VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
INFORMATIKOS KATEDRA

Klientų duomenų valdymas bankininkystėje

Client data management in banking

Magistro baigiamasis darbas

Atliko : Giedrius Žiupsnys
Darbo vadovas: prof. Leonidas Sakalauskas
Recenzentas: dr. Antanas Mitašiūnas

Vilnius – 2010

Santrauka

Darbas apima banko klientų kredito istorinių duomenų dėsningumų tyrimą. Pirmiausia nagrinėjamos banko duomenų saugyklos, siekiant kuo geriau perprasti bankinius duomenis. Vėliau naudojant banko duomenų imtis, kurios apima kreditų grąžinimo istoriją, siekiama įvertinti klientų nemokumo riziką. Tai atliekama adaptuojant algoritmus bei programinę įrangą duomenų tyrimui, kuris pradedamas nuo informacijos apdorojimo ir paruošimo. Paskui pritaikant įvairius klasifikavimo algoritmus, sudarinėjami modeliai, kuriais siekiama kuo tiksliau suskirstyti turimus duomenis, nustatant nemokius klientus. Taip pat siekiant įvertinti kliento vėluojamų mokėti paskolą dienų skaičių pasitelkiami regresijos algoritmai bei sudarinėjami prognozės modeliai.

Taigi darbo metu atlikus numatytus tyrimus, pateikiami duomenų vitrinų modeliai, informacijos srautų schema. Taip pat nurodomi klasifikavimo ir prognozavimo modeliai bei algoritmai, geriausiai įvertinantys duotas duomenų imtis.

Raktiniai žodžiai: duomenų tyrimas, banko klientų istorinių duomenų apdorojimas, duomenų vitrina, kredito rizikos vertinimas, klasifikavimas, prognozavimas, kryžminis patikrinimas, nesutapimų matrica, tiesinė regresija, klasifikavimo taisyklė, sprendimų medis.

Summary

This work is about analysing regularities in bank clients historical credit data. So first of all bank information repositories are analyzed to comprehend banks data. Then using data mining algorithms and software for bank data sets, which describes credit repayment history, clients insolvency risk is being tried to estimate. So first step in analyzis is information preprocessing for data mining. Later various classification algorithms is used to make models wich classify our data sets and help to identify insolvent clients as accurate as possible. Besides clasiffication, regression algorithms are analyzed and prediction models are created. These models help to estimate how long client are late to pay deposit.

So when researches have been done data marts and data flow schema are presented. Also classification and regressions algorithms and models, which shows best estimation results for our data sets, are introduced.

Keywords: data mining, historical bank clients data preprocessing, data mart, credit risk estimation, classification, regression, cross validation, confusion matrix, linear regression, classification rule, desicion tree.

Turinys

1 Įvadas

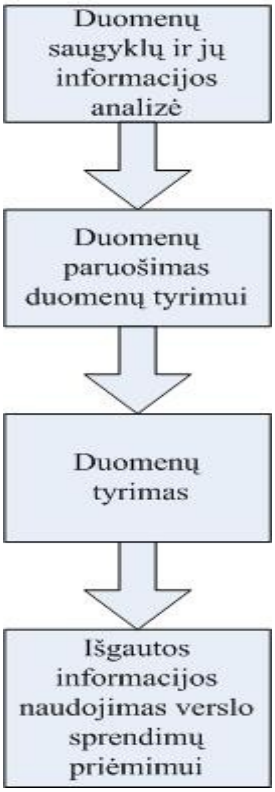
Klientų duomenų valdymas pastaruoju metu įgyja vis svarbesnę reikšmę bankininkystės sektoriuje. Kredito rizikos vertinimas yra viena iš populiariausių tyrimo sričių, nes itin didėja gyventojų kreditų skaičius. Taigi teikiant paskolą reikia įvertinti, ar klientas bus mokus, t.y. ar galės grąžinti paskolą. Visgi pageidaujančių paimti kreditą klientų srautas yra didelis ir įsigilinti į kiekvieno poreikius tampa gana sudėtinga, todėl populiarėja kredito rizikos vertinimas naudojantis iš anksto sudarytais modeliais. Jie kuriami pagal istorinius klientų duomenis, t.y. informacija, kuri parodo, ar klientas atidavė anksčiau išduotas paskolas. Tokiu būdu sprendžiama, ar išduoti kreditą naujam klientui.

Dar 1941 metais David Durant pirmasis aprašė standartinius modelius, pagal kuriuos galima atskirti, kurios paskolos bus grąžintos, o kurios sukels problemų. Tokiems pirmiesiems darbams įtaką padarė, kai finansų institucijos susidūrė su problemomis, susijusiomis su paskolomis. Tačiau nors poreikis ir augo, šiems tyrimams didesnės reikšmės nebuvo skirta, nes šios srities specialistų veikla buvo nukreipta su kariniais tikslais susijusioms problemoms spręsti. Bet 1960 metais, kai atsirado kreditinės kortelės ir žymiai išaugo paskolų poreikis, neliko kitos išeities, tik automatizuoti kredito išdavimo sistemą sudarinėjant standartinius kredito vertinimo modelius. Be to, buvo pastebėta, kad pasitelkiant šiuos modelius kliento kredito rizikos įvertinimas tampa net 50% tikslesnis [Tho00].

Šie tyrimai, taip pat yra svarbūs ir šiuolaikinėje bankininkystėje vertinant kreditų riziką. Tai galime pastebėti po finansų krizės, kuri Lietuvoje turėjo skausmingų padarinių – bankai patyrė daug nuostolių dėl neapdairiai skiriamų paskolų, žmonės, netekę pajamų ir neturėdami už ką grąžinti įmokų, prarado visą savo turtą ir liko skolingi bankams. Taigi šie modeliai turėtų padėti kuo tiksliau įvertinti klientų kredito riziką ir taip sumažinti banko patiriamus nuostolius bei padidinti jų pelną. Tačiau dauguma tokių modelių yra sudaryti naudojant užsienio valstybių bankinę informaciją, todėl jie nevisiškai atspindi Lietuvos kreditavimo duomenis, nes jų požymiai skiriasi. Kaip matome, kredito vertinimo modeliai yra gana priklausomi nuo turimų duomenų ir jų tikslumą gali įtakoti – tiriamas laikotarpis, valstybė, kurioje renkami duomenys, ar kiti specifiniai informacijos požymiai. Taigi mes tyrėme bei kūrėme kredito rizikos vertinimo modelius naudodami vieno Lietuvos banko duomenis, kurie apėmė 2005-2006 metų periodus.

Kredito rizika nėra vienintelė sritis, kurios analizė gali būti naudinga bankams. Kiti tyrimai taip pat gali būti reikšmingi: investicinio portfelio rizikos vertinimas, įvairių paslaugų pardavimo skatinimas ir kt. Kadangi mūsų turėti duomenys atspindėjo paskolos grąžinimo istoriją, kuri leido vertinti kliento mokumą, šiame darbe nagrinėjome kliento kredito rizikos vertinimą.

Taigi sudarinėjant bet kokį modelį pirmiausia reikia išanalizuoti turimus duomenis bei tinkamai juos paruošti. Šis procesas yra reikšmingas, nes duomenų kokybė gali įtakoti modelio tikslumą. 1 pav. pateikiami žingsniai [Tag], kurie turi būti atlikti, kuomet vykdomas duomenų tyrimas ir sudarinėjami banko informacijos vertinimo modeliai.



1 pav. duomenų tyrimo etapai

1. Duomenų saugyklų ir jų informacijos analizė atliekama siekiant kuo detaliau perprasti turimus duomenis:
   - pirmiausia tiriamos duomenų saugyklos bei jų architektūros;
   - sudarinėjami duomenų vitrinų modeliai;
   - tiriami duomenų srautai ir kt.

   Šiame etape taip pat siekiama įsisavinti duomenų formatą, kuris gaunamas iš banko duomenų saugyklų ir naudojamas duomenų paruošimo metu.
2. Duomenų paruošimas duomenų tyrimui, t.y. sugadintų reikšmių pašalinimas ar pataisymas ir kt. Šis etapas taip pat apima informacijos konvertavimą į failo formatą, kurį palaiko pasirinkta duomenų tyrimo programinė įranga, pvz. „Weka“ formatas arff.
3. Duomenų tyrimo algoritmų pritaikymas – naudojant paruoštus duomenis vykdomi duomenų tyrimo algoritmai, kurie padeda sudaryti kredito rizikos ar kitos srities vertinimo modelius.

4. Išgautos informacijos naudojimas verslo sprendimų priėmimui – paskutiniame žingsnyje imamasi atitinkamų veiksmų naudojant gautus vertinimo modelius siekiant pagerinti banko veiklą pvz., sprendžiama ar naudoti sudarytus kredito vertinimo modelius bankinėse sistemose

Darbe detaliau aptarti pirmieji trys žingsniai, nes paskutinysis etapas, verslo sprendimų priėmimas, dažniausiai atliekamas kompetentingų banko darbuotojų, o mūsų tikslas yra informacijos išskyrimas bei modelių sudarymas.

