifikavimo modelių bei lokalių ir globalių paaiškinimų kokybės įvertinimas. Lokalių paaiškinimų analizėje vertinamas jų patikimumas ir stabilumas. Globalūs paaiškinimai vertinami pagal jų gebėjimą atspindėti modelio sprendimų logiką – tiek atkuriant atributų svarbą, tiek įvertinant šios svarbos įtaką modelio elgsenai, bei pagal stabilumą. Žemiau pateikiamos taikomos metrikos pagal atskiras vertinimo grupes:

Klasifikavimo modelių našumo vertinimui:

- kryžminė entropija (angl. *logloss*) – prognozių patikimumo įvertinimui
- Tikslumas, preciziškumas, jautrumas *(*angl. *recall*), F1 reikšmė

Lokalių paaiškinimų vertinimui:

- Patikimumui – determinacijos koeficientas (R^2)
- Stabilumui – *Jaccard* panašumas tarp top-10 atributų

Globalių paaiškinimų vertinimui:

- Atributų atitikimas (tik logistinės regresijos modeliams). Vertinama, kiek gerai globalus svarbos pasiskirstymas atspindi tikrąjį atributų svarbumą.
    - Atributų eiliškumo koreliacija – *Spearmano* koreliacijos koeficientas
    - Atributų įtakos pasiskirstymo neatitikimas – *Jensen-Shannon* atstumas (JSD)
- Eiliškumo poveikis modeliui. Vertinama, kiek aiškinamų atributų svarba daro įtaką modeliui.
    - Tikslumo kritimo AUC
    - F1 reikšmės kritimo AUC
- Stabilumas
    - *Jaccard* panašumas tarp globalių svarbos sąrašų (top 10 atributų), lyginant skirtingus metodus ar duomenų pokyčius

2.3.1. Paaiškinimų stabilumo matavimas

Norint įvertinti paaiškinimų stabilumą S pritaikoma toliau pateikta metodika:

1. Duomenų atvejui sugeneruojami n nepriklausomi paaiškinimai. Kiekvienas i paaiškinimas grąžina top-k atributų sąrašą E_i kur:
   - E_i: top-k atributų sąrašas i-ojo nepriklausomo bandymo,
   - n : bendras bandymų, atliktų stabilumui įvertinti, skaičius
2. Porinis Jaccard panašumo skaičiavimas. Kiekvienai atributų sąrašų porai (E_i, E_j) kai $i < j$, skaičiuojamas Jaccard panašumas $J(E_i, E_j)$, apibrėžiamas:

$$J(E_i, E_j) = \frac{|\mathrm{E_i} \cap \mathrm{E_j}|}{|\mathrm{E_i} \cup \mathrm{E_j}|} \tag{16}$$

   kur $|E_i \cap E_j|$ - žymi abiem paaiškinimams bendrų savybių skaičių, o $|E_i \cup E_j|$ - bendras unikalių savybių skaičius abiejuose paaiškinimuose. Šis panašumo balas kiekybiškai įvertina atributų sąrašų sutapimą.

3. Stabilumas S apibrėžiamas kaip visų porinių Jaccard panašumų vidurkis, kuris atspindi bendrą atributų parinkimo nuoseklumą tarp bandymų:

$$S_k = \frac{2}{n(n-1)} \sum_{\{1 \le i < j \le n\}} J(E_i, E_j) \tag{17}$$

   čia:

   - S – stabilumo balas klasei L, su reikšmėmis nuo 0 (nėra stabilumo) iki 1 (visiškas stabilumas),

- $\frac{2}{n(n-1)}$ – normalizavimo koeficientas, užtikrinantis, kad vidurkis būtų apskaičiuotas visose unikaliose paaiškinimų porose.
- k – top atributų skaičius

Šis stabilumo matas suteikia apie atributų svarbos reitingų patikimumą, o aukštesnis balas rodo patikimesnį ir nuoseklesnį atributų pasirinkimą kartojant paaiškinimą.

2.3.2. Atributų atitikimo įvertinimas

Norint palyginti agreguotų LIME metodo pagrindu gautus globalius atributų svarbos paaiškinimus, buvo taikomi du papildomi panašumo matavimo metodai: Spirmano rangų koreliacija (angl. *Spearman rank correlation*) ir Jenseno–Shannono atstumas (angl. *Jensen-Shannon distance*).

Pirmiausia, siekiant įvertinti atributų išdėstymo pagal svarbą panašumą, buvo skaičiuojamas Spirmano rangų koreliacijos koeficientas ρ. Šis neparametrinis statistinis matas parodo monotonišką priklausomybę tarp dviejų atributų rangų. Duotoms atributų svarbos vektorių reikšmėms X ir Y, kur $R(X_i)$ ir $R(Y_i)$ žymi i-ojo atributo rangą:

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(R(X_i) - R(Y_i))^2}{n(n^2-1)} \tag{18}$$

čia n – atributų skaičius; $\rho = 1$ rodo visišką rangų atitikimą, $\rho = 0$ – nėra koreliacijos, o $\rho = -1$ – visiškai priešingi rangai.

Antra, siekiant įvertinti atributų svarbos reikšmių pasiskirstymo panašumą, buvo taikytas ADI literatūroje [20,22,30] dažnai naudojamas *Jenseno–Shannono* atstumas (angl. *JSDistance*). Tegul P ir Q – normalizuoti (t. y. sumuojasi iki 1) atributų svarbos vektoriai. Tuomet *Jenseno–Shannono* atstumas apibrėžiamas taip:

$$JSD(P,Q) = \sqrt{\frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M)} \tag{19}$$

2.3.3. Atributų eiliškumo poveikio modeliui įvertinimas

Šiame vertinime analizuojama, kiek modelio veikimas priklauso nuo atributų, kurie laikomi svarbiausiais pagal paaiškinimo metodą. Vietoj tiesioginio atributų pašalinimo jų įtaka modelio prognozei neutralizuojama – reikšmės pakeičiamos viso duomenų rinkinio vidurkiu, taip imituojant jų neprieinamumą. Tokiu būdu vertinama, ar paaiškinimas iš tiesų identifikuoja esminius atributus. Kuo didesnis gautas AUC (plotas po kreive), tuo stipresnė paaiškinimo atributų įtaka modeliui. Šie AUC skaičiuojami atskirai pagal skirtingus klasifikavimo veikimo rodiklius:

- **Tikslumo kritimo plotas po kreive (Tikslumo kritimo AUC)**: įvertina klasifikavimo tikslumo mažėjimą, neutralizuojant aukščiausiai reitinguotus atributus.

- **F1 kritimo plotas po kreive (F1 kritimo AUC)**: įvertina F1 rodiklio mažėjimą pagal tą pačią procedūrą. Ypač aktualu nesubalansuotiems duomenims, kur F1 rodiklis geriau atspindi tikrąjį modelio veikimą.

Bendrai AUC paskaičiuojamas šia formule:

$$AUC = \int_0^1 d(x)dx \approx \sum_{i=1}^{n-1} (x_{i+1} - x_i) \cdot \frac{d_i + d_{i+1}}{2} \quad (20)$$

čia:

- x_i - neutralizuotų atributų dalis,
- m_i – modelio našumas (tiklsumas arba F1) x_i žingsnyje
- $d_i = 1 - m_i$ – degradacijos vertė

AUC reikšmė artima 1 rodo, kad neutralizuoti atributai turėjo esminę reikšmę modelio veikimui; artima 0 – kad paaiškinimas buvo mažai informatyvus.

2.4. Eksperimentų eiga

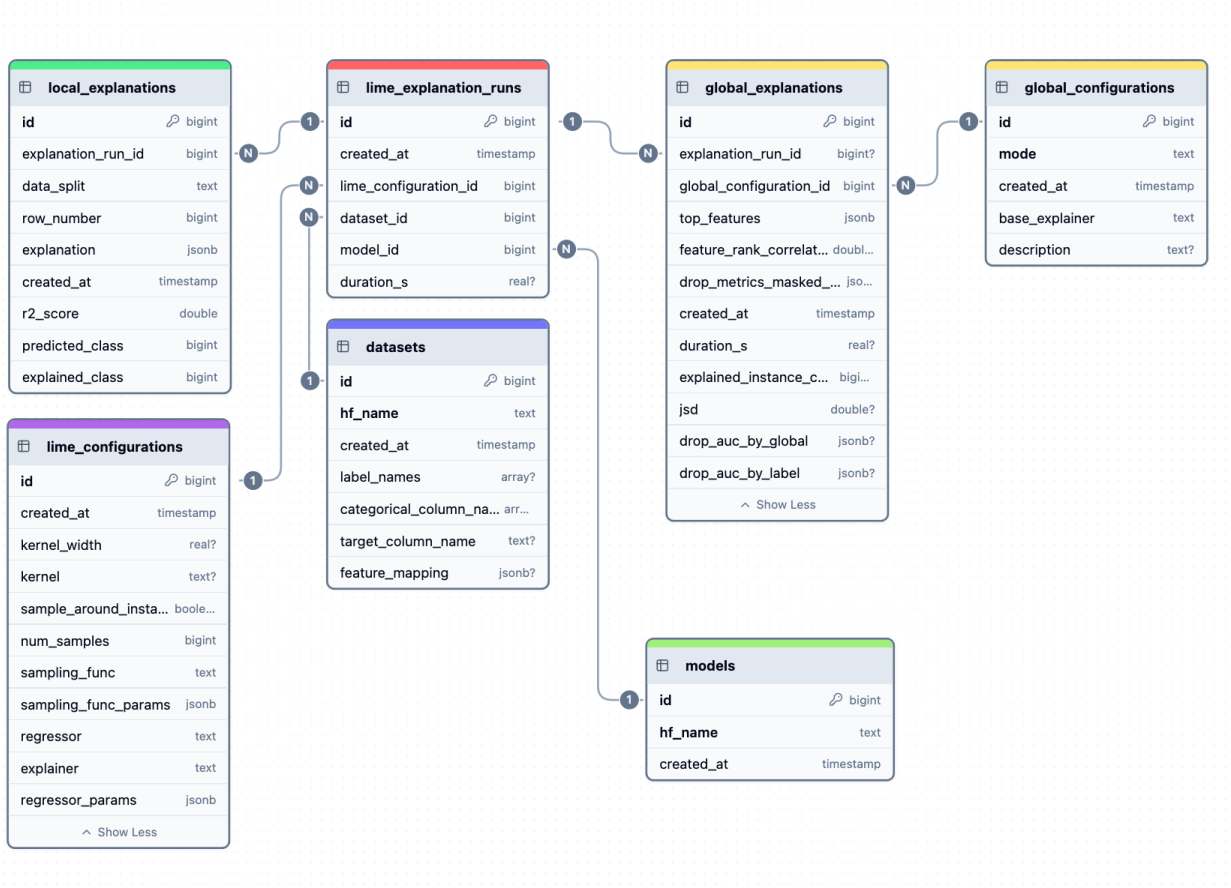
Eksperimentinis tyrimas buvo suskirstytas į keturis pagrindinius etapus:

1) **Modelių paruošimas.** Kiekvienam iš pasirinktų duomenų rinkinių buvo sukurti du klasifikavimo modeliai – logistinės regresijos (LR) ir „*XGBoost*“. Prieš modelių mokymą atliktas duomenų paruošimas (normalizacija, paskirstymas į testavimo, validacijos ir treniravimo imtis) ir hiperparametrų paieška, siekiant optimizuoti prognozavimo tikslumą.
2) **Lokalių paaiškinimų generavimas su LIME.** Iš kiekvieno duomenų rinkinio buvo atrinkta po 10 testavimo atvejų. Kiekvienas jų buvo aiškinamas 10 kartų naudojant skirtingas LIME parametrų konfigūracijas. Vertintas generuotų paaiškinimų patikimumas (R^2) ir stabilumas. Pasirinkus optimalias konfigūracijas, visai testavimo imčiai buvo sugeneruoti lokalūs paaiškinimai tolimesnei analizei.
3) **Agregavimo strategijų tyrimas.** Buvo tiriamos dvi pagrindinės patobulinimų kryptys:
   a) Paaiškinimų svėrimas pagal R^2. Esami globalūs metodai buvo papildyti svoriais pagal kiekvieno paaiškinimo determinacijos koeficientą. Vertintas poveikis globalių atributų svarbos atitikimui, eiliškumo įtakai modeliui ir stabilumui.
   b) KDE pagrįstų konfigūracijų analizė. Buvo tiriami skirtingi normalizavimo (N0, NR, NRMM) ir svėrimo (be svorio, su R^2) deriniai. Vertintas pasiskirstymo atitikimas (JSD), eiliškumo koreliacija ir rezultatų stabilumas. Išrenkama geriausia KDE konfigūracija.
4) **Metodų palyginimas.** Atrinkti geriausiai pasirodę metodai (R^2 papildytos versijos ir viena KDE konfigūracija) buvo lyginami su literatūros metodais pagal tris metrikų grupes:
   a) Atributų svarbos pasiskirstymo atitikimas (JSD),
   b) Eiliškumo poveikis modeliui (tikslumo/F1 kritimo AUC),
   c) Paaiškinimų stabilumas (Jaccard panašumas tarp 10 kartų sugeneruotų globalių paaiškinimų).

