Kauno technologijos universitetas

Informatikos fakultetas

Bankroto prognozavimo tyrimas naudojantis mašininio mokymosi metodus

Baigiamasis magistro projektas

Tomas Kanapickas

Projekto autorius

Doc. Dr. Andrius Kriščiūnas

Vadovas

Kaunas, 2025

Kauno technologijos universitetas

Informatikos fakultetas

Bankroto prognozavimo tyrimas naudojantis mašininio mokymosi metodus

Baigiamasis magistro projektas

Dirbtinio intelekto informatika (6211BX007)

Tomas Kanapickas

Projekto autorius

Doc. Dr. Andrius Kriščiūnas

Vadovas

Doc. Dr. Darius Naujokaitis

Recenzentas

Kaunas, 2025

Kauno technologijos universitetas

Informatikos fakultetas

Tomas Kanapickas

Bankroto prognozavimo tyrimas naudojantis mašininio mokymosi metodus

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.

Tomas Kanapickas

Patvirtinta elektroniniu būdu

Kanapickas, Tomas. Bankroto prognozavimo tyrimas naudojantis mašininio mokymosi metodus. Magistro baigiamasis projektas / vadovas doc. dr. Andrius Kriščiūnas; Kauno technologijos universitetas, Informatikos fakultetas.

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

Reikšminiai žodžiai: bankroto prognozavimas, duomenų apdorojimas, klasifikavimo modeliai, nebalansuoti duomenys, kintamųjų atranka, logistinė regresija, sprendimų medžiai, k artimiausių kaimynų metodas.

Kaunas, 2025. 67 p.

Santrauka

Bankroto prognozavimas vis didesnį susidomėjimą kelianti tema mokslinėje veikloje bei praktiniam taikymui. Ši problema yra sudėtinga, dėl nuolat kintančios verslo aplinkos, bėgant metams ekonominė situacija pasaulyje, tiek šalyse, tiek verslo sektoriuose keičiasi. Dėl šių priežasčių anksčiau sudaryti modeliai tampa mažiau patikimi ir norint užtikrinti kokybišką bankroto klasifikavimą reikia sudaryti naujus bankroto prognozių modelius. Tokios aplinkybės kelia būtinybę universalioms metodikoms, kurių pagalba galima būtų sudaryti patikimus ir kokybiškus bankroto prognozavimo modelius su skirtingomis verslo aplinkomis siekiant įvertinti įmonių bankroto riziką.

Spendžiant šią problemą lengvai interpretuojami modeliai yra labai gerai vertinami, kadangi tokie modeliai leidžia ne tik nustatyti bankroto riziką, bet ir paaiškinti kokie rodikliai įtakoją bankrotą. Dėl šių priežasčių tyrime buvo siekiama sudaryti bankroto prognozavimo modelius nenaudojant sudėtingų metodų, kaip dirbtiniai neuroniniai tinklai, bet pasitelkiant paprastesnius statistinius ir mašininio mokymosi metodus.

Atlikta duomenų analizė ir realizuota metodologija ištestuoti įvairių duomenų apdorojimo metodų sąveiką tarpusavyje. Trūkstamų reikšmių, išskirčių apdorojimo, duomenų standartizacijos ir balansavimo metodai buvo išanalizuoti ir identifikuoti, tiek individualūs metodai, kurie vidutiniškai pasirodė geriausiai, bet ir geriausios metodų kombinacijos.

Sudarytas reikšmingų kintamųjų atrankos metodas, pasitelkiantis žingsnine kintamųjų atranką. Taip pat pasitelkiantis kintamųjų reikšmingumo ir multikolinearumo patikrą siekiant užtikrinti, kad kiekvienas kintamasis modeliui suteiktų tik reikšmingą informaciją.

Pritaikius prieš tai minėtus metodus buvo apmokyti ir sudaryti bankroto prognozių modeliai pasitelkiantys logistinę regresiją, sprendimų medžius ir k artimiausių kaimynų metodą. Šie buvo ištestuoti su penkiomis skirtingomis duomenų imtimis ir pasiekti rezultatai palyginti su prieš tai atliktais tyrimais, kuriuose šios duomenų imtys buvo analizuotos.

Įvertinti reikšmingi finansiniai rodikliai tarp skirtingų šalių verslo aplinkų. Sulyginti bendri reikšmingi kintamieji, kurie buvo naudojami modeliuose, pasitelkiančiuose skirtingų šalių įmonių duomenų rinkinius. Nustatyta būtinybė sudaryti bankroto prognozių modelius skirtus konkrečioms verslo aplinkoms.

Pasiūlytas modelis pasiūlė stipriai interpretuojamą modelį, kuris suteikia patikimą ir kokybišką bankroto prognozavimą, kuris suderinimas su skirtingomis verslo aplinkybėmis.

Kanapickas, Tomas. A Study of Bankruptcy Prediction Using Machine Learning Methods. Master's Final Degree Project supervisor Assoc. Prof. Dr. Andrius Kriščiūnas; Faculty of Informatics, Kaunas University of Technology.

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

Keywords: bankruptcy prediction, data preprocessing, classification models, imbalanced data, feature selection, logistic regression, decision trees, k-nearest neighbors' method.

Kaunas, 2025. 67 p.

Summary

Bankruptcy prediction is an increasingly relevant topic in both academic research and practical applications. This issue is complex due to the constantly changing business environment — over time, economic conditions shift globally, nationally, and across business sectors. As a result, previously developed models become less reliable, and in order to ensure high-quality bankruptcy classification, new prediction models must be developed. These circumstances create a need for universal methodologies that can be used to build reliable and effective bankruptcy prediction models adaptable to different business environments and capable of accurately assessing corporate bankruptcy risk.

In addressing this problem, easily interpretable models are highly valued because they not only estimate bankruptcy risk but also help explain which indicators contribute to it. Therefore, this study aimed to build bankruptcy prediction models without using complex methods such as artificial neural networks and instead relied on simpler statistical and machine learning techniques.

A data analysis was conducted, and a methodology was developed to test the interaction of different data preprocessing techniques. Methods for handling missing values, outliers, data standardization, and class imbalance were analyzed, identifying both individually effective methods and the best-performing combinations.

A significant feature selection method was also proposed, incorporating stepwise selection along with relevance and multicollinearity checks to ensure that each variable included in the model provides meaningful information.

Using the aforementioned methods, bankruptcy prediction models were developed based on logistic regression, decision trees, and the k-nearest neighbors' algorithm. These models were tested on five different datasets, and the results were compared with those from previous studies that used the same datasets.

Key financial indicators across different countries' business environments were evaluated, and shared significant variables used in models across datasets were identified. The results highlighted the necessity of developing models tailored to specific business environments.

The proposed approach delivers a highly interpretable model that ensures reliable and high-quality bankruptcy prediction, while remaining adaptable to various business conditions.

Turinys

Lentelių sąrašas

Paveikslų sąrašas

Santrumpų ir terminų sąrašas

Santrumpos:

ANN – dirbtinis neuroninis tinklas (angl. Artificial Neural Network)

KNN – k artimiausių kaimynų metodas (angl. k-Nearest Neighbors)

Logit – Logistinė analizė (angl. Logistic regression)

DT – sprendimų medis (angl. Decision Tree)

AUC – plotas po ROC kreive (angl. Area Under the Curve)

RR – Eilučių šalinimas (angl. Row Removal): šalinami įrašai, kuriuose trūksta reikšmių virš nurodyto slenksčio.

RC – Kintamųjų šalinimas (angl. Column Removal): pašalinami kintamieji, kuriuose per daug trūkstamų reikšmių.

RRC – Eilučių, tuomet kintamųjų šalinimas: pirma pašalinamos eilutės, tada kintamieji.

RCR – Kintamųjų, tuomet eilučių šalinimas: pirma pašalinami kintamieji, tada eilutės.

Įvadas

Bankroto klasifikacija pastaruoju metu tapo vis didesnį susidomėjimą kelianti tema mokslinėje literatūroje [1]. Tradiciniai sprendimai rėmėsi statistiniais modeliais, tokiais kaip Altmano Z-Score [2], Ohlsono O-Score (Ohlson, 1980), Zmijevskio modelis [4] arba Biverio vienfaktorinis modelis [5], kurie buvo sukurti septintajame – devintajame dešimtmetyje. Šie metodai yra laikomi standartu ir vieni dažniausiai naudojamų sprendimų sprendžiant bankroto klasifikavimo problemą. Tačiau laikui bėgant šių metodų klasifikavimo sugebėjimai tapo mažiau patikimi, lyginant su naujais sudėtingesniais modeliais, tokiais kaip sprendimų medžiai (*angl.* Decision Trees), atraminių vektorių mašinos (*angl.* Support Vector Machines), dirbtiniai neuroniniai tinklai (*angl.* Artificial Neural Networks), XGBoost ir kiti mašininio mokymosi metodai [6,7,8]. Todėl mokslininkai ir specialistai ieško naujų būdų, kaip galima būtų pagerinti bankroto prognozavimo galimybės. Tikslus bankroto klasifikavimas yra labai svarbi dalis norint anksti nustatyti finansinius sunkumus ir priimti pagrįstus sprendimus, susijusius su paskolų teikimu, investicijomis bei rizikos valdymu. Tokie modeliai suteikia svarbių įvalgų apie įmonės finansinę būklę ir įspėti, apie bankroto riziką, kad galima būtų imtis veiksmų norint tai išvengti. Bankroto klasifikavimas yra sudėtinga problema dėl nuolat kintančios verslo aplinkos, todėl kyla sunkumų išlaikant modelio kokybę, nes po modelio apmokymo tikslumas laikui bėgant prastėja [9]. Toliau, šalyse galioja skirtingi įstatymai versle, nusakantys „įmonių žlugimo ir apskaitos taisykles“ (Veganzones ir Severin (2021), kaip cituojama [10]). Taip pat, priklausomai nuo įmonių verslo sektoriaus egzistuojančių ypatumų gali išsiskirti skirtingi bankrotui statistiškai reikšmingi kintamieji. Neseniai atliktame tyrime apie bankrotus statybos sektoriuje, autorius pabrėžė šio sektoriaus „ypatingas charakteristikas ir finansines rizikas“ (Tserng ir kt, (2011), kaip cituojama [10]). Šie iššūkiai pabrėžia universalaus, lengvai pritaikomos klasifikavimo sistemos būtinybę, kuri užtikrintų patikimus rezultatus įvairiose pramonės šakose, teisinėse sistemose bei finansinėse aplinkose.

Kuriant bankroto klasifikavimo modelį labai svarbu užtikrinti duomenų vientisumą, kas pasiekiama taikant parengiamuosius duomenų apdorojimo metodus. Šis etapas apima trūkstamų reikšmių, išskirčių, duomenų balanso problemų tvarkymą, bei duomenų standartizaciją. Šių problemų sprendimas yra labai svarbus užtikrinant duomenų vientisumą ir modelio patikimumą. Tinkamai neatlikus parengiamojo duomenų apdorojimo, gali nukentėti modelio klasifikavimo tikslumas. Trūkstamos reikšmės gali sukelti šališkumą, kuris neigiamai paveiktų klasifikavimo tikslumą ir apribotų duomenų apibendrinimo galimybes [11] – „Trūkstamų duomenų poveikis kiekybiniams tyrimams gali būti labai rimtas, nes dėl to parametrai įvertinami neobjektyviai, prarandama informacija, sumažėja statistinė galia, padidėja standartinės paklaidos ir susilpnėja rezultatų apibendrinamumas.“. Duomenyse esančios išskirtys sukelia triukšmą, kuris gali tūrėti statistiškai reikšmingą įtaką, ypač metodams kaip logistinė regresija, kuri yra ypatingai jautri išskirtims [12]. Kita labai svarbi problema – duomenų balansas. Kai modelis apmokomas naudojant nesubalansuotą duomenų rinkinį, jis tampa šališkas daugumos klasės atžvilgiu (Leevy ir kt. kaip cituojama [13]), nes dėl duomenų disbalanso modelis daug geriau išmoksta daugumos klasės tendencijas, o tai sukelia sunkumus klasifikuojant mažumos klasę. Duomenų standartizacija taip pat yra svarbi problema: duomenų masteliai retai būna vienodi, o kai modelis yra apmokomas naudojant skirtingo mastelio duomenis, modelio svoriai gali išsiskirti priklausomai nuo kintamųjų skalės. Toks elgesys iškraipo kintamųjų svarbą, todėl yra būtina standartizuoti arba normalizuoti duomenis iki vienodo mastelio, kad kintamųjų svarba būtų tinkamai parodyta [14].

Toliau, kuriant bankroto klasifikavimo modelį kintamųjų atrankos metodai yra itin svarbūs, nes padeda sumažinti statistiškai nereikšmingų kintamųjų skaičių, taip pagerindami modelio klasifikavimo galimybes (Guyon ir Elisseeff, 2003, kaip cituojama [15]). Taip pat, tai nėra vienintelis privalumas, nes tai pagerina modelio našumą, kaip veikimo laikų sutrumpinimas ir sprendžiant dimensijų problemas (Guyon ir Elisseeff, 2003, kaip cituojama [15]). Galiausiai, tai suteikia papildomų įžvalgų apie kintamuosius, kurie yra arbe nėra statistiškai reikšmingi bankroto klasifikavimui (Dash ir Liu, 1997, kaip cituojama [15]).

Kuriant bankroto prognozavimo modelius, tiriama daugybė mašininio mokymosi metodų, siekiant pasiekti kuo efektyvesnes klasifikavimo galimybės. Tiek tradiciniai, tiek sudėtingesni mašininio mokymosi metodai yra taikomi, ir kiekvienas jų turi specifinių privalumų priklausomai nuo sprendžiamos užduoties. Tyrėjai šiuos metodus suskirstė į dvi grupes: interpretuojamus ir neinterpretuojamus modelius (Mittelstadt ir kt. 2019; Cao ir kt. 2020, kaip cituojama [10]). Paprastesni mašininio mokymosi metodai leidžia lengviau paaiškinti kintamųjų įtaką modelio klasifikavimo rezultatams, o interpretuojamais modeliais galima laikyti tik logistinės regresijos bei medžių grindžiamus modelius [16]. Kita vertus, tokie sprendimai kaip dirbtiniai neuroniniai tinklai, kurie laikomi kaip neinterpretuojamais modeliais [10], gali atpažinti sudėtingesnius netiesinius ryšius tarp kintamųjų [17]. Dėl to šie metodai dažnai pasiekia geresnius klasifikavimo tikslumus nei logistinės regresijos ar medžiu pagrindu sukurti modeliai. Bankroto klasifikavimo problema reikalauja aukšto kintamųjų interpretavimo lygio, nes bankroto priežasties nustatymas gali būti ne mažiau svarbus nei pati prognozė, o tai yra itin reikšminga praktikoje, kur taikomos bankroto prognozavimas (Lundberg ir Lee 2017, kaip cituojama [10]).

Šio tyrimo tikslas – sukurti patikimą bankroto prognozavimo metodiką, siekiant aukšto klasifikavimo tikslumo įvairaus pobūdžio bei charakteristikų duomenų rinkiniuose. Atsižvelgiant, kad skirtingos duomenų imtys pasižymi įvairiomis tendencijomis ir struktūromis, nėra klasifikavimo modelio, kuris galėtų nuolat gerai apibendrinti duomenis įvairiuose duomenų rinkiniuose [18]. Dėl to mašininio mokymosi algoritmų veiksmingumas gali skirtis priklausomai nuo naudojamo duomenų rinkinio [18,19]. Ne tik modelio klasifikavimo tikslumas yra labai svarbus sprendžiant bankroto prognozės problemą, bet ir kintamųjų svarbos interpretavimas, nes supratimas, kaip konkretūs požymiai kaip finansiniai rodikliai veikia bankrotą, yra itin reikšmingas nustatant bankroto priežastis. Pasiūlyta metodika sprendžia šias problemas ir apima veiksmų eigą, kurią sudaro pagrindiniai etapai: trūkstamų reikšmių apdorojimas, išskirčių tvarkymas, duomenų balanso sprendimas, duomenų standartizavimas, statistiškai reikšmingų kintamųjų atrinkimas bei mašininio mokymosi metodo taikymas. Sukurtas modelis buvo skirtas norint užtikrinti aukštas klasifikavimo ir interpretavimo galimybes įvairiuose duomenų rinkiniuose.

Tyrimo uždaviniai:
1. atlikti duomenų analizę ir atrinkti tinkamiausius duomenų apdorojimo metodus;
2. sudaryti bankroto prognozavimo modelio metodiką;
3. įvertinti ir palyginti sudaryto modelio rezultatus su esamais sprendimais;
4. palyginti skirtingų duomenų rinkinių pagrindu sudarytų modelių atrinktus statistiškai reikšmingus kintamuosius, identifikuoti sutampančius kintamuosius ir įvertinti jų reikšmingumą.

1. Bankroto prognozavimo modelių ir duomenų apdorojimo metodų literatūros apžvalga

Bankroto prognozių modelio kūrimas reikalauja daug laiko ir pastangų. Tokie modeliai sudaromi naudojantis skirtingus mašininio mokymosi metodus, kadangi vienas modelis negali užtikrinti vienodo veiksmingumo, kuomet modelis naudojamas su skirtingomis duomenų imtimis. Šiuo metu dauguma tyrimų orientuojasi į modelio prognozių tikslumo tobulinimą taikant skirtingus mašininio mokymosi metodus arba sudarant hibridinius metodus. Tačiau, nėra nagrinėjama, kaip veiksmingai tarpusavyje veikia skirtingi duomenų apdorojimo, kintamųjų atrankos ir mašininio mokymosi metodai. Todėl universalus sprendimas, apjungiantis skirtingus šių metodų derinius, būtų itin naudingas – jis supaprastintų bankroto prognozių modelio kūrimo procesą ir padėtų veiksmingiau vertinti bankroto riziką.

1.1. Duomenų rinkiniai

Viešai prieinamų bankroto prognozavimo duomenų rinkinių yra nedaug, ypač tokių, kurie būtų tiriami mokslinėje literatūroje, nes dažniausiai pasitelkiami duomenys yra privatūs. Keletą duomenų imčių pavyko identifikuoti, kurios buvo viešai prieinamos ir buvo naudojami šio tyrimo metu kuriant ir testuojant modelius, siekiant pagerint šių tyrimų bankroto prognozavimo pasiekiamus tikslumus. Šie rinkiniai (žr. **lentelė 1**) apima Lenkijos įmonių duomenų rinkinį [20], Taivano įmones [21], JAV akcijų biržos [22], Slovakijos įmones [23], bei Lietuvos statybos sektoriaus privačių įmonių duomenis [10].

1 lentelė. Ankstesniuose tyrimuose naudojamų duomenų imčių specifikacijos ir pasiekti rezultatai

Duomenų rinkinys	**Aukščiausias pasiektas tikslumas, %**		**Geriausias metodas**	**Duomenų imties dydis**	**Kintamųjų skaičius**
Lenkiškos įmonės	Bendras	95.3	XGBoost + Dirbtinis Neuroninis Tinklas + Genetinio algoritmo optimizacija	43405	64
	Ne Bankroto	96.7		41314	
	Bankroto	75.2		2091	
Taivano įmonės	Bendras	87.27	Daugiasluoksnis parceptronas	6819	96
	Ne Bankroto	88.68		6599	
	Bankroto	86.61		220	
JAV akcijų biržos įmonės	Bendras	74.35	Dirbtinis Neuroninis Tinklas	78682	18
	Ne Bankroto	74.09		73462	
	Bankroto	82.18		5220	
Slovakijos įmonės	Bendras	-	Sprendimų medis	51407	63
	Ne Bankroto	-		51156	
	Bankroto	-		251	
Privačios Lietuvos statybos sektoriaus įmonės	Bendras	94.2	Logistinė regresija	2932	92
	Ne Bankroto	95.7		1775	
	Bankroto	91.7		1157	

Vienas reikšmingas požymis, kuris apima šiuos duomenų rinkinius, išskyrus Lietuvos statybos sektoriaus rinkinį, yra itin didelis duomenų disbalansas – nebankrutavusių įmonių klasė sudaro didžiąją dalį duomenų imties, bent per 10 kartų daugiau nei bankrutavusių. Disbalanso problema dažnai pasitaiko, kuomet susiduriama su bankroto prognozių problema, nes nebankrutavusių įmonių kiekis yra ženkliai didesnis nei bankrutavusių, o šių klasių santykis gali siekti nuo 100:1 net iki 1000:1 [24]. Šie tyrimai buvo pasirinkti, nes juose naudojami duomenų rinkiniai buvo viešai prieinami ir pateikė pakankamai išsamius rezultatus, leidžiančius atlikti reikšmingą palyginimą su šiame darbe siūlomos metodikos gaunamais rezultatais. Duomenų rinkiniai buvo naudojami modelių mokymui, testavimui ir gautų rezultatų palyginimui. Geriausi šių tyrimų rezultatai buvo pasirinkti pagal autorių nurodytą metodą, kuris geriausiai pasirodė, o jei toks nurodymas nebuvo pateiktas – pagal aukščiausią bei labiausiai subalansuotą prognozių tikslumą. Taip pat svarbu paminėti, kad Slovakijos įmonių duomenų rinkinys pateikė tik AUC vertes ir sudaryti modeliai buvo sudaryti naudojant duomenis iš konkrečių metų ir verslo sektorių. Dėl šios priežasties, rezultatai pasiekti su Slovakijos įmonių duomenimis buvo lyginami su kitame tyrime pasiektais rezultatais, tačiau jame jau buvo naudojama kita duomenų imtis, bet šis tyrimas buvo pasirinktas todėl, nes jo duomenys apėmė tą patį laiko periodą ir šiame duomenų rinkinyje buvo taip pat buvo Slovakijos įmonės. Toliau, beveik visos duomenų imtys tūrėjo didelį skaičių nepriklausomų kintamųjų – nuo 60 iki 100, tuo tarpu JAV akcijų biržos duomenų imtis, kuri tūrėjo tik 18 kintamųjų. Slovakijos, JAV akcijų biržos ir Lenkijos įmonių duomenų rinkiniai tūrėjo daugiausiai įrašų (apie 40 – 80 tūkst.), o Lietuvos ir Taivano įmonių duomenų rinkiniai apie 3000 ir 7000 įrašų atitinkamai.

1.2. Duomenų apdorojimas

1.2.1. Trūkstamos reikšmės

Trūkstamos reikšmės į duomenis įsineša šališkumą, kuris neigiamai paveikia modelio prognozavimo ir duomenų apibendrinimo galimybes [11]. Trūkstamos reikšmės gali būti apdorojamos dviem būdais – atliekant reikšmių šalinimą arba priskiriant apskaičiuotas reikšmės (*angl.* impute). Pasirenkamas metodas priklauso nuo trūkstamų reikšmių pobūdžio - atsitiktinai trūkstamos reikšmės (MCAR), sąlyginai atsitiktinos trūkstamos reikšmės (MAR), ar neatsitiktinai trūkstamos (MNAR) [25]. Vienas šių sprendimų – nepriklausomų kintamųjų šalinimas, kurių trūkstamų reikšmių kiekis siekia nuo 15 ar 20 % [26], kadangi tokie kintamieji yra mažai informatyvus ir laikomi nereikšmingais, todėl šiuos kintamuosius galimai pašalinti (Shah ir kiti., 2017, kaip cituojama [27]). Šie metodai yra taikomi tik tuomet kai kintamasis turi labai didelį kiekį trūkstamų reikšmių, nes norima užtikrina, kad reikšminga informacija nebūtų prarasta šalinant nepriklausomus kintamuosius. Tokie metodai dažniausiai taikomi kai tenkinama MCAR sąlyga arba duomenų imtis yra laikoma labai didele. Kitas trūkstamų reikšmių šalinimo būdas, kuomet patys duomenų rinkinio įrašai yra šalinami ir tai atliekama kuomet trūksta 5 % arba daugiau kintamųjų, nes tokiu atveju įrašas yra laikomas neišsamiu [27]. Tačiau šalinimo metodai nėra patys optimaliausi sprendimai, nes dėl jų galima prarasti reikšmingos informacijos ir tai gali neigiamai paveikti modelio kokybę, patikimumą bei tikslumą [25]. Reikšmių imputacija ir leidžia išvengti šios problemos, nes čia trūkstamos reikšmės yra apskaičiuojamos iš kitų įrašų esančių duomenų rinkinyje. Imputavimui gali būti taikomi įvairūs mašininio mokymosi metodai, tokie kaip artimiausių kaimynų metodas (KNN), sprendimų medžiai (DT), atsitiktinių miškų algoritmas (RF), atraminių vektorių mašinos (SVM), klasterizavimo metodai, pavyzdžiui, K vidurkių metodas (K-means) ir hierarchinis klasterizavimas. Taip pat, dažnai taikoma ir paprastesni būdai trūkstamų reikšmių imputavimui, kai trūkstamos reikšmės pakeičiamos kintamojo vidurkiu ar modos reikšme [27]. Papildomai, trūkstamų reikšmių apdorojimo metodų tin1kamumas gali būti įvertinamas

pasinaudojant kokybės metrikomis kaip vidutinė absoliuti paklaida (MAE), vidutinė kvadratinė paklaida (MSE), šaknies vidutinė kvadratinė paklaida (RMSE) ir kreivės po ROC plotu (AUC) [25].

1.2.2. Išskirtys

Duomenyse esančios išskirtys sukelia triukšmą, kas gali tūrėti reikšmingai neigiamą poveikį modelio rezultatams. Logistinė regresija yra ypač jautri išskirtims, todėl tokiuose modeliuose yra būtina tinkamai apdoroti duomenų rinkinyje esančias išskirtis [12,28]. Yang ir kt. [28] išskirčių apdorojimui siūlo taikyti naudoti standartinį nuokrypį (SD), medianos absoliučią paklaidą (MAD), tarpkvartilinį diapazoną (IQR) bei metodus kaip Clever SD ir Hotellingo T kvadrato metodas (T^2). Šie metodai leidžia nustatyti kintamųjų minimalių ir maksimalių ribinių reikšmių intervalus, pagal kuriuos aptiktos išskirtys gali būti šalinamos arba imputuojamos [29].

- Standartinis nuokrypis (SD)

$$\begin{aligned} &\min = \mu - \ a \cdot SD \\ &max = \mu + a \cdot SD \\ &kur\ \mu - vidurkis, a - konstanta, SD - standartinis\ nuokrypis \end{aligned} \quad (1)$$

- Medianos Absoliučioji Paklaida (MAD)

$$\begin{aligned} &\min = meadian(X) - \ a \cdot MAD \\ &max = meadian(X) + a \cdot MAD \\ &MAD = b \cdot median(|X - median(X)|) \\ &kur\ X - duomenų\ imties\ reikšmė, a - MAD\ normalizavimo\ skalės\ konstanta, b \\ &\qquad - diapazono\ skaičiavimo\ skalės\ faktorius \end{aligned} \quad (2)$$

- Tarpkvartilinis Diapazonas (IQR)

$$\begin{aligned} &\min = Q_1 - \ c \cdot IQR \\ &max = Q_3 + c \cdot IQR \\ &IQR = Q_3 - Q_1 \\ &kur\ c - konsanta\ (1.5), Q_1 - pirmas\ kvartilis, Q_3 - trečias\ kvartilis \end{aligned} \quad (3)$$

- Z-Score

$$\begin{aligned} &\mathrm{z} = \frac{\mathrm{x} - \ \mu}{\sigma} \\ &kur\ x - duomenų\ imties\ reikšmė, \mu - vidurkis, \sigma - standartinis\ nuokrypis \end{aligned} \quad (4)$$

- Clever SD. Išskirčių šalinimo sprendimas remiantis standartiniu nuokrypiu (SD), kur išskirtys šalinamos iteracijomis. Kiekvienos iteracijos metu pašalinama viena išskirtis, kuri yra ekstremaliausia. Po kiekvienos iteracijos standartinis nuokrypis yra perskaičiuojamas. Ši procedūra yra kartojama, kol duomenų rinkinyje nebelieka išskirčių [30].
- T^2 metodas. Šis metodas susideda iš dviejų žingsnių. Pirmiausia pašalinamos visos ekstremaliausios išskirtys, kur ribinės reikšmės apskaičiuojamos pagal santykinę reikšmę. Toliau taikomas Clever SD metodas, kur visos kitos išskirtys yra pašalinamos. Taip pat, nebūtina naudoti

standartinio nuokrypio apskaičiuojant išskirčių ribas, bet galima pasinaudoti ir tarpkvartilinius diapazonu (IQR), medianos absoliučiąja paklaida (MAD) bei kitomis metrikomis [28].