1.1 *Darbo tikslas*

Išnagrinėję duomenų analizės reikšmingumą bankiniams duomenims, nusprendėme savo darbe ištirti banko klientų kredito istorinių duomenų dėsningumus, adaptuojant duomenų tyrimo algoritmus bei programinę įrangą.

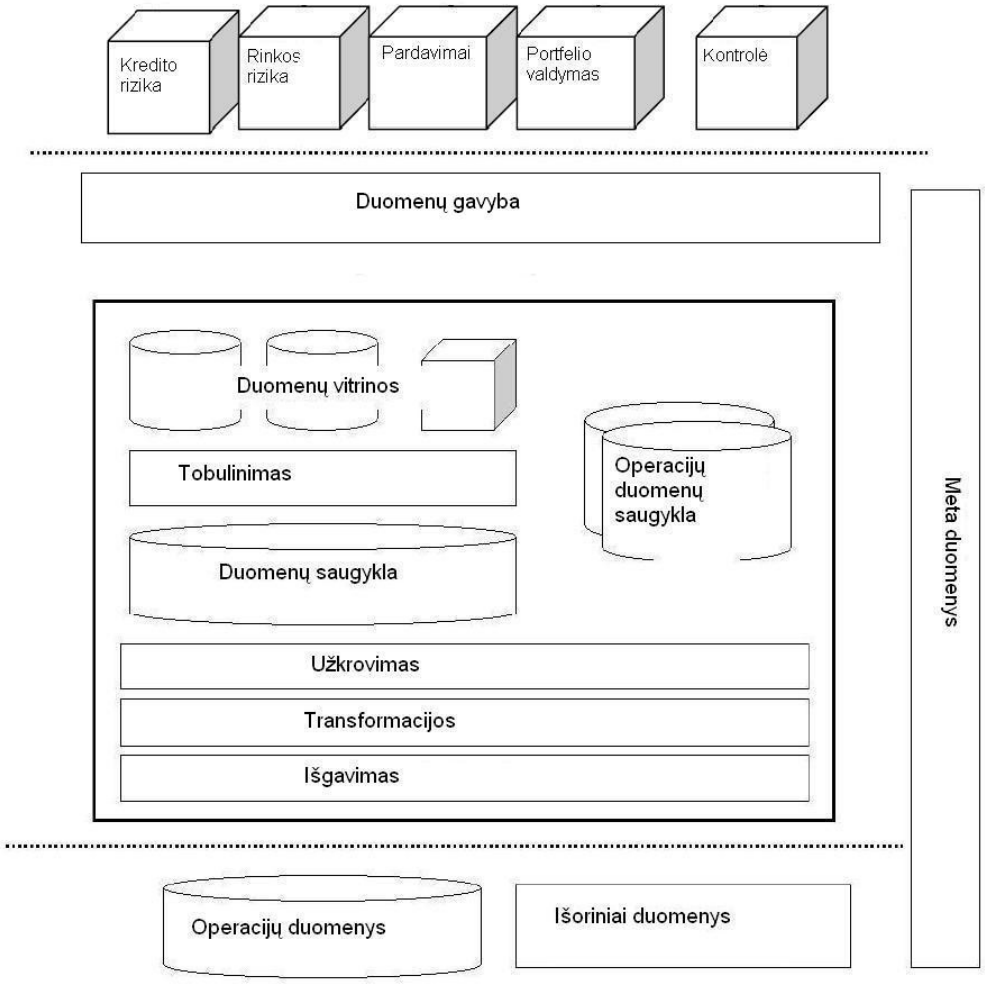
1.2 *Darbo uždaviniai*

Norėdami įgyvendinti šį tikslą nustatėme uždavinius:

- Ištirti duomenų saugyklų architektūrą, naudojamą banko informacinėse sistemose;
- Ištirti duomenų vitrinų sudarymo metodus ir sukurti jų modelius, turimoms duomenų struktūroms;
- Sukurti programinę įrangą,įgalinančią konvertuoti gautus klientų istorinis duomenis į analizei tinkamą formatą;
- Ištirti klasifikavimo ir prognozavimo algoritmų tinkamumą turimiems duomenims;
- Naudojant ištirtus algoritmus sukurti klasifikavimo bei prognozavimo modelius, kurie geriausiai įvertintų turimų duomenų dėsningumus, padedančius nustatyti nemokius klientus;
- Nustatyti banko klientų istorinių duomenų specifinius požymius;

2 Duomenų modelis

2 pav. [Das98] [BAB+01] pavaizduota banko duomenų saugyklų architektūra ir duomenų analizės modelis, vaizduojantis, kaip renkami, kaupiami ir analizuojami duomenys. Šiame skyriuje detaliau nagrinėjamas informacijos patalpinimas į duomenų saugyklą, turimų duomenų struktūra, jų pirminis paruošimas analizei bei trumpai apžvelgiamos duomenų tyrimo taikymo sritys.



2 pav. – duomenų analizės modelis

2.1 Pradiniai duomenys

Žemiausioje struktūroje matomi duomenų šaltiniai:

- Operacijų duomenys – įvairi informacija apie įvairias klientų operacijas, atliekamas bankuose, mūsų atveju tai būtų kliento kredito paėmimas bei jo grąžinimo istorija.
- Išoriniai duomenys – išorinė, gaunama ne iš banko operacijų, informacija pvz. įvairių apklausų ar tyrimų duomenys.

2.2 Duomenų transformacijos

Vėliau, talpinant duomenis į duomenų saugyklas, informacija yra susisteminama ir transformuojama. Ji yra pradinė ir apima duomenų išgryninimą pvz., draudžiama įterpti nevalidžius įrašus. Taigi duomenų transformacijos užtikrina, kad duomenys saugyklose bus validūs ir reikšmingi. Vėliau, prieš duomenų tyrimo vykdymą, dažnai atliekamas papildomas duomenų apdorojimas.

2.3 Metaduomenys

Metaduomenys suteikia papildomos informacijos apie informaciją, esančią duomenų saugyklose, pvz., kaip reikėtų interpretuoti datos formatą, jei turime datą 5/10/2009, t.y. ar ji atitinka 2009 spalio 5 ar 2009 gegužės 10. Vėliau pagal šiuos duomenis gali būti vykdoma duomenų patikrinimas pvz., visi datos laukai turi būti pateikti yyyy.MM.dd(metai.mėnuo.diena) formatu.

2.4 Duomenų vitrinos

Prieš pateikiant duomenis analizei vykdomas duomenų saugyklų informacijos tobulinimas, kuris apima [BAB+01]:

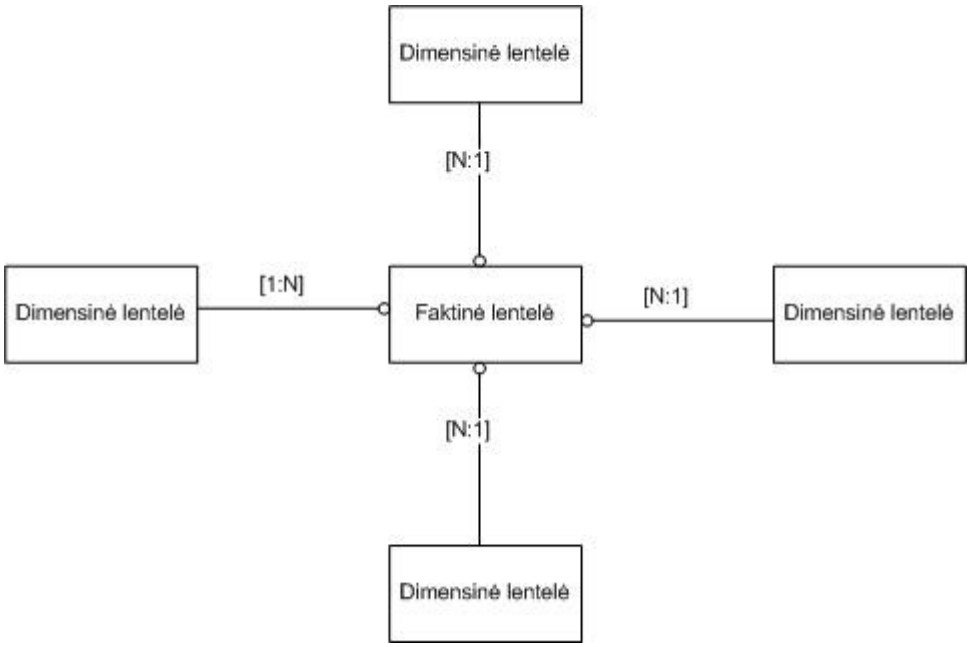
- Duomenų agregaciją – pvz., duomenų saugyklose laikomi kiekvienos dienos duomenys, o vitrinose pateikiamos tik savaitinės reikšmės.
- Duomenų apibendrinimas – pvz., duomenų saugyklose laikoma informacija apie gaunamą pelną iš kiekvienos parduotos prekės, o vitrinose pateikiami duomenys apie pelną, gautą iš prekių grupių.