Kiekvienam modeliui buvo sugeneruoti 10 globalių paaiškinimų, naudojant skirtingus lokalių paaiškinimų pakartojimus, siekiant įvertinti rezultatų vidurkį ir stabilumą.

2.5. Eksperimentų aplinka ir duomenų saugojimas

Eksperimentams atlikti naudojama *Python* programavimo kalba bei tokios pagrindinės bibliotekos kaip *scikit-learn* [43] - klasifikavimo modelių kūrimui, *optuna* [2] – hiperparametrų paieškai, *hydra* [40] – eksperimentų konfigūracijų valdymui, *lime* [34] - LIME metodo realizacijai, *pandas* - duomenų analizei ir apdorojimui. Modelių bei duomenų rinkinių versijavimui ir saugojimui pasitelkiama *HuggingFace* [41] platforma.



6 pav. Eksperimentų duomenų bazės klasių diagrama

Eksperimentinei darbo eigai užtikrinti buvo sukurta *PostgreSQL* [19] duomenų bazė, skirta valdyti ir struktūruoti visus duomenis, susijusius su lokalių ir globalių modelių paaiškinimų generavimu. Ši schema užtikrina atsekamumą, atkuriamumą ir lankstumą vertinant skirtingus modelius, duomenų rinkinius bei paaiškinimų konfigūracijas. 6 pav**.** pateikta šios duomenų bazės klasių diagrama. Ją sudaro šios lentelės:

- Lentelėje *„datasets”* saugoma informacija apie naudotus duomenų rinkinius: jų identifikatoriai, klasių pavadinimai ir atributų aprašai.
- Lentelė *„models”* aprašo išmokytus modelius: nurodomas susietas duomenų rinkinys, modelio identifikatorius ir sukūrimo laikas. Tai leidžia susieti paaiškinimus su konkrečiu modeliu.

- Lentelėje *„lime_configurations”* saugomos LIME metodo lokaliesiems paaiškinimams taikytos konfigūracijos: artumo fukcijos parametrai, perturbacinių taškų skaičius, perturbacinių taškų generavimo funkcijos ir surogatinio modelio parametrai.
- Kiekvienas paaiškinimo paleidimas registruojamas lentelėje *„lime_explanation_runs”*, kurioje nurodomas naudotas modelis, duomenų rinkinys ir LIME konfigūracija. Tai leidžia užtikrinti eksperimentų atkuriamumą ir lyginamumą.
- Lokalūs paaiškinimai saugomi lentelėje *„local_explanations”*, kurioje pateikiamas kiekvieno atvejo paaiškinimas JSON formatu bei papildomi duomenys, tokie kaip modelio spėta klasė, paaiškinta klasė ir patikimumo metrikos (determinacijos koeficiento) reikšmė.
- Globalių paaiškinimų metodams taikyti naudojama lentelė *„global_configurations”*, kurioje aprašomos globalaus agregavimo konfigūracijos.
- Galiausiai, lentelėje *„global_explanations”* saugomi globalių paaiškinimų rezultatai: atributų įtakos santraukos, metrikos (pvz., eiliškumo koreliacija, JSD) bei išvestiniai našumo rodikliai (pvz.: tikslumo kritimas).

3. Globalių paaiškinimų metodų tyrimo rezultatai

Eksperimentai buvo suskirstyti į keturis pagrindinius etapus: (1) klasifikavimo modelių paruošimą, (2) lokalių LIME paaiškinimų generavimą šiems modeliams, (3) lokalių paaiškinimų agregavimo į globalius paaiškinimus strategijų komponentų tyrimą bei naujų strategijų kūrimą ir galiausiai (4) naujai siūlomų metodų palyginimą su literatūroje aprašytais naujausiais globalaus LIME paaiškinamumo metodais.

3.1. Klasifikavimo modelių paruošimas

Tolimesniems paaiškinamumo eksperimentams reikalingi klasifikavimo modeliai buvo išmokyti naudojant trijų skirtingų duomenų rinkinių informaciją. Tam buvo naudojama modelių kūrimo seka: pirmiausia kiekvienas duomenų rinkinys buvo apdorotas, tada kiekvienam modeliui atlikta hiperparametrų optimizacija, ir galiausiai, atrinkus geriausius parametrus, išmokyti galutiniai klasifikatoriai. Šiame poskyryje aptarti šių žingsnių rezultatai.

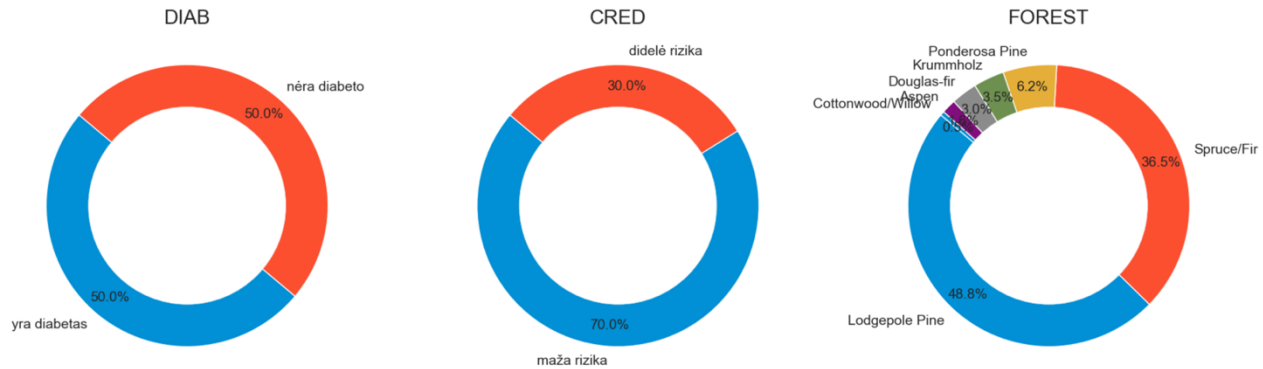
3.1.1. Duomenų rinkinių paruošimas

Eksperimentams buvo paruošti trys skaitinių duomenų rinkiniai**.** Duomenų rinkiniams paruošti buvo taikyta vienetinė koduotė (angl. *one-hot encoding*) metodas nominaliems atributams, o skaitiniai duomenys buvo standartizuoti. Paruoštų duomenų apibendrinimas pareiktas 3 lentelėje.

3 lentelė. Naudojamų duomenų rinkinių apibendrinimas

Duomenų rinkinys	Klasių skaičius	Atvejų skaičius	Atributų skaičius	Kategorinių atributų skaičius	Skaitinių atributų skaičius
Diabeto rodikliai (DIAB)	2	70 692	21	18	3
Kreditingumo rizika (CRED)	2	1000	35	28	7
Miško danga (FOREST)	7	581 012	54	44	10

3 lentelėje matoma, kad DIAB ir CRED rinkiniai turi daug kategorinių kintamųjų (atitinkamai 18 ir 28), o FOREST išsiskiria dydžiu ir didesne skaitinių atributų dalimi (10 iš 54). Tai rodo skirtingus iššūkius: DIAB ir CRED reikalauja efektyvios kategorinių atributų interpretacijos, o FOREST reikalauja gebėjimo apdoroti didelės apimties ir daugiaklasę informaciją.



7 pav. Duomenų rinkinių klasių pasiskirstymas

Visi duomenų rinkiniai buvo padalinti naudojant stratifikuotą atranką į mokymo (80 %), validacijos (10 %) ir testavimo (10 %) imtis, užtikrinant subalansuotą tikslinio kintamojo reprezentavimą visuose padalijimuose. 7 pav. pavaizduotas tikslinių klasių pasiskirstymas kiekviename duomenų rinkinyje, rodo, kad DIAB rinkinys yra visiškai subalansuotas, CRED – vidutiniškai nesubalansuotas, o FOREST – pasižymi itin stipriu klasių pasiskirstymo netolygumu.

3.1.2. Logistinės regresijos ir XGBoost hiperparametrų paieška

Hiperparametrų optimizacija naudojantis „*Optuna*“ [2] biblioteka ir atrenkami geriausi hiperparametrai pagal validacijos kryžminės entropijos (angl. *cross-entropy*, *log loss*) reikšmę.

Logistinės regresijos hiperparametrų paieška

Logistinės regresijos hiperparametrų derinimui reguliarizavimo stiprumas (C) buvo optimizuojamas skalėje nuo 10^{-5} iki 100. Buvo išbandyti skirtingi baudos tipai (*L1, L2*) kartu su sprendimo algoritmais, kurie užtikrina tinkamą baudos pritaikymą išvengiant persimokymo.

4 lentelė. Logistinės regresijos hiperparametrų paieškos rezultatai

Duomenų rinkinys	Hiperparametrai			Nuostolio funkcija	Validacijos imties nuostolis	Treniravimo imties nuostolis
	C*	Baudos tipas**	Sprendimo algoritmas***			
DIAB	0.152902443	L2	LBFGS	logloss	0.514736139	0.51117312
CRED	0.443532885	L2	LIBLINEAR	logloss	0.488224596	0.483099168
FOREST	11,6554723	L1	LIBLINEAR	mlogloss	0.661416086	0.65876267

*C – reguliarizavimo stiprumas; **baudos tipai: L1 (angl. *Lasso regularisation*), L2 (angl. *Ridge regression*); ***sprendimo optimizacijos algoritmai: LIBLINEAR(angl. *Coordinate Descent algorithm*) , LBFGS (angl. *Limited-memory Broyden-Fletcher-Goldfarb-Shanno*), SAGA (angl. *stochastic gradient-based optimization algorithm*)

4 **lentelė.** Logistinės regresijos hiperparametrų paieškos rezultatai lentelėje matoma, kad DIAB ir CRED duomenų rinkiniuose geriausi rezultatai pasiekti naudojant L2 baudos tipą su atitinkamais sprendimo algoritmais. FOREST rinkinyje geriausi rezultatai gauti taikant L1 baudą su LIBLINEAR algoritmu.

„XGBoost“ hiperparametrų paieška

„XGBoost“ hiperparametrų [44] derinimo eksperimentuose buvo keičiami medžių kūrimo parametrai: medžių kūrimo metodas, maksimalus medžių gylis, minimalus vaikų svoris (angl. *minimum child weight*), pavyzdžių atrankos (angl. *subsample*) koeficientas, mazgų kintamųjų parinkimas (angl. *colsample_bynode*), reguliavimo parametras λ ir mokymosi greitis η (angl. *learning rate*).

5 lentelė. „XGBoost” hiperparametrų paieškos rezultatai

Duomenų rinkinys	Hiperparametrai							Nuostolio funkcija	Validacijos imties nuostolis	Treniravimo imties nuostolis
	tree_ method	*max_ depth*	*min_ child_ weight*	*subsample*	*colsample_ bynode*	λ	η			
DIAB	*hist*	5	19	0.630538	0.721365	0.413478	0.029976	logloss	0.4997925	0.4885652
CRED	*approx*	8	12	0.761462	0.370136	1.650020	0.180929	logloss	0.4350932	0.3670878
FOREST	*hist*	12	10	0.825591	0.983103	0.114044	0.136165	mlogloss	0.0779827	0.0163900

Iš 5 lentelės matyti, kad kiekvienam duomenų rinkiniui buvo atrinkti skirtingi optimalūs hiperparametrai, atsižvelgiant į jų specifiką. Pavyzdžiui, DIAB rinkinyje geriausi rezultatai pasiekti naudojant *hist* medžių kūrimo metodą su mažesniu medžių gyliu, o CRED rinkinyje – *approx* metodą su didesniu gyliu.

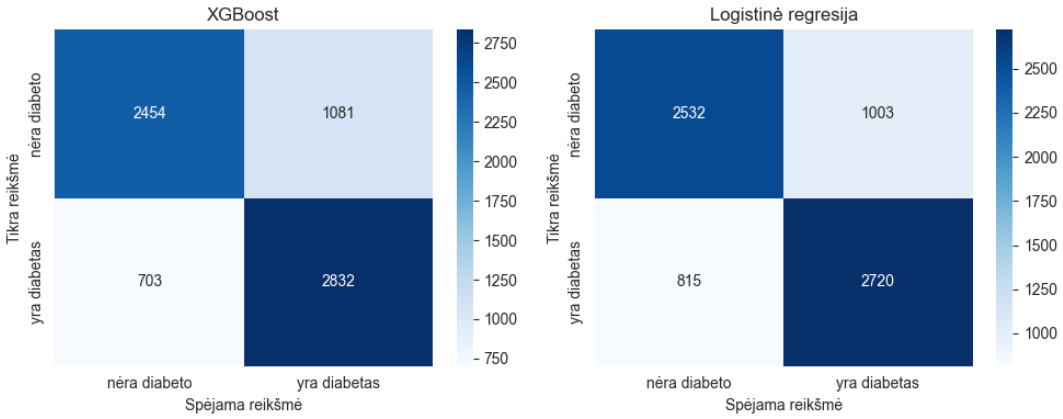
3.1.3. Galutiniai klasifikavimo modelių rezultatai

Po hiperparametrų optimizacijos, kiekvienam modeliui buvo atrinkti parametrai, pasiekiantys mažiausią validacijos imties nuostolį. Tuomet treniravimo ir validacijos duomenų imtys buvo sujungtos, ir galutinis modelis išmokytas iš naujo, naudojant visą prieinamą mokymo informaciją bei optimalius parametrus. Žemiau pateikiami galutiniai modelių klasifikavimo rezultatai, gauti testuojant su testavimo duomenų imtimi. Rezultatai apima pagrindinius klasifikavimo tikslumo rodiklius bei atitinkamas sumaišymo matricas.