1.2.3. Duomenų standartizacija

Duomenų standartizavimas yra kita problema su kuria susiduriama sudarant prognozių modelius, kadangi tarp nepriklausomų kintamųjų duomenų mastelis dažniausiai skiriasi. Kai modelis yra apmokamas su duomenimis, kurių mastelis nesutampa, modelio surasti svoriais gali skirtis, kuomet mažesnio mastelio kintamieji pasirodyti reikšmingesni, nei išties yra. Dėl šių priežasčių kintamųjų svarba tampa iškraipyta. Sprendžiant šią problemą reikia standartizuoti duomenis, kad visų kintamųjų masteliai sutaptų ir jų svarba būtų teisingai išreikšta [14]. Tokie metodai yra itin svarbūs kuomet naudojami metodai reikalaujantys duomenų standartizacijos, kaip KNN, kuris remiasi atstumo metrikomis kaip Euklidinis atstumas. Duomenų normalizacija ir standartizacija padeda padidinti įvairiais mašininio mokymosi algoritmais modelių klasifikavimo sugebėjimus. Viename iš ankstesnių tyrimų, lyginant su normalizacija, duomenų standartizacija pagerino modelių prognozavimo sugebėjimus per ~5% , kuomet buvo naudojami SVM, MLP, RT arba logistinė regresija. Tačiau normalizacija lyginant su standartizacija, pagerino modelio tikslumus per ~1% su Naiviu Bajeso metodu [31]. Nors ir duomenų standartizavimas rodo geresnius rezultatus su įvairiais mašininio mokymosi metodais, tačiau tokie pagerėjimai ne visuomet pasiekiami, nes kai kurių mašininio mokymosi metodų tikslumai sumažėja, kai naudojami tam tikri nesuderinami duomenų standartizavimo metodai. Ankstesniuose tyrimuose sprendžiant prognozavimo užduotis buvo naudojami duomenų standartizavimo metodai kaip standartinis, min-maks, maksimalaus absoliutaus dydžio, ir atsparusis standartizavimo metodas [32,33].

- Standartinis (*angl.* Standard Scaler)

$$\mathrm{X}' = \frac{X - \mu}{\sigma}$$
$$kur\ X' - standartizuota\ reikšmė, X - orginali\ reikšmė, \mu - duomenų\ imties\ vidurkis, \sigma - duomenų\ imties\ standartinis\ nuokrypis \quad (5)$$

- Min-Max, įprastas diapazonas [0,1] arba [-1,1]

$$\mathrm{X}' = \frac{X - X_{min}}{X_{max} - X_{min}}$$
$$\mathrm{X}' = \mathrm{a} + \frac{X - X_{min}}{X_{max} - X_{min}} \cdot (b - a)$$
$$kur\ X' - standartizuota\ reikšmė, X - orginali\ reikšmė, X_{min} - duomenų\ imties\ minimumas, X_{max} - duomenų\ imties\ maksimumas, a - diapazono\ pradžia, b - diapazono\ pabaiga \quad (6)$$

- Maksimalaus absoliutaus dydžio (*angl.* Max Absolute Scaler)

$$\mathrm{X}' = \frac{X}{|X_{max}|} \quad (7)$$
$$kur\ X' - standartizuota\ reikšmė, |X_{max}| - duomenų\ imties\ maksimali\ absoliuti\ reikšmė$$

- Atsparusis (*angl.* Robust Scaler)

$$\mathrm{X}' = \frac{X - median(X)}{IQR(X)} \quad (8)$$
$$kur\ X' - standartizuota\ reikšmė, IQR - tarpkvartilinis\ diapazonas$$

- L2 standartizacija

$$\mathrm{X}' = \frac{X}{\sqrt{\sum_{i=1}^{n} X_i^2}} \quad (9)$$
$$kur\ X' - standartizuota\ reikšmė, X - orginalus\ vektorius, X_i - kiekvienas\ elementas\ vektoriuje, \sqrt{\sum_{i=1}^{n} X_i^2} - L2\ normalizacija\ vektoriaus\ X\ (euklidinė\ normalizacija), n - kintamųjų\ skaičius\ (vektoriaus\ dimensijos)$$

1.2.4. Duomenų balansavimas

Duomenų balansas yra kita labai svarbi problema, kuomet kuriamas prognozių modelis. Kuomet modelis yra apmokamas su nebalansuotais duomenimis, modelis tampa šališkas daugumos klasės atžvilgiu (Leevy ir kt. kaip cituojama [13]) ir dėl to modelis tampa nepatikimas prognozuojant mažumos klasę. Duomenų balanso problema yra itin dažna sprendžiant bankroto prognozių problemą, kur ne bankroto įmonės apima didelę dalį duomenų imties, kurių proporcija su bankrutavusiomis įmonėmis yra 100:1 iki 1000:1 [24]. Ši problema sprendžiama arba duomenų rinkinio perdėtu ėmimu (*angl. Oversampling*), arba sumažintu ėmimu (*angl.* Undersampling). Sumažinto ėmimo metodas naikina įrašus iš daugumos klasės, kas gali būti žalingas būdas, nes gali lemti reikšmingos informacijos praradimą ir neigiamai paveikti galutinio modelio kokybę. Prieš tai buvo minėta, kad bankroto prognozavimo duomenų rinkiniai yra labai nesubalansuoti ir toks sprendimas nebus naudingas, kadangi ankstesni tyrimai rodo, kad perdėtas ėmimas yra itin naudingesnis kuomet taikomas ekstremaliai nesubalansuotiems duomenų rinkiniams [34]. Autorius Hussin [35] duomenų balansavimo sprendimus apibrėžia į dvi grupes: duomenų ir algoritmo lygio grupes. Duomenų lygio būdai duomenis apdoroja prieš paduodant juos į modelį, o algoritmo lygio būdai, modifikuoja algoritmą, kad jis būdų jautresnis mažumos klasei. Tačiau prieš tai minėtas būdas iš procesoriaus reikalauja daugiau sąnaudų skaičiavimams atlikti. Autorius Hussin [35] ištyrė kaip gerai veikia skirtingi duomenų balansavimo metodai kartu su skirtingais mašininio mokymosi metodais. Tyrime buvo parodyta, kad nėra vieno metodo, kuris geriausiai veikia, nes duomenų balansavimo metodai pasirodė geriausiai su skirtingais mašininio mokymosi metodais, todėl duomenų balansavimo metodas turi būti atrenkamas atsižvelgiant koks mašininio mokymosi metodas yra naudojamas. Tyrime analizuoti duomenų balansavimo metodai apėmė tik duomenų lygio sprendimus:

- Atsitiktinis perdėtas ėmimas (*angl.* Random Oversampling) – nauji duomenys generuojami kopijuojant jau egzistuojančius duomenis. Šis metodas gali sukelti permokymą.
- SMOTE – interpoliacija, kuomet nauji įrašai yra generuojami panašūs ir artimi jau egzistuojantiems įrašams mažumos klasėje. Šis metodas veiksmingiausias, kuomet naudojamas sprendimų medis. Tačiau kiti metodai nepagerėjo, kai šis duomenų balansavimo metodas buvo taikomas [36].
- K vidurkių klasterizavimo SMOTE (*angl.* KMeans SMOTE) – šis metodas sujungia K vidurkių klasterizavimą su SMOTE metodu, kuomet mažumos klasė yra padalinta į klasterius ir juose taikomas SMOTE.
- Ribinio atvejo SMOTE (*angl.* Borderline SMOTE) – SMOTE modifikacija, kuomet nauji įrašai kuriami ties ribiniais atvejais tarp abiejų klasių, siekiant pagerinti sudėtingesniu atvejų klasifikavimą. Šis metodas yra ypatingai naudingas, kuomet sprendimų riba nėra aiškiai apibrėžta ir gali reikšmingai pagerinti mažumos klasės klasifikavimą. Šis metodas pasiekė geriausius rezultatus, kuomet buvo taikomas su Gradiento stiprinimo (*angl.* Gradient Boosting) ir logistinės regresijos metodais [36].

1.2.5. Kintamųjų atrinkimas

Bankroto prognozavime reikšmingų kintamųjų atrinkimas yra itin svarbus norint pašalinti nereikšmingus kintamuosius ir sprendžiant dimensijų problemą (Guyon ir Elisseff, 2003, kaip cituojama [15]). Mašininio mokymosi metoduose kaip KNN ir Logit, ar net DT yra itin svarbu atrinkti reikšmingus kintamuosius ar jų kombinacijas, norint pagerinti modelių prognozavimo galimybės. Tam spręsti yra daugybė sprendimo būdų, pradedant pačių mašininio mokymosi metodų pritaikymu kaip DT ar Logit atliekant pakopinę atranką (*angl.* Stepwise selection), kuomet kiekvienas kintamasis pridedamas arba pašalinamas iš modelio iteraciniu būdu, kuomet kiekviena iteracija pagerina ir pasieka geriausią modelio tikslumą. Kiti būdai yra metrikos, kaip Džini ne grynumas (*angl.* Gini impurity) ar kitos, kurios nustato kintamųjų informatyvumo lygį. Dažnai kintamųjų atrankos metodai pagerina net mašininio mokymosi metodus, kurie jau tai atlieka patys, kaip DT, Naivus Bajesas, ar MLP. Toliau, kintamųjų reikšmingumo metrikos, kaip informacijos prieaugis, Chi-squared ar prieaugio santykis rodė vienus geriausių prognozių modelių pagerinimo lygius [37].

Taip pat metodai kaip LASSO yra dažnai naudojami, kurie vidutiniškai modelių tikslumą pagerindavo apie 1% ir geriausiais atvejais apie 3%. Toliau, mašininio mokymosi metodai kaip Logit, KNN, RBF, MLP, SVM, bei daugelis kitų parodė reikšmingus pagerėjimus kuomet buvo taikomas LASSO [38]. Į priekį atrankos metodai pasinaudojantys mašininio mokymosi metodus, iteraciniu būdu pridedant pačius reikšmingiausius kintamuosius iki tol kol modelio tikslumas nepagerėja, taip pat parodė teigiamą įtaką prognozavimo modelių kokybei ir tikslumui. Viename tyrime buvo nustatyta, jog modeliai kaip DT, Logit, Naivus Bajesas, KNN ir SVM buvo labiau patikimi ir parodė reikšmingą klasifikavimo tikslumo pagerėjimą apytiksliai nuo 2% iki 8% [39].

Toliau, kitas kintamųjų atrinkimo metodas yra kintamųjų mažinimas, kur itin koreliuojantys kintamieji yra pašalinami ir paliekami tik tie kintamieji, kurie kiekvienoje koreliuojančių kintamųjų grupėse lieka tik po vieną. Šis būdas dažniausiai naudojamas logistinėse regresijose, kadangi šie modeliai bandant įvertinti unikalius kintamųjų poveikius ir įvertinant kintamųjų reikšmingumą patiria sunkumų susidorojant su multikolinearumu. Šis sprendimas parodė logistinės regresijos klasifikavimo tikslumo pagerėjimą apytiksliai per 11%, kol kiti metodai sumažėjimą nuo 3.3% iki

net 34.8%. Tačiau logistinė regresija parodė neigiamos klasės prognozių tikslumo suprastėjimą, kol kiti metodai rodė pagerėjimą. Kintamieji, kurių koreliacija yra didesnė nei 0.7 arba mažesnė nei -0.7 yra laikomi kaip stipriai koreliuojantys, kur kintamieji diapazone |0.7, 1| laikomi stipriai koreliuojančiais [40].

1.3. Mašininio mokymosi metodai

1.3.1. Bankroto prognozių modeliai

Bankroto prognozavimo modeliai dažniausiai naudoja tradicinius statistinius modelius į kuriuos įeina daugybinė diskriminantinė analizė (*angl.* Multiple Discriminant Analysis - MDA), logistinė regresija, Altmano Z-Score [2], Ohlsono O-Score [3], Zmijewskio modelis [4] ar Beaverio vienmatis modelis [5]. Šie modeliai yra vieni pirmųjų bankroto prognozavimo modelių, kurie naudojo finansinius rodiklius, vėlesni modeliai pradėjo naudoti makroekonominius rodiklius ir istorinius įmonių duomenis [41]. Toliau, įvairūs kintamųjų atrinkimo metodai yra panaudojami, kaip koreliacinė analizė ir optimizavimo metodai, kaip genetinis algoritmas [35].

Mašininio mokymosi metodo pasirinkimas yra esminis žingsnis kuomet sudaromas bankroto prognozių modelis ir metodas priklauso nuo sprendžiamos problemos bei duomenų charakteristikų. Autorius Shi ir Li [42] apžvelgė 321 ankstesnį tyrimą, kurie analizavo bankroto prognozių problemas ir sudarė populiariausių ir dažniausiai naudotų metodų lentelės, kurie buvo naudoti per pastaruosius metus. Šie metodai buvo suskirstyti į dvi grupes: statistiniai ir mašininio mokymosi metodai.

2 lentelė. Statistinių modelių populiarumas pagal atliktų tyrimų skaičių. Šaltinis: pritaikyta pagal [42]

KLASIKINIAI STATISTINIAI MODELIAI		
Reitingas	**Metodas arba modelio pavadinimas**	**Tyrimų skaičius**
1	Logistinė regresija (Logit)	123
2	Diskriminantinė analizė	52
3	Daugybinė diskriminantinė analizė	33
4	Hazard	19
5	Logit ir Probit	7
6	Probit	6

3 lentelė. Mašininio mokymosi ir dirbtinio intelekto modelių populiarumas pagal atliktų tyrimų skaičių. Šaltinis: pritaikyta pagal [42]

MAŠININIO MOKYMOSI IR DIRBTINIO INTELEKTO MODELIAI		
Reitingas	**Metodas arba modelio pavadinimas**	**Tyrimų skaičius**
1	Neuroninis Tinklas	56
2	Atraminių vektorių mašina	32
3	Sprendimų medis	21
4	Genetinis algoritmas	20
5	Fuzzy	17
6	Rough set	13
7	Duomenų kasimas	11

4 lentelė. Kitų mašininio mokymosi modelių populiarumas pagal atliktų tyrimų skaičių. Šaltinis: pritaikyta pagal [42]

KLASIKINIAI STATISTINIAI MODELIAI		
Reitingas	**Metodas arba modelio pavadinimas**	**Tyrimų skaičius**
1	AdaBoost	7
2	Casc pagrįstas argumentavimas	6
3	Dalelių spiečiaus optimizacija	5
4	K artimiausi kaimynai	5
5	Atsitiktinis miškas	5
6	Naivaus Bajeso klasifikatorius	3

Lentelėse (žr. **lentelė 2**, **lentelė 3**, **lentelė 4**) galima pastebėti, kad logistinė regresija buvo dažniausiai naudojamas metodas, kas buvo numatyta, kadangi šis metodas yra laikomas kaip standartas sudarinėjant bankroto prognozių modelius (Lessmann ir kt.. 2012, kaip cituojama [10]), nes logistinė regresija suteikia „tinkama tikslumo, efektyvumo ir paaiškinamumo balansą“ (Crone ir Finlay, 2012; Nikolic ir kt., 2013, kaip cituojama [10]). Neuroniniai tinklai, diskriminantinė analizė, daugybinė diskriminantinė analizė, pagalbinių vektorių mašinos ir sprendimų medžiai buvo sekantys penki dažniausiai naudojami metodai bankrotų prognozei. Viename tyrime MDA, logistinė regresija ir ANN buvo palyginti, kur šie pasiekė 68-78%, 71-77% ir ~85% vidutinius tikslumus atitinkamai (Becerra-Vicario ir kt., 2020, kaip cituojama [10]). Nors ANN parodė geresnius prognozavimo sugebėjimus, negu dažnai naudojami statistiniai modeliai, tačiau ANN yra laikoma ne interpretuojamu modeliu [10], kas yra vienas pagrindinių ANN trūkumų (Figini ir kt., 2017; Horak ir kt., 2020, kaip cituojama [10]). Statistiniai modeliai ir DT yra lengvai interpretuojami modeliai ir tokie modeliai turi didelį pranašumą prieš neinterpretuojamus modelius, kadangi tokie modeliai gali paaiškinti kas lemia įmonės bankrotą ir todėl galimybė interpretuoti modelio kintamųjų reikšmingumą yra tiek svarbus, kiek aukšto klasifikavimo tikslumo pasiekimas (Lundberg ir Lee, 2017, kaip cituojama [10]).

1.3.2. Mašininio mokymosi metodai

Nors plačiausiai naudojami bankroto prognozavimo metodai apima ANN, logistinę regresiją bei kiti statistiniai metodai, tačiau šiai problemai spręsti taip pat naudojami metdai kaip SVM, DT, KNN, GA bei daugelis kitų [42].

Neseniai atliktas tyrimas [20], bandė sudaryti bankroto prognozių modelį Lenkijos įmonių duomenų rinkiniui. Čia pasiūlė sprendimą, kuris buvo sudarytas iš XGBoost ir ANN su genetinio algoritmo parametrų optimizacija. Geriausias pasiektas rezultatas pasiekė 95.3%, klasifikavimo tikslumą, bankroto klasifikavimo tikslumą su 75.2% ir sveikų įmonių klasifikavimo tikslumą su 96.7%. Šio modelio pasiekti rezultatai yra tinkami, kadangi modelis geba klasifikuoti abi klases geriau negi atsitiktinai spėjant (50%), tačiau tai nėra itin gerai subalansuotas modelis, nes bankrutavusių įmonių klasė buvo klasifikuojama 20.1% prasčiau. Kiti modeliai kaip RF pasiekė klasifikavimo tikslumą siekiantį 97.7%, tačiau bankrutavusios įmonės buvo suklasifikuotos tik 42.9% tikslumu, kas jau negali būti tinkamas modelis. SVM pasiekė prasčiausius rezultatus kur bankrutavusias įmones sugebėjo suklasifikuoti tik 10.5% tikslumu, kas yra taip pat netinkamas modelis. Toliau, XGBoost be parametrų optimizacijos pasiekė 98.2% tikslumą, kuomet bankrotas buvo klasifikuotas 54.2% tikslumu, kas jau yra ribinis atvejis kuomet modelis gali prognozuoti šiek tiek geriau nei atsitiktiniu

būdu. Toks modelis gali būti laikomas tinkamu, tačiau reiktų ieškoti geresnių. Kiti, hibridiniai modeliai, kaip KNN + SVM, AP + SVM ir AP + Logit, kurių pasiektas AUC atitinkamai siekė 0.917, 0.917 ir 0.735. Nors kai kurie metodai pasiekė taip pat aukštus rezultatus, tačiau šiame tyrime pasiūlytas modelis pasinaudojantis XGBoost + ANN su genetinio algoritmo parametrų optimizacija pasiekė aukščiausius rezultatus, kurio AUC pasiekė 0.958. Svarbu paminėti, kad duomenų rinkinys buvo stipriai nesubalansuotas, kuomet 6756 įrašai buvo nebankrutavusių įmonių ir 271 bankrutavusių, kas galėjo paveikti klasių klasifikavimo tikslumų disbalansą.

Tyrime [21], kuriame buvo nagrinėjamas Taivano įmonių duomenų rinkinys, pasiekė itin subalansuotą tikslumą – bendras tikslumas siekė 87.27%, o bankrutavusių ir nebankrutavusių įmonių klasifikavimo tikslumas atitinkamai buvo 88.61% ir 86.61%. Šis rezultatas buvo pasiektas pasinaudojant MLP modelį kartu su nepriklausomų kintamųjų atrinkimo metodu, paliekant tik po vieną kintamąjį iš kiekvienos tarpusavyje multikolinearių kintamųjų grupių. Šis kintamųjų atrinkimo metodas užtikrino, kad duomenų rinkinyje būtų paliekami tik statistiškai reikšmingi duomenys. Tyrime, taip pat buvo ištirti metodai, kaip MLP be kintamųjų atrankos, KNN, CART ir SVM, kurių tikslumas atitinkamai pasiekė 79.1%, 76.5%, 78.4% ir 81.3%. Tokie rezultatai pademonstravo, kad kintamųjų atrankos metodai reikšmingai paveikė prognozių modelį, pagerindami modelio klasifikavimo galimybes.

Kitame tyrime [22], kuriame buvo ištirtas bankroto prognozavimas su Jungtinių Amerikos Valstijų akcijų biržos duomenų rinkiniu. Šio tyrimo metodu buvo taikomi metodai, kaip SVM, ANN, logistinė regresija, AdaBoost, RF, gradientų stiprinimas ir XGBoost. Vidutiniškai, SVM metodas pasiekė aukščiausią tikslumą – 76.9%, tačiau bankrotas buvo klasifikuojamas tikslumu mažesniu nei 50%, kas nėra laikoma tinkamu rezultatu, kadangi bankroto klasės klasifikavimas buvo prastesnis nei atsitiktinis pasirinkimas. Kiti metodai, parodė jau geresnius rezultatus kuomet abidvi klasės buvo klasifikuojamos didesniu tikslumu nei 60%. Geriausiai pasirodęs metodas taikė dirbtinius neuroninius tinklus, kuris pasiekė 73.77% vidutinį tikslumą, bankroto ir ne bankroto klasių vidutinis tikslumas atitinkamai siekė 81.58% ir 73.5%, kuomet dažniausiai pasiektas atskirų klasių klasifikavimas buvo subalansuotas. Prasčiausius rezultatus pasiekė logistinė regresija, kur vidutinis tikslumas siekė 61.3%, bankroto ir ne bankroto klasifikavimo tikslumas atitinkamai siekė 84.47% ir 60.52%. Nors šis modelis pasiekė prasčiausią bendrą klasifikavimo tikslumą, tačiau jis pademonstravo geriausią bankrutavusių įmonių klasifikavimo kokybę, kuri buvo pasiekta netaikant duomenų apdorojimo sprendimų. Geriausias modelis buvo pasiektas, kuomet buvo naudojami įmonės vienerių metų duomenys ir ANN. Šis modelis pasiekė 81.04% bendrą tikslumą, 76.19% bankroto klasifikavimo tikslumą ir 81.2% ne bankroto klasifikavimo tikslumą, kuomet buvo pasiektas labiausiai subalansuotas abiejų klasių klasifikavimo tikslumas, bei pasiektas aukštas prognozavimo tikslumas.

Tyrime [23], naudojusiame Slovakijos įmonių duomenų rinkinį, buvo panaudota tik AUC modelio validavimo metrika. Buvo sudaryti keletas klasifikavimo modelių, kurie panaudojo duomenų rinkinio poaibius suskirstytus pagal duomenų įrašų metus bei įmonių verslo sektorius. Geriausias modelis buvo pasiektas naudojantis 2015 metų ir prekybos sektoriaus duomenis, kurio AUC reikšmė siekė 0.938 – didesnis negu bet kuris kitas modelis. Visų modelių vidutinis AUC siekė 0.7738, kuris apėmė visus duomenų poaibius su skirtingais įmonių metais bei jų verslo sektoriais. Antras geriausias modelis naudojo duomenų poaibį apimantį 2016 metus ir statybos sektorių, kurio AUC reikšmė pasiekė 0.859.

Lietuvos statybos sektoriaus įmonių analizuojamame tyrime buvo sudaryti bankroto prognozių modeliai naudojantis finansinius, makroekonominius, mikroekonominius ir nefinansinius rodiklius. Keli klasifikavimo modeliai buvo sudaryti pasinaudojant modelius, kaip logistinė regresija, MARS ir dviejų stadijų hibridiniai modeliai su logistine regresija, kartu su MLP arba RBF tinklu. Logistinė regresija buvo taikoma kartu su kintamųjų atrinkimo metodais. Iš visų testuotų sprendimų MARS ir logistinė regresija pasiekė geriausius rezultatus, kurie pasiekė balansuotus klasifikavimo rezultatus viršijančius 90%. Prasčiausius rezultatus pasiekė hibridiniai modeliai naudojantys logistinę regresiją kartu su MLP arba RBF, kuomet jie apytiksliai pasiekė 70% ir 80% atitinkamai. Geriausius pasirodę modeliai naudojo logistinę regresiją su reikšmingų kintamųjų atrankos metodu. Šeši tokie modeliai buvo sudaryti, kurių tikslumas apytiksliai buvo nuo 80% iki 95%. Geriausias modelis pasiekė 94.2% tikslumą ir gerai klasifikavo 91.7% bankrutavusių ir 95.7% nebankrutavusių įmonių. Šis modelis naudojo tik finansinius ir makroekonominius kintamuosius, kurie buvo parinkti naudojantis kintamųjų atrankos metodą. Duomenų rinkinys buvo sudarytas iš 1157 bankrutavusių ir 1775 nebankrutavusių įmonių, kuomet šis duomenų rinkinys turėjo mažiausią klasių disbalansą tarp visų klasių.

1.4. Modelio validavimo metrikos

Modelių validacija yra itin svarbus etapas bankroto prognozių modelio sudarymo metu, kuomet įvertinamas modelio patikimumas bei kokybė. Vieni autoriai ankstesniuose tyrimuose [43,44] išanalizavo šias metrikas, kurios naudojamos įvertinti bankroto prognozių modelių klasifikavimo galimybes.

Klaidų matrica (*angl.* Confusion matrix) – pateikiami klasifikavimo rezultatai lentelės formatu, kur nurodomi visų teisingai ir neteisingai klasifikuotų klasių skaičiai (žr. **lentelė 5**).

5 lentelė. Bankroto klasifikavimo klaidų matricos lentelė

Tikroji Klasė	Prognozuota Klasė	
	Ne Bankrotas	Bankrotas
Ne Bankrotas	TN	FP
Bankrotas	FN	TP

Klasifikavimo ir balansuotas tikslumas (BACC) yra dažnai naudojamos vertinimo metrikos. Balansuotas tikslumas yra itin naudinga metrika, kuomet duomenų rinkiniai yra nebalansuoti, kadangi atsižvelgiama, tiek į jautrumą, tiek į specifiškumą suteikiant patikimesnį rodiklį modelio kokybei vertinti, nes abi klasės vertinamos atsižvelgiant į duomenų rinkinyje esančius abiejų klasių kiekius suteikiant didesnį svorį mažumos klasei.