Taigi susisteminti duomenys perkeliami į duomenų vitrinas, pagal [Inm02], vitrinose laikomi konkrečios veiklos srities duomenys, pvz. kreditų suteikimo statistika ir t.t.

[MKo00] nagrinėjama, kad duomenų saugyklos dažniausiai būna sudarytos iš lentelių, kurių struktūra atitinka bent 3 pirmąsias normines formas. Tai lemia, kad susidaro didžiulės duomenų bazių schemos, kurios informacijos analitikams dažnai būna sudėtingos ir sunkiai suprantamos, todėl jas siūloma supaprastinti iš vis nesilaikyti norminėmis formomis, reikalaujamų duomenų bazių struktūroms, ar bent jau laikytis daugiausia 2 - ąja. Įsidėmėtina, kad taip suprojektuotos schemos vadinamos dimensinėmis ir analitikams yra lengviau suprantamos. Būtent dėl to duomenų vitrinos dažnai projektuojamos taip, kad jos pasižymėtų dimensinių

duomenų bazių savybėmis. Šiems modeliams sudaryti dažniausiai naudojamos šios duomenų modelio schemos:

- Plokščia schema – pats paprasčiausias duomenų vitrinos sudarymo būdas. Visi susiję duomenys sudedami į vieną lentelę; galiausiai gaunamos kelios didelės lentelės, apimančios daug duomenų.
- Žvaigždės schema – loginė duomenų modelio schema, kuri nurodo vienintelį kelią tarp dviejų objektų. Šios schemos centre yra faktinių duomenų lentelė, turinti vidinius raktus, jungiančius ją su kitomis lentelėmis. 3 pav. pavaizduota žvaigždės tipo schema.



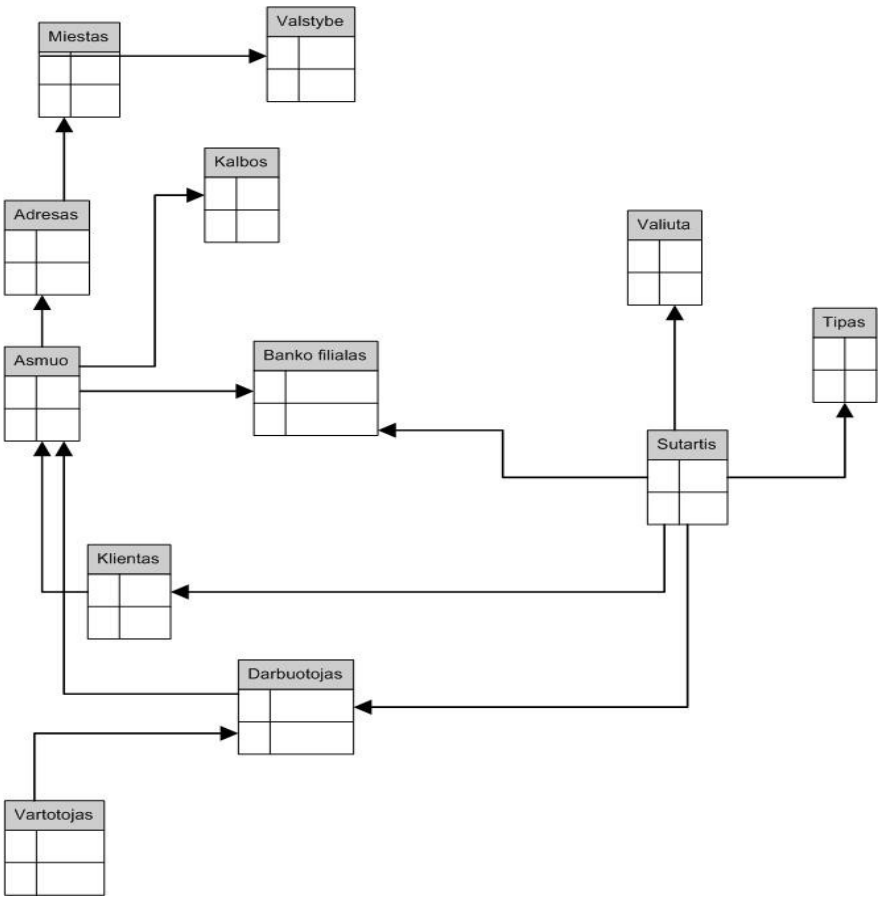
3 pav. žvaigždės tipo schema.

2.4.1 Duomenys apie vartojamąsias paskolas

Taigi buvo gautos dvi operacijų duomenų imtys apibūdinančios banko suteikiamas paskolas. Pirmąją sudaro duomenis apie vartojamąsias paskolas bei su jomis susijusi informacija:

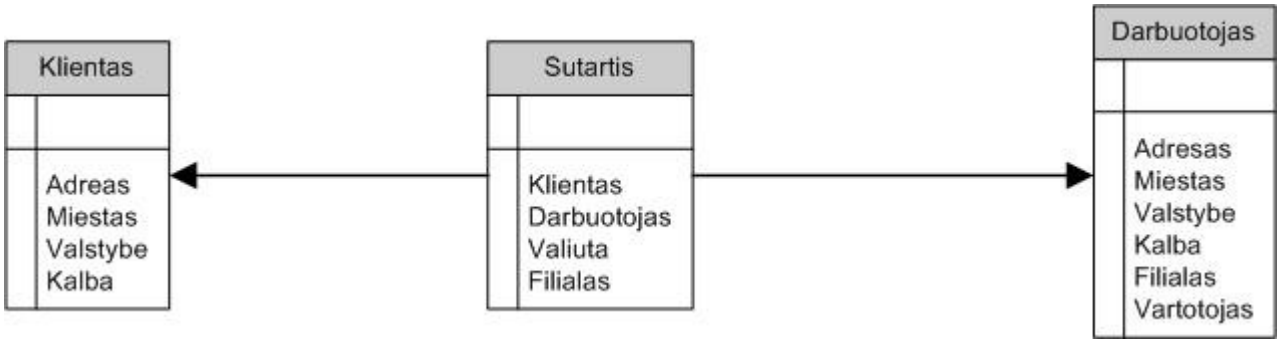
- asmuo sudaręs sutartį;
- darbuotojas sudaręs sutartį;
- banko filialas, kuriame buvo sudaryta sutartis;
- sutartis bei jos informacija;
- kiek dienų vėluoja paskolos grąžinimas.

Ši duomenų aibė apie yra agreguota ir apima 2005-2006 metų laikotarpį. Pradinė informacija yra gana sunkiai suprantama. 4 pav. pateikiamas reliacinis duomenų bazės modelis, kurį sudarėme remdamiesi banko gauta informacija.



4 pav. vartojamųjų paskolų duomenų modelis

Kaip matome modelis yra sudėtingas jis apima 11 lentelių, kurios dar turi papildomų atributų, todėl buvo sukurta žvaigždės tipo duomenų vitrina, kuri supaprastina modelį, tačiau ji turi perteklinių duomenų. Schema susideda iš vienos pagrindinės lentelės – „sutartis“, kuri saugo aktualiausius duomenis. Kitos schemos papildomos lentelės buvo „darbuotojas“ bei „klientas“. Šią žvaigždės tipo schemą matome 5 paveiksle.



5 pav. Vartojamųjų paskolų duomenų žvaigždės schema

Taigi ji patogi analitikui norint paprasčiau perprasti turimus duomenis.

Tuo tarpu duomenų paruošimui ir jų analizei patogiausia plokščia duomenų schema, kurioje visi duomenys pateikiami vienoje lentelėje (6 pav.). Žinoma, tuomet duomenų stulpelių atsiranda daug daugiau ir ši forma analitikui tampa mažiau suprantama. Taigi būtent tokius duomenis analizei pateikė bankas, o informacija buvo pateikta „Microsoft Excel“(*.xls) formato

rinkmenoje. Dėl to, kad duomenys buvo pateikti plokščios schemos formoje, užteko programos, kuri nuskaito vienintelį failą ir konvertuoja duomenis į analizei reikiamą formą.