6 lentelė. Galutinių klasifikavimo modelių rezultatai DIAB

Modelis	Tikslumas	Preciziškumas	Jautrumas	F1
XGB_{DIAB}	0.74767	0.72374	0.80113	0.76047
LR_{DIAB}	0.74286	0.73059	0.76945	0.74952

6 lentelėje pateikti DIAB duomenų rinkinio klasifikavimo modelių rezultatai. Abu modeliai pasiekė panašų tikslumą (~0.74) ir F1 rodiklį (~0.75–0.76). „XGBoost" pasiekė aukštesnį jautrumą (0.801), o LR – šiek tiek didesnį preciziškumą (0.731). Tai rodo, kad „XGBoost" geriau atpažįsta teigiamus atvejus, o LR labiau linkęs vengti klaidingų teigiamų prognozių. Bendras rezultatas subalansuotas, tačiau „XGBoost" kiek lenkia pagal F1.



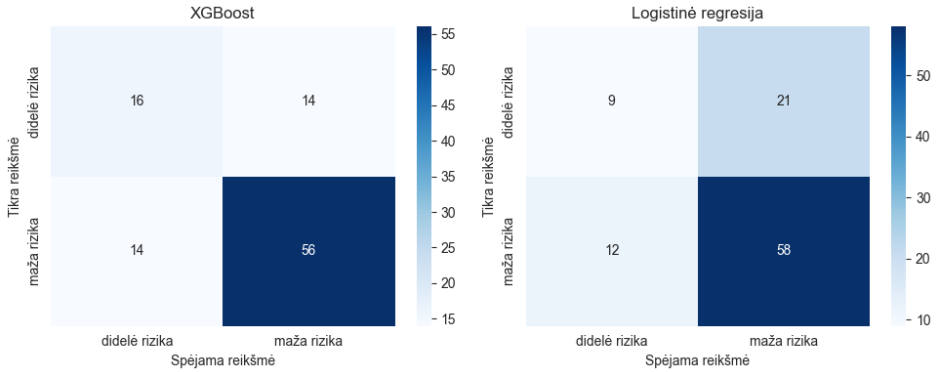
8 pav. DIAB Galutinių klasifikavimo modelių sumaišymo matricos

8 pav. pateiktos sumaišymo matricos patvirtina metrikų skirtumus: „XGBoost" klaidingai klasifikavo mažiau teigiamų atvejų nei LR, bet LR padarė mažiau klaidų klasifikuodamas neigiamus atvejus.

7 lentelė. Galutinių klasifikavimo modelių rezultatai CRED

Modelis	Tikslumas	Preciziškumas	Jautrumas	F1
XGB_{CRED}	0.7200	0.8000	0.8000	0.8000
LR_{CRED}	0.6700	0.7342	0.8286	0.7785

7 lentelėje, kurioje pateikiami CRED duomenų rinkinio klasifikavimo modelių rezultatai, „XGBoost" aiškiai pranoksta LR modelį pagal tikslumą (0.720 vs. 0.670) ir F1 (0.800 vs. 0.779). Tačiau LR modelis pasiekė didesnį jautrumą (0.829), nors kartu sumažėjo preciziškumas (0.734). Tai rodo, kad LR linkęs labiau „perklasifikuoti" teigiamus atvejus, o „XGBoost" geriau išlaiko balansą.



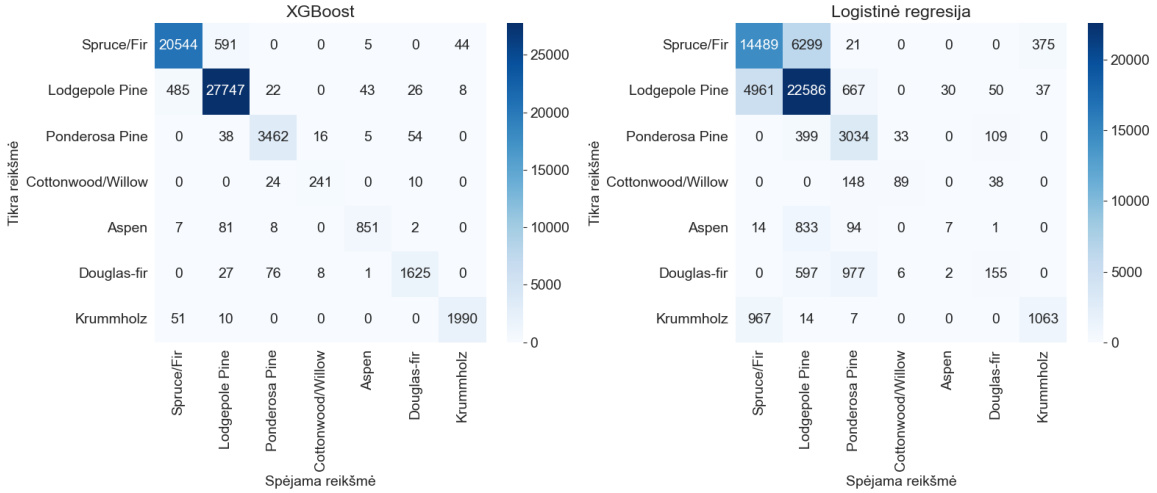
9 pav. CRED Galutinių klasifikavimo modelių sumaišymo matricos

9 pav. CRED duomenų rinkinio klasifikavimo modelių sumaišymo matricose matoma, kad „XGBoost" padarė mažiau klaidų abiem klasėms: tiek klaidingų teigiamų, tiek klaidingų neigiamų, palyginti su LR. Matricos rodo, kad XGBoost labiau subalansuotas, o LR šiek tiek šališkas teigiamos klasės („maža rizika") atžvilgiu.

8 lentelė. Galutinių klasifikavimo modelių rezultatai FOREST

Modelis	**Tikslumas**	**Makro**			**Svertinis**	
		Preciziškumas	**Jautrumas**	**F1**	**Preciziškumas**	**F1**
XGB_{FOREST}	0.97174	0.95465	0.94235	0.94837	0.97170	0.97170
LR_{FOREST}	0.71294	0.58456	0.46691	0.48290	0.69949	0.69868

8 lentelėje pateiktuose klasifikavimo modelių rezultatuose, gautuose naudojant gauti naudojant FOREST daugiaklasį duomenų rinkinį, „XGBoost" iš esmės lenkia LR visais kriterijais. Makro F1 skirtumas yra labai ryškus (0.948 vs. 0.483), o svertinis – 0.972 vs. 0.699. Tai rodo, kad „XGBoost" ne tik gerai prognozuoja dažnas klases, bet ir palaiko našumą retesnėse.



10 pav. FOREST Galutinių klasifikavimo modelių sumaišymo matricos

Pateiktose sumaišymo matricose (10 pav.) matoma, kad „XGBoost" klasifikuoja beveik visus atvejus teisingai visose klasėse. Tuo tarpu LR modelyje klaidų kiekis žymiai didesnis – klasės susipina, ypač tarp *Spruce/Fir, Lodgepole Pine* ir *Douglas-Fir* klasių. Tai rodo, kad LR nesugeba tinkamai atskirti artimų klasių, o XGBoost tai daro itin tiksliai.

Šiame skyriuje aprašyti galutiniai klasifikavimo modeliai, įvertinti testavimo metrikomis, bus naudojami kaip pagrindas tolimesniuose paaiškinamumo eksperimentuose.

3.2. Lokalūs LIME paaiškinimai

Šiame poskyryje siekiama paruošti patikimus lokalius LIME paaiškinimus tolimesnei globalių paaiškinimų analizei. Pirmiausia atliekama LIME metodo parametrų įtakos analizuotų modelių paaiškinimų patikimumui ir stabilumui analizė, o tuomet, remiantis atrinktais parametrais, sugeneruojami galutiniai paaiškinimai visiems testavimo imties duomenų atvejams.

3.2.1. Lokalių LIME paaiškinimų parametrų paieška

Lokalių paaiškinimų eksperimente analizuojami trys pagrindiniai LIME parametrai, siekiant įvertinti jų poveikį paaiškinimų tikslumui ir stabilumui:

- Parametras **branduolio plotis** kontroliuoja kaimynystės dydį aplink aiškinamą pavyzdį, nulemiantį, kaip svertiniai perturbuoti taškai vertinami pagal jų atstumą.
- Generuojamų **perturbuotų taškų kiekis** lemia paaiškinimo stabilumą ir skaičiavimo sąnaudas.
- **Surogatinio modelio** parametras apibrėžia surogatinį interpretuojamą modelį, kuris naudojamas vietinei „juodosios dėžės“ modelio sprendimų ribai aproksimuoti.

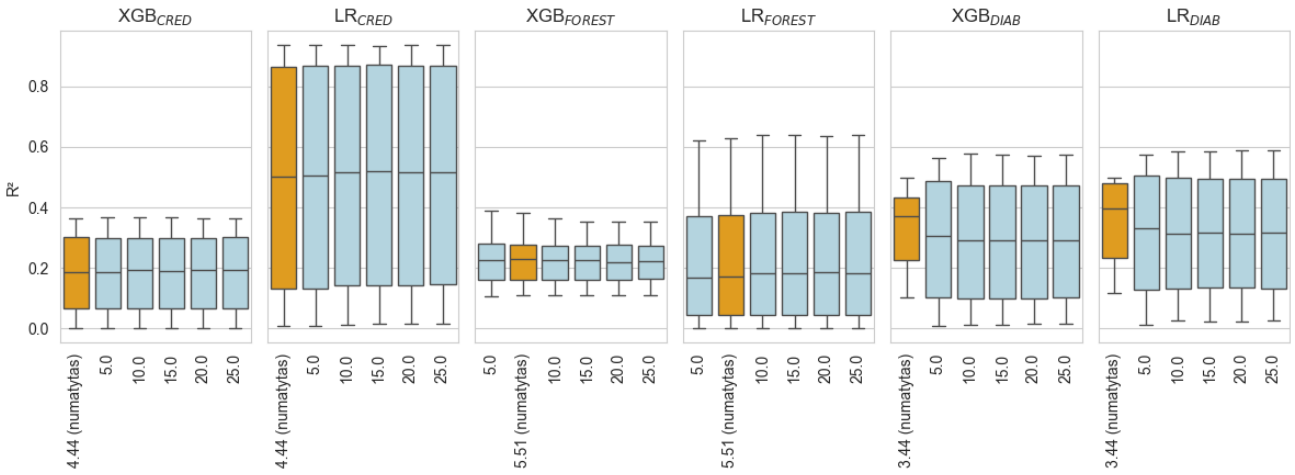
Kiekvieno parametrų derinio atveju buvo sugeneruoti 10 lokalių paaiškinimų 10-čiai atsitiktinai pasirinktų duomenų pavyzdžių. Kiekvienam paaiškinimui apskaičiuotas surogatinio modelio patikimumas (R^2). Vėliau šie rezultatai buvo agreguoti, apskaičiuojant kiekvieno parametro lygio vidutinę R^2 reikšmę ir jos standartinį nuokrypį atskirai kiekvienam modeliui, ir pateikti 9 lentelėje.