ROC kreivė, kur AUC nusako plotą po ROC kreive. Kitos metrikos naudojamos bankroto prognozių modelių kokybės vertinimui yra MCC, F metrika, jautrumas, specifiškumas, preciziškumas, kurie apskaičiuojami iš TP, TN, FP ir FN įrašų kiekių.

$$\text{Tikslumas} = \frac{TP + FN}{FN + TP + TN + FP} \tag{10}$$

$$\text{Balansuotas tikslumas (BACC} - angl.\,Balanced\ \text{accuracy)} = \frac{1}{2} \cdot \left(\frac{TP}{FN + TP} + \frac{TN}{TN + FP}\right) = \frac{TPR + TNR}{2} \tag{11}$$

$$\text{Precizišk}umas = \frac{TP}{TP + FP} \tag{12}$$

$$\text{Jautrumas} = TPR = \frac{TP}{TP + FN} \tag{13}$$

$$\text{Specifišk}umas = \frac{TN}{TN + FP} \tag{14}$$

$$\text{F}_\beta = \frac{(1 + p)^2 \cdot TP}{(1 + p)^2 \cdot TP + p^2 \cdot FN + FP} \tag{15}$$

$$\text{F}_1 = \frac{Preciziškumas \cdot Jautrumas}{Preciziškumas + Jautrumas} \tag{16}$$

$$\text{FNR} = \frac{FN}{FN + TP} \tag{17}$$

$$\text{FPR} = \frac{\text{FP}}{TN + FP} \tag{18}$$

$$\text{G}_{\text{mean}} = \sqrt{TNR + TPR} \tag{19}$$

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{20}$$

- Tikslumas (ACC – *angl.* Accuracy) – Proporcija tarp visų ir teisingai klasifikuotų įrašų.
- Balansuotas tikslumas (BACC – *angl.* Balanced Accuracy) – Vidurkis tarp jautrumo ir specifiškumo, suteikiantis patikimą metriką įvertinti modelio kokybę su nebalansuotais duomenimis.
- Preciziškumas – Teisingai klasifikuotų teigiamų įrašų proporcija.
- Jautrumas – Teisingai klasifikuotų tikrųjų teigiamų įrašų proporcija.
- Specifiškumas – Teisingai klasifikuotų tikrųjų neigiamų įrašų proporcija.
- F_β metrika (generalizuota F metrika) – Svertinis harmoninis vidurkis, balansuojantis preciziškumą ir jautrumą, kur β nusako jautrumo svorį.
- F_1 metrika – Preciziškumo ir jautrumo vidurkis, suteikiantis labiau subalansuotą metriką, modelio kokybei su ne balansuotais duomenimis nusakyti.
- Klaidingai neigiamų klasifikacijų dažnis (FNR) – Proporcija tarp tikrųjų teigiamų, kurie neteisingai klasifikuoti kaip neigiami ir visų tikrųjų teigiamų, nusako kiek tikrųjų teigiamų neteisingai klasifikavo.
- Klaidingai teigiamų klasifikacijų dažnis (FPR) – Proporcija tarp tikrųjų neigiamų, kurie neteisingai klasifikuoti kaip teigiami ir visų tikrųjų neigiamų, nusako kiek tikrųjų neigiamų neteisingai klasifikavo.
- G-Mean – Geometrinis vidurkis tarp jautrumo ir specifiškumo, nusako kiek gerai klasifikacija yra subalansuota tarp tikrųjų teigiamų ir neigiamų.
- Plotas po kreive (AUC - *angl.* Area Under Curve) – Tikimybė, kad modelis atsitiktinai parinktą teigiamą atvejį įvertins aukščiau nei atsitiktinį neigiamą.
- Matthew koreliacijos koeficientas (MCC – *angl.* Matthews Correlation Coefficient) – Subalansuota metrika, kuri nustato ryšį tarp prognozuotų ir tikrųjų klasių, kuomet įvertinamas tiek teisingų ir neteisingos prognozės, teigiamiems ir neigiamiems įrašams.

Klasifikavimo modelis tūrėtų klasifikuoti geriau negu atsitiktinis pasirinkimas su abejomis klasėmis, kas yra virš 50% tikslumo, kitu atveju modelis yra netinkamas [45]. Taip pat, autorius Mileris [46] nusakė pirmojo ir antrojo tipo klaidas (žr. **lentelė 6**) bankroto prognozių užduoties problemos atvejui.

Pirmojo tipo klaida yra laikoma svarbesnė klaidą iš jų abiejų, kuri įvyksta kai bankrutavusi įmonė neteisingai klasifikuota kaip nebankrutavusi. Priešingai, antrojo tipo klaida įvyksta kai ne bankrotas yra klaidingai klasifikuojamas kaip bankrotas.

6 lentelė. Bankroto klasifikavimo klaidų matricos lentelė, I ir II tipo klaidos. Šaltinis: pritaikyta pagal [46]

Tikroji Klasė	Prognozuota Klasė	
	Ne Bankrotas	Bankrotas
Ne Bankrotas	TN	FP (I tipo klaida)
Bankrotas	FN (II tipo klaida)	TP

Mileris [46] aiškina, kad jei klientas bus neteisingai klasifikuotas kaip bankrotas (II tipo klaida), bankas praranda tik palūkanas. Tačiau, kai įvyksta pirmojo tipo klaida, kuomet klientas neteisingai klasifikuojamas, kaip nebankrutavęs, bankas praranda ne tik palūkanas, bet ir visą paskolintą sumą. Dėl šios priežasties pirmojo tipo klaida laikoma svarbesne ir kritiškesne. Altman‘as [47] taip pat teigia, jog didžiausi praradimai yra asocijuojami su pirmojo tipo klaida, kad dažniausiai prirituojama klaida sprendžiant bankroto identifikavimo užduotis.

2. Bankroto prognozių metodologija ir modelių sudarymas

2.1. Naudotos technologijos

Šiame tyrime buvo naudota *Python* programavimo kalba, kuri yra plačiai naudojama duomenų analizės ir mašininio mokymosi srityse dėl jos paprastumo ir gausios įrankių bibliotekos. *Python* suteikia galime pasinaudoti įvairiais įrankiais, skirtais mašininio mokymosi, statistinių bei duomenų apdorojimo metodų įgyvendinimui. Keletas bibliotekų buvo naudojamos norint realizuoti šį tyrimą, kurios atlieka skirtingas funkcijas, kaip *Sccikit-learn* – duomenų apdorojimui, klasifikacijai bei įvertinimui, *SciPy* – statistinės funkcijos, bei validacijos metrikos, *Pandas* – lanksti ir efektyvi duomenų struktūra, kuri dažnai naudojama su prieš tai minėtomis bibliotekomis.

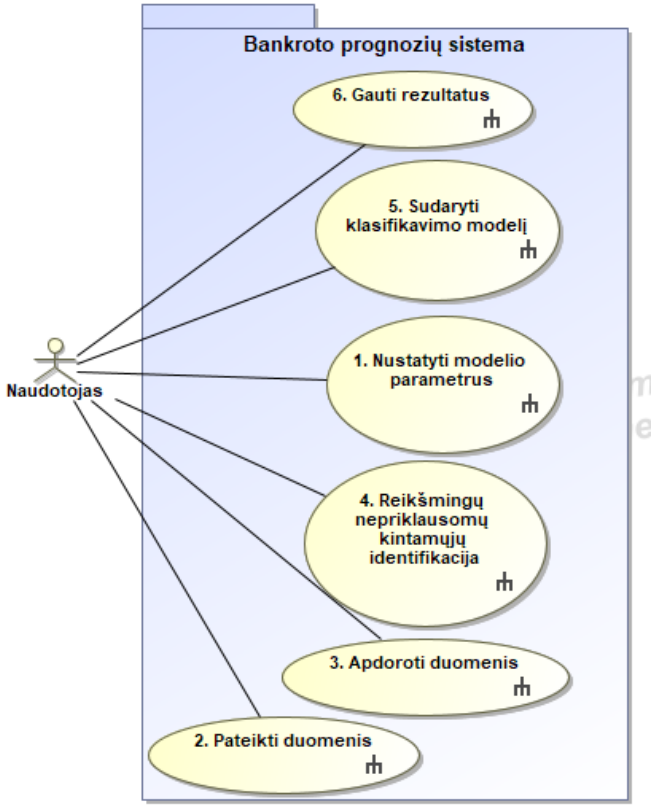
2.2. Duomenų rinkiniai

Šiame tyrime buvo siekiami sudaryti universalų bankroto prognozių modelį, kurį galimą būtų panaudoti su skirtingais duomenų rinkiniais, kurie pasižymi įvairiomis charakteristikomis. Keletas duomenų rinkinių iš ankstesnių tyrimų buvo pasirinkti, kad galima būtų nustatyti ir palyginti rezultatus, kad galima būtų išsiaiškinti ar pasiūlytas modelis pasiekia aukštą klasifikavimo tikslumą demonstruotoja duomenų apibendrinamumą, leidžiant tiesiogiai palyginti gautus rezultatus. Keturi viešai prieinami rinkiniai buvo panaudoti, kiekvienam esant iš skirtingų anksčiau atliktų tyrimų ir vienas privatus duomenų rinkinys. Šios duomenų imtys atspindi įvairių šalių ir laikotarpių įmones, įskaitant Lenkijos, Slovakijos, Taivano, JAV akcijų biržos ir privačius Lietuvos statybos sektoriaus įmonių duomenų rinkinius (žr. **lentelė 7**).

7 lentelė. Ankstesniuose tyrimuose naudojamų duomenų imčių rezultatai ir naudoti sprendimai.

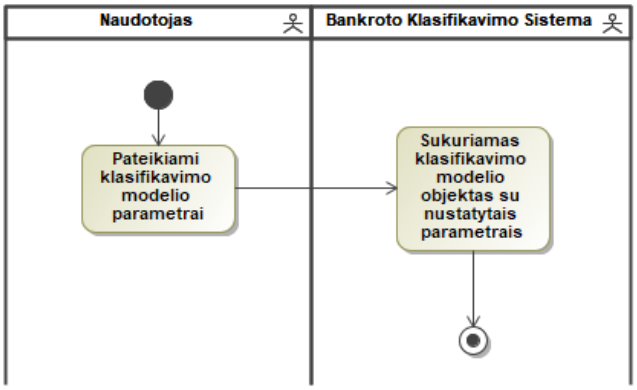
Duomenų rinkinys	**Aukščiausias pasiektas tikslumas, %**		**Geriausias metodas**
Lenkiškos įmonės	Bendras	95.3	XGBoost + Dirbtinis Neuroninis Tinklas + Genetinio algoritmo optimizacija
	Ne Bankroto	96.7	
	Bankroto	75.2	
Taivano įmonės	Bendras	87.27	Daugiasluoksnis parceptronas
	Ne Bankroto	88.68	
	Bankroto	86.61	
JAV akcijų biržos įmonės	Bendras	74.35	Dirbtinis Neuroninis Tinklas
	Ne Bankroto	74.09	
	Bankroto	82.18	
Slovakijos įmonės	Bendras	-	Sprendimų medis
	Ne Bankroto	-	
	Bankroto	-	
Privačios Lietuvos statybos sektoriaus įmonės	Bendras	94.2	Logistinė regresija
	Ne Bankroto	95.7	
	Bankroto	91.7	

2.3. Sistemos funkcijos



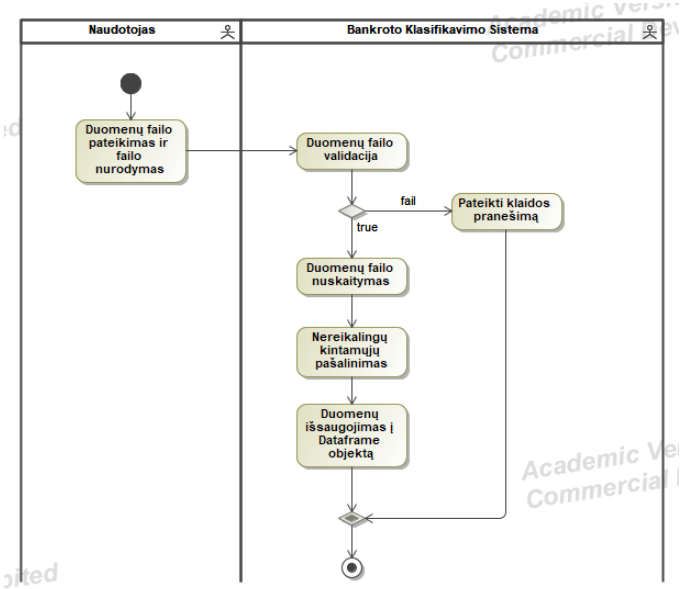
1 pav. Bankroto prognozių sistemos funkcijos

Sistema turi šešias funkcijas (žr. **1 pav1 pav.** Bankroto prognozių sistemos funkcijos): duomenų pateikimas, duomenų apdorojimas, reikšmingų kintamųjų atrinkimas, modelio parametrų nustatymas, klasifikavimo modelio sudarymas bei rezultatų gavimas.



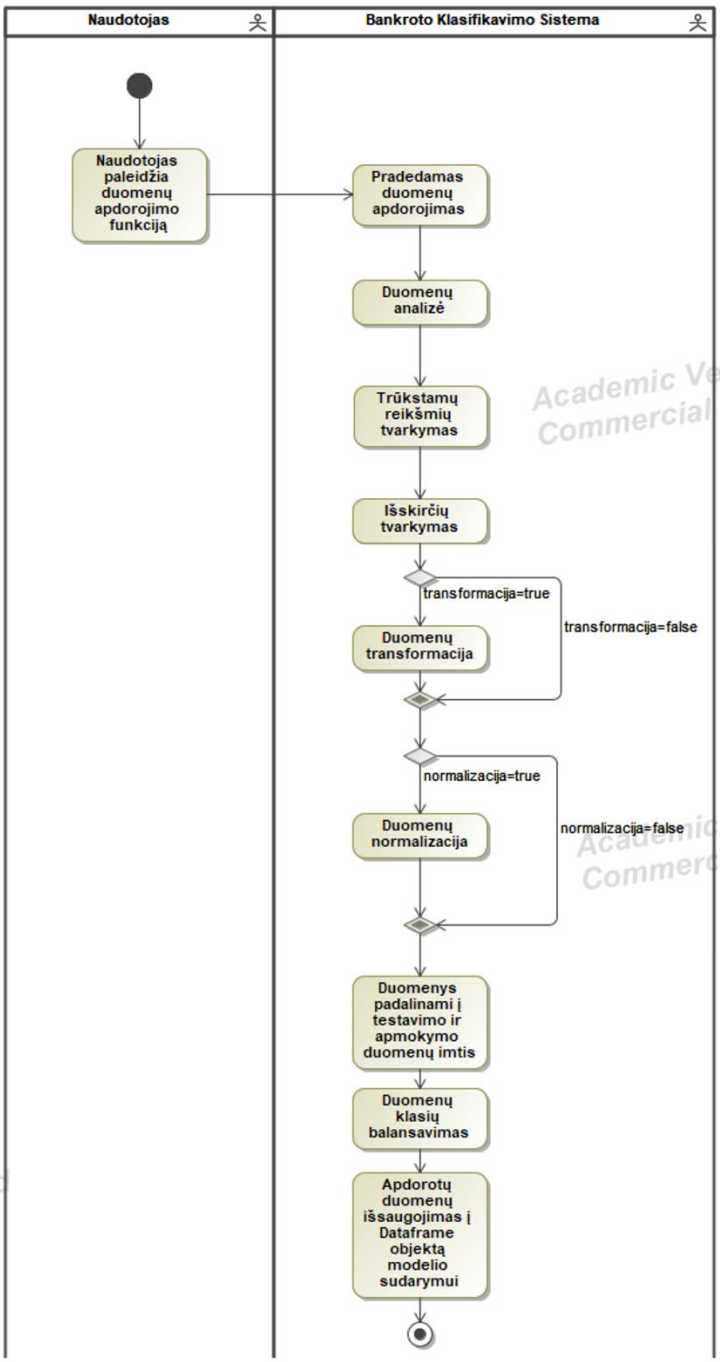
2 pav. Panaudojimo atvejis: nustatyti modelio parametrus UML diagrama.

Sistemoje pirmiausia bus nustatomi klasifikavimo modelio parametrai, pagal kuriuos bus sukuriamas klasifikavimo modelio objektas bei išsaugomi parametrai (žr. **2 pav**).



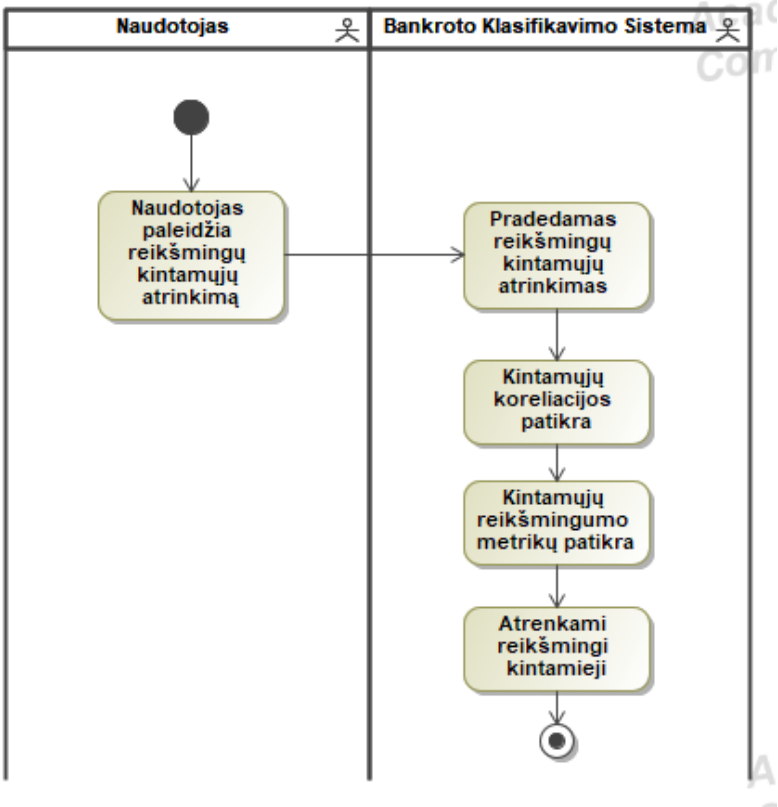
3 pav. Panaudojimo atvejis: pateikti duomenis UML diagrama.

Sistemoje (žr. **3 pav**) nurodomas failų sistemos kelias iki failo. Pirmiausia atliekama failo validacija, kuomet jei netinkama failo struktūra yra pateikiama, gaunamas klaidos pranešimas. Jei failo validacija sėkmingai praėjo, pirmiausia įvykdomas failo nuskaitymas, kuomet nurodyti kintamieji transformuojami į kategorinio tipo kintamuosius. Atitinkamos kintamųjų transformacijos nurodomos modelio parametruose. Toliau, pašalinami kintamieji, kurie nurodyti modelio parametruose, kaip kintamieji šalinimui. Galiausiai duomenys įrašomi į *Dataframe* tipo objektą.



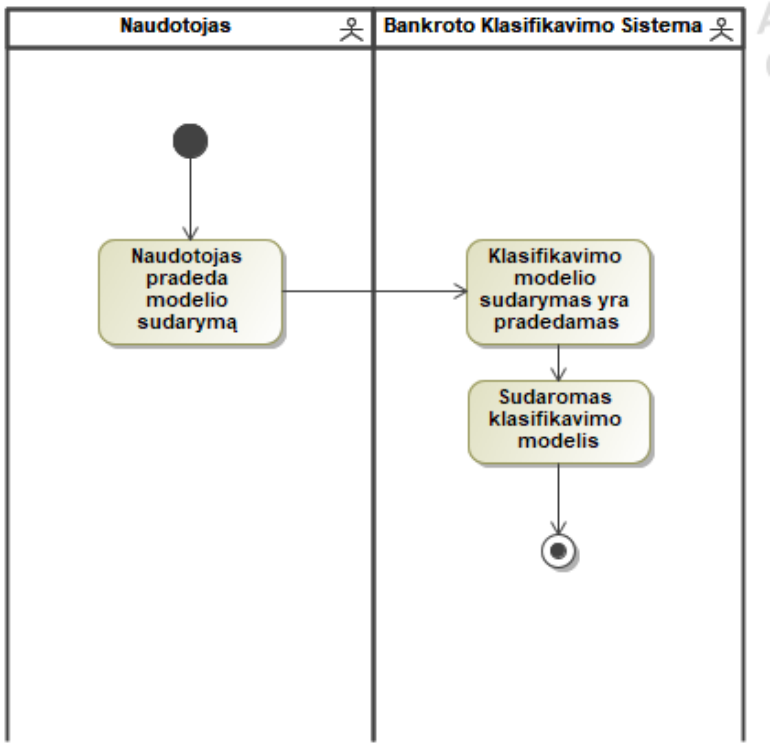
4 pav. Panaudojimo atvejis: apdoroti duomenis UML diagrama.

Duomenų apdorojimo etape (žr. **4 pav**), naudotojui nurodžius apdoroti duomenis, pirmiausia atliekama duomenų analizė ir gauti rezultatai išsaugomi tolimesniam naudojimui, kol modelio objekto gyvavimo laikotarpiu. Toliau pagal nurodytus parametrus pritaikomi, trūkstamų reikšmių, išskirčių apdorojimo metodai, atitinkamos duomenų transformacijos ir standartizacija. Po šių metodų duomenų imtis padalinama į apmokymo ir testavimo duomenų imtis. Galiausiai taikomas duomenų balansavimo metodas pagal nurodytus parametrus ir išsaugomas į naujus *Dataframe* objektus, kurie bus taikomi apmokyti ir testuoti sudarytą modelį.



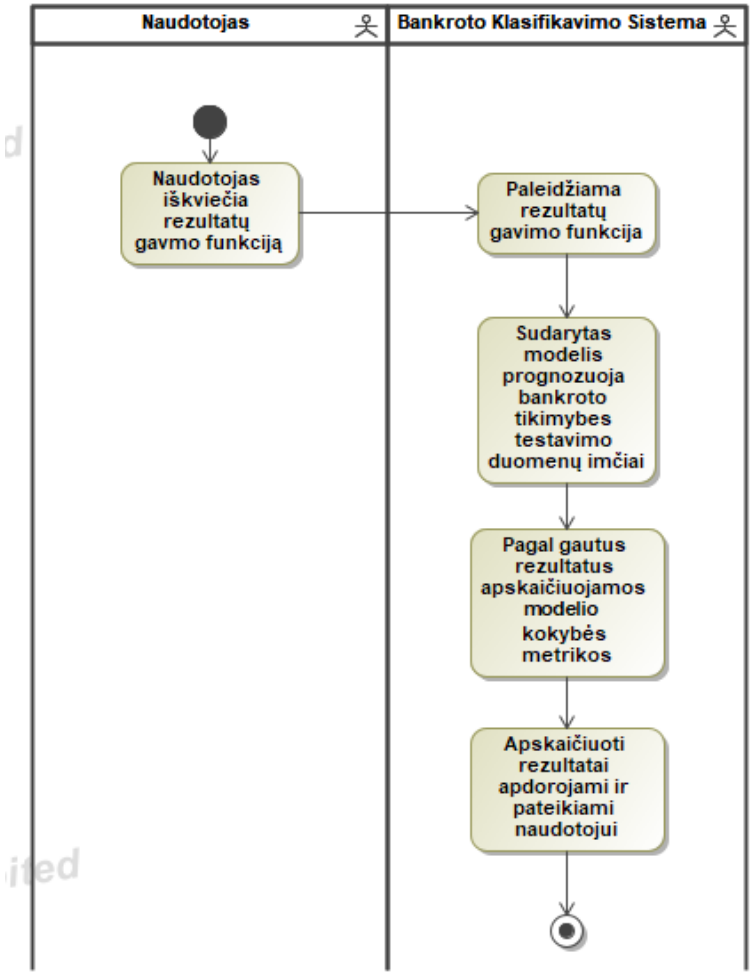
5 pav. Panaudojimo atvejis: reikšmingų nepriklausomų kintamųjų atrinkimas UML diagrama.

Naudotojui nurodžius, paleidžiamas reikšmingų kintamųjų atrinkimų procesas, kuomet atliekama kintamųjų koreliacijos patikra, bei statistiniai testai kintamųjų reikšmingumo nustatymui. Vadovaujantis šiais testų rezultatais atliekamas reikšmingų kintamųjų atrinkimas (žr. **5 pav**).



6 pav. Panaudojimo atvejis: sudaryti klasifikavimo modelį UML diagrama.

Pagal nurodytus parametrus, inicijuojamas atitinkamas modelis, bei atliekama *Grid-Search* hiperparametrų radimas. Rasti parametrai įstatomi ir apmokamas bankroto prognozių modelis (žr. **6 pav**).



7 pav. Panaudojimo atvejis: rezultatų gavimas UML diagrama.

Naudotojui nurodžius prognozuojami testavimo imties įrašai. Gavus klasifikavimo rezultatus, apskaičiuojamos modelio kokybės metrikos ir apskaičiuotos metrikos ir kiti modelio rezultatai pateikiami naudotojui, išvedami į failą (žr. **7 pav**).

8 lentelė. PA – 1: Nustatyti modelio parametrus.

PA – 1: Nustatyti modelio parametrus	
Aktorius	Naudotojas
Prieš sąlyga	Paleista aplinka, kurioje yra įrašyti reikalingi įrankiai sistemos naudojimui.
Iškvietimo sąlyga	Modelio klasės objekto iškvietimas su nurodytais parametrais.
Po sąlyga	Modelio klasės objektas yra sukurtas.

9 lentelė. PA – 2: Pateikti duomenis.

PA – 2: Pateikti duomenis	
Aktorius	Naudotojas
Prieš sąlyga	Sukurtas modelio klasės objektas.
Iškvietimo sąlyga	Iškviečiama duomenų nuskaitymo funkcija su reikalingais parametrais
Po sąlyga	Išsaugomi duomenys modelio klasės objekte.

10 lentelė. PA – 3: Apdoroti duomenis.

PA – 3: Apdoroti duomenis	
Aktorius	Naudotojas
Prieš sąlyga	Sukurtas modelio klasės objektas ir duomenys yra nuskaityti.
Iškvietimo sąlyga	Modelio duomenų apdorojimo funkcijų iškvietimas.
Po sąlyga	Klasės objektui duomenys yra apdoroti ir įkopijuoti į apmokymo ir testavimo duomenų rinkinius.

11 lentelė. PA – 4: Reikšmingų nepriklausomų kintamųjų atrinkimas.

PA – 4: Reikšmingų nepriklausomų kintamųjų atrinkimas	
Aktorius	Naudotojas
Prieš sąlyga	Sukurtas modelio klasės objektas ir testavimo, bei apmokymo duomenų imtys yra paruoštos.
Iškvietimo sąlyga	Modelio reikšmingų kintamųjų atrankos metodo iškvietimas.
Po sąlyga	Atrinktas reikšmingų kintamųjų sąrašas.

12 lentelė. PA – 5: Sudaryti klasifikavimo modelį.

PA – 5: Sudaryti klasifikavimo modelį	
Aktorius	Naudotojas
Prieš sąlyga	Sukurtas modelio klasės objektas ir testavimo, bei apmokymo duomenų imtys yra paruoštos.
Iškvietimo sąlyga	Modelio apmokymo funkcijos iškvietimas.
Po sąlyga	Sudarytas apmokytas bankroto prognozių modelis.

13 lentelė. PA – 6: Rezultatų gavimas.

PA – 6: Rezultatų gavimas	
Aktorius	Naudotojas
Prieš sąlyga	Sukurtas modelio klasės objektas ir paruoštas apmokytas bankroto prognozių modelis.
Iškvietimo sąlyga	Modelio rezultatų gavimo funkcijos iškvietimas.
Po sąlyga	Modelio rezultatai ir kokybės metrikos pateikiamos naudotojui ir išsaugomos faile.

2.4. Funkciniai ir Nefunkciniai reikalavimai

2.4.1. Funkciniai reikalavimai

- Sistema turi apdoroti trūkstamas reikšmes bei sugadintus įrašus, išskirtis. Kadangi sprendžiama bankroto prognozių problemą reikia užtikrinti duomenų kokybę, o įmonių duomenys ne visada būna geri, nes dažnai pasitaiko trūkstamos reikšmės ar sugadintos reikšmės, kas gali paveikti modelio sudaromo modelio rezultatus.
- Sistema turi apdoroti duomenų balanso problemą. Sprendžiant bankroto prognozės problemą dažniausiai pasitaiko stipriai nebalansuoti duomenys, kuomet nebankrutavusios įmonės apima didžiąją dali duomenų imties. Todėl reikia suvienodinti įrašų skaičių tarp klasių, kitu atveju modelis bus permokintas daugumos klasių atžvilgiu ir modelis bus linkęs klasifikuoti daugumos klasę geriau.
- Sistema turi priimti tik *Excelio* formato failus.
- Negali egzistuoti duomenų nutekėjimas tarp apmokymo ir testavimo duomenų imčių. Duomenys turi būti subalansuoti prieš duomenų imties padalinimą į testavimo ir apmokymo imtis, kad būtų užtikrinta, kad įrašai iš kurių buvo generuoti sintetiniai įrašai neegzistuotų abiejose klasėse.