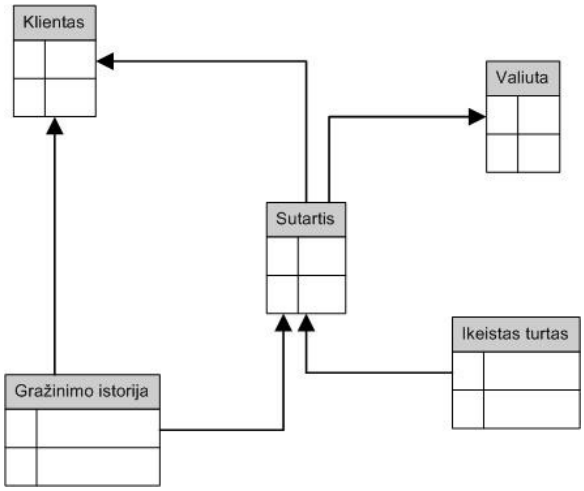
6 pav. Vartojamųjų paskolų duomenų plokščia schema

2.4.2 Duomenys apie būsto paskolas

Antroji gauta duomenų imtis apima informaciją apie būsto paskolas (nuo vartojamųjų paskolų jos skiriasi tuo, kad joms gauti dažniausiai keliami papildomi reikalavimai, pvz., reikalaujamas užstatas. Informacija apie šias sutartis yra panaši, kuri buvo pateikta apie vartojamąsias sutartis:

- asmuo paėmęs sutartį;
- sutartis bei su ja susijusi informacija;
- informacija apie grąžinimus.

Ši informacija apima 2007 metų 4 ketvirčio – 2008 metų 1 ketvirčio laikotarpį. Šių duomenų modelis buvo paprastesnis nei vartojamųjų paskolų, pilnas modelis pateiktas 7 pav.



7 pav. būsto paskolų duomenų modelis

Šis modelis yra nesudėtingas, todėl žvaigždės schemos nuspręsta nedaryti.

Kaip ir duomenys apie vartojamąsias paskolas, informacija apie būsto paskolas, gauta iš banko, buvo plokščios schemos pavidalu (8 pav.), ji patalpinta *.xls formato rinkmenoje. Be to, duomenis apie grąžinimą atitiko kelios įrašo reikšmės, t.y. kokį ilgiausią laikotarpį vėlavo paskolos ar palūkanų mokėjimas bei kiek delspinigių buvo sumokėta už paskolą ar palūkanas per visą paskolos laikotarpį.

8 pav. Vartojamųjų paskolų duomenų plokščia schema

2.5 Duomenų laukai

Kaip buvo minėta anksčiau buvo gautos dvi duomenų imtys, kurios atitinka vartojamąsias bei būsto paskolas. Taigi duomenis apie vartojamąsias paskolas apibūdino 32 atributai, būsto – 18. Duomenų atributai, jų tipai bei pavyzdinės reikšmės pateiktos 1 lentelėje.

Vartojamosios paskolos			**Būsto paskolos**		
Lauko pavadiniams	Lauko tipas	Pavyzdinė reikšmė	Lauko pavadiniams	Lauko tipas	Pavyzdinė reikšmė
Kliento kodas	Skaitinė	546235	Sutarties kodas	Skaitinis	726950
Ar klientas aktyvus	Skaitinis(0, 1)	1	Sutarties numeris	Simbolins	VART-2007-xxxxx-02
Požymis bazėje (identifikacija)	Simbolinis	vardenisp	Kliento sąskaita	Simbolinis	LTxxxxxxxxxxxxxxxxxx
Vardas pavardė	Simbolinis	Vardenis Pavardenis	Valiuta	Simbolinis	EUR
Kliento adresas	Simbolinis	Laukų 17-5	Paskolos suma	Skaitinis	10000
Kliento miesto kodas	Skaitinis	2629	Sutarties pradžios data	Data	2007.11.06
Kliento šalies kodas	Simbolinis	Lt	Sutarties pabaigos data	Data	2011.11.05
	Data	2002.01.01	Kliento vardas, pavardė	Simbolinis	Vardenis Pavardenis
Kliento filialas	Simbolinis	SB	Kliento alga (pingai vienam asmeniui šeimoje)	Skaitinis	700
Kliento kalba	Simbolinis	LIT	Įkeisto turto vertė	Skaitinis	11000
Kliento statusas	Simbolinis	A - aktyvus	Delspinigiai paskolai(%)	Skaitinis	0,05
Sutarties būsena	Simbolinis	EXECUTE	Delspinigia	Skaitinis	0,02

		- vykdoma	palūkanoms(%)		
Sutarties numeris	Skaitinis	472696	Vėluoja palūknos(pinigais)	Skaitinis	0
Sutarties pavadinimas	Simbolinis	VART-2005-XXXX-00	Vėluoja paskola(pinigais)	Skaitinis	0
Sutarties sudarymo filialas	Simbolinis	SB	Kiek dienų vėluoja palūkanų grąžinimas	Skaitinis	0
Kliento alga(pingai vienam asmeniui šeimoje)	Skaitinis	1600	Kiek dienų vėluoja paskolos grąžinimas	Skaitinis(0,1)	0
Pakolos dydis	Skaitinis	5000			
Paskolos dydis nacionaline valiuta	Skaitinis	5000			
Suatrties valiuta	Simbolinis	LTL			
Sutarties palūkanų norma	Skaitinis	7,9			
Sutarties pradžia	Data	2005.07.13			
Sutarties pabaiga	Data	2007.07.13			
Suatrties periodas mėnesiais	Skaitinis	36			
Sutarties periodas dienomis	Skaitinis	730			
Suadarymo data	Data	2005.07.13			
Sutarties darbuotojo valstybės kodas	Simbolinis	Lt			
Sutarties darbuotojo gimimo data	Data	1977.06.09			
Sutarties darbuotojo kalba	Simbolinis	LIT			
Sutarties darbuotojo miestas	Simbolinis	Šiauliai			
Kiek dienų vėluoja paskolos grąžinimas	Skaitinis(1,0)	0			

1 lentelė banko klientų vartojamųjų ir būsto paskolų duomenų laukai

Taigi gauti duomenys „Microsoft Excel“ formato rinkmenoje. Visi duomenys buvo pateikiami eilutėmis, o stulpeliai atitiko įrašo reikšmes. Duomenys apie vartojamąsias paskolas sudarė 90 įrašai bei 148 – būsto paskolas.

Be to, dėl saugumo sumetimų nemažai pradinių laukų reikšmių nebuvo įvardintos, pvz., tokios reikšmės, kaip vartotojų ar darbuotojų vardai nebuvo pateikiamos, todėl šių atributų buvo atsisakyta. Taigi buvo išmesti šių duomenų laukai:

- vartojamųjų paskolų – kliento kodas, požymis bazėje (identifikacija), kliento vardas ir pavardė, adresas, sutarties numeris, sutarties pavadinimas;
- būsto paskolų – sutarties numeris, kliento vardas ir pavardė, jo sąskaita, sutarties kodas.

2.6 Duomenų įverčiai

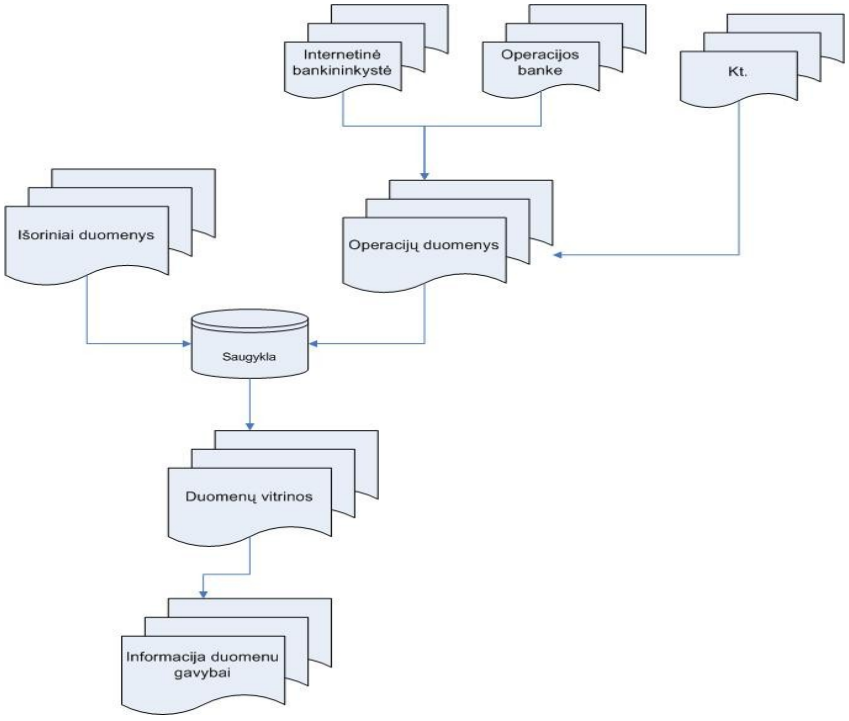
Vėliau, įvairius duomenų tyrimo algoritmus pritaikius informacijai, esančiai duomenų vitrinose, apdoroti, gauti rezultatai naudojami įvairiems įverčiams:

- Įvertinti kredito riziką – t.y., ar galima klientui suteikti kreditą, ar jis bus pajėgus jį grąžinti;
- Įvertinti rinkos riziką – įvertinti ar verta investuoti, ar geriau palaukti, nes šiuo metu investicijos neatneš norimo pelno;
- Pardavimų valdymui – kontroliuoti ir didinti paslaugų pardavimus;
- Portfelio valdymui – efektyviai valdyti finansinį portfelį, taip sumažinti riziką ir padidinti pelno galimybes;
- Kontrolei – kontroliuoti bei gerinti ryšius su klientais.