9 lentelė. LIME parametrų paieškos rezultatai pagal patikimumą (vidutinė R^2 reikšmė)

Parametras	Reikšmė	Aiškinamas modelis					
		LR_{CRED}	**LR_{DIAB}**	**LR_{FOREST}**	**XGB_{CRED}**	**XGB_{DIAB}**	**XGB_{FOREST}**
Perturbuotų taškų skaičius	1000	0.4905	0.3174	0.241	0.1805	0.2961	0.2217
	2000	0.4873	0.3122	0.2388	0.1754	0.2899	0.2146
	500	**0.4941**	**0.3237**	**0.2453**	**0.1941**	**0.305**	**0.2338**
Branduolio plotis	10.0	0.4925	0.3103	0.2435	0.1834	0.2887	0.2224
	15.0	0.4931	0.3114	0.2434	0.1835	0.2888	0.2227
	20.0	0.4932	**0.3115**	0.2443	**0.1837**	0.2888	0.2213
	25.0	**0.4934**	0.3107	**0.2446**	0.1835	0.2891	0.2215
	5.0	0.4864	0.3102	0.2363	0.1828	0.2904	0.2265
	Numatyta*	0.4852	0.3524	0.2382	0.1829	**0.3363**	**0.2257**
Surogatinis modelis	*ElasticNet*(α=0.1)	0.2043	0.1888	0.0652	0.1009	0.1696	0.1843
	Lasso(α=0.01)	0.825	0.4642	0.3019	0.2886	0.43	0.2563
	Lasso(α=0.1)	0.0169	0.0579	0.0006	0.0202	0.0601	0.1265
	Ridge(α=1)	**0.9164**	**0.5601**	**0.5992**	**0.3235**	**0.5283**	**0.3265**

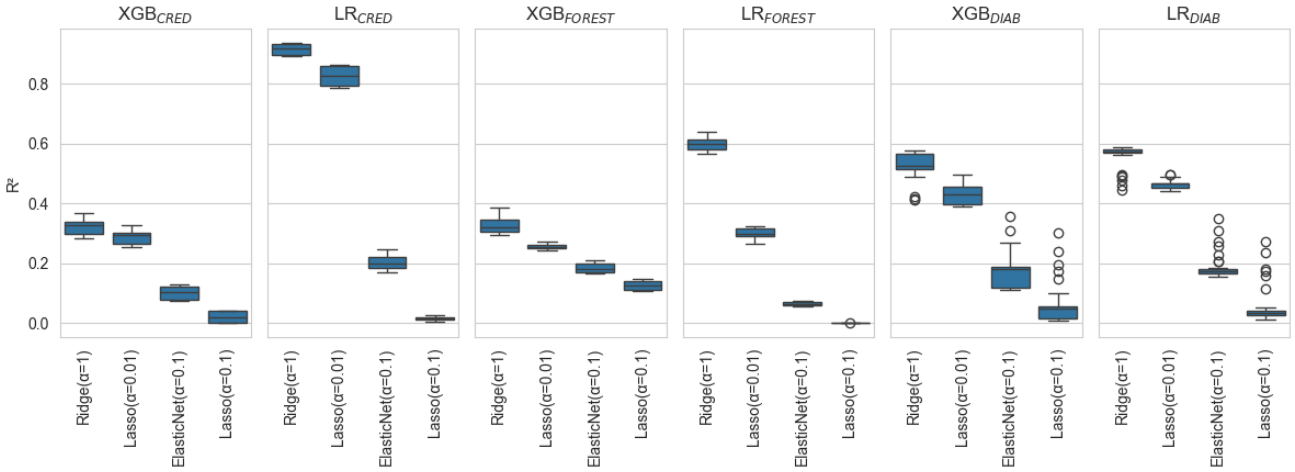
* $\sqrt{}$(atributų skaičius) × 0.75

Perturbuotų taškų kiekio (500, 1000, 2000) įtaka aproksimacijos tikslumui yra nedidelė – skirtumai tarp vidutinių R^2 reikšmių nėra ryškūs. Vis dėlto, daugelyje atvejų mažesnė imtis (500) pasiekia aukštesnį vidurkį nei didesnės imtys. Pavyzdžiui, XGB_{CRED} modelyje R^2 padidėja nuo 0.1754 (2000) iki 0.1941 (500), o LR_{DIAB} atveju – nuo 0.3122 (2000) iki 0.3237 (500). Tai rodo, kad 500 taškų ne tik yra pakankama imtis patikimam aproksimavimui, bet ir dažnai duoda geresnius rezultatus. Be to, mažesnė imtis sumažina skaičiavimo sąnaudas, todėl ši reikšmė buvo pasirinkta kaip numatytoji tolesniems eksperimentams.



11 pav. Branduolio pločio įtaka paaiškinimų patikimumui (R^2) pagal modelį

Kaip matoma 9 lentelėje ir *boxplot* diagramoje (11 pav.) branduolio pločio reikšmės (5.0–25.0) generuoja labai artimas R^2 reikšmes daugumai modelių. Tačiau numatytoji reikšmė, apskaičiuojama pagal √(atributų skaičiaus) × 0.75 formulę, ypač XGB modeliuose, pasiekia aukštesnį vidurkį (pvz., XGB_{DIAB} – 0.3363, XGB_{CRED} – 0.1829). Šis rezultatas rodo, kad adaptuotas branduolio plotis gali būti veiksmingesnis už fiksuotas reikšmes tam tikrose duomenų struktūrose.



12 pav. Surogatinio modelio įtaka paaiškinimų patikimumui (R^2) pagal modelį

Pagal 9 lentelėę ir *boxplot* diagramą (12 pav.) matoma, kad *Ridge* regresorius (α = 1) nuosekliai pateikia aukščiausius R^2 rezultatus visuose aiškinamuose modeliuose (vidurkiai viršija 0.5–0.9), tuo tarpu *Lasso* (α = 0.1) pasiekia žemiausius rezultatus, kai kuriose kombinacijose beveik nulinį patikimumą. *Lasso* (α = 0.01) ir *ElasticNet* (α = 0.1) pasiekia tarpinį tikslumą, bet su pastebimu rezultato nestabilumu tarp modelių.

Atsižvelgiant į gautus rezultatus, tolesniuose eksperimentuose naudojami šie parinkti parametrai: numatytasis branduolio plotis, 500 perturbuotų taškų bei Ridge(α = 1) surogatinis modelis. Šie nustatymai užtikrina optimalų balansą tarp paaiškinimų patikimumo, skaičiavimo efektyvumo ir pritaikomumo skirtingiems modeliams.

3.2.2. Lokalių paaiškinimų rezultatai

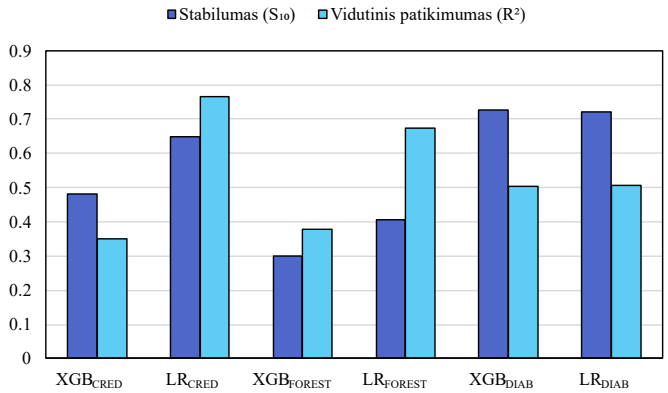
Po LIME parametrų analizės ir parinktais tinkamiausiais, kiekvienam duomenų atvejui iš testavimo duomenų imties buvo sugeneruota po 10 lokalių LIME paaiškinimų. Kiekvienam atvejui apskaičiuotas stabilumas S_{10} (viena reikšmė per atvejį, iš 10 paaiškinimų) bei vidutinis R^2 (vidutinis surogatinio modelio patikimumas R^2 per 10 paaiškinimų). Galiausiai, rezultatai buvo apibendrinti modelio lygmenyje, apskaičiuojant šių metrikų vidurkius ir standartinius nuokrypius. Rezultatai pateikiami 10 lentelėje.

10 lentelė. Galutiniai visų modelių lokalių paaiškinimų rezultatai

Aiškinamas modelis	Aiškinamų atvejų skaičius	Stabilumas (S_{10})	Vidutinis patikimumas (R^2)
XGB_{CRED}	1 000	0.48219 ± 0.04066	0.35103 ± 0.0695
LR_{CRED}	1 000	0.64848 ± 0.05267	0.76668 ± 0.0828
XGB_{FOREST}	10 000	0.29996 ± 0.04717	0.37884 ± 0.07377
LR_{FOREST}	10 000	0.40537 ± 0.06836	0.67341 ± 0.07345
XGB_{DIAB}	7 070	0.72627 ± 0.08183	0.50312 ± 0.18904
LR_{DIAB}	7 070	0.72196 ± 0.08329	0.50719 ± 0.18992

Reikšmės pateikiamos kaip vidurkis ± standartinis nuokrypis, apskaičiuoti iš visų duomenų atvejų, kiekvienam atvejui sugeneravus 10 paaiškinimų.

Tarp visų testuotų derinių, didžiausias vidutinis patikimumas (R^2 = 0.76668) užfiksuotas naudojant LR modelį su CRED duomenų rinkiniu, o žemiausia vidutinė reikšmė (R^2 = 0.35103) – taikant XGB modelį tam pačiam rinkiniui. Kai kuriuose modeliuose (ypač XGB_{DIAB} ir XGB_{FOREST} modeliuose) pastebimas didesnis R^2 standartinis nuokrypis, o tai reiškia, kad dalis lokalių paaiškinimų yra nepatikimi. Todėl tolimesniame globalių paaiškinimų konstravime, siekiant tikslesnių rezultatų svarbu taikyti R^2 kaip lokalių paaiškinimų svorio koeficientą arba pašalinti paaiškinimus su žemu R^2, kad būtų išvengta triukšmo globalioje agregacijoje.

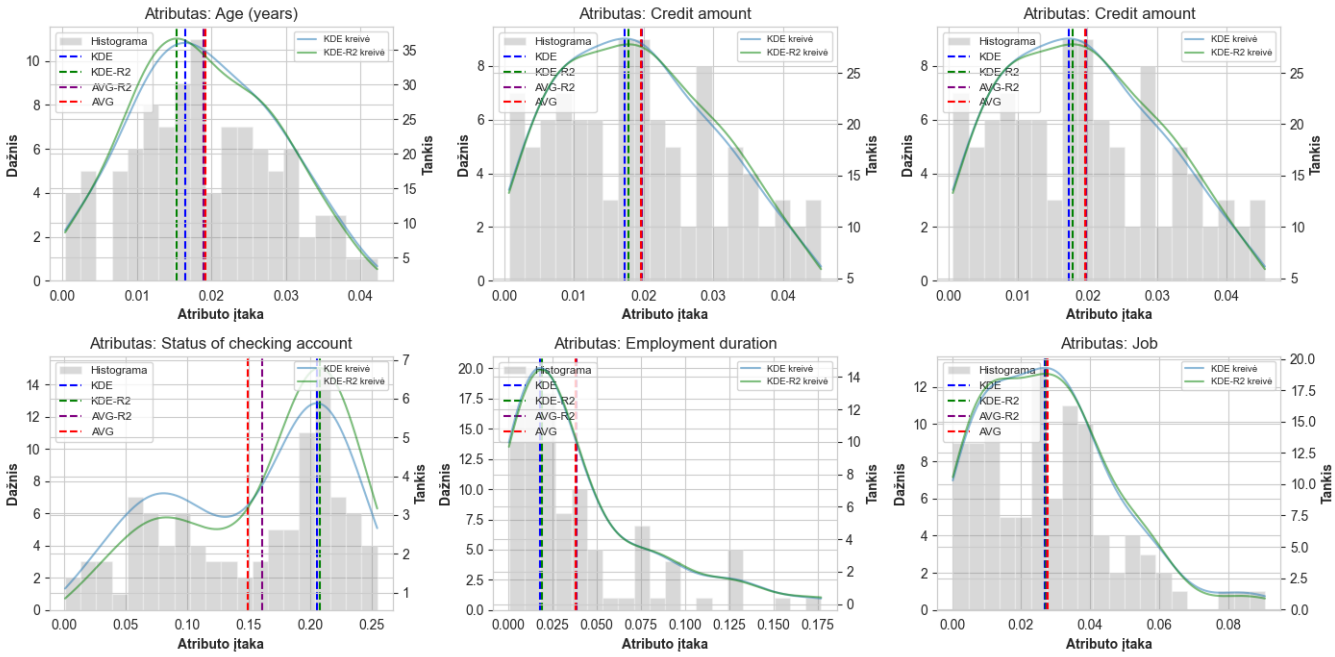


13 pav. Galutinių lokalių paaiškinimų stabilumo ir patikimumo palyginimas tarp modelių

Stabilumo ir patikimumo palyginimas taip pat pateikiamas 13 pav., kuriame matoma, kad visais duomenų rinkiniais interpretuojamas modelis – logistinė regresija (LR) – generavo patikimesnius ir stabilesnius LIME paaiškinimus nei analogiškas XGBoost (XGB) modelis. Pavyzdžiui, LR_{CRED} pasiekė vidutinį R^2 = 0.76668 ir stabilumą S_{10} = 0.64848, o XGB_{CRED} – atitinkamai 0.35103 ir 0.48219. FOREST ir DIAB duomenų rinkiniuose išlaikoma tokia pati tendencija, rodanti, kad interpretuojami modeliai yra geriau aproksimuojami LIME surogatinio modelio, todėl jų paaiškinimai yra patikimesni.