2.4.2. Nefunkciniai reikalavimai

- Klasifikuojamų duomenų klasės turi būti klasifikuojamos ne mažesniu nei 50 procentų tikslumu. Sudaromas modelis turi abi klases klasifikuoti geriau nei atsitiktiniu būdu, priešingu atveju negalima pasitikėti modeliu, kuris negali užtikrinti klasifikacijos, geresnės nei atsitiktinis spėjimas.
- Sistemos sudarytas modelis turi užtikrinti aukštą interpretaciją. Sprendžiant bankroto prognozių problemą yra itin svarbu užtikrinti ne tik aukštą klasifikavimo tikslumą, bet ir galimybes interpretuoti modelio rezultatus. Modelis turi paaiškinti nepriklausomų kintamųjų įtaką bankrotui, nusakančią kaip kiekvienas atskiras kintamasis paveikia bankrotą.
- Sistema turi užtikrinti aukštą balansą tarp modelio klasifikuojamų klasių tikslumų.

2.5. Modelio kokybės įvertinimo metrikos

- Tikslumas –parodo bendrą modelio tikslumą, bet negalima šito kriterijaus užtikrintai naudoti, nes atskirų klasių tikslumas gali stipriai paveikti šį rezultatą, nes viena klasė gali prognozuoti 1/10 santykiu tai prognozavus mažesnė klasę su 10 % tikslumu, o didesnę klasę su 90 % tikslumu bendras tikslumas gausis apie 86 % ir toks klasifikavimo tikslumas yra nekokybiškas, nes prognozuojant bankrotą negalima leisti aukštą prognozavimą daugumos klasės atžvilgiu kas lemia aukštą bendrą klasifikavimo tikslumą, tačiau prastą mažumos klasės klasifikavimą. Toks modelis yra nekokybiškas.
- Klaidų matrica (**5 lentelė.** Bankroto klasifikavimo klaidų matricos lentelė) – modelio vertinimas pagal tikruosius teigiamus (TP), klaidingus teigiamus (FP), tikruosius neigiamus (TN), klaidingus neigiamus (FN) rezultatus siekiant užtikrinti klasifikavimo tikslumą ir pademonstruoja kaip gerai bankroto bei ne bankroto klasės yra klasifikuojamos.
- ROC kreivė ir AUC – Šie kriterijai nustato modelio sugebėjimus atskirti klases. Šis kriterijus yra patikimas ir labai svarbus klasifikavimo problemoje, nes gali parodyta, kad modelis gali atskirti klases geriau negu atsitiktiniu būdu (AUC – 0.5 yra lygus atsitiktiniam klasifikavimui, o 0.9 puikų atskyrimą kaip abidvi klasės yra puikiai atskiriamos tarpusavyje).
- Preciziškumas - nustato bankroto klasifikacijų procentą, kurie iš tikrųjų buvo bankrotai.
- Jautrumas – teisingai klasifikuotų bankroto įrašų dalis iš visų įrašų, kurie yra bankrotai.

- Specifiškumas – teisingai klasifikuotų ne bankroto įrašų dalis iš visų įrašų, kurie yra ne bankrotai.
- F1 statistika –metrika, kuri subalansuoja preciziškumo ir jautrumo metrikas ir tai yra naudinga, kadangi naudojami duomenys visuomet bus nebalansuoti. Ši metriką suteikia svorį preciziškumo ir jautrumo metrikoms ir jas galima geriau įvertinti balanso atžvilgiu.

2.6. Testavimo planas

Sistemos testavimas bus vykdomas žingsniais kuomet funkcijos bus testuojamas po kiekvieno pridėto metodo norint užtikrinti, kad viskas teisingai veikia tarpusavyje:

- Sudarytas baziniai modeliai taikant skirtingus mašininio mokymosi metodus įvertinant modelio pradinę kokybę.
- Pridedami duomenų apdorojimo metodai, kurie apima trūkstamų reikšmių, išskirčių apdorojimą, duomenų standartizacija, duomenų balansavimą ir duomenų padalinimą į apmokymo ir testavimo imtis. Kiekvienas duomenų apdorojimo metodas galėjo taikyti skirtingus metodus, kurie parenkami pagal parametrus. Šie metodai po kiekvieno apdorojimo metodo sukūrimo buvo ištestuoti ar gerai veikia su prieš tai realizuotais metodais. Jei bent vienas metodas gaudavo klaidą, jie buvo atitinkamai pataisyti, kol visi metodai buvo realizuoti.
- Toliau buvo atliekami duomenų apdorojimo ir mašininio mokymosi metodų kombinacijų testavimas. Kuomet visos kombinacijos buvo ištestuotos ir rezultatai išvesti.
- Sudarius rezultatus geriausių metodų kombinacijos ir iš kiekvieno geriausio duomenų apdorojimo metodo bus sudarytos papildomos kombinacijos kiekvienam mašininio mokymosi metodui. Kiekvienai kombinacijai bus papildomai pritaikytas kintamųjų atrankos metodas.
- Šie metodai buvo ištestuoti su visomis duomenų imtimis ir išvesti rezultatai, kurie buvo įvertinti ir lyginami su prieš tai buvusių tyrimų rezultatais, kuriuose šie duomenų rinkiniai buvo naudojami.

2.7. Metodologija

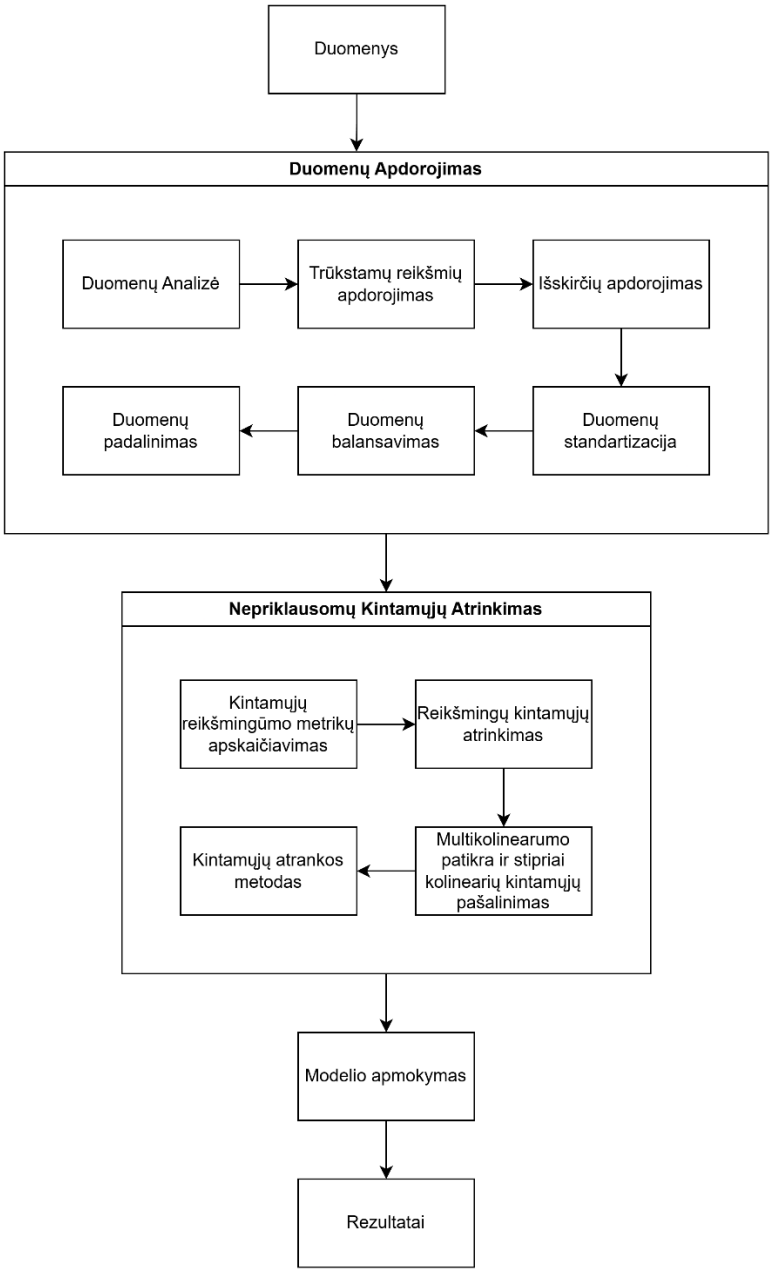
2.7.1. Metodinio proceso seka

Bankroto prognozių modeliui sudaryti buvo pasiūlyta struktūrizuota metodologinė eiga, sudaryta iš trijų pagrindinių etapų: duomenų analizė ir apdorojimas, reikšmingų kintamųjų atrinkimas ir modelio apmokymas. Metodologinės eigos diagrama () iliustruoja visą procesą, pradedant nuo duomenų įvesties ir išskiriant pagrindines procedūras kiekviename etape, kurie galiausiai lydi iki galutinio modelio rezultatų.

Duomenų apdorojimo etape, atliekamos pagrindinės užduotys, kaip duomenų analizė, trūkstamų reikšmių ir išskirčių apdorojimas, duomenų standartizavimas, duomenų balansavimas bei duomenų padalinimas į apmokymo ir testavimo imtis. Šis etapas užtikrina duomenų vientisumą ir patikimumą.

Reikšmingų kintamųjų atrankos etapas apima statistinių metrikų apskaičiavimus reikalingus nustatyti kintamųjų reikšmingumą, multikolinearių kintamųjų pašalinimas ir kintamųjų atrinkimo metodų panaudojimas. Šis etapas padeda pagerinti modelio paaiškinamumą ir kokybę, sumažinant dimensijų problemą ir pašalinant nereikšmingus kintamuosius.

Galiausiai, atrinkti kintamieji yra panaudojami klasifikavimo modelio apmokymui. Papildomai, tinklelio paieškos metodas taikomas modelio hiperparametrams optimizuoti, kad geriausia modelio konfigūracija būtų nustatyta.



8 pav. Tyrimo metodologijos procesų sekos diagrama.

Šiame tyrime buvo pasiūlytas detali metodologija (žr. **8 pav**) su kuria buvo sudarytas bankroto prognozių modelis apimantis duomenų apdorojimo, kintamųjų atrankos ir mašininio mokymosi metodų kombinaciją siekiant sudaryti bankroto prognozių modelį. Ši metodologija buvo taikoma sistemiškai testuojant įvairias šių metodų kombinacijas.

Pavyzdžiui, viena kombinacija galėjo apimti tik SMOTE balansavimo, min-max standartizacijos ir sprendimų medžių metodų kombinaciją, kol kita kombinacija galėjo apimti mean-mode trūkstamų reikšmių imputaciją su atsitiktiniu duomenų balansavimo metodais ir logistine regresija. Kiekvienas šis derinys buvo įvertinamas atskirai ir jų klasifikavimo kokybė buvo lyginama pasitelkiant įvairias modelių validavimo ir įvertinimo metrikomis.

Po visų kombinacijų įvykdymo, visi rezultatai buvo išanalizuoti, norint išsiaiškinti, kurios kombinacijos pasiekė aukščiausius rezultatus. Kuomet buvo identifikuotos geriausios metodologijos,

jos buvo pritaikomos su kitomis duomenų imtimis iš ankstesnių tyrimų, norint įvertinti, ar pasiūlytas modelis pagerino šiuos ankstesnių tyrimų rezultatus.

2.7.1.1. Duomenų apdorojimas

Pirmiausia apibrėžiamas bazinis modelis – šiame tyrime bazinis modelis laikomas prognozavimo modelis, kuriame netaikomi jokie duomenų apdorojimo ir kintamųjų atrankos metodai. Šis modelis yra naudojamas kaip atskaitos taškas, siekiant įvertinti metodo slenksčių ar koeficientų įtaką gaunamiems rezultatams, lyginant jų rezultatus izoliuotoje aplinkoje.

2.7.1.1.1. Trūkstamų reikšmių apdorojimas

Trūkstamos reikšmės buvo tvarkomos, taikant reikšmių šalinimą arba imputaciją, pritaikant įvairius metodus:

- Vidurkių ir modos (Mean-Mode) imputacija – dažnai taikoma strategija kuomet skaitinės reikšmės pakeičiamos vidurkiu ir kategoriniai kintamieji pakeičiami su moda.
- Tiesinės Regresijos (LR) imputacija – Trūkstamos reikšmės imputuojamos su prognozuotomis reikšmėmis taikant tiesinę regresiją.
- Sprendimų Medžio (DT) imputacija – Trūkstamos reikšmės imputuojamos su prognozuotomis reikšmėmis taikant sprendimų medį.
- K artimųjų kaimynų (KNN) imputacija – Trūkstamos reikšmės imputuojamos su prognozuotomis reikšmėmis taikant KNN.
- Eilučių pašalinimas (RR) – Įrašai, kurie viršijo nurodytą trūkstamų reikšmių slenkstį, buvo šalinami. Numatytasis slenkstis buvo 5% [27]. Optimalus slenkstis buvo nustatytas, jį didinant po 5%, kol mažiau nei 20% įrašų būtų pašalinti arba kol modelio rezultatai nebebus pagerinti bazinio modelio atžvilgiu. Taip pat optimalus slenkstis buvo nustatomas, jį mažinant po 5% iki tol, kol jis pasiekė 1%, kol modelio rezultatai nebebus pagerintas bazinio modelio atžvilgiu. Slenkstis, kuris pasiekė geriausią klasifikacijos tikslumą ir užtikrino, kad mažiau nei 20% reikšmių bus pašalintos, buvo atrinktas kaip pats optimaliausias slenkstis įrašų pašalinimui.
- Kintamųjų pašalinimas (RC) – Kintamieji, kurie viršijo nurodytą trūkstamų reikšmių slenkstį, buvo šalinami. Numatytasis slenkstis buvo 20% [26]. Optimalus slenkstis buvo nustatytas, jį didinant po 5%, kol modelio rezultatai nebebus pagerinti bazinio modelio atžvilgiu. Taip pat optimalus slenkstis buvo nustatomas, jį mažinant po 5% iki tol, kol jis pasiekė 1%, kol modelio rezultatai nebebus pagerintas bazinio modelio atžvilgiu. Slenkstis, kuris pasiekė geriausią klasifikacijos tikslumą, buvo atrinktas kaip pats optimaliausias slenkstis kintamųjų pašalinimui.
- Eilučių, tuomet kintamųjų šalinimas (RRC) – Sudėtinė strategija, kuomet įrašų pašalinimas (RR) yra taikomas pirmiausia, po to sekant kintamųjų šalinimu (RC). Optimalūs slenksčiai buvo nustatomi atskirai, pritaikant tokią pačią strategiją kaip RR ir RC atvejais.
- Kintamųjų, tuomet eilučių šalinimas (RRC) – Sudėtinė strategija, kuomet kintamųjų pašalinimas (RC) yra taikomas pirmiausia, po to sekant eilučių šalinimu (RR). Optimalūs slenksčiai buvo nustatomi atskirai, pritaikant tokią pačią strategiją kaip RR ir RC atvejais.
- Netaikyta – trūkstamų reikšmių apdorojimas nebuvo pritaikytas. Tačiau, kai kurie duomenų balansavimo ir mašininio mokymosi metodai reikalavo, kad trūkstamos reikšmės neegzistuotų duomenų rinkinyje, tokiais atvejai Mean-Mode imputacijos metodas buvo pritaikytas.

Papildomai, imputacijos metodai buvo apjungiami kartu su šalinimo metodais, kadangi eilučių ir kintamųjų šalinimo metodai nepašalindavo visų trūkstamų reikšmių. Norint užtikrinti, kad po

šalinimo metodų pritaikymo duomenų rinkinyje neliktų trūkstamų reikšmių buvo taip pat pritaikyti duomenų imputacijos metodai.

2.7.1.1.2. Išskirčių apdorojimas

Testavimo metu išskirčių apdorojimo metodai pasinaudojantys įrašų šalinimo strategijas, nepavyko pasiekti optimalaus sprendimo. Kiekvienu atveju modelis pasirodė daug prasčiau nei bazinis modelis arba pašalino didelį skaičių įrašų viršijantį 20% visos duomenų imties. Tačiau, rezultatai skyrėsi kuomet buvo taikomas Z-Score būdas, kuomet buvo atliekama išskirčių imputacija rezultatai buvo prastesni nei bazinis modelis, bet kai buvo taikomas išskirčių šalinimui buvo pasiekti rezultatai geresni nei bazinis modelis. Dėl šių rezultatų buvo taikomi tik išskirčių imputacijos metodai, bei Z-Score išskirčių šalinimo metodai, atliekant išskirčių apdorojimą.

- IQR imputacija – Reikšmės, kurios pateko už tarpkvartilinio (IQR) diapazono buvo imputuotos, su atitinkamomis minimumo ir maksimumo IQR ribomis. Numatytasis koeficientas – 1.5.
- Standartinio nuokrypio (SD) imputacija – Reikšmės, kurios pateko už standartinio nuokrypio (SD) diapazono buvo imputuotos, su atitinkamomis minimumo ir maksimumo SD ribomis. Numatytasis koeficientas – 1.5.
- Medianos absoliutaus nuokrypio (MAD) imputacija - Reikšmės, kurios pateko už medianos absoliutaus nuokrypio (MAD) diapazono buvo imputuotos, su atitinkamomis minimumo ir maksimumo MAD ribomis. Numatytasis koeficientas – 1.5.
- Z-Score šalinimas – Išskirtys buvo pašalintos remiantis įrašo Z-Score reikšme. Visi duomenų rinkinio įrašai buvo pašalinti, kurie viršijo koeficientą. Numatytasis koeficientas – 3.
- Z-Score iteracinis šalinimas – Modifikuotas *SD Clever* metodas [30], kuris naudojasi Z-Score, kad identifikuotų reikšmes, kurias reikia pašalinti. Šiame metode išskirtys buvo šalinamos iteraciniu būdu: kiekvieną iteraciją buvo pašalintas vienas įrašas, kuris pasiekė aukščiausią Z-Score ir viršijo nurodytą koeficientą. Po kiekvienos iteracijos Z-Score buvo perskaičiuojami ir šis procesas buvo kartojamas tol, kol duomenų rinkinyje neliko reikšmių, kurios viršija koeficientą. Numatytasis koeficientas – 3.
- Z-Score pakartotinis šalinimas (*fitter*) – Supaprastinta iteracinio Z-Score šalinimo versija, kuomet kiekvienoje iteracijoje buvo pašalinamos visos išskirtys viršijančios koeficientą, o ne tik viena pati ekstremaliausia išskirtis. Procesas kartojamas tol, kol duomenų rinkinyje nebelieka trūkstamų išskirčių. Numatytasis koeficientas – 3.
- Netaikyta – Išskirčių šalinimas netaikomas ir duomenys rinkinys liko nepakeistas.

Taip pat, buvo taikomas koeficientų optimizavimas, kuomet jie buvo didinami 0.5 žingsniu, kol nebuvo pastebėtas modelio pagerėjimas. Išskirčių šalinimo atveju, pašalintų išskirčių dalis negalėjo viršyti 20% viso duomenų rinkinio. Kiekvienas koeficientas buvo pasirinktas, kuomet konkretus koeficientas pasiekė aukščiausią tikslumą ir modelis buvo geresnis, bazinio modelio atžvilgiu.

2.7.1.1.3. Duomenų standartizacija

Duomenų standartizacija yra itin svarbus žingsnis kuriant bankroto prognozių modelį, ypač kuomet naudojami metodai kaip K artimiausių kaimynų (KNN), kuris yra jautrus skirtingiems įvesties kintamųjų masteliui. Be standartizacijos, kintamieji su didesniu masteliu gali neproporcingai paveikti

modelio rezultatus, kadangi KNN klasifikavimui naudojasi Euklidinį atstumą. Ši problema buvo sprendžiama pasitelkiant šiuos metodus:

- Standartinė (Z-Score) standartizacija.
- Atsparioji standartizacija.
- Maksimalaus absoliutaus dydžio standartizacija.
- Min-Max standartizacija.
- L2 normalizacija.

2.7.1.1.4. Duomenų padalinimas

Duomenys buvo padalinti į apmokymo ir testavimo duomenų imtis, siekiant padalinti duomenis skirtus modelio apmokymui ir modelio testavimui. Buvo naudojamas 80/20 santykis, kuomet 80% duomenų imties buvo skirta modelio apmokymui ir 20% duomenų imties buvo naudojami modelio testavimui. Toks santykis tarp apmokymo ir testavimo duomenų imčių yra dažnai naudojamas sprendžiant prognozavimo užduotis. Stratifikuotas ėmimas buvo pritaikytas, siekiant išlaikyti mažumos ir daugumos klasių proporcijas, tiek apmokymo, tiek testavimo duomenų imtyse.

2.7.1.1.5. Duomenų balansavimas

Duomenų rinkinio balanso problema buvo sprendžiant pasitelkiant duomenų ėmimo metodus, kaip perdėtas (*angl.* oversampling) ir sumažintas (*angl.* undersampling) ėmimas Bankroto prognozavime, turi ekstremalų klasių disbalansą ir sumažintas ėmimas gali lemti didelį kiekį prarastos reikšmingos informacijos iš daugumos klasės. Esant tokiam dideliam disbalansui šis sprendimo būdas yra nėra laikomas naudingu [34]. Dėl šios priežasties, šiame tyrime buvo taikomi tik perdėto ėmimo duomenų balansavimo metodai. Šie metodai buvo taikomi sprendžiant balanso problemą:

- Atsitiktinis perdėtas ėmimas (*angl.* Random oversampling) – Atsitiktinai kopijuoja įrašus iš mažumos klasės, siekiant subalansuoti klasių pasiskirstymą.
- SMOTE, sintetinis mažumos klasės perdėtasis ėmimas – generuoja sintetinius įrašus mažumos klasei, pasitelkiant interpoliaciją tarp jau egzistuojančių įrašų ir artimiausių kaimynų.
- SVM SMOTE, atramos vektorių mašinos pagrindu veikiantis SMOTE – generuoja sintetinius mažumos klasės įrašus remiantis SVM identifikuotais atraminiais vektoriais, kuris sugeneruoja sintetinius įrašus artimus sprendimų ribai, kurie padeda geriau apibrėžti sprendimų ribą tarp mažumos ir daugumos klasių.
- Ribinis SMOTE – SMOTE metodo variacija, kuomet sintetiniai mažumos klasės duomenys yra generuojami artimi klasių ribai, ypač tuos įrašus, kuriuos yra rizika klaidingai suklasifikuoti.
- ADASYN (adaptyvusis sintetinis ėmimas) – sintetiniai duomenys yra generuojami mažumos srityse, kuriose duomenys yra labai išsisklaidę, kuomet didžiausias dėmesys yra skiriamas sudėtingesniems, sunkiau išmokstamiems duomenims, kurie yra sunkiau klasifikuojami.
- KNN SMOTE (K artimiausių kaimynų SMOTE) – mažumos klasei generuojami sintetiniai įrašai, taikant interpoliacija tarp mažumos klasės įrašų ir jų k-artimiausių kaimynų, kurie yra identifikuojami pasitelkiant KNN metodą.

Duomenų balansavimo metodai buvo pritaikomi tik mažumos klasės modelio apmokymo duomenims, siekiant užtikrinti vienodą klasių pasiskirstymą modelio apmokymo procese.

2.7.1.2. Reikšmingų kintamųjų atrinkimas

Reikšmingų kintamųjų atranka buvo vykdoma trimis etapais: statistiškai reikšmingų kintamųjų nustatymas taikant hipotezių tikrinimą, nustatant multikolinearių kintamųjų grupes atliekant koreliacinę analizę ir kintamųjų atrankos metodų taikymas.

2.7.1.2.1. Metrikos ir statistikos nusakančios kintamųjų reikšmingumą

Pirmasis etape buvo identifikuoti statistiškai nereikšmingi kintamieji remiantis hipotezių tikrinimu. Trys statistiniai testai buvo naudojami tam atlikti: Komogorov-Smirnov testas, Mann-Whitney U testas ir nepriklausomų dviejų imčių t testas.

2.7.1.2.2. Kolmogorov-Smirnov (KS) testas

Procesas buvo pradedamas pasinaudojant Kolmogorov-Smirnov (KS) testą, kuris nustatė ar kintamasis pasiskirstęs pagal normalųjį skirstinį. Šis žingsnis buvo būtinas, norint nustatyti kuris testas tūrėtų būti naudojamas, norint nustatyti kintamųjų reikšmingumą, ar Mann-Whitney U testas (nenormalusis skirstinys), ar nepriklausomas dviejų imčių t testas (apytiksliai normalusis skirstinys). Mann-Whitney U testo hipotezė nusakoma taip:

- H0: duomenys populiacijoje pasiskirstę pagal normalųjį skirstinį.
- Ha: duomenys populiacijoje pasiskirstę pagal ne normalųjį skirstinį.

$$\begin{aligned} &\mathrm{D} = \sup_x |F_1(x) - F_2(x)| \\ &kur\ D - Kolmogorov - Smirnov\ D\ statistika, F_1(x) \\ &- duomenų\ įrašo\ empirinė\ kaupiamoji\ pasiskirstymo\ funkcija\ (ECDF), F_2(x) \\ &- lyginamojo\ skirstinio\ kaupiamasis\ pasiskirstymo\ funkcija\ (CDF), sup_x \\ &- didžiausias\ skirtumas\ visame\ x\ reikšmių\ intervale\ (t.y. maksimalus\ skirtumas\ tarp \\ &\qquad dviejų\ kaupiamų\ pasiskirstymo\ funkcijų) \end{aligned} \tag{21}$$

Jei nulinė hipotezė buvo atmesta, alternatyvi hipotezė buvo priimta, tai rodė, kad duomenys nėra normaliai pasiskirstę ir Mann-Whitney U testas buvo naudojamas, nustatant kintamųjų reikšmingumą. Jei nulinė hipotezė buvo neatmesta, kintamasis parodė statistiškai reikšmingą pasiskirstymą pagal normalųjį skirstinį ir nepriklausomas dviejų imčių t testas buvo naudojamas nustatant kintamųjų reikšmingumą.

2.7.1.2.3. Mann-Whitney U testas

Mann-Whitney U testas yra neparametrinis metodas naudojamas nustatyti ar dvi nepriklausomos imtys priklauso tam pačiam skirstiniui. Šis metodas buvo taikomas kuomet kintamojo skirstinys nėra pasiskirstęs pagal normalųjį skirstinį. Mann-Whitney U testo hipotezė nusakoma taip:

- H0: Kintamojo pasiskirstymas tarp abiejų klasių yra vienodas.
- Ha: Kintamojo pasiskirstymas tarp abiejų klasių skiriasi.

$$\begin{aligned} \mathrm{U}_2 &= n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - \mathrm{R}_1 \\ \mathrm{U}_1 &= n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - \mathrm{R}_2 \end{aligned} \tag{22}$$

$$kur\ U - Mann - Whitney\ U\ statistika, n_1 - bankroto\ klasės\ dydis, n_2 - nebankroto\ klasės\ dydis, R_1 - bankroto\ klasės\ įrašo\ rangų\ suma, R_2 - nebankroto\ klasės\ įrašo\ rangų\ suma$$

Jei nulinė hipotezė buvo atmesta, alternatyvi hipotezė buvo priimta ir kintamasis parodė statistiškai reikšmingą skirtumą tarp abiejų klasių skirstinių ir kintamasis buvo paliktas tarp reikšmingų kintamųjų. Jei nulinė hipotezė buvo priimta, kintamasis netūrėjo statistiškai reikšmingo skirtumo tarp abiejų klasių skirstinių ir kintamasis buvo pašalintas ir modelio.

2.7.1.2.4. Nepriklausomas dviejų imčių t testas.

Kintamiesiems, kurie netūrėjo apytiksliai normalaus pasiskirstymo, nepriklausomas dviejų imčių t testas buvo naudojamas įvertinti vidurkių skirtumą tarp dviejų klasių. T testo hipotezė nusakoma taip:

- H0: Kintamojo vidurkiai tarp abiejų klasių yra vienodi.
- Ha: Kintamojo vidurkiai tarp abiejų klasių skiriasi.

$$\mathrm{t} = \frac{\widehat{X}_1 - \widehat{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{23}$$

$$kur\ t - statistika\ t, \widehat{X}_{1,2} - abiejų\ klasių\ vidurkiai, n_{1,2} - abiejų\ klasių\ dydžiai, s_p - bendrasis\ standartinis\ nuokrypis$$

Jei nulinė hipotezė buvo atmesta, alternatyvi hipotezė buvo priimta ir kintamasis parodė statistiškai reikšmingą skirtumą tarp abiejų klasių vidurkių ir kintamasis buvo paliktas tarp reikšmingų kintamųjų. Jei nulinė hipotezė buvo priimta, kintamasis netūrėjo statistiškai reikšmingo skirtumo tarp abiejų klasių vidurkių ir kintamasis buvo pašalintas ir modelio.