Kaip minėjome mes sudarinėjome kredito rizikos vertinimo modelius, nes turimi duomenys apibūdino kredito grąžinimo istoriją. Jų sudarymas detaliau aprašytas 3 skyriuje.

2.7 Duomenų srautų schema

Siekdami vizualiai parodyti, kaip duomenys migruoja nuo pradinių iki naudojamų duomenų tyrime, sudarėme duomenų srautų schemą 9 pav.



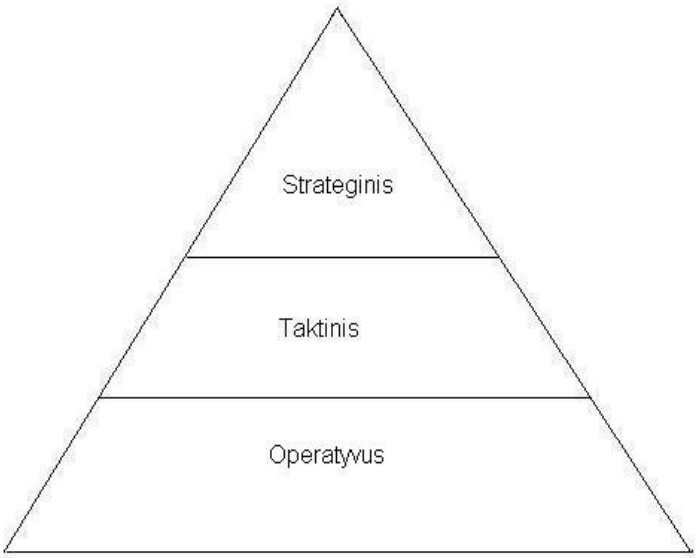
9 pav. duomenų srautai

Šią schemą reikėtų nagrinėti iš viršaus į apačią. Taigi pirmiausia turimi pradiniai duomenys, kurie dažniausiai būna popierinėje ar internetinėje formoje ir dar nėra patalpinti duomenų bazėje. Tačiau užbaigus operaciją visi duomenys patalpinami centralizuotoje duomenų saugykloje. Kaip buvo aprašyta 2.4 skyriuje, šioje saugykloje yra laikomi visi istoriniai duomenys, todėl jų susidaro didelės apimtys bei jų tarpusavio sąryšiai yra sudėtingi bei sunkiai suprantami analitikams. Todėl siekiant, duomenų struktūrų paprastumo kuriamos duomenų vitrinos, kuriose pateikiama apibendrinta, supaprastintos struktūros informacija. Vėliau ji įvairiais formatais perduodama duomenų analizei (mūsų atveju xls formato rinkmena).

3 Duomenų tyrimas

Duomenų tyrimas – tai procesas, kuomet išgaunama anksčiau nežinoma, reikšminga informacija iš didelių duomenų saugyklų, kuri vėliau naudojama svarbiems verslo sprendimams priimti[CCK+99].

Paprastai galima išskirti tris pagrindinius duomenų analizės lygius, jie pateikiami 10 pav.



10 pav – duomenų analizės lygiai.

Operatyvus lygis – analizuojami kiekvienos dienos duomenys. Priimamiems sprendimams taikomos iš anksto suformuotos taisyklės (pvz., negali būti pažeistas kliento kortelės limitas).

Taktinis lygis – duomenys naudojami priimti taktiniams sprendimams (pvz., ar buvo įvykdytas skyriaus mėnesio vartojimo paskolų išdavimo planas).

Strateginis lygis – duomenys naudojami strateginiams sprendimams priimti (pvz., bankas turi peržiūrėti kreditavimo politiką, nes pagal seną modelį išduodamos paskolos neduoda pelno).

Dažniausiai pirmieji du lygiai būna apibrėžti iš anksto nustatytais šablonais, o trečiasis lygis paprastai reikalauja papildomų tyrimų: duomenų šaltiniai analizuojami tam, kad padėtų surasti paslėptus šablonus ir pagerinti banko veiklą. Savo darbe detaliau nagrinėsime šį lygį, bandydami surasti paslėptą informaciją, esančią duomenų saugyklose.

3.1 *Duomenų paruošimas duomenų tyrimui*

Kaip buvo rašyta 1.2 skyriuje, duomenų gryninimas yra atliekamas prieš vykdant duomenų tyrimą, po pirmojo informacijos tvarkymo, kuris vykdomas prieš duomenis talpinant į duomenų

saugyklas. Šis etapas apima papildomą informacijos apdorojimą, nes duomenų tyrime naudojami įrašai turi būti kuo tikslesni. Taigi duomenų paruošimas yra svarbus dėl trijų aspektų [ZZY03]:

1. Realūs duomenys būna su trūkumais:
   - nepilni – trūksta kai kurių naudingų atributų reikšmių;
   - triukšmingi – duomenys gali turėti klaidų ar pašalinių duomenų, kurie yra pertekliniai ir nereikalingi;
   - prieštaringi – duomenų reikšmės neatitinka teiginio;
2. Efektyviam duomenų tyrimui reikalingi kokybiški duomenys. Šiam tikslui dažniausiai naudojamas aktualiausių duomenų išrinkimas:
   - pagal tinkamiausius duomenų atributus, pvz., pagal kelis informatyviausius atributus;
   - anomalijų pašalinimas;
   - besidubliuojančių įrašų pašalinimas.

Atlikus šiuos veiksmus dažniausiai gaunama mažesnė duomenų imtis nei turėtoji pradinė, todėl duomenų tyrimas, naudojant šią imtį, yra efektyvesnis.

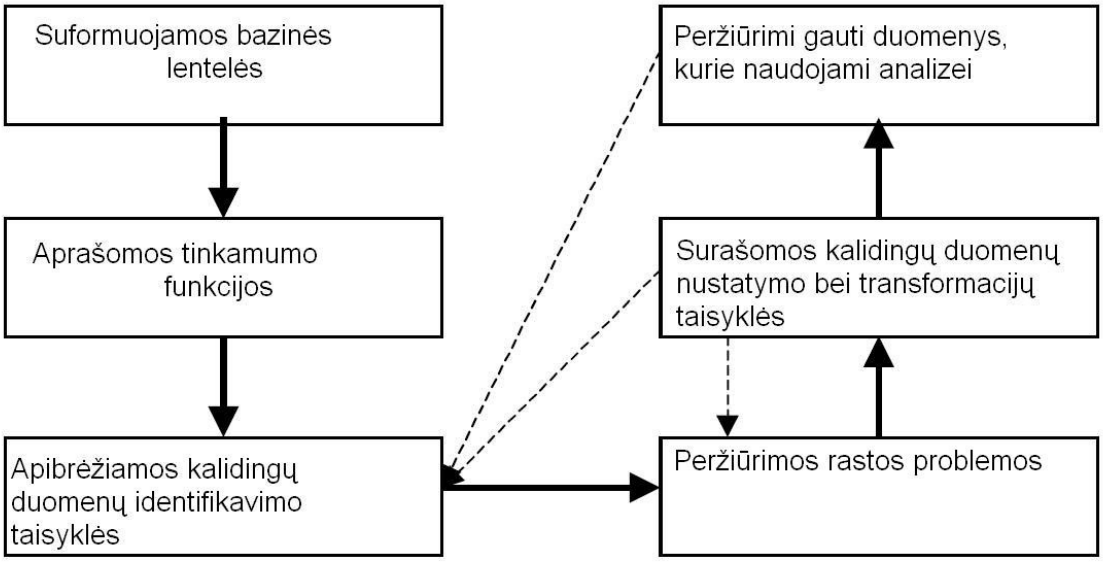
3. Kokybiški duomenys, leidžia gerai atlikti duomenų tyrimą ir gauti patikimus duomenų įverčius. Šis aspektas apibūdina duomenų trūkumų pašalinimą, dažniausiai atliekami šie veiksmai:
   - pataisomi nepilni duomenys – užpildomos trūkstamos atributų reikšmės;
   - duomenys išgryninami – ištaisomos klaidos, pašalinami pertekliniai duomenys, duomenų prieštaringumas;

Taigi, kaip matome, duomenų paruošimo etapas yra labai svarbus, nes jis užtikrina, kad duomenų tyrimas bus efektyvus ir gauti įverčiai bus kuo tikslesni.