3.3. Globalaus LIME paaiškinimo metodų komponentų tyrimas ir naujų strategijų kūrimas

Prieš pereinant prie kiekybinių palyginimų, atlikta vizualinė analizė, siekiant iliustruoti, kaip skirtingos agregavimo strategijos apibrėžia globalaus paaiškinimo reikšmę, remiantis lokalių paaiškinimų pasiskirstymais. 14 pav. pateikti šešių pasirinktų atributų (XGB_{CREDIT} modelis) paaiškinimų pasiskirstymai – histogramos bei KDE tankio įverčiai. Pavaizduotos keturios agregavimo strategijos: paprastas vidurkis (AVG), R^2 svertinis vidurkis (AVG-R2), KDE maksimumas (KDE) ir R^2 svertinis KDE maksimumas (KDE-R2). Praktikoje stebima, kad šios reikšmės nesutampa – jų skirtumai aiškiai matomi, o jų tarpusavio atstumas didėja esant išreikštam šališkumui ar asimetrijai pasiskirstyme. Tai rodo, kad pasirinktas agregavimo metodas, ypač R^2 svorio taikymas ir KDE tipo skaičiavimai, gali reikšmingai paveikti globalų paaiškinimą, kai lokalūs paaiškinimai nėra normaliai pasiskirstę. Dėl šių pastebėtų skirtumų atlikta tolimesnė analizė.



14 pav. XGB_{CREDIT} pasirinktų atributų absoliučios įtakos pasiskirstymas tarp lokalių paaiškinimų. Metodai: AVG – vidurkis, AVG-R2 svorinis vidurkis, KDE ir KDE-R2

Šiame poskyryje toliau tiriami du pagrindiniai agregavimo komponentai – svoris (lokalaus paaiškinimo patikimumas, R^2) ir agregavimo funkcija (KDE) – siekiant įvertinti jų individualų ir bendrą poveikį globaliam paaiškinimui. Be to, kuriami ir vertinami nauji metodų variantai, jungiantys šiuos komponentus su skirtingomis normalizavimo strategijomis, siekiant pagerinti globalių paaiškinimų tikslumą, stabilumą ir interpretacinę vertę.

3.3.1. Svorio vaidmuo: patikimumo (R^2) reikšmės pritaikymo poveikis

Siekiant įvertinti, ar lokalaus patikimumo (R^2) įtraukimas į LIME paaiškinimų agregavimo procesą pagerina globalių paaiškinimų kokybę, buvo atlikta kiekybinė analizė. Šio eksperimento metu esami globalaus LIME paaiškinimo metodai buvo papildyti R^2 svorio komponentu, siekiant įvertinti jo poveikį rezultatų kokybei.

11 lentelė. Globalių paaiškinimų neatitikimas (JSD↓) metodus papildant paaiškinimo patikimumo (R^2) svoriu

Modelis	LR_{DIAB}		LR_{FOREST}		LR_{CRED}		Visi LR	
Modifikacija / Metodas	-	Su R^2	-	Su R^2	-	Su R^2	-	Su R^2
GALE-AVG	0.03670	**0.03438**	0.15576	**0.15548**	0.06919	**0.06896**	0.08722	**0.08627**
OLEA	0.03670	**0.03438**	0.15576	**0.15548**	0.06919	**0.06896**	0.08722	**0.08627**
NORM-LIME	0.05450	**0.04785**	0.17781	**0.17771**	**0.10287**	0.10368	0.10101	**0.09971**
ILIME-10	0.08396	**0.07737**	**0.14496**	0.14806	0.07413	**0.07370**	0.11173	**0.10975**
SP-LIME-ALL	0.10102	**0.10001**	0.15415	**0.15387**	0.09156	**0.09134**	0.11558	**0.11507**
SP-LIME-5	**0.10338**	0.10343	0.15457	**0.15457**	**0.09879**	0.09910	**0.11891**	0.11903
GALE-H	0.20680	**0.20665**	0.16313	**0.16304**	0.09590	**0.09547**	0.15528	**0.15505**

Melsvas fonas - modifikacija pagerino metriką.

11 lentelėje pateikiami LR modelių globalių paaiškinimų ir tikrojo atributų svarbos pasiskirstymo neatitikimo rezultatai, įvertinti JSD metrika. JSD rezultatai rodo, kad daugumoje atvejų R^2 svorio taikymas sumažino skirtumą tarp agreguotų atributų svarbų ir tikrojo pasiskirstymo. Ryškiausias pagerėjimas matomas NORM-LIME metode su DIAB duomenų rinkiniu (iš 0.05450 į 0.04785), rodančiu, kad svėrimas pagal R^2 ypač veiksmingas triukšmingesniuose atvejuose. Panašus, nors ir mažesnis efektas pastebimas ILIME-10 metode su FOREST rinkiniu. Priešingai, SP-LIME-5 metodas išlieka beveik nepakitęs, kas tikėtina susiję su jų specifiniu optimizaciniu pavyzdžių atrinkimu, kuris mažiau jautrus svorio įvedimui. Tuo tarpu metodai kaip GALE-AVG, kuriuose nėra nei atrankos, nei vidinės svėrimo logikos, pagerėja visuose modeliuose. Tai patvirtina, kad R^2 svoris gali pagerinti globalų pasiskirstymo atitikimą, kai paaiškinimų kokybė yra kintama ir nėra iš anksto filtruojama atrenkant paaiškinimus agregacijai.

12 lentelė. Globalių paaiškinimų tikslumo kritimo AUC metodus papildant paaiškinimo patikimumo (R^2) svoriu

Modelis	LR_{DIAB}		XGB_{DIAB}		LR_{FOREST}		XGB_{FOREST}		LR_{CRED}		XGB_{CRED}	
Modifikacija / Metodas	-	Su R^2	-	Su R^2	-	Su R^2	-	Su R^2	-	Su R^2	-	Su R^2
GALE-AVG	0.41137	**0.41512**	0.45895	0.45895	0.47741	0.47741	0.44801	**0.44804**	**0.30294**	0.30283	**0.45129**	0.44751
GALE-H	0.36927	**0.36938**	0.44992	0.44992	**0.48393**	0.48226	**0.46387**	0.46385	0.30540	**0.30551**	**0.45146**	0.45037
ILIME-10	0.39154	**0.39242**	0.45147	**0.45187**	0.46486	**0.47242**	0.45001	**0.45003**	**0.30529**	0.30514	0.42563	0.42563
NORM-LIME	0.41445	**0.41548**	0.45895	0.45895	**0.47599**	0.46881	0.44678	**0.44688**	**0.30260**	0.30240	**0.43826**	0.43643
OLEA	0.41137	**0.41512**	0.45895	0.45895	0.47741	0.47741	0.44801	**0.44804**	**0.30294**	0.30283	**0.45129**	0.44751
SP-LIME-5	0.39726	**0.40042**	0.45865	**0.45896**	0.47658	0.47683	0.34512	0.34512	**0.30700**	0.30697	0.34006	0.34006
SP-LIME-ALL	0.41137	**0.41512**	0.45895	0.45895	0.47741	0.47741	0.44801	**0.44804**	**0.30294**	0.30283	**0.45129**	0.44751

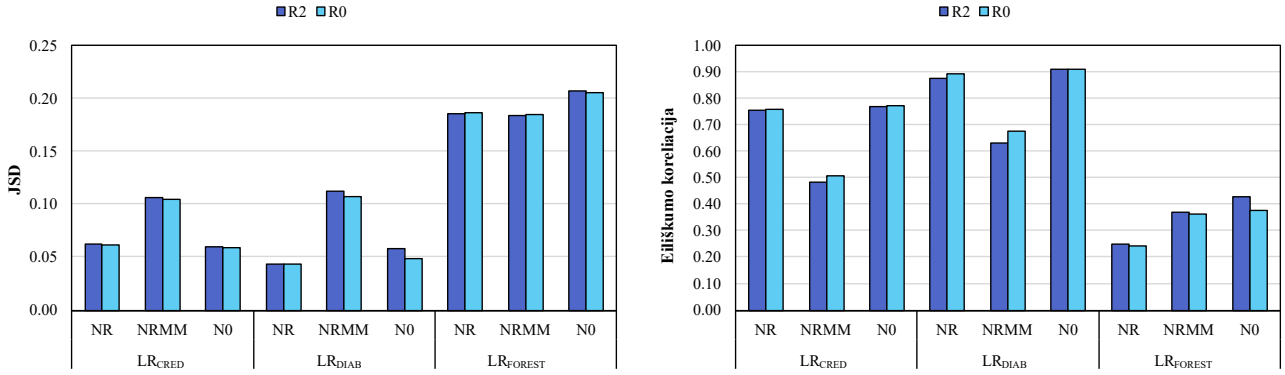
Melsvas fonas - modifikacija pagerino metriką.

12 lentelėje pateikiamos tikslumo kritimo AUC reikšmės, vertinančios, ar R^2 svorio taikymas turi įtakos globalių paaiškinimų gebėjimui atkurti modelio elgseną. Rezultatai rodo, kad daugeliu atvejų R^2 svorio taikymas turi labai nedidelį poveikį arba jo visai neturi – AUC reikšmės keičiasi minimaliai.

Nors R^2 svorio taikymas reikšmingai nepakeičia atributų svarbos eiliškumo ir nedaro esminės įtakos tikslumo kritimo metrikoms, jis pagerina atributų svarbos pasiskirstymo atitikimą – distribucija tampa artimesnė tikrajam modelio elgesiui, kas gali padidinti globalių paaiškinimų patikimumą ir interpretacinę vertę.

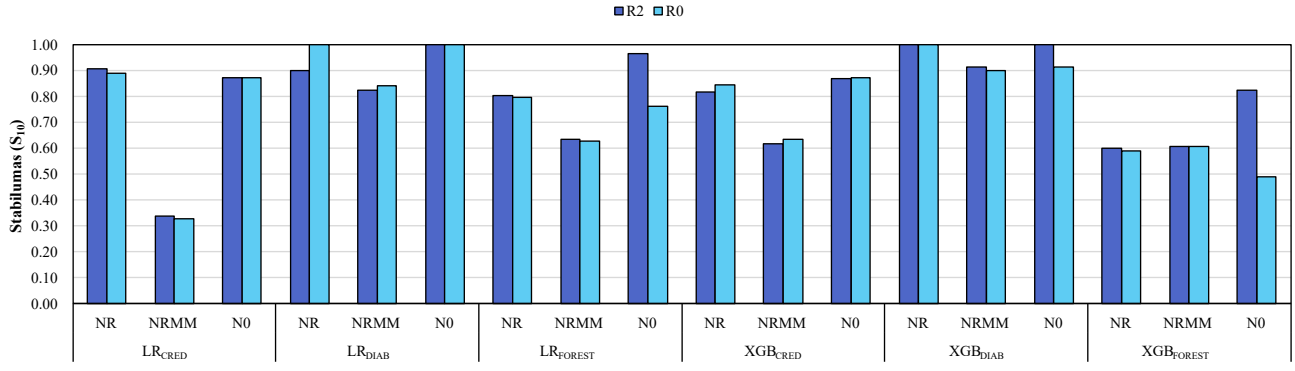
3.3.2. KDE pagrįstų agregavimo konfigūracijų analizė

Šio skyrelio tikslas – išbandyti įvairias KDE taikymo konfigūracijas ir identifikuoti vieną patikimą bei stabilų metodą, tinkamą tolesnei analizei. Eksperimento metu sistemingai keisti du pagrindiniai parametrai: normalizacijos ir svėrimo strategijos. Naudotos trys normalizavimo parinktys: be normalizavimo (N0), eilučių (per paaiškinimą) L1 normalizacija (NR) ir jos kombinacija su stulpeline (per atributą) min–maks normalizacija (NRMM). Taip pat išbandytos dvi svėrimo strategijos: be svėrimo ir atvejo paaiškinimo svėrimas pagal jų lokalaus modelio R^2 reikšmes (R2).



15 pav. KDE pagrindo globalių paaiškinimų atitikimo tikriems atributams įvertinimas. Metrikos: pasiskirstymo neatitikimas (JSD↓) ir eiliškumo koreliacija (↑).

Pirmiausia buvo tikrinamas metodų globalių paaiškinimų atitikimas tikriems atributams: reikšmių pasiskirstymo neatitikimas (JSD) ir atributų eiliškumo koreliacija. Šie rezultatai (15 pav.) rodo, geriausią rezultatų nuoseklumą visose LR modeliuose demonstravo metodai, kuriuose nebuvo taikoma jokia normalizacija (N0). Šie metodai išsiskyrė tiek didesniu pasiskirstymo atitikimu, tiek didesne eiliškumo koreliacija. Taip pat pastebėta, kad R2 svėrimas dažniausiai gerino rezultatus, ypač kai buvo derinamas su N0 arba paprasta L1 normalizacija (NR). Tuo tarpu dvigubos normalizacijos strategija (NRMM) buvo mažiausiai efektyvi ir dažniausiai blogino paaiškinimų atitikimą tikram atributų svarbos pasiskirstymui.