2.7.1.2.5. Nepriklausomų kintamųjų multikolinearumas

Multikolinearumo nustatymas tarp nepriklausomų kintamųjų yra vienas iš daugelio būdų naudojamas kintamųjų atrinkimui, kuris užtikrina, kad tik kintamieji suteikiantys reikšmingos informacijos būtų atrinkami į modelį. Multikolinearūs kintamieji parodo, kad kintamieji turi perteklinę arba nusako tą pačią informacijos naudą modeliui ir pasirinkimas pasirenkant tik vieną iš šių kintamųjų, negu pasirenkant kelis tokius kintamuosius, gali pagerinti modelio paaiškinamumą ir kokybę.

Šiame tyrime, grupinės koreliacijos esančios tarp visų kintamųjų. Kintamieji demonstruojanti aukštą teigiamą (≥ 0.7) arba neigiamą (≥ -0.7) koreliacijas buvo laikomi kaip multikolinearūs ir sugrupuoti atitinkamai. Kiekviena grupė nusakė kintamųjų grupę, kuri modeliui suteikė panašią informaciją. Šios grupės buvo naudojamos kintamųjų atrinkime siekiant išvengti tarpusavyje multikolinearių kintamųjų prognozavimo modelyje.

2.7.1.2.6. Kintamųjų atrankos metodai

Egzistuoja daugybe kintamųjų atrankos metodų, tokių kaip *LASSO,* žingsninė atranka į priekį (*angl.* Forward-step selection), žingsninė atranka atgal (*angl.* Backward-step selection), ir kiti. Testavimo etape buvo identifikuota, jog *Lasso* ir žingsninė atranka atgal pasirodė daug prasčiau nei bazinis modelis ir šie metodai nebuvo taikomi kintamųjų atrankos etape. Tačiau žingsninė atranka į priekį

pasiekė geriausius rezultatus. Todėl buvo taikoma žingsninė atranka į priekį, kuri naudojosi vienu iš pasirinktų mašininio mokymosi metodų, kaip KNN, DT ir logistinė regresija. Šis metodas buvo modifikuotas, kad galėtų taip pat ir šalinti kintamuosius iš jau atrinktų kintamųjų sąrašo, bet tik tuo atveju jei modelio kokybė yra pagerinama. Kintamųjų atrankos procesas buvo nusakytas taip:

- Kintamųjų atrinkimo algoritmas kiekvieną iteraciją atrenka kintamąjį, kuris pasiekia aukščiausią balansuotą klasifikavimo tikslumą, kuomet kintamasis yra pridedamas į modelį.
- Kuomet, modelis turi jau du atrinktus kintamuosius, kiekvienos iteracijos pabaigoje, kuomet jau yra atrinktas naujas kintamasis, modelis buvo pervertinamas laikinai pašalinant po vieną anksčiau pridėta kintamąjį. Jei konkretų kintamąjį pašalinus, modelio subalansuotas klasifikavimo tikslumas pagerėjo, tas kintamasis iš modelio buvo pašalintas. Toliau, kintamasis buvo pašalintas tik tas, kuris tūrėjo reikšmingiausią modelio pagerėjimą.
- Šis procesas buvo kartojamas tol kol pridedamas arba pašalinamas kintamasis nebepagerina modelio subalansuoto tikslumo.
- Kintamųjų atrankos apribojimai:
    - Kintamieji, kurie buvo identifikuoti kaip statistiškai nereikšmingais, nebuvo naudojami kintamųjų atrankos procese.
    - Kintamieji, kurie tūrėjo aukštą koreliaciją su bent vienu kintamuoju, kuris jau buvo atrinktas į modelį, nebuvo naudojami kintamųjų atrankos procese.

Kintamųjų atranka buvo taikoma tik galutiniame modelių testavimo etape, kuomet geriausiai pasirodę modeliai jau buvo atrinkti, kadangi kintamųjų atrankos procesas reikalavo daug skaičiavimo sąnaudų.

2.7.1.3. Mašininio mokymosi metodai

Šiame tyrime buvo atrinkti trys mašininio mokymosi metodai: sprendimų medis, logistinė regresija ir K artimiausių kaimynų metodas. Ankstesniuose tyrimuose sprendimų medis nuolat demonstravo aukštą klasifikavimo tikslumą ir yra laikomi vienais iš geriausių modelių atliekančių klasifikavimo užduotis. Logistinė regresija yra plačiai naudojama bankroto prognozavimo tyrimuose ir yra laikomas vienu geriausių sprendimų sprendžiant šią problemą, dėl šio metodo prognozavimo galimybių patikimumo ir aukšto paaiškinamumo per modelio koeficientus. Abu DT ir Logit modeliai yra laikomi, kaip stipriai interpretuojami metodai, kas leidžia identifikuoti statistiškai reikšmingiausius kintamuosius ir paaiškinti jų įtaką bankroto prognozėms. Priešinai, KNN metodas buvo rečiau naudojamas metodas ir dažniausiai nepasiekdavo aukšto klasifikavimo tikslumo, lyginant su kitais bankroto prognozavimo modeliais. Tačiau, kai kuriais atvejais, KNN metodas parodė geresnes prognozavimo galimybes. Nors ir KNN yra mažiau interpretuojamas modelis, nei DR ir LR, bet jis suteikia kitokį modelio interpretavimo būdą, kuomet jis paaiškinamas, identifikuojant kitus įrašus, kurie yra patys panašiausi prognozuojamam įrašui. Toks rezultatų interpretavimas gali būti itin naudingas tam tikrame sprendimų priėmimo kontekste. Taip pat, KNN yra ne parametrinis, duomenų įrašais grįstas mokymosi būdas, kuris suteikia išsamesnį palyginimą su kitais labiau struktūrizuotais DT ir LR metodais. Šie trys metodai leidžia atlikti išsamų palyginimą, apimantį tiek modelių kokybę ir jų paaiškinamumą.

Šiame tyrime, dirbtinių neuroninių tinklų metodai nebuvo naudojami, kadangi jie nėra interpretuojami modeliai. Šios metodų charakteristikos yra pagrindinis trūkumas (Figini ir kt., 2017; Horak ir kt., 2020, kaip cituojama [10]), ypač kai sprendžiama bankroto prognozių problema. Nors,

praėjusiuose tyrimuose, šie metodai parodė geresnes prognozavimo galimybes (Becerra-Vicario ir kt., 2020, kaip cituojama [10]), tačiau šie metodai yra komplikuoti ir reikalauja sudėtingos neuroninių tinklų konfigūracijos, hiperparametrų derininimo ir reikalauja daugiau laiko apmokant šiuos modelius. Šie ANN metodų trūkumai buvo pagrindinė priežastis, kodėl šiame tyrime tokie metodai nebuvo naudojami, ypač kai kintamųjų reikšmingumo paaiškinimas yra laikomas ne mažiau svarbiu nei aukšto klasifikavimo tikslumo pasiekimas (Lundberg ir Lee, 2017, kaip cituojama [10]).

2.7.1.4. Galutinė tyrimo metodologija

Šis tyrimas apėmė įvairių metodų įvertinimą ir palyginimą skirtinguose bankroto prognozių modelio sudarymo etapuose. Įvairiose kombinacijose metodiškai testuojami ir lyginami trūkstamų reikšmių, išskirčių, duomenų standartizacijos, balansavimo, kintamųjų atrankos ir mašininio mokymosi metodai. Rezultatai buvo išanalizuoti, norint identifikuoti individualius metodus pasiekiančius geriausius rezultatus, kiekvienoje duomenų apdorojimo ir modeliavimo etape. Galiausiai, buvo nustatytos pačios optimaliausios metodų kombinacijos.

3. Bankroto prognozavimo modelių rezultatų analizė

3.1. Metodų rezultatų palyginimas

3.1.1. Trūkstamų reikšmių apdorojimas

Trūkstamų reikšmių apdorojimas buvo pirmas žingsnis duomenų apdorojimo procesų sekoje ir rezultatai skyrėsi priklausomai nuo taikytino metodo (žr. **14 lentelė**).

14 lentelė. Trūkstamų reikšmių apdorojimo metodų rezultatų vidurkiai

ML Metodas	Išskirčių Apdorojimo Metodas	Tikslumas	Specifiškumas	Jautrumas	F1	AUC	TP	TN	FP	FN	Laikas
DT	NA	**0.9505**	**0.9623**	0.9323	0.9372	0.9473	210	**332**	**13**	15	8.993
DT	RC	0.9502	0.9593	0.9361	**0.9372**	**0.9477**	211	331	14	14	8.881
DT	RCR	0.9486	0.9525	**0.9424**	0.9348	0.9474	209	327	16	**12**	8.533
DT	RR	0.9419	0.9496	0.9301	0.9260	0.9398	206	326	17	15	8.644
DT	RRC	0.9389	0.9516	0.9192	0.9216	0.9354	203	326	16	17	8.913
DT	RR + MeanMode	0.9350	0.9372	0.9316	0.9315	0.9344	294	312	21	20	10.551
DT	RC + MeanMode	0.9326	0.9350	0.9285	0.9293	0.9318	296	313	21	21	10.193
DT	RRC + MeanMode	0.9320	0.9328	0.9302	0.9294	0.9315	296	313	22	21	10.432
DT	RR + LR	0.9296	0.9372	0.9205	0.9261	0.9289	295	317	21	24	17.862
DT	LR	0.9295	0.9282	0.9298	0.9266	0.9290	297	311	24	21	16.595
DT	RC + LR	0.9284	0.9285	0.9273	0.9252	0.9279	297	312	23	22	15.677
DT	RRC + LR	0.9282	0.9321	0.9240	0.9254	0.9280	292	309	22	23	16.858
DT	RR + KNN	0.9280	0.9301	0.9247	0.9248	0.9274	287	303	23	22	12.220
DT	RCR + MeanMode	0.9275	0.9253	0.9290	0.9247	0.9271	295	309	25	22	10.765
DT	KNN	0.9273	0.9279	0.9253	0.9234	0.9266	292	308	24	22	10.564
DT	MeanMode	0.9271	0.9273	0.9249	0.9224	0.9261	289	308	23	22	10.081
DT	RRC + KNN	0.9264	0.9241	0.9282	0.9239	0.9261	291	305	24	22	11.468
DT	RRC + DT	0.9253	0.9260	0.9237	0.9227	0.9249	292	307	24	23	25.634
DT	RR + DT	0.9243	0.9274	0.9203	0.9210	0.9239	287	304	23	24	27.228
DT	RC + KNN	0.9243	0.9229	0.9247	0.9210	0.9238	295	309	25	22	11.795
DT	RCR + KNN	0.9236	0.9208	0.9253	0.9208	0.9230	291	304	26	22	12.107
DT	RCR + LR	0.9228	0.9195	0.9250	0.9196	0.9222	**297**	311	27	23	15.701
DT	RC + DT	0.9217	0.9228	0.9197	0.9183	0.9213	292	307	25	24	27.761
DT	RCR + DT	0.9211	0.9175	0.9241	0.9181	0.9208	293	307	27	23	26.480

ML Metodas	Išskirčių Apdorojimo Metodas	Tikslumas	Specifišku mas	Jautrumas	F1	AUC	TP	TN	FP	FN	Laikas
DT	DT	0.9167	0.9127	0.9202	0.9130	0.9164	288	300	28	24	53.432
LR	MeanMode	0.8204	0.8085	0.8256	0.8043	0.8171	262	268	63	51	**3.515**
KNN	RC + LR	0.8201	0.7762	0.8599	0.8187	0.8181	277	260	75	41	6.976
KNN	KNN	0.8195	0.7741	0.8592	0.8167	0.8166	273	256	75	41	3.681
LR	RC + MeanMode	0.8194	0.8066	0.8268	0.8056	0.8167	265	270	64	52	3.533
KNN	RR + MeanMode	0.8192	0.7598	0.8756	0.8200	0.8177	278	252	80	36	3.511
LR	KNN	0.8192	0.8077	0.8245	0.8033	0.8161	262	267	64	52	3.667
KNN	**RC + MM**	0.8186	0.7753	0.8575	0.8167	0.8164	276	259	75	42	3.423
KNN	RR + KNN	0.8186	0.7623	0.8706	0.8187	0.8165	273	247	78	37	2.940
LR	RR + MeanMode	0.8185	0.7926	0.8395	0.8052	0.8160	267	263	69	47	3.368
KNN	LR	0.8176	0.7714	0.8601	0.8170	0.8157	277	258	77	41	9.621
KNN	RC + KNN	0.8163	0.7674	0.8611	0.8156	0.8142	276	257	78	41	2.819
KNN	RRC + LR	0.8162	0.7623	0.8671	0.8171	0.8147	276	252	79	39	7.520
KNN	RRC + MeanMode	0.8154	0.7583	0.8700	0.8165	0.8141	279	254	81	38	3.595
LR	RC + KNN	0.8152	0.7995	0.8257	0.8012	0.8126	265	268	67	53	3.216
LR	RCR + KNN	0.8144	0.7935	0.8311	0.8028	0.8123	264	262	68	50	3.036
KNN	RR + DT	0.8141	0.7552	0.8693	0.8146	0.8123	274	248	81	37	130.42
LR	RCR + DT	0.8139	0.7960	0.8269	0.7992	0.8114	265	267	67	52	18.672
KNN	RR + LR	0.8130	0.7602	0.8620	0.8127	0.8111	279	257	81	41	7.429
KNN	RRC + KNN	0.8129	0.7549	0.8679	0.8145	0.8114	275	248	81	38	2.632
LR	RR + KNN	0.8125	0.7866	0.8337	0.8005	0.8102	260	256	70	49	3.070
LR	RRC + MeanMode	0.8124	0.7889	0.8320	0.8002	0.8105	267	264	71	50	3.599
LR	RR + DT	0.8123	0.7932	0.8259	0.7950	0.8096	260	259	68	51	20.379
LR	RCR + MeanMode	0.8120	0.7894	0.8307	0.8011	0.8100	265	263	70	51	3.583
KNN	DT	0.8113	0.7663	0.8508	0.8083	0.8085	270	254	77	43	129.02
LR	RRC + KNN	0.8106	0.7868	0.8299	0.7986	0.8084	262	259	70	50	3.132
KNN	MV - R C+R + KNN	0.8092	0.7569	0.8580	0.8099	0.8074	272	249	81	41	2.693
KNN	**RC + DT**	0.8080	0.7606	0.8513	0.8069	0.8060	275	256	81	44	32
LR	DT	0.8073	0.7943	0.8152	0.7910	0.8048	257	261	67	55	61.037

ML Metodas	Išskirčių Apdorojimo Metodas	Tikslumas	Specifiškumas	Jautrumas	F1	AUC	TP	TN	FP	FN	Laikas
KNN	RCR + MeanMode	0.8073	0.7568	0.8538	0.8070	0.8053	273	252	81	43	3.605
LR	RRC + DT	0.8071	0.7837	0.8260	0.7913	0.8048	263	260	71	52	18.759
LR	RC + DT	0.8065	0.7901	0.8184	0.7917	0.8043	261	263	70	55	20.014
LR	LS	0.8065	0.7922	0.8155	0.7919	0.8039	262	265	70	56	8.778
KNN	RCR + LR	0.8029	0.7531	0.8480	0.8018	0.8005	275	255	83	45	6.694
LR	RC + LR	0.8014	0.7868	0.8109	0.7868	0.7988	261	264	72	58	11.577
LR	RRC + LR	0.8013	0.7791	0.8187	0.7889	0.7989	261	258	73	54	6.806
LR	RR + LR	0.7950	0.7718	0.8124	0.7829	0.7921	263	260	77	57	7.533
LR	RCR + LR	0.7924	0.7682	0.8111	0.7808	0.7897	263	260	78	57	6.345
KNN	MeanMode	0.7829	0.7511	0.8083	0.7774	0.7797	257	248	82	57	5.309

Pirmiausia buvo identifikuota, kad visi sprendimai, kurie taikė tik reikšmių šalinimo metodus buvo atmesti, kaip netinkami, kadangi šie metodai parodė, didelį kiekį prarastų duomenų, viršijantį 20% prarastų įrašų. Sprendimų medžių klasifikatorius, pasirodė geriausiai, kuomet bent kokia trūkstamų reikšmių šalinimo strategija buvo taikoma kartu su Mean-Mode imputacija. Tik RCR + MeanMode ir MeanMode parodė prasčiausią kokybę, lyginant su kitais metodais. Taip pat, reikšmių šalinimo metodai buvo taikomi su LR imputacija ir buvo sekantys pagal gerumą modeliai, išskyrus kuomet buvo taikomas kartu su RCT šalinimo metodu, kuris pasiekė vieną prasčiausių vidutinišką tikslumą. Abu KNN ir DT imputacijos metodai, naudojant kartu su šalinimo metodais pasiekė prasčiausius rezultatus. Šiems metodams buvo vienintelės išimtys, kuomet buvo naudojami KNN ir RR + KNN strategijos, kurie pasiekė šiek tiek geresnius rezultatus, nei keli prasčiausi MeanMode metodai. Nors skirtumas tarp šių modelių kokybės yra pastebimas, tačiau skirtumas tarp prasčiausių ir geriausių pasiektų vidutinių tikslumų apytiksliai siekė 1.3%, kas nėra labai reikšmingas skirtumas. DT pasiekė geriausius rezultatus kai buvo naudojamas RR + MeanMode trūkstamų reikšmių apdorojimo metodas.

Logistinė regresija pasiekė geriausius rezultatus, kuomet buvo taikoma tik MeanMode imputacija. Metodai, kurie taikė reikšmių šalinimą kartu su KNN, DT ir Mean-Mode imputacijos metodais, modelio kokybė buvo labiau išsibarstę ir nebuvo identifikuota, kuris metodas pasirodė geriausiai. Tačiau LR pademonstravo prasčiausią kokybę tarp visų modelių. Skirtumas tarp prasčiausiai ir geriausiai pasirodžiusių metodų buvo apytiksliai 2.8%, kuris buvo šiek tiek geresnis nei DT imputacija, kas parodo, kad trūkstamų reikšmių apdorojimo metodai turi daug didesnę įtaką, kuomet jie buvo taikomi kartu su logistine regresija.

KNN klasifikacijos rezultatai žymiai skyrėsi nuo prieš tai nagrinėtų mašininio mokymosi metodų. Čia modelių kokybė buvo stipriau išsibarsčiusi ir nei vienas imputacijos metodas neparodė statistiškai reikšmingą pranašumą prieš kitus. Geriausiai pasirodęs metodas buvo RC + LR, kuomet kiti du mašininio mokymosi metodai pasirodė prasčiau su logistinės regresijos imputacija. Toliau skirtumas tarp prasčiausiai ir geriausiai pasirodžiusių metodų buvo apytiksliai 3.7%, kas parodė, kad trūkstamų reikšmių apdorojimo metodai turi statistiškai reikšmingesnę įtaką nei abu DT ir LR metodai.

3.1.2. Išskirčių apdorojimas

Išskirčių apdorojimas buvo antras žingsnis duomenų apdorojimo procesų sekoje ir rezultatai skyrėsi priklausomai nuo taikytino metodo (žr. **15 lentelė**).

15 lentelė. Išskirčių apdorojimo metodų rezultatų vidurkiai

ML Metodas	Išskirčių Apdorojimo Metodas	Tikslumas	Specifiškumas	Jautrumas	F1	AUC	TP	TN	FP	FN	Laikas
DT	IQR	**0.9311**	0.9294	**0.9320**	**0.9284**	**0.9307**	**309**	326	24	22	17.776
DT	SD	0.9308	**0.9307**	0.9300	0.9279	0.9304	308	**326**	24	22	18.026
DT	NA	0.9306	0.9306	0.9297	0.9278	0.9302	308	326	24	22	18.387
DT	MAD	0.9290	0.9275	0.9299	0.9262	0.9287	308	325	25	22	15.944
DT	ZSCORE FITTER	0.9269	0.9253	0.9273	0.9232	0.9263	303	323	26	23	18.287
DT	ZSCORE ITTER	0.9259	0.9238	0.9270	0.9223	0.9254	304	324	26	23	18.751
DT	ZSCORE	0.9156	0.9257	0.9027	0.9092	0.9142	193	212	**17**	**20**	14.668
LR	MAD	0.8617	0.8820	0.8367	0.8354	0.8594	281	309	41	53	**9.155**
LR	IQR	0.8550	0.8424	0.8639	0.8467	0.8531	290	295	55	44	9.410
KNN	MAD	0.8367	0.7617	0.9119	0.8426	0.8368	306	267	83	28	20.713
KNN	IQR	0.8282	0.7616	0.8940	0.8326	0.8278	300	267	83	33	20.470
KNN	SD	0.8117	0.7558	0.8648	0.8138	0.8103	291	265	85	42	20.389
KNN	ZSCORE	0.8083	0.7807	0.8268	0.7992	0.8037	179	177	49	34	19.710
KNN	ZSCORE FITTER	0.8067	0.7689	0.8383	0.8021	0.8036	281	270	81	50	19.338
KNN	ZSCORE ITTER	0.8053	0.7672	0.8371	0.8007	0.8022	281	270	82	51	20.285
LR	ZSCORE	0.8036	0.8033	0.7957	0.7838	0.7995	172	182	44	42	9.201
LR	SD	0.8011	0.7317	0.8641	0.7961	0.7979	292	256	94	41	11.363
LR	ZSCORE FITTER	0.8009	0.7859	0.8112	0.7868	0.7985	269	274	74	60	9.948
LR	ZSCORE ITTER	0.7982	0.7829	0.8091	0.7837	0.7960	269	274	76	61	10.599
KNN	NA	0.7901	0.7406	0.8355	0.7906	0.7881	281	260	91	52	20.689
LR	NA	0.7490	0.7071	0.7877	0.7404	0.7474	265	248	102	69	15.091

Pirmiausia buvo identifikuota, kad visi sprendimai, kurie taikė tik Z-Score metodą buvo atmesti, kaip netinkami, kadangi šie metodai parodė, didelį kiekį prarastų duomenų, viršijantį 20% prarastų įrašų. Sprendimų medžių klasifikatorius, pasirodė geriausiai, kuomet IQR strategija buvo taikoma apdoroti išskirtis. Kiti metodai, kaip SD, MAD ir išskirčių neapdorojimas, demonstravo labai panašius rezultatus su IQR, kuomet vidutinio tikslumo skirtumas buvo intervale nuo 0.03% iki 0.21%. Kiti Z-

Score metodai, kurie yra išskirčių šalinimo metodai, parodė prasčiausius rezultatus, kuomet vidutinis tikslumas buvo mažesnis intervale nuo 0.4% iki 1.6%, kol standartinis Z-Score metodas turėjo prastesnį vidutinį tikslumą apytiksliai per 1.5%. Tačiau, skirtumas tarp prasčiausio ir geriausio metodų buvo apytiksliai 1.6%, kas negali būti laikoma kaip statistiškai reikšmingas skirtumas. SD ir IQR teigiamai paveikė modelio tikslumą, kadangi šie metodai pagerino vidutinį tikslumą per 0.02%, lyginant su metodu netaikančio išskirčių apdorojimo.

Išskirčių apdorojimas turėjo didžiausią įtaką logistinės regresijos metodui, kuomet bet kuris taikytas metodas pagerino modelio vidutini klasifikavimo tikslumą apytiksliai per 5% iki 12%. Toliau MAD ir IQR metodai, turėjo reikšmingiausią įtaką, kuomet vidutinis tikslumas apytiksliai pagerėjo per 11% ir 12% atitinkamai. Visi Z-Score metodai, kartu su SD pasiekė panašius rezultatus ir šie metodai turėjo taip pat teigiamai reikšmingą įtaką modelio tikslumui, kuris apytiksliai siekė 6%. MAD buvo identifikuotas, kaip pats geriausias metodas, kuris pasiekė apytiksliai 11% tikslumo pagerėjimą, lyginant kuomet nėra taikomas išskirčių apdorojimas.

KNN panašius rezultatus, kaip logistinė regresija, kuomet visi išskirčių apdorojimo metodai turėjo teigiamą įtaką modeliams, nors poveikis buvo kiek silpnesnis. Šie metodai pagerino vidutinį tikslumą apytiksliai intervale nuo 1% iki 4.5%. MAD ir IQR pademonstravo reikšmingiausią poveikį, su 3.8% ir 4.7% pagerėjimais, atitinkamai. SD turėjo šiek tiek mažesnį poveikį, su apytiksliu 2.2% pagerėjimu, nors ir Z-Score strategijos turėjo labai panašius rezultatus, su pagerėjimo siekiančių apytiksliai 1.6%. MAD buvo identifikuotas kaip geriausiai pasirodęs išskirčių apdorojimo metodas, kuris siekė apytiksliai 4.7% pagerėjimas lyginant su baziniu modeliu. Svarbu paminėti visų metodų reliatyvus efektyvumas buvo panašus su tais, kurie buvo pastebėti su Logit modeliu. Abu klasifikatoriai turėjo tuos pačius metodus, kurie pasirodė geriausiai (IQR ir MAD), kol SD ir Z-Score metodai pasirodė prasčiau, tačiau visos išskirčių apdorojimo strategijos pagerino rezultatus lyginant su baziniu modeliu.

Apibendrinant, DT, kuomet buvo taikomas išskirčių apdorojimo metodas neparodė statistiškai reikšmingo poveikio, o tai rodo, kad šis klasifikatorius yra labiau atsparus išskirtims. Nedideli vidutinio tikslumo skirtumai gali būti paaiškinami triukšmu. Priešingai, logistinė regresija buvo labiausiai jautri išskirtims, kol KNN metodas buvo šiek tiek mažiau paveiktas. Skirtingų išskirčių metodų reliatyvus efektyvus buvo panašus tarp KNN ir LR, nes abu metodai pademonstravo prastesni veikimą su tais pačiais išskirčių apdorojimo metodais, nors poveikio mastas skyrėsi.

3.1.3. Duomenų standartizacija

Duomenų standartizacija buvo trečias žingsnis duomenų apdorojimo procesų sekoje ir rezultatai skyrėsi priklausomai nuo taikytino metodo (žr. **16 lentelė**).

DT pasirodė geriausiai kai buvo naudotas Z-Score standartizavimo metodas. Tačiau, šio metodo ir kitų metodų poveikis nebuvo reikšmingas, kadangi jis turėjo vidutinio tikslumo pagerėjimą siekiantį 0.12% lyginant su baziniu modeliu, kas taip pat pademonstravo vidutinio tikslumo pagerėjimo intervalą, kurį apėmė visų metodų pagerėjimą lyginant su baziniu modeliu. Viena išimtis buvo L2 metodas, kuris turėjo vidutinį tikslumo sumažėjimą siekiantį 15%.

Logistinė regresija parodė visiškai kitokią situaciją, kuomet geriausias metodas pademonstravo vidutinio tikslumo pagerėjimą apytiksliai siekiantį 25%, lyginant su baziniu modeliu, kuomet buvo taikoma atsparioji standartizacija. Šis modelio pagerėjimas pademonstravo, jog logistinei regresijai

duomenų standartizacija, turi didžiulę svarbą ir duomenų standartizavimo metodas privalo būti tinkamai parinktas. Kiti metodai taip pat parodė reikšmingą poveikį, nors poveikio mastas buvo mažesnis, tačiau jie leido pasiekti vidutinio tikslumo pagerėjimą nuo apytiksliai 10% iki 14%. Išskyrus, kuomet buvo taikomas L2 standartizavimo metodas, kuris neigiamai paveikė modelį ir sumažino tikslumą per 5%, lyginant su baziniu modeliu.