[JDR99] pateikia procesą apie duomenų paruošimo vykdymą, kuris pavaizduotas 11 pav.: pajuodintos vientisos linijos rodo eigos kryptį, kurią vykdant gaunami išgryninti duomenys, o punktyrinės linijos galimybę grįžti į ankstesnius etapus, nes aprašytas procesas yra iteracinis. Taigi procesas susideda iš 6 etapų:

1. iš turimų duomenų suformuojamos bazinės lentelės, kurios atitinka pradinę, dar netvarkytą duomenų imtį;
2. apibrėžiamos tinkamumo funkcijos – šios funkcijos nurodo, kokie duomenys yra tinkami, pvz., jei duomenyse apie asmenis nenurodytas asmens kodas, šio įrašo į galutinę imtį įtraukti nereikia, nes jis gali neigiamai įtakoti duomenų tyrimo tikslumą;
3. pagal tinkamumo funkcijas sudaromos taisyklės, kurios padeda nustatyti netinkamus įrašus;
4. pagal 3 etape apibrėžtas taisykles peržiūrimos rastos problemos;

5. surašomos anksčiau apibrėžtos klaidų identifikavimo ir transformacijų taisyklės, kurios aprašo būdus, kaip taisyti klaidingus įrašus. O jei transformacijos taisyklės netenkina, galima grįžti iš šio etapo į 3 ar 4 žingsnį ir kartoti, kol pasiekiamas pageidaujamas rezultatas;
6. peržiūrimi gauti duomenys, jei jie netenkina grįžtama į 3 etapą ir vykdomas tolesnis tobulinimas;



11 pav. procesas duomenų paruošimui.

Be to, jei duomenų tyrimas vykdomas pasitelkiant jau sukurtą programine įranga, kuri naudoja savitą duomenų formatą, kuris transformuojamas iš turimos informacijos(ši transformacija atliekama paruošimo žingsnyje pvz., jei pasirenkama sistema „Weka“, informacija privalo būti transformuota i arff formatą).

3.2 Duomenų apdorojimas

Taigi prieš atliekant duomenų tyrimą buvo sudarytos tinkamumo taisyklės bei vykdomas duomenų patikrinimas remiantis jomis. Tad pradžioje buvo iškelti keli paprasti reikalavimai duomenims:

- datos formatas – yyyy.MM.dd HH:mm:SS (metai.mėnuo.diena valandos:minutės:sekundės) ar yyyy.MM.dd, o jei datos negalima transformuoti įrašoma tuščia reikšmė;
- leistinos valiutos eurai bei litai, o jei valiuta kita įrašoma tuščia reikšmė;
- draudžiamos neigiamos reikšmės – visos skaitinės reikšmės turi būti didesnės ar lygios 0, jei ji mažesnė už nulį įrašomas 0;

Kaip matome, visais duomenų apdorojimo atvejais duomenų įrašai nėra pašalinami iš imties, nes informacijos kiekis yra gana mažas ir kiekvienas įrašas yra svarbus.

Taip pat prieš vykdydami duomenų tyrimą turėjome apdoroti duomenis taip, kad galėtume juos naudoti klasifikavimo ir prognozavimo modeliams sudaryti.

Klasifikavimui naudojami duomenys turėjo būti suskirstyti į klases. Buvo pasirinktos dvi klasės:

1. Blogos paskolos – tos, kurioms vėluojama mokėti įmokas. Jos nustatomos pagal šias taisykles:
    a. vartojamosioms paskoloms – jei grąžinti paskolą vėluojama daugiau nei 31 dieną;
    b. būsto paskoloms – jei gražinti paskolą ar palūkanas vėluojama daugiau nei 40 dienų;
2. Geros paskolos – kai mokėti įmokas vėluojama mažesnį dienų skaičių nei nurodyta 1 punkte arba įmoka sumokama laiku.

Taigi, informacijos imtyje blogųjų paskolų buvo: aštuonios vartojamųjų paskolų ir dvylika būsto.

Prognozavimui naudojami duomenys nebuvo skirstomi į klases, o buvo paliktas skaitinis vėluojamų duomenų skaičius.

3.3 Duomenų analizės įrankis

Prieš pradedant duomenų tyrimą reikėjo pasirinkti, kaip ji bus vykdoma. Buvo nuspręsta nekurti naujo duomenų analizės įrankio, nes jau yra sukurta nemažai patogių bei nemokamų programų, skirtų informacijos analizei. Vienas tokių įrankių yra „Weka“ Ši programa turi nemažai privalumų:

- atvirojo kodo, parašyta Java kalba, todėl jei atsirastų poreikis patobulinti ar praplėsti programą, tai būtų galima nesunkiai įgyvendinti;
- ji suteikia galimybę pasinaudoti standartiniais duomenų filtrais, klasifikavimo ir prognozavimo algoritmais;
- darbas su programa yra nesudėtingas, nes ji turi grafinę sąsają.

Žinoma, įrankis turi ir trūkumų:

- darbui su programa reikia nemažai operatyviosios atminties;
- mažiau galimybių nei didesniuose įrankiuose – pvz., Statistica;

Visgi trūkumai nėra tokie žymūs. Operatyvioji atmintis šiais laikais nėra didelė problema, o trūkstamas funkcijas (jei tokios reikalingos) galima susiprogramuoti patiems.

3.3.1 ARFF formatas

Duomenų formatas, kurį „Weka“ programa naudoja duomenų užkrovimui yra *.arff (angl. Attribute-Relation File Format) [1]. Šis formatas susideda iš antraštės bei duomenų dalies. Antraštėje aprašomas duomenų rinkinio pavadinimas, prieš jį rašomas raktažodis @RELATION, bei duomenų atributų tipai su raktažodžiu @ATTRIBUTE. Jie bus naudojami informacijos dalyje. Po antraštės rašomi raktinis žodis @data bei duomenys, pateikti eilutėmis ir atskirti kableliais. Eilutėje turi būti tiek įrašų, kiek buvo aprašyta atributų antraštės skyriuje.

3.3.2 Duomenų transformavimas

Kaip jau buvo paminėta norint vykdyti duomenų tyrimą naudojant įrankį „Weka“ reikėjo duomenis iš xls formato transformuoti į arff, kuris reikalingas „Weka“ programai. Šiam informacijos apdorojimui buvo parašyta programa, kuri nuskaitinėjo pradinius xls rinkmenos duomenis ir pagal tipą juos konvertuodavo į atitinkamą arff formato failą. Ji buvo sukurta naudojant Java programavimo kalbą. Xls failų nuskaitymui buvo panaudota atvirojo kodo Apache Poi biblioteka, skirta darbui su Microsoft produkto rinkmenomis. Taigi sukurta programinė įranga turi nesudėtingą grafinę sąsają, kurios pagalba galima užkrauti vieną iš dviejų tipų xls rinkmeną (vartojamąsias ar būsto paskolas). Vėliau programa pagal nurodytą tipą apdoroja šį failą bei grąžina suformuotą arff failą pasirinktai duomenų imčiai. Žemiau pateiktame pavyzdyje pavaizduotas sugeneruotas arff failas būsto paskolų klasifikavimui su antrašte ir viena informacijos eilute:

```
@Relation bustoPaskolos
@ATTRIBUTE valiuta  STRING
@ATTRIBUTE paskolosSuma  NUMERIC
@ATTRIBUTE paskolosPrData DATE "yyyy.MM.dd"
@ATTRIBUTE paskolosPabData DATE "yyyy.MM.dd"
@ATTRIBUTE periodasDienomis INTEGER
@ATTRIBUTE periodasMenesiais INTEGER
@ATTRIBUTE klientoAlga NUMERIC
@ATTRIBUTE imoka NUMERIC
@ATTRIBUTE uzstatas NUMERIC
@ATTRIBUTE baudaUzPaskola NUMERIC
@ATTRIBUTE baudaUzPalukanas NUMERIC
@ATTRIBUTE blogaPaskola {0,1}

@Data
"EUR",10000.0,"2007.12.22","2012.12.21",1826.0,60.0,1100.0,166.67,11542.04,0.05,0.01,0
```

Tuo tarpu arff rinkmena apie vartojamąsias paskolas atrodo taip:

```
@Relation vartojamosios
```

```
@ATTRIBUTE klientoMiestoKodas  STRING
@ATTRIBUTE klientoSaliesKodas STRING
@ATTRIBUTE klientoRegistracijosData DATE "yyyy.MM.dd"
@ATTRIBUTE klientoFilialas STRING
@ATTRIBUTE klientoKalba STRING
@ATTRIBUTE klientoStatusas {A}
@ATTRIBUTE klientoAlga NUMERIC
@ATTRIBUTE sutartiesBusena {EXECUTE}
@ATTRIBUTE sutartiesFilialas STRING
@ATTRIBUTE paskolosDydis NUMERIC
@ATTRIBUTE paskolosDydisNacionVal NUMERIC
@ATTRIBUTE paskolosValiuta STRING
@ATTRIBUTE paskolosPalukanuNorma NUMERIC
@ATTRIBUTE paskolosPradzia DATE "yyyy.MM.dd"
@ATTRIBUTE paskolosPabaiga DATE "yyyy.MM.dd"
@ATTRIBUTE paskolosPeriodaisMen INTEGER
@ATTRIBUTE paskolosPeriodaisDienomis INTEGER
@ATTRIBUTE imoka INTEGER
@ATTRIBUTE paskolosSudarymoDats DATE "yyyy.MM.dd HH:mm:ss"
@ATTRIBUTE sutDarbuotSaliesKodas STRING
@ATTRIBUTE sutDarbuotGimData DATE "yyyy.MM.dd"
@ATTRIBUTE sutDarbuotKalba STRING
@ATTRIBUTE sutDarbuotMiestas STRING
@ATTRIBUTE blogaPaskola {0,1}

@Data
2629.0,"LT","2002.01.01","SB","LIT","A",900.0,"EXECUTE","SB",2000.0,2000.0,"LTL",0.0,
"2005.07.13","2007.07.13",24.0,730.0,83.33,"2005.07.13     15:14:31","LT","1977.06.09","LIT",
"Šiauliai" ,0
```