16 pav. Stabilumas KDE konfigūracijų

Stabilumo analizėje (16 pav. Stabilumas KDE konfigūracijų) matoma, kad be normalizacijos (N0) taikyti metodai dažniausiai pasiekė aukščiausią stabilumą, ypač DIAB duomenų rinkinio modeliuose,

kur stabilumas buvo maksimalus visais atvejais. NR strategija (L1 normalizacija) taip pat užtikrino gerą stabilumą, ypač su LR modeliais. Priešingai, NRMM normalizacijos strategija pasirodė silpniausia – tiek LR, tiek XGB modeliuose šios konfigūracijos rodė žemiausius stabilumo rezultatus. R^2 svėrimas (R2) turėjo nedidelį poveikį, tačiau kai kuriais atvejais, pavyzdžiui XGB_{FOREST}, prisidėjo prie stabilumo pagerėjimo, ypač derinant su N0.

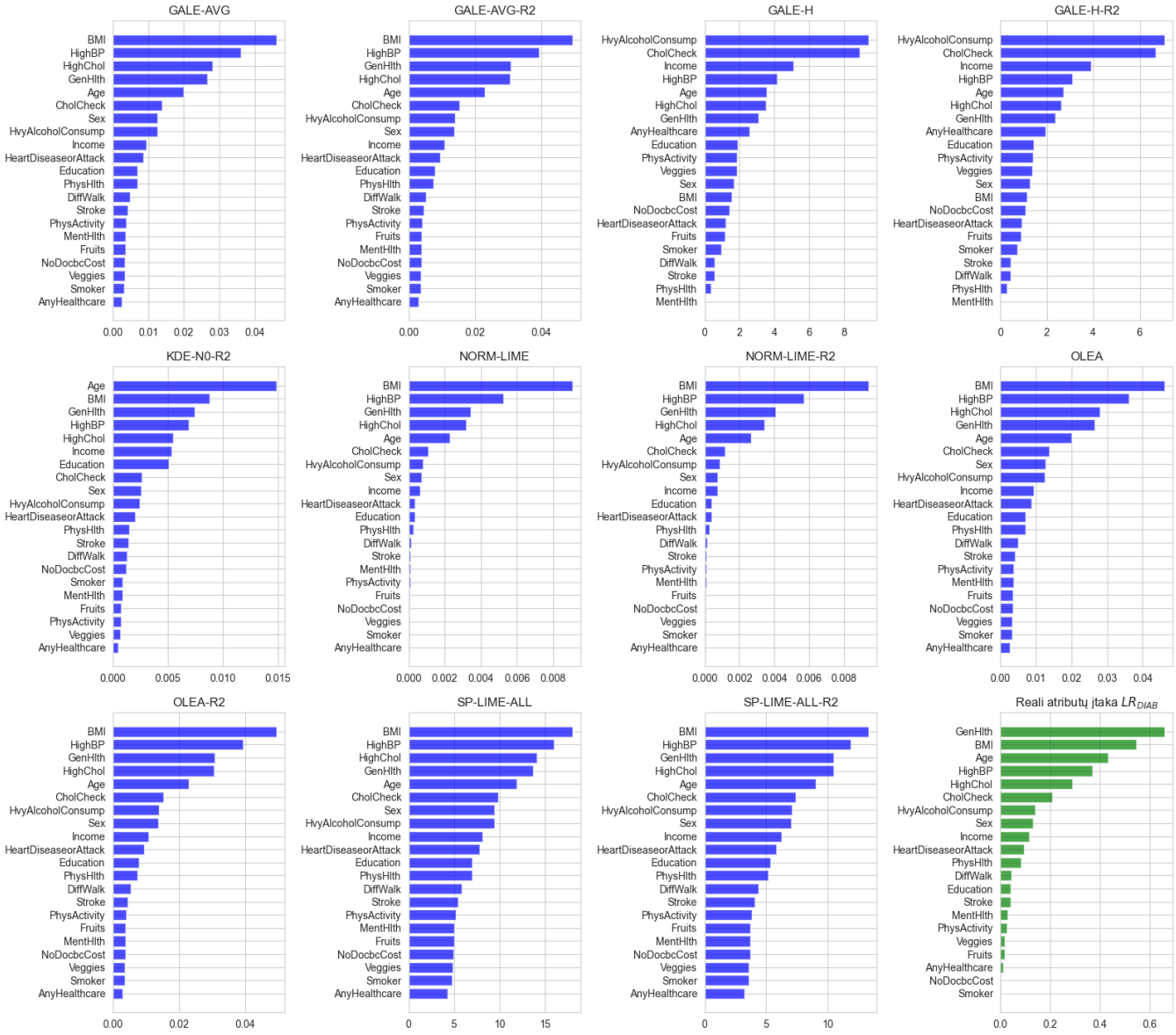
13 lentelė. KDE bendra lentelė su ranku

Modelio tipas	**LR**					**XGB**			**Vidutinis metodo rangas tarp metrikų (↓)**
Metodas	**JSD (↓)**	**Eiliškumo koreliacija (↑)**	**Tikslumo kritimo AUC (↑)**	**F1 kritimo AUC (↑)**	**S_{10} (↑)**	**Tikslumo kritimo AUC (↑)**	**F1 AURC (↑)**	**S_{10} (↑)**	
KDE-N0-R2	0.10865	**0.70278**	**0.4134**	**0.44602**	**0.94478**	0.44713	0.578	**0.89802**	2.625
KDE-NR-R0	**0.09683**	0.63295	0.40787	0.43645	0.8954	0.45887	0.5978	0.81123	2.875
KDE -N0-R0	0.10442	0.68642	0.41038	0.44313	0.87779	0.45319	0.58801	0.75986	3.25
KDE-NR-R2	0.09711	0.62671	0.40906	0.43476	0.87003	0.45719	0.59397	0.80622	3.5
KDE-NRMM-R0	0.1321	0.51557	0.40806	0.42819	0.59991	**0.47662**	**0.61499**	0.71427	4
KDE-NRMM-R2	0.13427	0.49533	0.40891	0.42451	0.59866	0.47656	0.61333	0.71324	4.75

Apibendrintoje lentelėje pateikiamos visos atributų atitikimo, paaiškinimo svarbos ir stabilumo metrikos modelio tipo lygmenyje. KDE pagrindu grįstas globalus paaiškinimo metodas geriausiai veikė taikant N0-R2 konfigūraciją: be normalizacijos ir į svorį įtraukiant paaiškinimų patikimumą (R^2). Ši konfigūracija pasiekė geriausius rezultatus daugumoje vertinimo metrikų. Išimtis buvo kai kurios XGB modelių atributų svarbos metrikos (AUC kritimo vertinimas), kur geriau pasirodė NRMM normalizacija, tačiau ši konfigūracija ženkliai nusileido kitose srityse, ypač stabilumo ir atitikimo metrikose.

Atsižvelgiant į bendrą rezultatų balansą, KDE-N0-R2 konfigūracija parenkama kaip siūlomas naujas globalaus paaiškinimo metodas, kuris toliau lyginamas su literatūroje aprašytais metodais.

3.4. Patobulintų LIME globalaus paaiškinimo metodų palyginimas su esamais metodais



17 pav. Lyginamų metodų sugeneruoti globalūs paaiškinimai ir tikra LR_{DIAB} modelio atributų įtaka.

Toliau lyginami LIME pagrindo globalaus paaiškinimo metodai tarpusavyje:

- 5 literatūroje esami metodai GALE-AVG, OLEA, SP-LIME-ALL, GALE-H, NORM-LIME
- 5 papildyti R^2 svoriu (OLEA-R2, GALE-AVG-R2, SP-LIME-ALL-R2, NORM-LIME-R2, GALE-H-R2)
- vienas išrinktas KDE metodas: KDE-N0-R2

Sugeneruotų globalių paaiškinimų su kiekvienu metodu pavyzdys, pateiktas 17 pav. iliustruoja, kaip atrodo atributų svarbos pasiskirstymas viename konkrečiame modelyje (LR_{DIAB}) vieno eksperimento pakartojimo metu. Paskutiniame grafike pateikti tikrieji modelio koeficientai leidžia vizualiai įvertinti atributų svarbos artimumą modeliui tiek pagal pasiskirstymą, tiek pagal jų eiliškumą.

3.4.1. Svarbiausių atributų atitikimo įvertinimas

Įvertintas globalių metodų atitikimas tikrosioms atributų svarboms LR modeliuose, kurių koeficientai laikomi etalonu. Naudotos dvi metrikos: pasiskirstymo neatitikimas (JSD) ir atributų eiliškumo koreliacija.

14 lentelė. Globalių paaiškinimų ir LR koeficientų vidutinis pasiskirstymo neatitikimas (JSD, n=10)

Metodas	Aiškinamas modelis			Vidutinis rangas tarp modelių
	LR_{FOREST}	LR_{DIAB}	LR_{CRED}	
OLEA-R2	0.15548	**0.03438**	0.06896	2.00
GALE-AVG-R2	0.15548	**0.03438**	0.06896	2.33
GALE-AVG	0.15576	0.03670	0.06919	4.00
OLEA	0.15576	0.03670	0.06919	4.33
SP-LIME-ALL-R2	**0.15387**	0.10001	0.09134	5.00
KDE-N0-R2	0.20735	0.05836	**0.06023**	6.00
SP-LIME-ALL	0.15415	0.10102	0.09156	6.33
GALE-H-R2	0.16304	0.20665	0.09547	8.33
NORM-LIME-R2	0.17771	0.04785	0.10368	8.33
NORM-LIME	0.17781	0.05450	0.10287	8.67
GALE-H	0.16313	0.20680	0.09590	9.33

14 lentelėje matoma, kad mažiausius neatitikimus (mažiausią JSD) rodo metodai su patikimumo (R^2) svėrimu: OLEA-R2 ir GALE-AVG-R2. KDE-N0-R2 parodė vidutinį rezultatą, aplenkdamas kai kuriuos klasikinius metodus, bet nusileisdamas geriausiai veikusiems R^2 svoriu pagrįstiems sprendimams.

15 lentelė. Globalių paaiškinimų ir LR koeficientų vidutinė eiliškumo koreliacija (*Spearman*, n=10)

Metodas	Aiškinamas modelis			Vidutinis rangas tarp modelių
	LR_{FOREST}	LR_{DIAB}	LR_{CRED}	
KDE-N0-R2	**0.42998**	0.90961	**0.76874**	3.67
GALE-AVG-R2	0.31762	0.97468	0.75314	4.00
OLEA-R2	0.31762	0.97468	0.75314	4.00
SP-LIME-ALL-R2	0.31762	0.97468	0.75314	4.00
GALE-AVG	0.31516	0.97390	0.75549	5.00
OLEA	0.31516	0.97390	0.75549	5.00
SP-LIME-ALL	0.31516	0.97390	0.75549	5.00
NORM-LIME	0.29798	**0.98078**	0.70011	6.67
NORM-LIME-R2	0.30568	0.97519	0.69196	7.00
GALE-H-R2	0.39026	0.47974	0.67042	7.33
GALE-H	0.39010	0.47455	0.66989	8.33

15 lentelėje pateiktuose eiliškumo atitikimo rezultatuose, KDE-N0-R2 pasiekė aukščiausią vidutinę eiliškumo koreliaciją tarp visų metodų ir rodo stiprų gebėjimą atkurti teisingą atributų svarbos tvarką. Tai buvo geriausias metodas tiek LR_{FOREST} tiek LR_{CRED} modeliams. Artimiausi rezultatai (GALE-AVG-R2, OLEA-R2 ir SP-LIME-ALL-R2) taip pat naudojo R^2 svėrimą. Tai rodo, kad šis svoris svarbus atributų eiliškumo įvertinime.

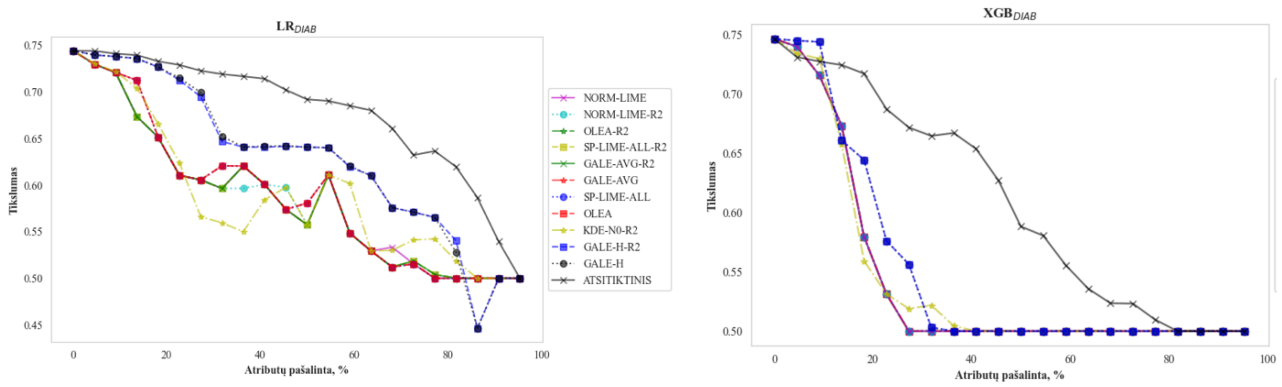
3.4.2. Globalių paaiškinimų eiliškumo poveikio modeliui įvertinimas

Toliau vertinamas skirtingų globalių paaiškinimų eiliškumo poveikis modelio veikimui, neutralizuojant svarbiausius atributus ir stebint klasifikavimo tikslumo bei F1 metrikos pokyčius.