16 lentelė. Duomenų standartizacijos metodų rezultatų vidurkiai

ML Metodas	Išskirčių Apdorojimo Metodas	Tikslumas	Specifišku mas	Jautrumas	F1	AUC	TP	TN	FP	FN	Laikas
DT	Zscore	**0.9412**	**0.9409**	0.9406	**0.9385**	**0.9407**	295	314	19	18	17.354
DT	Robust	0.9408	0.9402	0.9404	0.9380	0.9403	295	313	19	18	17.579
DT	MaxAbs	0.9406	0.9391	0.9412	0.9379	0.9402	295	313	20	17	17.242
DT	MinMax	0.9404	0.9393	0.9407	0.9377	0.9400	295	313	**20**	17	17.204
DT	NA	0.9400	0.9400	0.9390	0.9372	0.9395	294	**313**	19	**18**	17.565
LR	Robust	0.9316	0.8917	**0.9715**	0.9342	0.9316	**341**	313	38	10	13.933
DT	L2	0.8579	0.8641	0.8491	0.8503	0.8566	269	287	45	46	17.491
KNN	Robust	0.8325	0.7884	0.8734	0.8311	0.8309	278	262	70	37	20.151
KNN	Zscore	0.8178	0.7666	0.8646	0.8174	0.8156	276	255	78	39	20.264
LR	Zscore	0.8134	0.7493	0.8775	0.8246	0.8134	308	263	88	43	3.587
KNN	NA	0.8083	0.7539	0.8591	0.8087	0.8065	274	251	82	41	20.298
KNN	MaxAbs	0.8073	0.7543	0.8554	0.8069	0.8048	274	251	82	42	20.273
KNN	L2	0.8058	0.7603	0.8477	0.8037	0.8040	271	253	80	45	20.145
KNN	MinMax	0.8027	0.7507	0.8500	0.8023	0.8003	272	250	83	44	20.238
LR	MaxAbs	0.7821	0.7236	0.8405	0.7941	0.7821	295	254	97	56	2.403
LR	MinMax	0.7735	0.7179	0.8291	0.7854	0.7735	291	252	99	60	**2.362**
LR	NA	0.6738	0.5812	0.7664	0.7014	0.6738	269	204	147	82	3.155
LR	L2	0.6282	0.7436	0.5128	0.5797	0.6282	180	261	90	171	2.268

Duomenų standartizacija parodė mažiau reikšmingą poveikį KNN metodui, kuomet atsparusis ir Z-Score metodai pagerino modelio vidutinį tikslumą apytiksliai per 3.3% ir 1.8% atitinkamai. Tačiau kiti metodai turėjo neigiamą poveikį, kurio prarastas vidutinis tikslumas buvo intervale nuo 0.1% iki 0.6%. Tai, parodė, kad KNN yra labiau atsparus nuo parinkto duomenų standartizavimo metodo. KNN parodė reikšmingesnį modelio kokybės pagerėjimą, kuomet buvo naudojamas atsparusis duomenų standartizavimo metodas.

DT ir Logit metodai buvo neigiamai paveikti kai duomenų standartizavimo metodas buvo netinkamai parenkamas. Naudojant duomenų standartizavimo metodus, logistinė regresija buvo labiausiai paveikta, lyginant kaip veikė su kitais mašininio mokymosi metodais, kadangi buvo pastebėtas

pagerėjimas siekiantis 25%. KNN ir DT buvo atspariausi metodai, kadangi jie buvo mažiausiai paveikti, priklausant nuo parinkto duomenų standartizavimo metodo.

3.1.4. Duomenų balansavimas

Duomenų balansavimas buvo ketvirtas žingsnis duomenų apdorojimo procesų sekoje ir rezultatai skyrėsi priklausomai nuo taikytino metodo (žr. **17 lentelė**).

DT pasirodė geriausiai kuomet buvo taikomas RO arba SVM SMOTE duomenų balansavimo metodai, kurie pagerino vidutinį tikslumą apytiksliai per 1.2% ir 0.1% atitinkamai. Tačiau kiti balansavimo metodų poveikis buvo neigiamas kuomet vidutinis tikslumo sumažėjimas siekė 0.03% ir 1%. Šie rezultatai parodė reikšmingą poveikį modelio tikslumui. K vidurkių metodas turėjo didžiausią neigiamą poveikį, kuomet vidutinis tikslumas sumažėjo apytiksliai per 1%.

Logistinė regresija pademonstravo panašius rezultatus, kuomet duomenų balansavimas rezultatus paveikė šiek tiek mačiau. Visi metodai išskyrus K vidurkių pagerino vidutinį tikslumą intervale apytiksliai nuo 1.3% iki 1.8%, kuomet K vidurkių metodas sumažino vidutinį vidurkį apytiksliai per 0.5%. Geriausiai pasirodęs duomenų balansavimo metodas buvo ADASYN, kuris pasiekė tikslumo pagerėjimą siekiantį 1.8%.

17 lentelė. Duomenų balansavimo metodų rezultatų vidurkiai

ML Metodas	Išskirčių Apdorojimo Metodas	Tikslumas	Specifišku mas	Jautrumas	F1	AUC	TP	TN	FP	FN	Laikas
DT	RO	**0.9396**	0.9349	0.9443	**0.9399**	**0.9396**	**314**	311	21	18	17.336
DT	SVM Smote	0.9286	0.9212	0.9361	0.9291	0.9286	312	307	**26**	20	18.906
DT	NA	0.9273	**0.9385**	0.9094	0.9068	0.9240	196	**314**	20	19	13.862
DT	Adasyn	0.9270	0.9209	0.9332	0.9275	0.9270	311	306	26	21	18.226
DT	BorderlineSMOTE	0.9255	0.9229	0.9280	0.9257	0.9255	309	307	25	23	18.406
DT	Smote	0.9254	0.9251	0.9257	0.9254	0.9254	308	308	24	**24**	18.160
DT	KNN	0.9166	0.9272	0.9060	0.9148	0.9166	302	309	24	30	18.037
KNN	RO	0.8301	0.7570	0.9031	0.8414	0.8301	301	252	81	32	20.246
KNN	Adasyn	0.8293	0.7195	**0.9392**	0.8462	0.8293	313	239	93	20	19.963
KNN	Smote	0.8189	0.7529	0.8849	0.8300	0.8189	295	251	82	38	20.297
LR	Adasyn	0.8175	0.7551	0.8800	0.8282	0.8175	293	251	81	39	10.516
KNN	SVM Smote	0.8174	0.7443	0.8905	0.8296	0.8174	297	248	85	36	20.478
KNN	BorderlineSMOTE	0.8160	0.7159	0.9161	0.8327	0.8160	305	238	95	27	20.313
LR	SVM Smote	0.8158	0.7729	0.8587	0.8196	0.8158	286	257	75	46	10.737
LR	RO	0.8153	0.7808	0.8498	0.8182	0.8153	283	260	73	49	**10.506**
LR	Smote	0.8136	0.7798	0.8475	0.8165	0.8136	282	259	73	50	10.996

ML Metodas	Išskirčių Apdorojimo Metodas	Tikslumas	Specifiškumas	Jautrumas	F1	AUC	TP	TN	FP	FN	Laikas
LR	BorderlineSMOTE	0.8128	0.7631	0.8624	0.8184	0.8128	287	254	79	45	10.536
LR	NA	0.8000	0.8616	0.7033	0.7046	0.7825	151	286	46	62	10.547
LR	KNN	0.7943	0.8220	0.7665	0.7675	0.7943	256	273	59	76	10.930
KNN	NA	0.7923	0.8389	0.7169	0.7262	0.7779	154	279	54	58	19.846
KNN	KNN	0.7829	0.8079	0.7580	0.7756	0.7829	254	269	64	79	20.453

Duomenų balansavimo metodai stipriau paveikė KNN rezultatus. Visi metodai kaip KNN pagerino vidutinį tikslumą intervale apytiksliai nuo 2.4% iki 3.8%, kuomet K vidurkių metodas sumažino tikslumą apytiksliai siekiantį 1%. Šie rezultatai pademonstravo, kad duomenų balansavimas KNN metodą paveikė labiausiai. Atsitiktinio ėmimo duomenų balansavimo strategija pasiekė geriausią vidutinį modelio tikslumą siekiantį 3.8%.

Svarbu paminėti, kad testavimo etape visos metodų kombinacijos buvo testuojamos tik su šiek tiek ne balansuota duomenų imtimi. Dėl šios priežasties duomenų balansavimo metodų reikšmingumas galėjo būti nepakankamai įvertintas, kadangi kitos duomenų imtys buvo pasižymėjo stipriu arba net ekstremaliu disbalansu.

3.1.5. Mašininio mokymosi metodai

Analizuojant modelio rezultatus DT metodas buvo identifikuotas, kaip geriausiai pasirodęs modelis. Jis pasiekė aukščiausią tikslumą (92.71%), F1-Score (0.924), AUC (0.927), jautrumą (0.927) ir specifiškumą (0.929). kas indikuoja aukštas klasifikavimo galimybes su balansuota klasifikacija, abiejų teigiamų ir neigiamų klasių su mažu klaidų dažniu (žr. **18 lentelė**).

18 lentelė. Rezultatų vidurkiai pasiekti naudojantis skirtingus mašininio mokymosi metodus

ML Metodas	Tikslumas	Specifiškumas	Jautrumas	F1	AUC	TP	TN	FP	FN	Laikas
DT	**0.9271**	**0.9287**	**0.9268**	**0.9236**	**0.9265**	**291**	**310**	**24**	**23**	17.406
KNN	0.8117	0.7746	0.8576	0.8109	0.8096	275	254	80	42	19.458
LR	0.8099	0.7879	0.8240	0.7961	0.8074	263	263	70	53	**10.674**

KNN pademonstravo silpnesnę modelio kokybę, su vidutiniu tikslumu siekiančiu 81.17%, kas buvo mažiau per apytiksliai 11%, lyginant su DT. Tačiau gebėjo pasiekti palyginus aukštą jautrumą, kas indikuoja kokybišką bankroto klasės klasifikavimą. Tačiau KNN, turėjo žemą specifiškumą, kuris nusakė, kad modelis prasčiau klasifikavo ne bankrotą.

Logistinė regresija, tūrėjo žemiausią tikslumą (80.99%), tačiau pasiekė labiau subalansuota klasifikavimą, kuomet jautrumą ir specifiškumą pasiekė 0.824 ir 0.788 atitinkamai. Toliau, logistinė regresija parodė geriausius modelio apdorojimo laikus su vidurkiu siekiančiu 10.67 sekundes. Kas buvo beveik du kartus greičiau, taip pat buvo pastebėta, kad logistinė regresija buvo jautriausia duomenų apdorojimo metodams, ypač duomenų standartizacijai. Logit galėtų varžytis su DT modeliu jei tinkami duomenų apdorojimo metodai būtų pritaikomi.

Apibendrinant, DT buvo pats patikimiausias ir nuosekliausias modelis. Tačiau, Logit parodė didžiausią potencialą modelių pagerinimui. KNN parodė prasčiausius rezultatus ir tūrėtų būti naudojami tik konkrečiose situacijose, kuomet šio modelio paaiškinamumas pagal panašių įrašų identifikavimą galėtų būti panaudotas praktiniam pritaikymui.

3.2. Geriausi modeliai

Iš visų sudarytų metodologijų kombinacijų, kurios buvo ištestuotos, penki modeliai buvo atrinkti, kurie apėmė Logit, DT arba KNN modelius (žr. **19 lentelė**). KNN pasirodė prasčiausiai, kas buvo tikėtasi, kadangi šis metodas neišmoksta konkrečių ryšių tarp kintamųjų. Vietoj to, šis metodas lygina Euklidinį atstumą tarp įrašų ir klasę priskiria remiantis pagal daugiausia esančių tos klasės artimiausių kaimynų. Geriausias KNN modelis buvo 10% ne tikslesnis negu geriausi Logit ir DT modeliai.

19 lentelė. Penki modeliai pasiekę geriausius rezultatus su DT, KNN, Logit metodais.

ML Metodas	Trūkstamų Reikšmių Apdorojimas	Išskirčių Apdorojimas	Duomenų Balansavimas	Duomenų Standartizavimas	Tikslumas	Specifiškumas	Jautrumas	AUC	TP	TN	FP	FN
Logit	RC (0.05) MeanMode	Zscore	RO	Robust	**0.9811**	**0.9790**	0.9832	**0.9811**	234	233	**5**	4
DT	RR (0.1) MeanMode	Zscore	SVMSMOTE	MinMax	0.9762	0.9619	0.9905	0.9762	208	202	8	2
Logit	RCR (0.05, 0.1) MeanMode	MAD	SVMSMOTE	Zscore	0.9759	0.9545	**0.9972**	0.9759	**351**	336	16	**1**
Logit	RR (0.1) MeanMode	MAD	SVMSMOTE	Robust	0.9758	0.9544	**0.9972**	0.9758	350	335	16	**1**
Logit	RCR (0.05, 0.1) MeanMode	MAD	KNN	MinMax	0.9744	0.9659	0.9830	0.9744	346	340	12	6
Logit	RCR (0.05, 0.1) LS	MAD	ADASYN	MaxAbs	0.9744	0.9631	0.9858	0.9744	347	339	13	5
DT	RR (0.1) LS	SD		MaxAbs	0.9726	0.9744	0.9698	0.9721	225	**342**	9	7
DT	RR (0.1) LS			Robust	0.9726	0.9716	0.9741	0.9728	226	**342**	10	6
DT	RRC (0.1, 0.05) DT	ZscoreFitter		MaxAbs	0.9716	0.9770	0.9630	0.9700	208	340	8	8
DT	RC (0.05) DT		RO	MinMax	0.9715	0.9601	0.9829	0.9715	345	337	14	6
KNN	RR (0.1) KNN	Zscore	RO		0.8708	0.7943	0.9474	0.8708	198	166	43	11
KNN	RCR (0.05, 0.1) DT	MAD	RO	Robust	0.8679	0.7841	0.9517	0.8679	335	276	76	17
KNN	KNN	Zscore	RO		0.8641	0.7742	0.9539	0.8641	207	168	49	10
KNN	RC (0.05) MeanMode	SD (0.5)	ADASYN		0.8632	0.7778	0.9487	0.8632	333	273	78	18
KNN	RCR (0.1, 0.05) DT	ZscoreItter	RO	Robust	0.8624	0.8034	0.9213	0.8624	328	286	70	28

Geriausi DT ir Logit modeliai parodė labai panašius rezultatus, kur abu klasifikatoriai pasiekė aukštą tikslumą. Logit pasiekė patį geriausią modelį, kurio tikslumas siekė 98.11%, kuomet DT sekė su artimu tikslumu siekiančiu 97.62%. Logistinė regresija buvo pati efektyviausia klasifikuojant tiek bankrutavusias, tiek nebankrutavusias įmones, kurio specifiškumas siekė 0.979 ir buvo pasiektas 0.997 jautrumas.

Svarbu paminėti, kad pirmas logistinės regresijos modelis ir pirmi keturi geriausi DT modeliai tūrėjo didelį skaičių pašalintų įrašų iš duomenų imties, kas lemia didelį kiekį prarastos informacijos. Dėl

šios priežasties trečias geriausias modelis (Logit) ir dešimtas geriausias (DT) ir dvyliktas geriausias (KNN) modeliai buvo identifikuoti kaip patys geriausi modeliai, kadangi jie parodė pačius patikimiausius ir aukščiausias metrikas įvertinančias modelio kokybę.

Papildomai Logit modelis pasiekė aukščiausią tikslumą iš šių trijų metodų ir tūrėjo trumpiausius skaičiavimo laikus, todėl yra tinkamas laukui jautriose situacijose. Modelis pasiekė tiek aukštą jautrumą (0.9972), bei pasiekė šiek tiek mažesnį specifiškumą (0.955), kas pažymi modelio kokybės kompromisą tarp bankroto ir ne bankroto klasių klasifikavimo. Tačiau DT modelis pasiekė aukščiausią specifiškumą, kol tūrėjo šiek tiek aukštesnį jautrumą. Logit modelis galėjo geriau klasifikuoti bankrutavusias įmones, kol DT galėjo geriau klasifikuoti ne bankrutavusias įmones, kuomet abudu modeliai turi kompromisą tarp vienos ar kitos klasės. KNN pasiekė gan aukštą jautrumą (0.952), tačiau pasiekė žymiai mažesnį specifiškumą (0.784), kas parodė nebalansuotą klasių klasifikavimo tikslumą skirsiantį per 18% bankroto klasės naudai. Tai parodė, kad šis modelis yra tinkamas praktiniam naudojimui kuomet bankroto klasės teisingas klasifikavimas yra svarbesnis. DT ir Logit turi kompromisus jų veikimo efektyvume vienos ar kitos klasės naudai, tačiau skirtumas yra nereikšmingas ir Logit buvo laikomas kaip geresnis modelis.

3.3. Galutiniai modeliai

Galutiniam eksperimentų etapui buvo atrinkta dvylika skirtingų metodologijų (žr. **20 lentelė**), kaip galutiniai metodai. Jie apėmė keturis po modelius kiekvienam mašininio mokymosi algoritmui (Logit, DT ir KNN), atitinkančius arba geriausią sudarytą metodologiją testavimo etape, arba individualiai sudarytą metodologiją, remiantis geriausiai pasirodžiusiais duomenų apdorojimo metodais, kurie pasiekė geriausius rezultatus. Kiekvienam, modeliui buvo atskirai pritaikyta kintamųjų atranka ir gauti modeliai buvo vertinami kaip atskiri variantai siekiant nustatyti kintamųjų atrankos poveikį modelio pasiekiamiems rezultatams.

20 lentelė. Galutiniai keturi modeliai kiekvienam DT, Logit ir KNN metodui

Pavadinimas	ML Metodas	Trūkstamų Reikšmių Apdorojimas	Išskirčių Apdorojimas	Duomenų Balansavimas	Duomenų Standartizavimas	Kintamųjų atranka
LR-1	LR	RCR + Mean Mode	MAD	SVMSMOTE	Zscore	-
LR-1S	LR	RCR + Mean Mode	MAD	SVMSMOTE	Zscore	Taip
LR-2	LR	Mean Mode	MAD	ADASYN	Robust	
LR-2S	LR	Mean Mode	MAD	ADASYN	Robust	Taip
DT-1	DT	RC + LS	SD	SVMSMOTE	MaxAbs	
DT-1S	DT	RC + LS	SD	SVMSMOTE	MaxAbs	Taip
DT-2	DT	RR + Mean Mode	IQR	SVMSMOTE	ZScore	
DT-2S	DT	RR + Mean Mode	IQR	SVMSMOTE	ZScore	Taip
KNN-1	KNN	RCR + K-Means	MAD	ADASYN	Robust	
KNN-1S	KNN	RCR + K-Means	MAD	ADASYN	Robust	Taip
KNN-2	KNN	RC + LS	MAD	ADASYN	Robust	
KNN-2S	KNN	RC + LS	MAD	ADASYN	Robust	Taip

Kiekvienas modelis buvo pavadintas pagal tam tikras taisykles: mašininio mokymosi inicialai (LR, DT arba KNN), po kurių eina skaičius „1“, kuris nusako geriausius rezultatus pasiekusią metodologiją arba „2“, nusakantis metodologiją sudarytą iš geriausiai pasirodančių duomenų apdorojimo metodų. Raidė „S“ nurodo, kad modelis taiko kintamųjų atrinkimo metodą.

Nors atsitiktinis perdėtasis ėmimas (RO) pasirodė vidutiniškai geriausiai, tačiau ADASYN ir SVM SMOTE buvo pasirinkti, KNN ir DT metodams atitinkamai. Toks pasirinkimas buvo atliktas kadangi mažumos klasė visuose duomenų rinkiniuose tūrėjo gan mažą skaičių įrašų ir RO metodas šiems modeliams buvo mažiau tinkamas kadangi įrašų kopijavimas, modeliui nesuteiktų naujos informacijos iš kurios galėtų mokintis.

Ne visi geriausi metodai buvo pasirinkti, kadangi buvo taikoma pirmenybė tiems metodams, kurie išsaugojo didžiąją dalį duomenų rinkinio, kadangi aukščiausią tikslumą pasiekę modeliai parodė didelį kiekį pašalintų įrašų, dėl ko modeliai galėjo prarasti didelį kiekį informacijos. Dėl šios priežasties modelių kokybė galėtų būti itin neigiamai paveikti įrašų apibendrinimą arba derinant su kitais metodais.

Kiekvienas modelis naudojantis kintamųjų atrankos metodus, kurie naudojosi žingsnine atranka. Šis metodas iteraciniu būdu pridėdavo arba pašalindavo jau esamus, tol kol modelio AUC buvo pagerintas. Kintamųjų atrinkimui buvo taikomi tam tikri apribojimai, kuomet naujai įtraukiami kintamieji tūrėjo būti statistiškai reikšmingi arba ne multikolinearūs su jau atrinktais kintamaisiais.

3.4. Galutinių modelių rezultatai

3.4.1. Lietuvos statybos sektoriaus įmonių duomenys

Tyrime buvo panaudotas [10] duomenų rinkinys, apimantis Lietuvos statybos sektoriaus įmones, rezultatų palyginimui. Šis duomenų rinkinys taip pat buvo naudojamas įvertinti visas duomenų apdorojimo ir mašininio mokymosi metodų kombinacijas. Galiausiai, pasirinkti modeliai atspindi geriausius rezultatus pasiekiančias modelių konfigūracijas (žr. **21 lentelė**).

Tarp modelių LR-1S pasiekė aukščiausią specifiškumą (0.983) ir preciziškumą (0.973), kas parodo, kad šis modelis efektyviausiai identifikuoja nebankrutavusias įmones. LR-2S pasiekė aukščiausią jautrumą (0.991), demonstruojantį beveik tobulą bankrutavusių įmonių klasifikavimą. KNN-1S pasiekė aukščiausią AUC (0.977), tikslumą (0.976) ir F1-Score (0.970), kas parodo, kad šis metodas buvo tiksliausias ir labiausiai subalansuoto klasių tikslumo.

21 lentelė. Modelių rezultatai su Lietuvos statybos sektoriaus duomenimis

Sprendimas	AUC	Tikslumas	Specifiškumas	Jautrumas	Preciziškumas	F-1 Score
MARS	0.947	0.939	0.939	0.938	0.92	-
LR-1	0.964	0.967	0.980	0.948	0.969	0.958
LR-1S	0.972	0.974	**0.983**	0.961	**0.973**	0.967
LR-2	0.963	0.962	0.957	0.969	0.938	0.953
LR-2S	0.975	0.972	0.960	**0.991**	0.942	0.966
DT-1	0.937	0.938	0.944	0.931	0.915	0.923

Sprendimas	AUC	Tikslumas	Specifiškumas	Jautrumas	Preciziškumas	F-1 Score
DT-1S	0.956	0.957	0.960	0.952	0.940	0.946
DT-2	0.937	0.938	0.941	0.935	0.911	0.923
DT-2S	0.953	0.952	0.946	0.961	0.921	0.941
KNN-1	0.806	0.811	0.825	0.788	0.746	0.766
KNN -1S	**0.977**	**0.976**	0.972	0.983	0.958	**0.970**
KNN -2	0.798	0.799	0.800	0.797	0.723	0.758
KNN -2S	0.965	0.966	0.966	0.966	0.949	0.957

Visi modeliai išskyrus KNN-1 ir KNN-2, pranoko ankstesniame tyrime pasiektus rezultatus, kuris naudojo MARS modelį, pagerindami tiek bendrą ir klasių klasifikavimo kokybę. Svarbu paminėti, kad visi modeliai naudojantys kintamųjų atrankos metodus pademonstravo pagerintą tikslumą, jautrumą ir specifiškumą. Ypač LR-1S ir LR-2S parodė nedidelius bet reikšmingus ne bankrutavusių įmonių tikslumo pagerinimus su maždaug 1% jautrumo padidėjimu. Nors šie pokyčiai buvo nedideli, tačiau jie buvo statistiškai ir praktiškai reikšmingi.

DT-1S ir DT-2S modeliai pagerėjo panašiai, kaip jų atitikmenys be kintamųjų atrankos. Tačiau, reikšmingiausias pagerėjimas buvo pastebėtas su KNN modeliais, kuomet KNN-1S ir KNN-2S tikslumas, jautrumas ir specifiškumas pagerėjau apytiksliai per 16-17%, kas patvirtina, kad kintamųjų atranka tūrėjo reikšmingiausią poveikį atstumu grįstiems modeliams.

LR-1S modelis pagerino AUC, tikslumą, specifiškumą ir jautrumą, lyginant su MARS modeliu, kuris buvo sudarytas ankstesniame tyrime. LR-1S modelis pagerino šias metrikas apytiksliai per 0.25, 3.5%, 4.4% ir 2.5% atitinkamai. Šie patobulinimai parodė, kad modelis yra labiau patikimas identifikuojant bankrotą.

3.4.2. Taivano įmonių duomenys

Tyrime buvo panaudotas [21] duomenų rinkinys, apimantis Taivano įmones, rezultatų palyginimui (žr. **22 lentelė**).

Tarp modelių LR-1S pasiekė aukščiausią AUC (0.903) ir jautrumą (0.909), kas parodo, kad šis modelis buvo pats optimaliausias ir efektyviausiai identifikavo bankrutavusias įmones. LR-2S pasiekė tokį patį jautrumą, tačiau AUC buvo žemesnis, kas parodo, kad šis modelis buvo prastesnis identifikuojant tarp abiejų klasių. KNN-1S pasiekė aukščiausią tikslumą (0.988) ir specifiškumą (0.992, tačiau pasiekė prasčiausią jautrumą (0.200), indikuojant modelio nebalansuotą abiejų klasių tikslumą ir nepavyko efektyviai klasifikuoti bankrutavusias įmones, kuomet pasirodė prasčiau nei atsitiktinis spėjimas.

22 lentelė. Modelių rezultatai su Taivano įmonių duomenimis

Sprendimas	AUC	Tikslumas	Specifiškumas	Jautrumas	Preciziškumas	F-1 Score
MLP	-	0.8727	0.8868	0.8661	**0.9417**	-
LR-1	0.8587	0.9391	0.9447	0.7727	0.3178	**0.4503**
LR-1S	**0.9030**	0.8974	0.8970	**0.9091**	0.2273	0.3636

Sprendimas	AUC	Tikslumas	Specifiškumas	Jautrumas	Preciziškumas	F-1 Score
LR-2	0.8758	0.8871	0.8879	0.8636	0.2043	0.3304
LR-2S	0.8826	0.8578	0.8561	**0.9091**	0.1739	0.2920
DT-1	0.7201	0.9472	0.9629	0.4773	0.3000	0.3684
DT-1S	0.8068	0.8812	0.8864	0.7273	0.1758	0.2832
DT-2	0.7443	0.9516	0.9659	0.5227	0.3382	0.4107
DT-2S	0.8220	0.9106	0.9167	0.7273	0.2254	0.3441
KNN-1	0.5960	**0.9884**	**0.9921**	0.2000	0.1059	0.1385
KNN -1S	0.8978	0.9287	0.9289	0.8667	0.0541	0.1018
KNN -2	0.7413	0.9457	0.9598	0.5227	0.3026	0.3833
KNN -2S	0.8519	0.8409	0.8402	0.8636	0.1526	0.2594

Tiktais LR-1S ir KNN-1S modeliai pagerino ankstesnio tyrimo rezultatus. Nors daugelis modelių pasiekė aukštesnį tikslumą, tačiau jautrumas buvo itin mažesnis. Nebalansuotose duomenų imtyse, tikslumo pagerėjimas nėra pakankamas, norint paskelbti modelį kaip geresnį, todėl abiejų klasių klasifikavimo tikslumas turi būti subalansuotas. LR-1S buvo identifikuotas kaip geriausias modelis, pasiekiant aukščiausią AUC ir siūlantis stiprų balansą tarp jautrumo ir specifiškumo.

Visi modeliai pritaikantys kintamųjų pasirinkimo metodus, pademonstravo pagerintą jautrumą. Tačiau, šie pagerėjimai buvo pasiekti su kompromisu, kuomet specifiškumas sumažėjo ir kai kuriais atvejais bendras tikslumas. Pagerėjimas buvo reikšmingiausias su KNN ir DT modeliais, kurių jautrumas pradžioje buvo intervale nuo 0.200 iki 0.523. Po kintamųjų atrankos pritaikymo, jautrumas pagerėjo į intervalą nuo 0.727 iki 0.867, o tai parodo itin reikšmingą pagerėjimą. Tuo tarpu, specifiškumas sumažėjo nuo originalaus intervalo 0.960, 0.992 iki 0.840, 0.929. Šis sumažėjimas buvo laikomas mažiau reikšmingas, nei mažumos klasė pagerėjimas.