3.3.3 Duomenų konvertavimas ir papildymas

Vėliau, prieš pradedant duomenų tyrimą, informaciją reikėjo papildomai apdoroti, nes programa „Weka“ nepriima simbolinių kintamųjų reikšmių, kadangi sudarinėjant įvairius modelius jų negalima nei diskretizuoti, nei žinoti, kiek skirtingų reikšmių gali įgyti kintamieji. Taigi norint tyrime naudoti simbolinius atributus reikėjo juos transformuoti į nominaliąsias reikšmes. „Weka“ suteikia šios transformacijos galimybę panaudojant filtrą „simbolinė eilutė į nominalias reikšmes“ (angl. StringToNominal). Šis filtras simbolinį kintamąjį konvertuoja į atributą su visomis galimomis reikšmėmis, pvz.:

pradinės reikšmės – @ATTRIBUTE valiuta STRING

reikšmės po filtravimo – @ATTRIBUTE valiuta {EUR,LIT}

Taigi kintamasis „valiuta“ gali įgyti 2 reikšmes, kurios yra apibrėžiamos kaip nominaliosios. Be to, taip apdoroti duomenys tinkami naudoti klasifikavime, bet sudarinėjant skaitinės prognozės modelius šiuos atributus reikėjo pašalinti, nes visos reikšmės turi būti skaitinės.

Šiame etape taip pat buvo nuspręsta papildyti pradinius duomenis papildomais atributais, kurie gali suteikti reikšmingos informacijos konstruojant įvairius klasifikavimo bei prognozavimo modelius. Be to, nauji atributai neiškreips duomenų – jei bus nenaudingi nebus naudojami modelių formavimo metu. Taigi, duomenų šaltiniai buvo papildyti šiais atributais:

- duomenys apie būsto paskolas:
  - trukmė dienomis bei mėnesiais;
- duomenys apie vartojamąsias ir būsto paskolas:
  - preliminari įmoka – paskolos suma padalinta iš mėnesių skaičiaus;
  - algos dalis procentais, kuri tenka paskolai – kokią dalį (%) gaunamos algos sudaro įmoka apskaičiuojama pagal formulę: įmoka*100/alga ;
  - kokia suma lieka kitoms išlaidoms – kokia pinigų suma lieka sumokėjus mėnesinę paskolos įmoką, ši reikšmė lygi: alga – įmoka;

3.3.1 Atributų ir klasės tarpusavio informacija

Pirmiausia prieš sudarinėjant klasifikavimo modelius buvo ieškoma aktualiausių atributų. Šiam tikslui buvo paskaičiuota turimų duomenų atributų bei klasės kintamojo tarpusavio informacija, t.y. kiek kiekvienas iš duomenų atributų susijęs su klasės kintamuoju. Šis dydis lygus:

$I(A,B) = H(A) - H(A \mid B)$, čia H(A) – atributo A entropija, o H(A|B) – sąlyginė A entropija B atžvilgiu.

Taigi, atributų tarpusavio informacija su duomenų klasėmis buvo paskaičiuota pasinaudojant programa „Weka“. Gauti rezultatai pateikti 2 lentelėje, o atributų, kurie nepateikti, tarpusavio informacija lygi nuliui.

Būsto paskolos		**Vartojamosios paskolos**	
Tarpusavio inf.	**atributas**	**Tarpusavio inf.**	**atributas**
0.1581	algosDalis	0.217	algosDalis
0.1547	uzstatas	0.1119	imoka
0.1486	paskolosSuma	0.0816	paskolosPalukanuNorma
0.1177	imoka	0.0193	klientoFilialas
0.1113	liekaPragyvenimui	0.0193	sutDarbuotMiestas
0.0717	klientoAlga	0.0193	sutartiesFilialas
0.0587	periodasDienomis		
0.0257	baudaUzPalukanas		

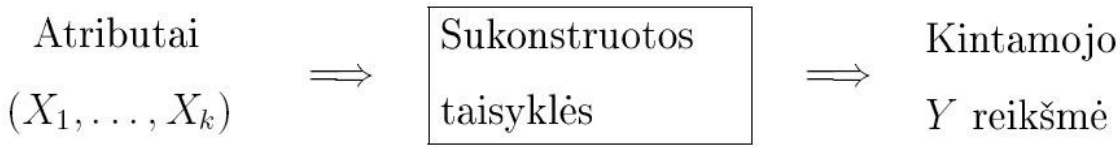
2 lentelė atributų ir klasių tarpusavio informacija.

Taigi pagal 2 lentelėje pateiktus duomenis galėtume teigti, kad mūsų sukurti atributai bus gana svarbūs klasifikavime, nes daugumos jų ir klasės tarpusavio informacija nėra lygi nuliui. Be to

matome, kad atributo „algosDalis“ tarpusavio informacija yra didžiausia, todėl galima daryti prielaidą, kad jis labiausiai įtakos sudaromus modelius. Šiuos teiginius bandysime patikrinti tolimesniuose skyriuose.

3.4 Duomenų klasifikavimas

Duomenų klasifikavimas – tai duomenų suskirstymas į turimas klases naudojant susidarytą klasifikavimo modelį. Jei turima duomenų imtis susideda iš nepriklausomų kintamųjų aibės X ir visų galimų klasių aibės Y, tai klasifikavimo tikslas yra sudaryti funkciją X→Y, kuri vadinama klasifikavimo modeliu, pvz., banko klientai pagal tam tikrus atributus, tokius kaip gaunamos pajamos ir šeimos sudėtis, suskirstomi į pelningus, potencialius investuotojus ar nevertingus. Vėliau imamasi atitinkamų veiksmų norint padidinti banko pelną. 12 pav. pateikiama schema, kuri vaizduoja klasifikavimo eigą [Mac07].



12 pav. Klasifikavimo taisyklių sudarymo schema

Taigi pasinaudodami klasifikavimo modeliu galime spręsti dviejų tipų uždavinius:

- Turimų duomenų suskirstymas į klases. Gautos taisyklės naudojamos kaip priemonė, leidžianti turimus objektus priskirti vienai iš turimų klasių.
- Naujai gaunamų duomenų priskyrimą vienai iš klasių. Turėdami naujo objekto reikšmes, pasinaudodami modeliu, sudarytu pagal turimus duomenis, įrašą galima priskirti vienai iš turimų klasių.

3.5 Klasifikatoriaus tikrinimas

Aktualu įvertinti klasifikatorius, nes sudaryti klasifikavimo modeliai geriausiai tinka tik tai duomenų imčiai, kuria naudojantis jie buvo sukurti. Todėl sudarinėti juos reikėtų naudojantis viena duomenų imtimi (mokymo), o tikrinti – kita (testavimo), kuri nebuvo naudojama kuriant modelį. Dažniausiai mokymui naudojami du trečdaliai duomenų, o testavimui likęs trečdalis [Mac07].

O jei duomenų nedaug, išskaidyti juos į kelias imtis yra gana sudėtinga, pvz.: turimi duomenys apie vartojamąsias paskolas turi tik 8 blogų paskolų įrašus ir jie visi gali pakliūti į

mokymo ar testavimo imtį, todėl dažnai naudojamas kitas algoritmo tikrinimo būdas – kryžminis tikrinimas:

- visa turima imtis skaidoma į k nepersikertančių dalių;
- viena dalis tampa testavimo, o likusios k-1 mokymo imtimis;
- naudojant mokymo imtis sudaromas modelis, o testavimo imtis naudojama modelio tikslumui įvertinti ieškant klasifikavimo klaidos;
- grįžtama į antrąjį žingsnį pakeičiant testavimo bei mokymo imtis, procesas baigiamas po k žingsnių praėjus visas skirtingas testavimo bei mokymo imtis.

Galutinis algoritmo modelio klaidos įvertis gaunamas suskaičiavus visų k modelių klaidų vidurkį. Taigi naudodami šį metodą bandysime rasti geriausią algoritmą – kurio modelių vidutinė klasifikavimo klaida yra mažiausia.