16 lentelė. Tikslumo kritimo AUC

Metodas	Aiškinamas modelis						Vidutinis rangas tarp modelių
	LR_{DIAB}	XGB_{DIAB}	LR_{FOREST}	XGB_{FOREST}	LR_{CRED}	XGB_{CRED}	
OLEA-R2	0.41512	**0.45895**	0.47741	0.44804	0.30283	0.44751	4.17
GALE-AVG-R2	0.41512	**0.45895**	0.47741	0.44804	0.30283	0.44751	4.33
SP-LIME-ALL-R2	0.41512	**0.45895**	0.47741	0.44804	0.30283	0.44751	4.33
GALE-H	0.36927	0.44992	0.48393	**0.46387**	0.30540	0.45146	4.83
GALE-AVG	0.41137	**0.45895**	0.47741	0.44801	0.30294	0.45129	4.83
OLEA	0.41137	**0.45895**	0.47741	0.44801	0.30294	0.45129	4.83
SP-LIME-ALL	0.41137	**0.45895**	0.47741	0.44801	0.30294	0.45129	5.50
GALE-H-R2	0.36938	0.44992	0.48226	0.46385	**0.30551**	0.45037	6.25
KDE-N0-R2	0.41178	0.45814	**0.52466**	0.44613	0.30377	0.43711	6.83
NORM-LIME-R2	**0.41548**	**0.45895**	0.46881	0.44688	0.30240	0.43643	7.67
NORM-LIME	0.41445	**0.45895**	0.47599	0.44678	0.30260	0.43826	7.83
Atsitiktinis	0.31914	0.39008	0.51603	0.34239	0.31754	0.33926	-

16 lentelėje pateikiami rezultatų vidurkiai visiems metodams šešiuose skirtinguose modeliuose. Didesnes AUC reikšmes daugumoje atvejų pasiekė metodai su paaiškinimo patikimumo (R^2) svoriu: OLEA-R2, GALE-AVG-R2 ir SP-LIME-ALL-R2. Kitų metodų rezultatai buvo artimi, išskyrus GALE-H, GALE-H-R2 ir NORM-LIME variantus, kurių kai kuriais atvejais užfiksuoti tikslumo kritimo AUC rodikliai buvo mažesni.



18 pav. Tikslumo kritimas LR_{DIAB}, XGB_{DIAB} modeliuose, išimant svarbiausius atributus pateiktus pagal metodą.

18 pav. pateikiamas tikslumo kritimo grafikas, kai LR_{DIAB} ir XGB_{DIAB} modeliuose neutralizuojami kiekvieno metodo identifikuoti svarbiausi atributai. Šis grafikas iliustruoja, kaip skirtingi metodai paveikia modelio tikslumą individualiuose scenarijuose. XGB_{DIAB} atveju visų metodų kreivės artimos – skirtumai minimalūs, nors GALE-H-R2 kiek atsilieka nuo kitų. Tuo tarpu LR_{DIAB} modelyje matomi ryškesni skirtumai: GALE-H ir GALE-H-R2 pasižymi mažiausiu tikslumo kritimu, o kitų metodų poveikis modelio veikimui stipresnis.

17 lentelė. F1 kritimo AUC

Metodas	Aiškinamas modelis						Vidutinis rangas tarp modelių
	LR_{DIAB}	XGB_{DIAB}	LR_{FOREST}	XGB_{FOREST}	LR_{CRED}	XGB_{CRED}	
GALE-AVG	0.59341	0.85931	0.54202	0.51066	0.18293	0.36871	3.83
OLEA	0.59341	0.85931	0.54202	0.51066	0.18293	0.36871	3.83
SP-LIME-ALL	0.59341	0.85931	0.54202	0.51066	0.18293	0.36871	3.83
GALE-H	0.45876	0.83380	0.55498	**0.54761**	**0.18590**	**0.38464**	4.33
GALE-H-R2	0.46042	0.83380	0.55320	0.54760	0.18581	0.38244	4.83
KDE-N0-R2	0.55168	0.85513	**0.60288**	0.51275	0.18348	0.36613	5.17
OLEA-R2	0.59450	0.85931	0.54191	0.51025	0.18289	0.36332	5.50
GALE-AVG-R2	0.59450	0.85931	0.54191	0.51025	0.18289	0.36332	5.67
SP-LIME-ALL-R2	0.59450	0.85931	0.54191	0.51025	0.18289	0.36332	5.67
NORM-LIME-R2	**0.59821**	0.85931	0.52889	0.50935	0.18294	0.34612	6.50
NORM-LIME	0.59225	0.85931	0.53558	0.50931	0.18299	0.34915	7.33
Atsitiktinis	0.35668	0.63217	0.55572	0.39016	0.20062	0.24525	-

17 lentelėje pateikiami rezultatai rodo ryškesnius skirtumus, ypač duomenų rinkiniuose su nesubalansuotomis klasėmis. Čia išsiskyrė trys metodai: GALE-AVG, OLEA ir SP-LIME-ALL, kurie pasiekė aukščiausias AUC reikšmes beveik visuose modeliuose. Šie metodai pasižymėjo ir stabilumu tarp skirtingų rinkinių, ir didesniu jautrumu modelio sprendimams retesnėse klasėse, ypač CRED atveju. Priešingai, NORM-LIME metodai mažai skyrėsi nuo atsitiktinio atributų pašalinimo. KDE-N0-R2 metodas pasiekė aukščiausią AUC reikšmę su FOREST (0.60288), bet bendrai liko vidutiniškas.

3.4.3. Globalių paaiškinimų stabilumo įvertinimas

Toliau pateikiamas globalių paaiškinimų metodų stabilumo įvertinimas, siekiant įvertinti, ar skirtinguose eksperimentuose generuojami paaiškinimai išlieka nuoseklūs ir patikimi.

18 lentelė. Stabilumas

Metodas	Aiškinamas modelis						Vidutinis rangas tarp modelių
	LR_{DIAB}	XGB_{DIAB}	LR_{FOREST}	XGB_{FOREST}	LR_{CRED}	XGB_{CRED}	
GALE-AVG-R2	**1.00000**	**1.00000**	**1.00000**	0.88687	**0.96364**	0.73458	2.33
OLEA-R2	**1.00000**	**1.00000**	**1.00000**	0.88687	**0.96364**	0.73458	2.33
SP-LIME-ALL-R2	**1.00000**	**1.00000**	**1.00000**	0.88687	**0.96364**	0.73458	2.33
GALE-AVG	**1.00000**	**1.00000**	0.96364	0.88687	**0.96364**	0.74499	2.50
OLEA	**1.00000**	**1.00000**	0.96364	0.88687	**0.96364**	0.74499	2.50
SP-LIME-ALL	**1.00000**	**1.00000**	0.96364	0.88687	**0.96364**	0.74499	2.50
NORM-LIME	**1.00000**	**1.00000**	**1.00000**	0.78570	0.90707	0.84916	4.00
KDE-N0-R2	**1.00000**	**1.00000**	0.96364	0.82418	0.87071	0.86988	4.50
NORM-LIME-R2	**1.00000**	**1.00000**	0.87475	0.80995	0.90707	0.87609	5.17
GALE-H-R2	0.79529	0.93535	0.96364	0.90034	0.66744	0.40424	8.00
GALE-H	0.79731	0.93535	0.93535	**0.92795**	0.66744	0.40424	8.50

18 lentelėje pateiktuose stabilumo palyginimo rezultatuose matoma, kad aukščiausias stabilumo reikšmes pasiekė GALE-AVG-R2, OLEA-R2 ir SP-LIME-ALL-R2 jų stabilumo indeksai siekė 1 aiškinant visus logistinės regresijos modelius, taip pat ir XGB_{DIAB}, o kitose konfigūracijose reikšmės išliko aukštos. Mažiausiu stabilumu pasižymėjo GALE-H ir GALE-H-R2, kurių reikšmės ženkliai sumažėjo aiškinant LR_{CRED} ir XGB_{CRED} modelius (žemiausios reikšmės siekė 0.40424). Šie metodai užėmė paskutines vietas pagal vidutinį rangą.

3.4.4. Globalių paaiškinimų metodų palyginimo apibendrinimas

Apibendrinant, geriausius rezultatus pagal visas vertintas metrikas demonstravo R^2 svoriu papildyti metodai – OLEA-R2, GALE-AVG-R2 ir SP-LIME-ALL-R2. Jie pasiekė mažiausią pasiskirstymo neatitikimą (JSD) su LR modelio koeficientais (pvz., OLEA-R2: 0.03438 LRDIAB modelyje), aukštą atributų eiliškumo koreliaciją (pvz., GALE-AVG-R2: 0.97468), mažą klasifikavimo tikslumo ir F1 kritimą (AUC), taip pat aukščiausią stabilumą tarp eksperimentų (stabilumo indeksas – 1).

Tarp nemodifikuotų metodų geriausiai pasirodė OLEA, GALE-AVG ir SP-LIME-ALL, tačiau visose metrikose nusileido savo R^2 versijoms. Prasčiausi rezultatai matomi paaiškinimus agreguojant GALE-H ir NORM-LIME metodams – jie pasižymėjo didžiausiu JSD (pvz., GALE-H: 0.20680 LRDIAB), mažu stabilumu (pvz., GALE-H: 0.40424 XGB_{CRED}), ir silpnesniu modelio tikslumo išsaugojimu.

KDE-N0-R2 metodas pasiekė aukščiausią atributų eiliškumo koreliaciją (0.90961 LR_{DIAB}), tačiau pagal JSD, tikslumo kritimą ir stabilumą nusileido R^2 svoriu papildytiems metodams, todėl laikytinas tinkamu, kai svarbiausia – atributų eiliškumas, o ne bendras paaiškinimo patikimumas ar modelio jautrumo išlaikymas.

Išvados

1. **Esami globalūs LIME paaiškinimo metodai retai sistemingai lyginami tarpusavyje ir turi mažai nagrinėtų agregavimo komponentų.** Dauguma egzistuojančių globalių LIME metodų naudoja paprastas agregavimo strategijas – absoliučių verčių vidurkį ar kvadratinės šaknies sumą. Svorio komponentas praktikoje beveik netaikomas, o tai riboja galimybę atskirti patikimus paaiškinimus nuo triukšmingų. Be to, metodų vertinimas literatūroje dažnai apsiriboja dvejetainiais klasifikavimo modeliais, nesuteikiant įžvalgų apie veikimą aiškinant daugiaklasius klasifikavimo modelius.
2. **Tyrimo planas ir eksperimentinė aplinka sudarė sąlygas sistemingam globalių LIME paaiškinimo metodų, taikomų klasifikavimo modeliams, vertinimui.** Buvo apibrėžti aktualūs duomenų rinkiniai, modeliai, LIME metodo modifikacijos bei vertinimo metrikos lokalių ir globalių paaiškinimų kokybei įvertinti. Realizuota eksperimentinė aplinka, įskaitant duomenų bazę, užtikrino tyrimo atkuriamumą, lankstumą bei išsamų rezultatų sekimą.
3. **Paruošti lokalūs LIME paaiškinimai įvairaus sudėtingumo ir tikslumo klasifikavimo modeliams, sudarė įvairiapusį pagrindą globalių paaiškinimo metodų analizei, leidžiantį įvertinti jų veikimą skirtingų modelių architektūrų, klasių pasiskirstymo ir tikslumo sąlygomis.** Kiekvienam rinkiniui (subalansuotui (DIAB), nesubalansuotui (CRED) ir daugiaklasiui su stipria klasių disproporcija (FOREST)) buvo apmokyti logistinės regresijos (LR) ir „XGBoost“ modeliai, kurių kokybė ženkliai skyrėsi: „XGBoost“ pasiekė aukščiausią tikslumą DIAB (0.748), CRED (0.72) ir FOREST (0.972) duomenų rinkiniuose, o LR_{FOREST} pasirodė prasčiausiai (F1 = 0.483). Jiems atlikus lokalius paaiškinimus, geriausi rezultatai pasiekti aiškinant LR_{CRED} modelį ($R^2 \approx 0.77$; $S10 \approx 0.65$), o prasčiausi – XGB_{FOREST} modeliui ($R^2 \approx 0.38$; $S10 \approx 0.30$), atskleidžiant lokalių paaiškinimų ribas sudėtingose, daugiaklasėse užduotyse.
4. **Modifikuotų lokalių paaiškinimų agregavimo strategijų eksperimentinis įvertinimas parodė, kad R^2 svorio įtraukimas pagerino paaiškinimo kokybę.** Pasitvirtino, kad R^2 svorio įtraukimas į klasikinius agregavimo metodus (pvz., GALE-AVG-R2, OLEA-R2) sumažino paaiškinimų neatitikimą (JSD), padidino jų stabilumą (iki 1.0) ir kai kuriais atvejais pagerino atributų eiliškumo tikslumą. KDE pagrindu atliekamas agregavimas iš dalies pasitvirtino – KDE-N0-R2 konfigūracija pasiekė aukščiausią atributų eiliškumo koreliaciją (Spearman = 0.91). Ši strategija, pasiekusi geriausią kompromisą tarp rezultatų, pasirinkta kaip reprezentacinė KDE metodo versija tolesnei analizei.
5. **Palyginamoji analizė parodė, kad modifikuoti globalūs LIME paaiškinimo metodai dažnai lenkia literatūroje pagal paaiškinimų atitikimą klasifikavimo modelio atributų svarboms, poveikį modelio veikimui ir paaiškinimų stabilumą.** Metodai su R^2 svoriais (GALE-AVG-R2, OLEA-R2) pasiekė mažesnį JSD (atitinkamai 0.198 ir 0.207), aukštesnį stabilumą (1.0) bei mažesnį degradacijos AUC tiek tikslumo, tiek F1 atžvilgiu. Tuo tarpu KDE-N0-R2 konfigūracija, nors ir pasiekė aukščiausią atributų eiliškumo koreliaciją (Spearman = 0.91) LR_{FOREST} modeliui (mažiausio tikslumo), pagal kitus aspektus – paaiškinimų atitikimą (JSD = 0.215), stabilumą (0.88) bei degradacijos AUC – nusileido R^2 svertiniams metodams.