LR-1S modelis pagerino MLP modelio tikslumą, specifiškumą ir jautrumą apytiksliai siekiantį 2.5%, 1% ir 3.4% atitinkamai. Šie pagerėjimai parodė, kad modelis yra labiau patikimas identifikuojant bankrutavusias įmones, kas yra labai svarbu prognozuojant bankrotą. Taip pat pasiekė aukštesnį specifiškumą, tačiau svarbu paminėti, kad LR-1S parodė prastesnį preciziškumą (0.227), atspindintį kompromisą tarp modelio galimybių identifikuoti bankrutavusias įmones, tačiau su padidėjusiu skaičiumi klaidingai identifikuotų nebankrutavusių įmonių.

3.4.3. Slovakijos įmonių duomenys

Tyrime buvo panaudotas [23] duomenų rinkinys, apimantis Slovakijos įmones, rezultatų palyginimui (žr. **23 lentelė**).

23 lentelė. Modelių rezultatai su Slovakijos įmonių duomenimis

Sprendimas	AUC	Tikslumas	Specifiškumas	Jautrumas	Preciziškumas	F-1 Score
MTE (Best - 2015 Retail)	0.9380	-	-	-	-	-
MTE (2nd Construction	0.8590	-	-	-	-	-

Sprendimas	AUC	Tikslumas	Specifiškumas	Jautrumas	Preciziškumas	F-1 Score
- 2016 Retail)						
MTE (Average)	0.7738	-	-	-	-	-
MDA	-	0.7347	0.6693	0.9310	-	-
LR-1	0.8199	0.9788	0.9803	0.6596	0.1390	0.2296
LR-1S	0.9317	0.9063	0.9060	0.9574	0.0468	0.0892
LR-2	0.8600	0.8996	0.9000	0.8200	0.0385	0.0736
LR-2S	**0.9342**	0.8690	0.8684	**1.0000**	0.0358	0.0691
DT-1	0.6668	0.9895	0.9926	0.3409	**0.1829**	**0.2381**
DT-1S	0.7425	0.9376	0.9395	0.5455	0.0418	0.0777
DT-2	0.6334	**0.9906**	**0.9941**	0.2727	0.1818	0.2182
DT-2S	0.8228	0.9174	0.9183	0.7273	0.0413	0.0781
KNN-1	0.5960	0.9884	0.9921	0.2000	0.1059	0.1385
KNN -1S	0.5345	0.9539	0.9578	0.1111	0.0122	0.0220
KNN -2	0.6755	0.9879	0.9910	0.3600	0.1636	0.2250
KNN -2S	0.8543	0.9478	0.9487	0.7600	0.0675	0.1240

Šis duomenų rinkinys apėmė kelių metų ir verslo sektorių Slovakijos įmones. Geriausi rezultatai buvo pasiekti naudojantis 2015 metų mažmeninės prekybos duomenis, kuomet AUC siekė 0.938, kuomet antras geriausias siekė 0.859 su statybos sektoriaus 2016 metų duomenimis. Tačiau vidutinis AUC siekė tik 0.774 visiems sektoriams ir laikotarpiams. Du modeliai pasiekė AUC viršijantį 0.9. Tačiau nei vienas modelis pasiekė geresnį rezultatą nei MTE 2015 metų mažmeninės prekybos duomenims sudarytas modelis. Geriausio modelio AUC buvo geresnis nei vidutinis AUC ir antras geriausias modelis. Šis rezultatas leido padaryti prielaidą, kad pasiūlytas modelis yra labiau apibendrinamas ne MTE modelis, kuris nesugebėjo pasiekti aukšto rezultato kaip pasiūlytas modelis. Tokios išvados buvo atliktos kadangi vidutinis AUC nepasiekė tokio kaip pasiūlyto modelio AUC. Galiausiai pasiūlytas modelis sugebėjo geriau apibendrinti visus duomenis geriau nei beveik visos metų ir verslo sektorių po grupės.

LR-2S pasiekė aukščiausią AUC (0.934) ir jautrumą (1), indikuojantis, kad modelis sugebėjo optimaliausiai suklasifikuoti bankrotą. DT-2 pasiekė aukščiausią tikslumą (0.991) ir specifiškumą (0.994). Tačiau jautrumas buvo itin prastas, kuris siekė tik 0.273, kas parodė, kad modelis yra itin šališkas daugumos klasės atžvilgiu ir duomenų imtis yra itin nebalansuota, kadangi jautrumas buvo labai mažas, o tikslumas buvo labai artimas specifiškumui. LR-1S pasiekė antra geriausią rezultatą, kurio AUC siekė 0.932, kurio skirtumas buvo apytiksliai 0.03 ir tūrėjo labiausiai subalansuotą tikslumą tarp abiejų klasių. Šis modelis greičiausiai labiau tinkamas nei LR-2S prognozuojant bankrotą, kuomet bankroto klasė yra svarbesnė nei ne bankroto.

Ankstesnis tyrimas pateikė tik AUC reikšmės rezultatus ir nebuvo pakankamai duomenų efektyviai palyginti modelį. Dėl šios priežasties kitas tyrimas [47] su palyginamais rezultatais, kas pasiūlytas

modelis galėtų būti detaliau sulygintas. Šis tyrimas buvo pasirinktas kadangi jame naudojami Slovakijos įmonės, kurios apima tą patį periodą, norint išlaikyti finansinę aplinką ir sudarė sąlygas ekonomiškai pagrįstam palyginimui. Visi KNN ir DT modeliai nepagerino abiejų klasių tikslumo, o LR-1 ir LR-2 negalėjo būti laikomi, kaip ankstesnio tyrimo modelio pagerinimas. LR-1S pagerino tiek tikslumą, specifiškumą ir jautrumą pagerinimą siekiantį 0.27, 0.24 ir 0.03 atitinkamai.

Visi modeliai su kintamųjų atrankos metodais išskyrus KNN-1S sugebėjo pagerinti jautrumą ir AUC, tačiau buvo prastesnis specifiškumas, galiausiai sumažinant bendrą tikslumą, kadangi duomenų rinkinys yra stipriai nesubalansuotas.

3.4.4. Lenkijos įmonių duomenys

Tyrime buvo panaudotas [48] duomenų rinkinys, apimantis Lenkijos įmones, rezultatų palyginimui (žr. **24 lentelė**).

24 lentelė. Modelių rezultatai su Lenkijos įmonių duomenimis

Sprendimas	AUC	Tikslumas	Specifiškumas	Jautrumas	Preciziškumas	F-1 Score
XGBoost + ANN + GA	0.9480	0.9530	0.9670	0.7520	-	-
LR-1	0.7872	0.8813	0.8831	0.6914	0.0530	0.0985
LR-1S	**0.8516**	0.8496	0.8495	0.8536	0.0510	0.0912
LR-2	0.7745	0.7806	0.7807	0.7683	0.0323	0.0620
LR-2S	0.8009	0.7133	0.7116	**0.8902**	0.0286	0.0554
DT-1	0.5977	**0.9783**	**0.9856**	0.2099	**0.1214**	0.1538
DT-1S	0.7562	0.9046	0.9074	0.6049	0.0583	0.1063
DT-2	0.5915	0.9780	0.9854	0.1975	0.1135	0.1441
DT-2S	0.7375	0.8676	0.8701	0.6049	0.0422	0.0790
KNN-1	0.5585	0.9733	0.9812	0.1358	0.0640	0.0870
KNN -1S	0.7559	0.8314	0.8328	0.6790	0.0370	0.0702
KNN -2	0.5448	0.9715	0.9798	0.1098	0.0492	0.0679
KNN -2S	0.7555	0.8627	0.8648	0.6463	0.0436	0.0817

Tarp visų modelių LR-1S pasiekė aukščiausią AUC (0.903), kas parodė, kad šis modelis buvo pats optimaliausias klasifikuojant bankrotą, kadangi ši metrika sumažina daugumos klasės įtaką įvertinant modelio rezultatus, kuomet duomenų rinkinys yra itin nesubalansuotas. LR-2S pasiekė aukščiausią jautrumą (0.890) indukuojantis, kad šis modelis geba geriausiai klasifikuoti bankrutavusias įmones. DR-1 pasiekė aukščiausią tikslumą (0.978) ir specifiškumą (0.986), tačiau tūrėjo vieną prasčiausių jautrumų (0.210), o tai parodė, kad modelio gebėjimas aptikti bankrutuojančias įmones yra silpnas, kadangi nepranoko geresnio klasifikavimo nei atsitiktinis pasirinkimas.

Nei vienas iš modelių sugebėjo pagerinti abiejų klasių klasifikavimą. Modeliai su aukštu jautrumu, kentėjo nuo žemo specifiškumo, ir atvirkščiai. Todėl optimalus modelis tūrėjo užtikrinti subalansuotą tikslumą. Modeliai, kurių specifiškumas viršijo 0.9, geriausiu atveju pasiekė tik 0.6 jautrumą, kas

nėra geras rezultatas, nes tai reiškia, kad šie modeliai nesugebėjo suklasifikuoti bankroto geriau nei atsitiktinis spėjimas arba tik nežymiai geriau. Dėl šios priežasties LR-1S buvo laikomas kaip geriausias modelis, nes jis sugebėjo pasiekti apie 85% tikslumą abiejų klasių atžvilgiu.

Tik LR-1S ir KNN-1S modeliai pagerino tam tikrų klasių klasifikavimo gebėjimus, kurie buvo pasiekti ankstesniam tyrime, kuris naudojo šį duomenų rinkinį. Nors beveik pusė modelių pasiekė aukštesnį bendrą tikslumą, tačiau jų jautrumas buvo žymiai mažesnis. Nebalansuotose imtyse vien pagerintas tikslumas nėra pakankamas įvertini modelį kaip geresnį, abiejų klasių klasifikavimas tūrėtų būt siekiamas, kuo labiau subalansuotas. LR-1S buvo identifikuotas kaip geriausias modelis, kuris pasiekė aukščiausią AUC ir pademonstravo stiprų balansą tarp jautrumo ir specifiškumo. LR-2S pasiekė antrą aukščiausią rezultatą, tačiau LR-1S dvipusius rezultatus kuomet teigiamos klasės klasifikavimas pagerėjo per 4% tačiau neigiamos klasės klasifikavimas suprastėjo per 13%. LR-1S nepasiekė tikslumo (0.953 – 0.850) ir specifiškumo (0.967 – 0.85), lyginant su rezultatais pasiektais ankstesniame tyrime [48], su skirtumu siekiančiu 10% ir 11.8% atitinkamai. Tačiau jautrumas reikšmingai pagerėjo, kuomet jautrumas pagerėjo nuo 0.752 iki 0.854, kurio pagerėjimas buvo 10.2%.

Toliau, visi modeliai, kurie taikė kintamųjų atranką jautrumas pagerėjo, kol specifiškumas sumažėjo, lyginant su modeliais be kintamųjų atrankos taikymu. Tačiau buvo pastebėta, kad visų modelių specifiškumas nukrito.

3.4.5. JAV akcijų biržos duomenys

Tyrime buvo panaudotas [22] duomenų rinkinys, apimantis JAV akcijų biržos įmones, rezultatų palyginimui (žr. **25 lentelė**).

25 lentelė. Modelių rezultatai su JAV akcijų biržos įmonių duomenimis

Sprendimas	AUC	Tikslumas	Specifiškumas	Jautrumas	Preciziškumas	F-1 Score
ANN	0.8720	0.7435	0.7409	0.8218	0.1220	-
LR-1	0.7376	0.8193	0.8207	0.6545	**0.0291**	0.0556
LR-1S	**0.7799**	0.7244	0.7235	0.8364	0.0242	0.0470
LR-2	0.7476	0.7228	0.7224	0.7727	0.0223	0.0434
LR-2S	0.7533	0.6627	0.6612	**0.8455**	0.0200	0.0392
DT-1	0.5156	0.9781	0.9858	0.0455	0.0255	0.0327
DT-1S	0.6793	0.8111	0.8132	0.5455	0.0234	0.0448
DT-2	0.5519	**0.9786**	**0.9856**	0.1182	0.0631	0.0823
DT-2S	0.6876	0.6575	0.6570	0.7182	0.0169	0.0330
KNN-1	0.5561	0.9779	0.9849	0.1273	0.0645	**0.0856**
KNN -1S	0.6470	0.7022	0.7032	0.5909	0.0161	0.0313
KNN -2	0.5561	0.9779	0.9849	0.1273	0.0645	**0.0856**
KNN -2S	0.6470	0.7022	0.7032	0.5909	0.0161	0.0313

Tarp sudarytų modelių LR-1S pasiekė aukščiausią AUC (0.780), kas indikuoja, kad šis modelis buvo pats optimaliausias sprendimas sprendžiant bankroto prognozių problemą, kadangi ši metrika

sumažina daugumos klasės įtaką įvertinant modelio rezultatus, kuomet duomenų rinkinys yra itin nesubalansuotas. LR-2S pasiekė aukščiausią jautrumą, kas indikuoja, kad modelis geriausiai klasifikuoja bankrutavusias įmones. DT-2 pasiekė aukščiausią tikslumą (0.979) ir specifiškumą (0.986), tačiau šis modelis turėjo vieną mažiausių jautrumų (0.1182), kas parodo bankroto identifikavimą, kas nėra geriau nėra atsitiktinis spėjimas.

Nei vienas modelis negalėjo pagerinti tiek bankrutavusių ir nebankrutavusių įmonių klasifikavimą. Modeliai su aukštu jautrumu, kentėjo nuo žemo specifiškumo ir atvirkščiai. Todėl, optimaliausias modelis turėjo parodyti balansuotą klasių klasifikavimą. Modeliai pasiekė specifiškumą aukštesnį nei 0.8, sugebėjo pasiekti tik 0.6 jautrumą geriausiu atveju, kas buvo prastas rezultatas, kadangi visi modeliai, negalėjo klasifikuoti geriau nei atsitiktinis spėjimas arba tik šiek tiek geriau. LR-1S buvo laikomas kaip geriausias modelis, kuris pasiekė apytiksliai 84% tikslumą klasifikuojant bankrutavusias įmones ir 72% tikslumą klasifikuojant ne bankrutavusias.

LR-1S modelis buvo vienintelis modelis, kuris pasiekė arčiausius rezultatus, kaip ANN modelis pasiekė ankstesniame tyrime, kadangi nei vienas modelis užtikrintai nepagerino rezultatų. Šis modelis pagerino bankrutavusių įmonių klasifikavimą per 1.5%. Tačiau pasiekė prastesnį tikslumą (2%), ne bankrutavusių įmonių klasifikavimo tikslumą (1.7%) ir AUC (0.09). Bendras tikslumas kentėjo labiausiai, kadangi duomenų rinkinys buvo stipriai nebalansuotas, nors abiejų klasių klasifikavimas turėjo kompromisų, kurie buvo beveik identiški. Tokie rezultatai parodė, kad pasiūlytas modelis pasiekė labai artimus rezultatus, kurie buvo pasiekti naudojant ANN. AUC tyrime pasiekė daug aukštesnius rezultatus, kas yra netikėta, kadangi visos kitos metrikos yra labai panašios. Toks neatitikimas galėjo buvo įtakotas, dėl skirtingų skaičiavimo būdų, naudojamų skirtingose programavimo kalbose, kaip *Python, R, MATLAB* arba kiti.

Visi modeliai, kurie naudojo kintamųjų atranką pagerino jautrumą, lyginant su modeliais, kurie nenaudojo kintamųjų atrankos, tačiau kompromisas buvo specifiškumo suprastėjimas.

3.5. Kintamųjų reikšmingumo analizė

Nustačius LR-1S modelį, kaip reikšmingiausią šių modelių kintamųjų reikšmingumas buvo pateiktas lentelėse, su kiekvieno kintamojo pavadinimu ir jo koeficientu logistinėje regresijoje. Kintamųjų įtaką, kuomet reikšmė pakinta per 1 nebuvo analizuojama, kadangi buvo atlikta duomenų standartizacija ir būtų tekę pateikti kiekvienam kintamajam atskirą formulę norint apskaičiuoti, bankroto tikimybės pokytį.

26 lentelė. Lietuvos įmonių duomenų rinkinio modelio kintamųjų reikšmingumas

Nr.	Sprendimas	Koeficientas
1	Apyvartinio kapitalo ir viso turto santykis	6.407463
2	(Grynųjų) trumpalaikių įsipareigojimų ir viso turto santykis	5.429118
3	Ilgalaikio turto apyvartumas	1.960909
4	I.2.20	1.260184
5	Pinigų ir trumpalaikių įsipareigojimų santykis (be atsargų)	0.913083
6	Pinigų ir jų ekvivalentų dalis nuo viso turto	0.638701

Nr.	Sprendimas	Koeficientas
7	Materialaus nuosavo kapitalo ir materialiojo turto (be žemės, grynųjų ir nematerialaus) santykis	-0.721461
8	Trumpalaikio turto apyvartumas	-0.779596
9	IV.3.2	-1.072777
10	Greitasis likvidumo rodiklis	-1.414814
11	Trumpalaikio ir viso turto santykis	-8.060095

Su Lietuvos įmonių duomenų rinkiniu (žr. **26 lentelė**) buvo identifikuoti trys kintamieji, kurie buvo reikšmingiausi. Pats reikšmingiausias kintamasis buvo trumpalaikio iš viso turto santykis, kuomet jo padidėjimas įtakojo, kad bankroto tikimybė nukris. Sekantys du reikšmingiausi kintamieji buvo apyvartinio kapitalo ir viso turto santykis ir (Grynųjų) trumpalaikių įsipareigojimų iš viso turto santykis, kuomet šių kintamųjų padidėjimas įtakojo bankroto tikimybę pakilti. Visi kiti kintamieji tūrėjo vidutinį reikšmingumą.

27 lentelė. Taivano įmonių duomenų rinkinio modelio kintamųjų reikšmingumas

Nr.	Sprendimas	Koeficientas
1	Turto grąža	0.888201
2	Turto ir BVP kainos santykis (retesnis rodiklis)	0.33618
3	Pinigų srautas / Visi įsipareigojimai	0.273544
4	Ilgalaikio turto apyvartumas	0.153279
5	Pinigų srautas iš veiklos / Visas turtas	0.150299
6	Grynasis pelno marža (po mokesčių)	0.100153
7	Veiklos sąnaudų santykis su pardavimais	0.031513
8	Pelnas prieš mokesčius vienai akcijai	-0.208072
9	Atsargų ir trumpalaikių įsipareigojimų santykis	-0.291702
10	Nepaskirstyto pelno ir viso turto santykis	-0.758664
11	Veiklos pinigų srautas / Trumpalaikiai įsipareigojimai	-0.853614
12	Pinigų ir viso turto santykis	-1.570758
13	Pelnas prieš mokesčius / Įstatinis kapitalas	-2.910119

Su Taivano įmonių duomenų rinkiniu (žr. **27 lentelė**) buvo identifikuoti trys kintamieji, kurie buvo reikšmingiausi. Pats reikšmingiausias kintamasis buvo pelnas prieš mokesčius iš įstatinio kapitalo, kuomet jo padidėjimas įtakojo, kad bankroto tikimybė nukris. Kintamieji 3 iki 9 tūrėjo mažą reikšmingumą, kol visi kiti tūrėjo vidutinį.

28 lentelė. Slovakijos įmonių duomenų rinkinio modelio kintamųjų reikšmingumas

Nr.	Sprendimas	Koeficientas
1	Greitasis likvidumo rodiklis, prieš 1 metus	1.540837

Nr.	Sprendimas	Koeficientas
2	Turto grąža	1.241401
3	Pinigų santykis, prieš 1 metus	0.990081
4	Darbo sąnaudų ir pajamų santykis, prieš 1 metus	0.05958
5	Turto apyvartumo dienos, prieš 1 metus	-0.050307
6	Dabartinis likvidumo rodiklis, prieš 1 metus	-0.485516
7	Pinigų santykis	-0.782998
8	Turto apyvartumas, prieš 1 metus	-1.376961
9	Atsargų apyvartumo dienos, prieš 1 metus	-2.411786
10	Atsargų apyvartumo dienos, prieš 2 metus	-2.800375
11	Greitasis likvidumo rodiklis, prieš 2 metus	-4.507191

Su Slovakijos įmonių duomenų rinkiniu (žr. **28 lentelė**) buvo identifikuoti trys kintamieji, kurie buvo reikšmingiausi. Patys reikšmingiausi kintamieji buvo greitasis likvidumo rodiklis prieš 2 metus, atsargų apyvartos dienos prieš 2 metus, atsargų apyvartos dienos prieš 1 metus, kuomet jo padidėjimas įtakojo, kad bankroto tikimybė nukris. Visi kintamieji tūrėjo vidutinį reikšmingumą, išskyrus 4 ir 5 kintamieji, kurie tūrėjo itin mažą reikšmingumą.

29 lentelė. Lenkijos įmonių duomenų rinkinio modelio kintamųjų reikšmingumas

Nr.	Sprendimas	Koeficientas
1	Mokėtinų sąskaitų padengimo trukmė (dienomis)	10.12328
2	Veiklos sąnaudų ir trumpalaikių įsipareigojimų santykis	2.275788
3	Veiklos sąnaudų ir visų įsipareigojimų santykis	0.359088
4	Koreguotas likvidumo efektyvumo rodiklis (individualus rodiklis)	0.041423
5	3 metų bruto pelnas / visas turtas	-0.008317
6	Ilgalaikio turto apyvartumas	-0.092167
7	Atsargų laikymo trukmė (dienomis)	-0.245818
8	Atsargų apyvartumo rodiklis	-0.32537
9	Pardavimų augimo tempas (metinis)	-0.39134
10	Greitasis likvidumo rodiklis	-0.559362
11	Pelningumo ir įsipareigojimų santykis	-1.693904
12	Pardavimų ir trumpalaikių įsipareigojimų santykis	-1.734098
13	Alternatyvus mokėtinų sąskaitų padengimo trukmės rodiklis (dienomis)	-9.322231

Su Lenkijos įmonių duomenų rinkiniu (žr. **29 lentelė**) buvo identifikuoti trys kintamieji, kurie buvo reikšmingiausi. Patys reikšmingiausi kintamieji buvo 1 ir 13 kintamieji, kuomet tryliktojo kintamojo padidėjimas įtakojo, kad bankroto tikimybė nukris. Toliau pirmojo kintamojo padidėjimas įtakojo, kad bankroto tikimybė pakils. Šie du kintamieji buvo itin reikšmingi. Antrasis kintamasis tūrėjo

aukštą reikšmingumą. Taip pat 4 – 7 kintamieji tūrėjo žemą reikšmingumą, kol visi kiti tūrėjo vidutinį reikšmingumą.

30 lentelė. Jungtinių Amerikos Valstijų akcijų biržos įmonių duomenų rinkinio modelio kintamųjų reikšmingumas

Nr.	Sprendimas	Koeficientas
1	Ilgalaikės skolos	-0.190135
2	EBITDA (pelnas prieš palūkanas, mokesčius, nusidėvėjimą ir amortizaciją)	-0.256047
3	Atsargos	-0.282345
4	Iš viso turtas	-0.339301
5	Trumpalaikis turtas	-0.675348
6	Grynasis pelnas	-0.89795
7	Rinkos vertė	-1.57501

Su JAV akcijų biržos įmonių duomenų rinkiniu (žr. **30 lentelė**) buvo identifikuotas vienas kintamasis su vidutiniu reikšmingumu. Šis kintamasis buvo pats reikšmingiausias, kuris buvo rinkos vertė, kuomet jo padidėjimas įtakojo bankroto tikimybės sumažėjimą. 1 – 3 kintamieji tūrėjo labai mažą reikšmingumą, kol visi kiti tūrėjo mažą reikšmingumą.

Turto grąžos kintamasis pasirodė su koeficientu 0.888 ir 1.241 Taivano ir Slovakijos duomenų rinkinių modeliuose, kas parodo, jog šie kintamieji yra stiprūs prognozuojant bankrotą. Šis kintamasis keliose šalyse parodo vidutinį reikšmingumą ir parodo, kad jis turi įtaką bankrotui. Kuomet šių kintamųjų reikšmė didėja, didėja tikimybė, kad įmonei gresia bankrotas. Šis rodiklis parodė, kad didesnis pelningumas sumažina bankroto riziką.

Ilgalaikio turto apyvarta parodė aukštą įtaką (1.96) bankrotui Lietuvos įmonėms. Priešingai Taivano (0.15) ir Lenkijos (-0.09) įmonėms parodė labai mažą reikšmingumą. Šis kintamasis reiškią, kad efektyviai naudojant ilgalaikį turtą, bankroto rizika yra mažinama. Dėl to Lietuvos verslo sektoriuje efektyvus ilgalaikio turtas turi labai reikšmingą įtaką bankrotui, kol Taivano ir Lenkijos įmonėms šis rodiklis turi itin mažą įtaką.

Greitasis likvidumo rodiklis turėjo reikšmingą įtaką prognozuojant bankrotą. Lietuvos, Slovakijos prieš 2 metus ir Lenkijos situacijoje, kuomet likvidumas yra didesnis bankroto rizika mažėja. Tačiau kuomet Slovakijos greitasis likvidumas prieš 1 metus padidėja bankroto rizika didėja. Šis kintamasis tūrėjo aukštą įtaką bankrotui, kol Lenkijos įmonėms tūrėjo vidutinį reikšmingumą. Kuomet Slovakijos likvidumas prieš 2 metus buvo didesnis tūrėjo didžiausią įtaką bankroto prognozavimui ir jos padidėjimas bankroto rizika mažino. Toliau, didesnis likvidumas prieš 1 metus bankroto rizika didino. Taip pat galima teigti, kad verslo aplinka šiose šalyse yra panaši, nes šie kintamieji gan stipriai įtakojo bankrotą. Toliau, kadangi šios įmonės yra Europos, galima būtų teigti, jog likvidumas turi didesnę įtaką bankrotui, nei ne Europinės šalys. Šių koeficientų rezultatai reiškia, kad įmonės galinčios greitai padengti trumpalaikius įsipareigojimus yra finansiškai stabilesnės. Priešingu atveju įmonės turi didesnė bankroto riziką.

Toliau pinigų santykis buvo -1.57, -0.78 ir 0.99, Taivano ir Slovakijos esamų ir prieš 1 metus koeficientai atitinkamai. Abiejų šalių variantu šis kintamasis turėjo vidutinį reikšmingumą. Taivano

įmonių atveju parodė, kad didesnis pinigų kiekis sumažino bankroto riziką ir šios įmonės yra stabilesnės. Slovakijos atveju, jog jei prieš vienerius metus buvo perteklinių pinigų bankroto rizika iškildavo, tačiau jei einamaisiais metais buvo perteklinių pinigų bankroto rizika mažėjo.

Galiausiai galima teigti, kad europinės šalys turi kelis vienodus finansinius rodiklius, kurie turi tam tikrą riziką bankrotui. Tačiau rodikliai įtakojantys kelių šalių įmonių bankrotą ne visuomet sutapo. Dėl šių priežasčių, galima teigti, kad skirtingose šalyse finansinės aplinkos nesutampa ir vieni rodikliai neturi tos pačios įtakos, kaip kiti. Taip pat nei vienas reikšmingas JAV įmonės rodiklis nesutapo su kitų modelių rodikliais, todėl JAV ekonominė aplinka stipriai skiriasi nuo kitų šalių. Todėl negalima remtis vienodais finansiniais rodikliais skirtingose šalyse — kiekvienos šalies ekonominė aplinka reikalauja individualaus modelio pritaikymo. JAV rodiklių nesutapimas su kitų šalių modeliais tik dar labiau tai patvirtina.

3.6. Rezultatų apžvalga

Buvo išanalizuoti visų dvylikos modelių rezultatai su penkiais skirtingais duomenų rinkiniais ir geriausius rezultatus pasiekę modelis buvo atrinktas kaip geriausias. Tarp modelių, kurie buvo panaudoti, kartu su Lietuvos statybos sektoriaus duomenimis, tiek LR-1S ir LR-2S modeliai pasiekė aukščiausias prognozavimo galimybes, kuomet LR-2S tūrėjo nedidelį pranašumą. Tačiau LR-1S pasirodė geriausiai su kitais duomenų rinkiniais, todėl buvo pripažintas labiausiai prisitaikančiu ir optimaliausiu modeliu. Šis modelis naudojo logistinę regresiją, kas patvirtino, kad šis metodas yra veiksmingesnis ir stipriau interpretuojamas negu DT ir KNN modeliai. Trūkstamos reikšmės buvo vykdomas pirmiausia pašalinant kintamuosius kuomet kintamasis turi didelį kiekį trūkstamų reikšmių ir toliau pašalinant įrašus, kurie turi didelį kiekį trūkstamų reikšmių per didelį kiekį kintamųjų. Galiausiai visoms likusioms trūkstamoms reikšmėms, Mean-Mode imputacija. Išskirtys buvo apdorotos taikant MAD slenkstinę reikšmė imputuojant visas išskirtis. SVM SMOTE buvo pritaikytas, siekiant subalansuoti duomenų rinkinį, kuomet buvo sugeneruoti sintetiniai įrašai, kas leido modeliui geriau atskirti mažumos nuo daugumos klasės. Z-Score standartizacija užtikrino, kad kintamųjų masteliai būtų vienodi. Galiausiai buvo pritaikytas kintamųjų atrankos metodas naudojant žingsninę atranką.

KNN ir DT modeliai be kintamųjų atrankos pasiekė aukštą klasifikavimo kokybę. Tačiau abu modeliai nesugebėjo klasifikuoti bankrutavusių įmonių, kuomet buvo pasiektas jautrumas apytiksliai intervale nuo 4% iki 52%, kas nebuvo tinkamas klasifikavimas, kadangi abi klasės tūrėjo klasifikuoti bent geriau nei atsitiktinis spėjimas (50%), išskyrus su Lietuvos įmonėmis, kuomet klasifikacija buvo daug geresnė. Retais atvejais šis modelis pasiekė apytiksliai 50% tikslumą, kas yra geresnis rezultatas, tačiau vis vien prastesnis rezultatas negu geriausi modeliai. Kintamųjų atrankos pritaikymas KNN ir DT modeliams, reikšmingai pagerino jautrumą iki tikslumo intervale nuo 54% iki 86%. Tačiau dėl šios priežasties specifiškumas nukrito nuo intervale 95% - 99% iki intervale 65% - 95%. Šis modelio suprastėjimas buvo reikšmingas, kadangi modelis suprastėjo iki 34%, kuomet klasifikuojamas ne bankrotas. Tačiau prarastas klasifikavimo patikimumas nebankrutavusioms įmonėms buvo gan didelis, bet pagerėjimas klasifikuot bankrotą buvo žymiai didesnis, siekiantis 60%, beveik du kartus didesnis nei prarastas ne bankroto klasifikavimo galimybės. Tokie rezultatai yra reikšmingas pagerėjimas, kuomet modeliai tapo tinkamais naudoti bankroto prognozėms, kuomet jie tapo pajėgus klasifikuoti abi klases geriau nei atsitiktinis spėjimas. Buvo pastebėta, kad nebalansuotas abiejų klasių tikslumas buvo daug reikšmingesnis, kuomet duomenų rinkiniai buvo didesni ir labiau nebalansuoti. Dėl šių faktorių modelio galimybės atpažinti mažumos klasės duomenų tendencijas ir ryšius tarp

kintamųjų. Lietuvos statybos sektoriaus duomenų rinkinys pasižymėjo mažesniais kompromisais ir disbalansu tarp abiejų klasių tikslumo, kadangi ši duomenų imtis buvo mažiausia ir labiausiai subalansuota. Šiam duomenų rinkiniui, abu DT ir Logit modeliai pasirodė panašiai, kol KNN tūrėjo daug reikšmingesnį pagerėjimą nuo kintamųjų atrankos metodų. Logit, DT ir KNN modeliai tūrėjo pagerėjimus specifiškume ir jautrume apytiksliai siekiančius (0.3%, 1.5%), (1%, 2%) ir (15.5%, 17%) atitinkamai. Tai parodė, kad artimai subalansuotos duomenų imtys beveik neturi kompromisų, kuomet kintamųjų atrankos metodai yra pritaikomi, tačiau ši situacija yra reta, kadangi bankroto prognozių problemose, duomenų rinkiniai yra stipriai nebalansuoti.

LR-1S pademonstravo pačius geriausius rezultatus su visomis duomenų imtimis ir buvo identifikuota kaip geriausia metodologija. Šis modelis pagerino tyrimo [10] rezultatus, kuriame buvo analizuotas Lietuvos statybos sektoriaus duomenų rinkinys. Šis modelis pagerino AUC (0.974), tikslumą (0.974), specifiškumą (0.983) ir jautrumą (0.961), lyginant su MARS modeliu per 0.025, 0.035, 0.044 ir 0.023 atitinkama. Čia buvo pagerintas klasifikavimo tikslumas, bankrutavusių ir ne bankrutavusių įmonių klasifikavimo tikslumai per 3.5%, 4.4% ir 2.3% atitinkamai. Taip pat, modelis pagerino tyrimo [21] rezultatus, kuriame buvo analizuotas Taivano įmonių duomenų rinkinys. Šis modelis pagerino tikslumą (0.897), specifiškumą (0.897) ir jautrumą (0.909), lyginant su MLP modeliu per 0.025, 0.010 ir 0.043 atitinkama. Čia buvo pagerintas klasifikavimo tikslumas, bankrutavusių ir ne bankrutavusių įmonių klasifikavimo tikslumai per 2.5%, 1% ir 4.3% atitinkamai. Toliau, modelis pagerino tyrimo [23] rezultatus, kuriame buvo analizuotas Slovakijos įmonių duomenų rinkinys. Tačiau, šiame tyrime modelio rezultatai buvo pateikiami tik AUC ir tik atskiriems metų, bei verslo sektorių poaibiams. Čia du geriausi modeliai buvo pasiekti naudojant 2015 metų mažmeninės prekybos ir 2016 metų statybos sektorių duomenų poaibius. Čia buvo pasiektas AUC 0.938 ir 0.859 atitinkamai. Pasiūlytas modelis pasiekė rezultatus panašius kaip tyrimo geriausio poaibio. Šis modelis pagerino AUC per 0.16 nei vidurkis per visus duomenų poaibių rezultatus. Tai parodė, kad modelis sugebėjo geriau apibendrinti duomenis geriau per visus verslo sektorius ir metus. Kito tyrimo [47] rezultatus taip pat pagerino, kuriame buvo taip pat naudojamas Slovakijos įmonių duomenų rinkinys. Šis modelis pagerino tikslumą (0.906), specifiškumą (0.906) ir jautrumą (0.957), lyginant su MDA modeliu per 0.172, 0.237 ir 0.026 atitinkama. Čia buvo pagerintas klasifikavimo tikslumas, bankrutavusių ir ne bankrutavusių įmonių klasifikavimo tikslumai per 17.2%, 23.7% ir 2.3% atitinkamai. Toliau, šis modelis pagerino tyrimo [48] rezultatus, kuriame buvo analizuotas Lenkijos įmonių duomenų rinkinys. Šis modelis pasiekė AUC (0.852), tikslumą (0.850), specifiškumą (0.850) ir jautrumą (0.854). Bendras tikslumas ir ne bankrutavusių įmonių klasifikavimo tikslumas suprastėjo per 10.3% ir 11.8% atitinkamai. Tačiau, bankroto klasės klasifikavimas pagerėjo per 10.2%. Pasiūlyto modelio klasifikavimo tikslumas tarp abiejų klasių buvo labiau subalansuotas ir tūrėjo beveik tokį patį pagerėjimą ir suprastėjimą tarp abiejų klasių klasifikavimo tikslumų lyginant su Lenkijos įmones analizuojančiu tyrimu. Tai parodė, kad modelis buvo pagerintas, ypač tuomet kai yra prioretizuojamas bankroto klasifikavimas. Toliau, bankroto klasė yra laikoma, kaip svarbesnė klasė, kas patvirtina, kad modelis buvo pagerintas, lyginant su ankstesniu tyrimu. Tačiau, ši duomenų imtis buvo stipriai nesubalansuota ir modelis pasiekė geresnį tikslumą su daugumos klase, kuomet atsiranda daugiau klaidingų klasifikacijų. Šis modelis tūrėtų būti naudojamas, kuomet praktiniame naudojime bankroto klasė yra laikoma kaip reikšmingesnė klasė. Toliau, tyrime pasiūlytas modelis naudojo XGBoost + ANN, kas yra neinterpretuojamas modelis. Priešingai, pasiūlytas modelis naudojo logistinę regresiją, kuris yra stipriai interpretuojamas, o tai laikoma, kaip pranašumas, kuomet sprendžiama bankroto prognozių problema. Šie sudėtingesni modeliai, kaip ANN pasiekia geresnes klasifikavimo galimybes, lyginant su paprastesniais modeliais, kaip statistinis logistinės regresijos modelis. Šis

modelis pasiekė žymiai geriau subalansuotą tikslumą, nei ANN. Tuo pačiu metu, net ir smulkūs modelio klasifikavimo galimybių ir interpretavimo pagerinimai buvo skaitomi, kaip reikšmingas rezultatų pagerinimas. Galiausiai, šis modelis pagerino tyrimo [22] rezultatus, kuriame buvo analizuotas JAV akcijų biržos duomenų rinkinys. Šis modelis pasiekė AUC (0.780), tikslumą (0.724), specifiškumą (0.724) ir jautrumą (0.836). Bendras tikslumas ir ne bankrutavusių įmonių klasifikavimo tikslumas suprastėjo per 1.9% ir 1.7% atitinkamai. Tačiau, bankroto klasės klasifikavimas pagerėjo per 1.5%. Pasiūlyto modelio klasifikavimo tikslumas tarp abiejų klasių buvo tūrėjo beveik tokį patį pagerėjimą ir suprastėjimą tarp abiejų klasių klasifikavimo tikslumų lyginant su ankstesniu tyrimu. Bankroto klasės klasifikavimas buvo pagerintas, o ši klasė yra reikšmingesnė sprendžiant bankroto problemą. Tačiau, bendras tikslumas suprastėjo, kadangi daugumos klasės klasifikavimas taip pat suprastėjo. Toliau, pasiūlytas modelis pasiekė labai panašius rezultatus, kuriuos pasiekė ANN modelis. Tuo pačiu, ANN yra neinterpretuojamas modelis, ir pasiūlytas logistinės regresijos sprendimas buvo stipriai interpretuojamas bei pasiekė labai panašius rezultatus. Dėl šių priežasčių modelis buvo laikomas kaip pagerinantis rezultatus, lyginant su ANN modeliu.

Šiame tyrime, modelis reikšmingai pagerino rezultatus, pasiektus tyrimuose, kurie naudojo Lietuvos statybos sektoriaus, Taivano ir Slovakijos duomenų rinkinius. Tačiau, rezultatai pasiekti naudojantis Lenkijos ir JAV akcijų biržos duomenų rinkinius pademonstravo kompromisus tarp pagerėjimo ir suprastėjimo jautrumo, bei specifiškumo metrikų. Čia, nebuvo galima vienareikšmiškai nustatyti pagerėjimo, kadangi ne bankroto klasė sudarė daugumą duomenų, o duomenų rinkinys buvo itin nebalansuotas, kas lėmė bendro tikslumo sumažėjimą. Tačiau, pasiūlyti modeliai ankstesniuose tyrimuose naudojo ANN metodus, kurie nėra interpretuojami, o modelis pasiūlytas šiame darbe naudojo logistinę regresiją, kas yra stipriai interpretuojamas metodas. Kintamųjų paaiškinamumas ir jų įtaka bankrotui yra viena reikšmingiausių charakteristikų, kuomet sprendžiama bankroto prognozių problema, kadangi tai suteikia įžvalgų, kas sukelia bankrotą. Dėl šių priežasčių rezultatai pasiekti su šiomis dviem duomenų imtimis buvo laikomi kaip rezultatų pagerėjimas, nes šie modeliai yra stipriai interpretuojami ir rezultatų kompromisas tarp dviejų klasių buvo labai panašus, nors ir suprastėjo bendras tikslumas. Taip pat JAV akcijų biržos duomenų rinkinio rezultatų pagerėjimas ir suprastėjimas buvo labai mažas, kas parodė, kad su šiuo duomenų rinkiniu rezultatai, net geresni, nes bendras tikslumas labai mažai sumažėjo.

Išvados

1. Tyrimo metu buvo įvertinti geriausi duomenų apdorojimo metodai, kurie pasiekė aukščiausius vidutinius tikslumus. Jie buvo identifikuoti testuojant skirtingų trūkstamų reikšmių, išskirčių, duomenų balansavimo ir standartizavimo kombinacijas. Rezultatai pademonstravo, kad naudojant vidutiniškai geriausius metodus, metodologija nepasiekė geriausių rezultatų. Geriausiai pasirodžiusi metodologija su visomis duomenų imtimis, buvo identifikuota metodologija, kuri pasiekė geriausią rezultatą kombinacijų testavimo etape. Šie rezultatai, patvirtino, kad metodų suderinamumas yra svarbesnis nei atskirų metodų kokybė.
2. Buvo sudaryta 12 metodologijų, kurias apėmė po keturis modelius kiekvienam mašininio mokymosi metodui (Logit, DT ir KNN), kuomet iš jų du modeliai taikė kintamųjų atrankos metodus ir vienas buvo sudarytas iš geriausių individualių duomenų apdorojimo metodų, kuomet kitas naudojo pačią geriausią metodologiją atrinktą metodų kombinacijų testavimo etape. Geriausia metodologija buvo įvertinta kaip LR-1S, kuri pasiekė geriausius rezultatus metodologijų kombinacijų testavimo etape ir naudojo kintamųjų atrankos metodiką. Ši metodologija apėmė procesų seką pradedant su trūkstamų reikšmių apdorojimu pasitelkiant kintamųjų, toliau eilučių šalinimu ir galiausiai likusios trūkstamos reikšmės buvo imputuojamos naudojant Mean Mode imputaciją. Toliau, buvo atliekama MAD išskirčių imputacija. Po to, buvo atliekama Z-Score standartizacija ir padalinimas į apmokymo, bei testavimo duomenų imtis. Tai sekė su SVM SMOTE duomenų balansavimu. Galiausiai buvo taikomas reikšmingų kintamųjų žingsninė atranka kuomet, buvo atrenkami tik reikšmingi kintamieji, užtikrinant, kad modelyje neegzistuotų multikolinearūs kintamieji. Modelis buvo apmokomas ir testuojamas naudojant logistinę regresiją.
3. Iš visų metodologijų LR-1S pasiekė geriausius rezultatus, kuomet buvo pagerinti tyrimų rezultatai, kurie apėmė Lietuvos statybos sektoriaus, Taivano ir Slovakijos duomenų rinkinius. Likusiųjų duomenų rinkinių rezultatai neparodė akivaizdaus rezultatų pagerinimo. Buvo pasiektas panašus bankrutavusių įmonių klasifikavimo tikslumo pagerinimas ir nebankrutavusių įmonių klasifikavimo tikslumo suprastėjimas. Tai galiausiai lėmė, sumažėjusį bendrą tikslumą. Tačiau, šių tyrimų modeliai buvo sudaryti pasinaudojant ANN metodus, kurie nėra interpretuojami, kol pasiūlytas modelis buvo sudarytas su logistine regresija, kuri yra stipriai interpretuojama. Taip pat, praktikoje teisinga bankroto klasifikacija yra laikoma svarbesnė nei ne bankroto klasifikacija. Dėl šių priežasčių šie modeliai buvo taip pat laikomi kaip pagerinti rezultatai. Galiausiai LR-1S buvo identifikuotas, kaip universali metodika, kurią galima pritaikyti skirtingoms duomenų imtims apimančioms skirtingas šalis bei verslo sektorius.
4. Buvo sulyginti modeliai sudaryti su skirtingų šalių duomenų rinkiniais ir buvo atrinkti statistiškai reikšmingi kintamieji. Atlikus analizę, identifikuoti kintamieji, kurie skirtingoms šalių įmonėms parodė reikšmingą įtaką bankrotui. Šiuos kintamuosius apėmė turto grąža, ilgalaikio turto apyvarta, greitasis likvidumo rodiklis ir pinigų santykis. Šių kintamųjų reikšmingumo įvertinimas, pademonstravo, kad kai kurie turėjo nuoseklią įtaką bankroto tikimybei, nepriklausomai nuo regiono, o kitų įtaką skyrėsi priklausomai nuo šalie ekonominės aplinkos ar verslo sektoriaus ypatybių. Šie rezultatai patvirtino, kad tam tikri finansiniai rodikliai gali būti naudojami kaip bankroto rizikos prognozės indikatoriai, tačiau taip pat išryškino būtinybę pritaikyti modelius konkrečios šalies ar verslo sektoriaus kontekste.

Literatūros sąrašas

1. KAUR, Ms.N. - RIZVI, Dr.S. Trends And Patterns In Bankruptcy Prediction: A Scopus-Based Bibliometric Review. In *Educational Administration: Theory and Practice* . 2024. p. 520–533. .
2. ALTMAN, E.I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. In *The Journal of Finance* [interaktyvus]. 1968. Vol. 23, no. 4, p. 589–609. Prieiga per internetą: <http://www.jstor.org/stable/2978933>.
3. OHLSON, J.A. Financial Ratios and the Probabilistic Prediction of Bankruptcy. In *Journal of Accounting Research* [interaktyvus]. 1980. Vol. 18, no. 1, p. 109–131. Prieiga per internetą: <http://www.jstor.org/stable/2490395>.
4. ZMIJEWSKI, M.E. Methodological Issues Related to the Estimation of Financial Distress Prediction Models. In *Journal of Accounting Research* [interaktyvus]. 1984. Vol. 22, p. 59–82. Prieiga per internetą: <http://www.jstor.org/stable/2490859>.
5. BEAVER, W.H. Financial Ratios As Predictors of Failure. In *Journal of Accounting Research* [interaktyvus]. 1966. Vol. 4, p. 71–111. Prieiga per internetą: <http://www.jstor.org/stable/2490171>.
6. CINDIK, Z. - ARMUTLULU, I.H. A revision of Altman Z-Score model and a comparative analysis of Turkish companies' financial distress prediction. In *National Accounting Review* [interaktyvus]. 2021. Vol. 3, no. 2, p. 237–255. Prieiga per internetą: <http://www.aimspress.com/article/doi/10.3934/NAR.2021012>.
7. GHOSH, A. - KAPIL, S. Is Altman's Model efficient in predicting bankruptcy? – A comparison among the Altman Z-score, DEA, and ANN models. In *Journal of Information and Optimization Sciences* . 2022. Vol. 43, no. 6, p. 1191–1207. .
8. JOHN JARVIS, P. - RAUF AHMAD, S. *A COMPARISON OF DIFFERENT METHODS FOR BANKRUPTCY PREDICTION*. 2024. .
9. BEADE, Á. ir kt. Business failure prediction models with high and stable predictive power over time using genetic programming. In *Operational Research* . 2024. Vol. 24, no. 3. .
10. KANAPICKIENĖ, R. ir kt. Bankruptcy Prediction for Micro and Small Enterprises Using Financial, Non-Financial, Business Sector and Macroeconomic Variables: The Case of the Lithuanian Construction Sector. In *Risks* . 2023. Vol. 11, no. 5. .
11. DONG, Y. - PENG, C.-Y.J. [interaktyvus]. .2013. Prieiga per internetą: <http://www.springerplus.com/content/2/1/222>.
12. SZÁNTÓ, T.K. Handling outliers in bankruptcy prediction models based on logistic regression. In *Public Finance Quarterly* . 2023. Vol. 69, no. 3, p. 89–103. .
13. ALTALHAN, M. ir kt. Imbalanced Data Problem in Machine Learning: A Review. In *IEEE Access* . 2025. Vol. 13, p. 13686–13699. .
14. VINAY, S. STANDARDIZATION IN MACHINE LEARNING. In [interaktyvus]. 2021. [žiūrėta 2025-05-20]. . Prieiga per internetą: <https://www.researchgate.net/publication/349869617_STANDARDIZATION_IN_MACHINE_LEARNING>.
15. TSAI, C.-F. Feature selection in bankruptcy prediction. In *Knowledge-Based Systems* [interaktyvus]. 2009. Vol. 22, no. 2, p. 120–127. Prieiga per internetą: <https://www.sciencedirect.com/science/article/pii/S0950705108001536>.
16. CAO, Y. ir kt. A two-stage Bayesian network model for corporate bankruptcy prediction. In *International Journal of Finance & Economics* [interaktyvus]. 2022. Vol. 27, no. 1, p. 455–472. Prieiga per internetą: <https://doi.org/10.1002/ijfe.2162>.

17. HASSANIPOUR, S. ir kt. Comparison of artificial neural network and logistic regression models for prediction of outcomes in trauma patients: A systematic review and meta-analysis. In *Injury* . 2019. Vol. 50, no. 2, p. 244–250. .
18. SHMUEL, A. ir kt. A Comprehensive Benchmark of Machine and Deep Learning Across Diverse Tabular Datasets. In [interaktyvus]. 2024. Prieiga per internetą: <http://arxiv.org/abs/2408.14817>.
19. UYSAL, E. - ÖZTÜRK, A. Comparison of machine learning algorithms on different datasets. In *2018 26th Signal Processing and Communications Applications Conference (SIU)* . 2018. p. 1–4. .
20. AINAN, U.H. ir kt. Advancing Bankruptcy Forecasting With Hybrid Machine Learning Techniques: Insights From an Unbalanced Polish Dataset. In *IEEE Access* . 2024. Vol. 12, p. 9369–9381. .
21. BRENES, R.F. ir kt. An intelligent bankruptcy prediction model using a multilayer perceptron. In *Intelligent Systems with Applications* . 2022. Vol. 16. .
22. LOMBARDO, G. ir kt. Machine Learning for Bankruptcy Prediction in the American Stock Market: Dataset and Benchmarks. In *Future Internet* . 2022. Vol. 14, no. 8. .
23. KANÁSZ, R. ir kt. Bankruptcy prediction using ensemble of autoencoders optimized by genetic algorithm. In *PeerJ Computer Science* . 2023. Vol. 9. .
24. VEGANZONES, D. - SÉVERIN, E. An investigation of bankruptcy prediction in imbalanced datasets. In *Decision Support Systems* [interaktyvus]. 2018. Vol. 112, p. 111–124. Prieiga per internetą: <https://www.sciencedirect.com/science/article/pii/S0167923618301088>.
25. EMMANUEL, T. ir kt. A survey on missing data in machine learning. In *Journal of Big Data* . 2021. Vol. 8, no. 1. .
26. LIN, W.-C. - TSAI, C.-F. Missing value imputation: a review and analysis of the literature (2006–2017). In *Artificial Intelligence Review* . 2020. Vol. 53, p. 1487–1509. .
27. DESIANI, A. ir kt. Handling Missing Data Using Combination of Deletion Technique, Mean, Mode and Artificial Neural Network Imputation for Heart Disease Dataset. In *Science and Technology Indonesia* . 2021. Vol. 6, no. 4, p. 303–312. .
28. YANG, J. ir kt. Outlier detection: how to threshold outlier scores? In *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing* [interaktyvus]. New York, NY, USA: Association for Computing Machinery, 2019. Prieiga per internetą: <https://doi.org/10.1145/3371425.3371427>.
29. SIMMONS, J.P. ir kt. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. In *Psychological Science* [interaktyvus]. 2011. Vol. 22, no. 11, p. 1359–1366. Prieiga per internetą: <https://doi.org/10.1177/0956797611417632>.
30. BUZZI-FERRARIS, G. - MANENTI, F. Outlier detection in large data sets. In *Computers and Chemical Engineering* . 2011. Vol. 35, no. 2, p. 388–390. .
31. DZIERŻAK, R. COMPARISON OF THE INFLUENCE OF STANDARDIZATION AND NORMALIZATION OF DATA ON THE EFFECTIVENESS OF SPONGY TISSUE TEXTURE CLASSIFICATION. In *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Srodowiska* . 2019. Vol. 9, no. 3, p. 66–69. .
32. BORKIN, D. ir kt. Impact of Data Normalization on Classification Model Accuracy. In *Research Papers Faculty of Materials Science and Technology Slovak University of Technology* . 2019. Vol. 27, no. 45, p. 79–84. .
33. AMORIM, L.B. V. DE ir kt. The choice of scaling technique matters for classification performance. In [interaktyvus]. 2022. Prieiga per internetą: <http://arxiv.org/abs/2212.12343>.

34. WONGVORACHAN, T. ir kt. A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. In *Information (Switzerland)* . 2023. Vol. 14, no. 1. .
35. HUSSIN ADAM KHATIR, A.A. - BEE, M. Machine Learning Models and Data-Balancing Techniques for Credit Scoring: What Is the Best Combination? In *Risks* . 2022. Vol. 10, no. 9. .
36. JAYANTA DATTAGUPTA, S. 2017. .
37. KARABULUT, E.M. ir kt. A comparative study on the effect of feature selection on classification accuracy. In *Procedia Technology* [interaktyvus]. 2012. Vol. 1, p. 323–327. Prieiga per internetą: <https://www.sciencedirect.com/science/article/pii/S2212017312000692>.
38. HASSAN, M.M. ir kt. A comparative assessment of machine learning algorithms with the Least Absolute Shrinkage and Selection Operator for breast cancer detection and prediction. In *Decision Analytics Journal* . 2023. Vol. 7. .
39. NOERSASONGKO, E. - FAJAR SHIDIK, G. CLASSIFICATION MODELS BASED FORWARD SELECTION FOR BUSINESS PERFORMANCE PREDICTION. In *Journal of Theoretical and Applied Information Technology* [interaktyvus]. 2014. Vol. 67, no. 2. Prieiga per internetą: <www.jatit.org>.
40. MUKAKA, M.M. Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research. In *Malawi Medical Journal* [interaktyvus]. 2012. Vol. 24, no. 3, p. 69–71. Prieiga per internetą: <www.mmj.medcol.mw>.
41. REŽŇÁKOVÁ, M. - KARAS, M. Bankruptcy Prediction Models: Can the Prediction Power of the Models be Improved by Using Dynamic Indicators? In *Procedia Economics and Finance* [interaktyvus]. 2014. Vol. 12, p. 565–574. Prieiga per internetą: <https://www.sciencedirect.com/science/article/pii/S2212567114003803>.
42. SHI, Y. - LI, X. An overview of bankruptcy prediction models for corporate firms: A systematic literature review. In *Intangible Capital* . 2019. Vol. 15, no. 2, p. 114–127. .
43. MISHRA, A. ir kt. AIRBP: Accurate identification of RNA-binding proteins using machine learning techniques. In *Artificial Intelligence in Medicine* . 2021. Vol. 113. .
44. ZELENKOV, Y. - VOLODARSKIY, N. Bankruptcy prediction on the base of the unbalanced data using multi-objective selection of classifiers. In *Expert Systems with Applications* . 2021. Vol. 185, p. 115559. [žiūrėta 2024-05-17]. . .
45. ČEKANAVIČIUS, V. - MURAUSKAS, G. *TAIKOMOJI REGRESINĖ ANALIZĖ SOCIALINIUOSE TYRIMUOSE*. . Vilnius: Vilniaus universiteto leidykla, 2014. ISBN 9786094593000.
46. MILERIS, R. Analysis of statistical credit risk estimation models efficiency. In *Economics and management* . 2009. .
47. SVABOVA, L. ir kt. Prediction of Default of Small Companies in the Slovak Republic. In *Economics and Culture* . 2018. Vol. 15, no. 1, p. 88–95. .
48. AINAN, U. ir kt. Advancing Bankruptcy Forecasting With Hybrid Machine Learning Techniques: Insights From an Unbalanced Polish Dataset. In *IEEE Access* . 2024. Vol. PP, p. 1. .