Literatūros sąrašas

1. AHERN, I. - NOACK, A. - GUZMAN-NATERAS, L. - DOU, D. - LI, B. - HUAN, J. [interaktyvus]. .[s.l.]: arXiv, 2019. [žiūrėta 2025-02-20]. arXiv:1909.04200 [cs]. Prieiga per internetą: <http://arxiv.org/abs/1909.04200>.

2. AKIBA, T. - SANO, S. - YANASE, T. - OHTA, T. - KOYAMA, M. [interaktyvus]. .2019. [žiūrėta 2025-05-19]. Prieiga per internetą: <https://github.com/optuna/optuna>.

3. APLEY, D.W. - ZHU, J. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. In *Journal of the Royal Statistical Society Series B: Statistical Methodology* . 2020. Vol. 82, no. 4, p. 1059–1086. .

4. BACH, S. - BINDER, A. - MONTAVON, G. - KLAUSCHEN, F. - MÜLLER, K.-R. - SAMEK, W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. In *PLOS ONE* . 2015. Vol. 10, no. 7, p. e0130140. .

5. BENZIE, A. - MONTASARI, R. Bias, Privacy and Mistrust: Considering the Ethical Challenges of Artificial Intelligence. In MONTASARI, R.Sud. *Applications for Artificial Intelligence and Digital Forensics in National Security* [interaktyvus]. Cham: Springer Nature Switzerland, 2023. p. 1–14. [žiūrėta 2024-12-16]. ISBN 978-3-031-40118-3Prieiga per internetą: <https://doi.org/10.1007/978-3-031-40118-3_1>.

6. BLACKARD, J. [interaktyvus]. .[s.l.]: UCI Machine Learning Repository, 1998. [žiūrėta 2025-05-21]. Prieiga per internetą: <https://archive.ics.uci.edu/dataset/31>.

7. BROTCKE, L. Time to Assess Bias in Machine Learning Models for Credit Decisions. In *Journal of Risk and Financial Management* . 2022. Vol. 15, no. 4, p. 165. .

8. CDC [interaktyvus]. .[s.l.]: UCI Machine Learning Repository, 2017. [žiūrėta 2025-01-18]. Prieiga per internetą: <https://archive.ics.uci.edu/dataset/891>.

9. CHEN, Z. - XIAO, F. - GUO, F. - YAN, J. Interpretable machine learning for building energy management: A state-of-the-art review. In *Advances in Applied Energy* . 2023. Vol. 9, p. 100123. .

10. DHURANDHAR, A. - CHEN, P.-Y. - LUSS, R. - TU, C.-C. - TING, P. - SHANMUGAM, K. - DAS, P. [interaktyvus]. .[s.l.]: arXiv, 2018. [žiūrėta 2024-12-18]. arXiv:1802.07623 [cs]. Prieiga per internetą: <http://arxiv.org/abs/1802.07623>.

11. DIEBER, J. - KIRRANE, S. [interaktyvus]. .[s.l.]: arXiv, 2020. [žiūrėta 2024-12-16]. arXiv:2012.00093 [cs]. Prieiga per internetą: <http://arxiv.org/abs/2012.00093>.

12. DWIVEDI, R. - DAVE, D. - NAIK, H. - SINGHAL, S. - OMER, R. - PATEL, P. - QIAN, B. - WEN, Z. - SHAH, T. - MORGAN, G. - RANJAN, R. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. In *ACM Comput. Surv.* . 2023. Vol. 55, no. 9, p. 194:1-194:33. .

13. ELSHAWI, R. - SHERIF, Y. - AL-MALLAH, M. - SAKR, S. ILIME: Local and Global Interpretable Model-Agnostic Explainer of Black-Box Decision. In WELZER, T. - EDER, J. - PODGORELEC, V. - KAMIŠALIĆ LATIFIĆ, A.Sud. *Advances in Databases and Information Systems* . Cham: Springer International Publishing, 2019. p. 53–68. .

14. FISHER, A. - RUDIN, C. - DOMINICI, F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. In *Journal of Machine Learning Research* . 2019. Vol. 20, no. 177, p. 1–81. .

15. FORMOSA, P. - ROGERS, W. - GRIEP, Y. - BANKINS, S. - RICHARDS, D. Medical AI and human dignity: Contrasting perceptions of human and artificially intelligent (AI) decision making in diagnostic and medical resource allocation contexts. In *Computers in Human Behavior* . 2022. Vol. 133, p. 107296. .

16. FRIEDMAN, J.H. Greedy function approximation: A gradient boosting machine. In *The Annals of Statistics* . 2001. Vol. 29, no. 5, p. 1189–1232. .

17. GLIKSON, E. - WOOLLEY, A.W. Human Trust in Artificial Intelligence: Review of Empirical Research. In *Academy of Management Annals* . 2020. Vol. 14, no. 2, p. 627–660. .

18. GOLDSTEIN, A. - KAPELNER, A. - BLEICH, J. - PITKIN, E. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. In *Journal of Computational and Graphical Statistics* . 2015. Vol. 24, no. 1, p. 44–65. .

19. GROUP, P.G.D. PostgreSQL. In *PostgreSQL* [interaktyvus]. 2025. [žiūrėta 2025-05-19]. Prieiga per internetą: <https://www.postgresql.org/>.

20. GUZMÁN-MARTÍNEZ, R. - ALAIZ-RODRÍGUEZ, R. Feature Selection Stability Assessment Based on the Jensen-Shannon Divergence. In GUNOPULOS, D. - HOFMANN, T. - MALERBA, D. - VAZIRGIANNIS, M.Sud. *Machine Learning and Knowledge Discovery in Databases* . Berlin, Heidelberg: Springer, 2011. p. 597–612. .

21. HOFMANN, H. [interaktyvus]. .[s.l.]: UCI Machine Learning Repository, 1994. [žiūrėta 2025-05-21]. Prieiga per internetą: <https://archive.ics.uci.edu/dataset/144>.

22. JIMÉNEZ-APARICIO, M. - RENO, M.J. - WILCHES-BERNAL, F. Shapley Additive Explanations for Traveling Wave-based Protection on Distribution Systems. In *2022 North American Power Symposium (NAPS)* [interaktyvus]. 2022. p. 1–6. [žiūrėta 2025-05-23]. Prieiga per internetą: <https://ieeexplore.ieee.org/document/10012195>.

23. LI, Q. - CUMMINGS, R. - MINTZ, Y. Optimal Local Explainer Aggregation for Interpretable Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence* . 2022. Vol. 36, no. 11, p. 12000–12007. .

24. LINARDATOS, P. - PAPASTEFANOPOULOS, V. - KOTSIANTIS, S. Explainable AI: A Review of Machine Learning Interpretability Methods. In *Entropy* . 2021. Vol. 23, no. 1, p. 18. .

25. LINDEN, I. Van der - HANED, H. - KANOULAS, E. [interaktyvus]. .[s.l.]: arXiv, 2019. [žiūrėta 2024-11-11]. arXiv:1907.03039. Prieiga per internetą: <http://arxiv.org/abs/1907.03039>.

26. LUNDBERG, S. - LEE, S.-I. [interaktyvus]. .[s.l.]: arXiv, 2017. [žiūrėta 2024-12-11]. arXiv:1705.07874 [cs]. Prieiga per internetą: <http://arxiv.org/abs/1705.07874>.

27. LUNDBERG, S.M. - ERION, G.G. - LEE, S.-I. [interaktyvus]. .[s.l.]: arXiv, 2019. [žiūrėta 2024-12-17]. arXiv:1802.03888 [cs]. Prieiga per internetą: <http://arxiv.org/abs/1802.03888>.

28. MI, J.-X. - LI, A.-D. - ZHOU, L.-F. Review Study of Interpretation Methods for Future Interpretable Machine Learning. In *IEEE Access* . 2020. Vol. 8, p. 191969–191985. .

29. MOLNAR, C. *8.1 Partial Dependence Plot (PDP) | Interpretable Machine Learning* [interaktyvus]. . 2019. .

30. MUNOZ, C. - COSTA, K. Da - MODENESI, B. - KOSHIYAMA, A. [interaktyvus]. .[s.l.]: arXiv, 2024. [žiūrėta 2025-02-26]. arXiv:2302.12094 [cs]. Prieiga per internetą: <http://arxiv.org/abs/2302.12094>.

31. PURIFICATO, E. - LORENZO, F. - FALLUCCHI, F. - DE LUCA, E.W. The Use of Responsible Artificial Intelligence Techniques in the Context of Loan Approval Processes. In *International Journal of Human–Computer Interaction* . 2023. Vol. 39, no. 7, p. 1543–1562. .

32. RIBEIRO, M.T. - SINGH, S. - GUESTRIN, C. Anchors: High-Precision Model-Agnostic Explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence* [interaktyvus]. 2018. Vol. 32, no. 1. [žiūrėta 2024-12-18]. . Prieiga per internetą: <https://ojs.aaai.org/index.php/AAAI/article/view/11491>.

33. RIBEIRO, M.T. - SINGH, S. - GUESTRIN, C. [interaktyvus]. .[s.l.]: arXiv, 2016. [žiūrėta 2024-11-20]. arXiv:1602.04938. Prieiga per internetą: <http://arxiv.org/abs/1602.04938>.

34. RIBEIRO, M.T.C. [interaktyvus]. .2025. [žiūrėta 2025-05-23]. Prieiga per internetą: <https://github.com/marcotcr/lime>.

35. SILVERMAN, B.W. *Density Estimation for Statistics and Data Analysis*. . New York: Routledge, 2018. 176 p. ISBN 978-1-315-14091-9.

36. STRICKLAND, E. IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. In *IEEE Spectrum* . 2019. Vol. 56, no. 4, p. 24–31. .

37. TKACH, I. - AMADOR, S. Towards addressing dynamic multi-agent task allocation in law enforcement. In *Autonomous Agents and Multi-Agent Systems* . 2021. Vol. 35, no. 1, p. 11. .

38. WACHTER, S. - MITTELSTADT, B. - RUSSELL, C. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. In *SSRN Electronic Journal* [interaktyvus]. 2017. [žiūrėta 2024-12-18]. . Prieiga per internetą: <https://www.ssrn.com/abstract=3063289>.

39. WANG, X. - WANG, Y. Analysis of trust factors for AI-assisted diagnosis in intelligent Healthcare: Personalized management strategies in chronic disease management. In *Expert Systems with Applications* . 2024. Vol. 255, p. 124499. .

40. Hydra | Hydra. In [interaktyvus]. [žiūrėta 2025-05-23]. Prieiga per internetą: <https://hydra.cc/>.

41. Hugging Face – The AI community building the future. In [interaktyvus]. 2024. [žiūrėta 2025-05-23]. Prieiga per internetą: <https://huggingface.co/>.

42. [Interaktyvus]. .2024. [žiūrėta 2024-12-16]. Legislative Body: CONSIL, EP. Prieiga per internetą: <http://data.europa.eu/eli/reg/2024/1689/oj/eng>.

43. scikit-learn: machine learning in Python – scikit-learn 1.6.1 documentation. In [interaktyvus]. [žiūrėta 2025-05-23]. Prieiga per internetą: <https://scikit-learn.org/stable/>.

44. XGBoost Parameters – xgboost 2.1.3 documentation. In [interaktyvus]. [žiūrėta 2025-01-19]. Prieiga per internetą: <https://xgboost.readthedocs.io/en/stable/parameter.html>.