Dar vienas aktualus atributas, įvertinantis modelį, yra nesutapimų matrica[05]. Ji parodo, kiek teisingai ar klaidingai įrašų buvo priskirta turimoms klasėms. Nesutapimų matrica, esant 2 klasėms, pateikiama 3 lentelėje, atitinkamai reikšmės:

a – teisingai klasifikuotų gerų paskolų įrašų skaičius;

c – teisingai klasifikuotų blogų paskolų įrašų skaičius;

c – neteisingai klasifikuotų geros paskolos įrašų skaičius, kai gera paskola klasifikuota kaip bloga;

d – neteisingai klasifikuotų blogos paskolos įrašų skaičius, kai bloga paskola klasifikuota kaip gera;

Klasifikuoti kaip			
Gera paskola	Bloga paskola		
a	b	Gera paskola	Tikroji reikšmė
d	c	Bloga paskola	

3 lentelė nesutapimų matrica

Kaip matome, naudodami nesutapimų matricą galime sužinoti ne tik kiek įrašų buvo klasifikuota neteisingai, bet ir koks jų skaičius buvo priskirtas netinkamoms klasėms. Šie duomenys mums gana aktualūs, nes blogąsias paskolas klasifikavus kaip gerąsias, galimi didesni praradimai nei priešingu atveju.

3.1 Klasifikavimo modelių sudarymo algoritmai

Šiame skyriuje apžvelgsime klasifikavimo algoritmus bei ištirsime, kuris yra tinkamiausias turimiems duomenims.

3.1.1 Klasifikavimo taisyklės

Klasifikavimo taisyklės – tai vienas paprasčiausių, tačiau gana efektyvių klasifikavimo modelių, kuris remdamasis viena ar keliomis klasifikavimo taisyklėmis, priskiria duomenis vienai iš turimų klasių. Klasifikatorius susideda iš vienos ar kelių taisyklių disjunktų, t.y., jei viena taisyklė netenkina, einama prie kitos:

R = (r1 $\bigcup$ r2....) tol, kol įrašas yra priskiriamas vienai iš klasių. Taisyklės išraiška: „jei ..., tai...“ ir susideda iš prielaidos P(r), kuri yra loginių duomenų atributų X1,X2... konjuktai, bei išvados y, kuri apibūdina vieną iš turimų klasių:

P(r)→y, P(R) = t1∩t2..., kur t1=X1 op x1, t2=X2 op x2, ... op yra iš aibės(=,≤,≥,<,>,≠), x1 turimų duomenų atributo reikšmės [Mac07].

Pvz., taisyklės, kurios nustato, ar išduoti klientui paskolą:

algosDalis < 22.805 → bloga paskola,

algosDalis >=22.805 ir < 24.82 → gera paskola,

algosDalis >= 24.82 → bloga paskola

...

Čia kyla klausimas, kokia tvarka reikėtų sudėti taisykles, kad gautume geriausius rezultatus. Keli taisyklių patikimumo kriterijų yra apimtis ir tikslumas. Apimtis a(r) nurodo, kiek iš mokymo duomenų apima pasirinkta taisyklė, o tikslumas t(r) nurodo, kokią dalį iš parinktų įrašų taisyklė klasifikuoja teisingai:

$$a(r) = \frac{P(r)}{|X|} \quad t(r) = \frac{P(r) \bigcup y}{|P(r)|},$$

P(r) – taisyklės apimtų įrašų skaičius,

P(r) $\bigcup$ y – teisingai klasifikuotų įrašų skaičius,

|X| - visų įrašų skaičius.

Taigi, remiantis šiais požymiais, klasifikatoriaus taisykles būtų galima sudėti pagal reitingą. Taip būtų pasiekta, kad pirmiausiai eitų tiksliausios ar turinčios didžiausią apimtį taisyklės. Be šių taisyklių reitingų klasifikatorius tūrėtų tenkinti dar kelis reikalavimus:

- kiekvieną įrašą turėtų apimti tik viena taisyklė, jei jos pasižymi šia savybe, tai sakoma, kad jos poromis nesutaikomos;
- bet kurį mokymo įrašą turėtų apimti bent viena taisyklė, jei klasifikatorius tenkina šį reikalavimą, tai klasifikatoriaus taisyklės sudaro pilną rinkinį.

3.1.1.1 1R algoritmas

1R (angl. One Rule) vienas iš paprasčiausių, tačiau gana veiksmingų klasifikavimo taisyklių sudarymo algoritmų. Algoritmas pagal turimus mokymo duomenis atrenka didžiausią klasifikavimo tikslumą turintį duomenų atributą ir pagal jį suformuoja klasifikatoriaus taisyklę.

Sudarinėjant modelį visos skaitinių atributų reikšmės yra diskretizuojamos – suskirstomos į intervalus. Diskretizavimo algoritmas yra gana nesudėtingas[Hol93], kurio žingsniai yra šie:

1. atributo reikšmių surūšiavimas;
2. sąrašo suskaldymas kiekvienoje vietoje, kurioje keičiasi klasės reikšmės;
3. gretimų intervalų, identifikuojančių tas pačias klases, suliejimas;
4. jei intervalai per smulkūs, vykdomas tikslumo pokyčio apskaičiavimas, suliejus intervalus su skirtingomis klasėmis ir priskyrus įrašus daugumos klasei;
5. intervalų suliejimas su mažiausiu tikslumo praradimu.

4 bei 5 žingsniai kartojami kol gaunamas norimas intervalų dydis, jis dažniausiai nustatomas pagal skaičių n_{disk}, kuris nurodo kiek intervale turi būti dažniausiai sutinkamos klasės įrašų. Rekomenduojama, kad kiekviename intervale šis skaičius būtų lygus šešiems, kai mokymo imties įrašų skaičius viršija 50, o jei imtis mažesnė – užtenka 3 reikšmių. Mes naudojome tik 4 klasės reikšmes, nes turėjome nedaug blogų paskolų duomenų, o naudojant didesnį skaičių galėjome negauti nė vienos blogų paskolų imties.

Taigi, 1R algoritmo struktūra[MHW03] yra tokia:

- kiekvienam duomenų atributui suformuoti taisyklę:
  - paimti skirtingas atributo reikšmes;
  - išrinkti iš mokymo duomenų visus įrašus, kurie įgyja pasirinktas atributo reikšmes;
  - paimti dažniausiai pasikartojančią klasę y_i
  - suformuoti taisyklę: jei įrašo atributas įgyja pasirinktą reikšmę, tai jis priskiriamas pasirinkai klasei y_i;
- atributo klasifikacijos tikslumo apskaičiavimas;

Galiausiai panaudojama taisyklė, kuri turi didžiausią klasifikacijos tikslumą.

Taigi, pasinaudoję programos „Weka“ galimybėmis 1R alogoritmo įgyvendinimui, gavome tokius rezultatus:

- vartojamųjų paskolų taisyklė:

  algosDalis < 22.805 → gera paskola,

  algosDalis >=22.805 ir < 24.82 → bloga paskola,

  algosDalis >= 24.82 → gera paskola

Ji teisingai klasifikuoja 86 iš 90 arba 95.56% įrašų pradinėje imtyje.

- būsto paskolų taisyklė:

imoka < 189.58	→ gera paskola
189.58<= imoka < 194.1	→ bloga paskola
194.1 <= imoka < 212.505	→ gera paskola
212.505<= imoka < 228.475	→ bloga paskola
imoka >= 228.475	→ gera paskola

Ji teisingai klasifikuoja 146 iš 148 arba 98.6% įrašų pradinėje imtyje.

Kaip matome, sudarytos taisyklės gana tiksliai klasifikuoja pradinę imtį. Tačiau jei atkreipsime dėmesį į taisykles logine prasme jos nėra labai patikimos(pvz., vartojamųjų paskolų taisyklėje gera paskola yra tada, kai algos dalis yra tam tikrame intervale), todėl didelė tikimybė, kad šie modeliai gana prastai klasifikuos naujus duomenis. Taigi siekdami pagerinti modelius nusprendėme padidinti dažniausiai sutinkamos klasės įrašų skaičių dvigubai bei gavome tokius rezultatus:

- vartojamųjų paskolų taisyklė nepakito;
- o būsto paskoloms buvo gauta tokia taisyklė:

algosDalis < 44.965	→ gera paskola,
algosDalis >= 44.965	→ bloga paskola,

Ji teisingai klasifikuoja 138 iš 148 arba 93.2% įrašų pradinėje imtyje.

Kaip matome, būsto paskolų tiesė pasidarė tinkamesnė naudojimui, nes pagal algos dalį mes galime spręsti ar paskola gera ar bloga, todėl būtų galima daryti prielaidą, kad ši tiesė bus patikimesnė nei pirmoji gauta naudojant 4 dažniausiai sutinkamo klasės kintamuosius. Tačiau vartojamųjų paskolų tiesė nepasikeitė, o toliau didinant klasės narių skaičių algoritmą visas paskolas priskiria gerosioms, todėl galima teigti, kad vartojamųjų paskolų duomenų klasifikavimui šis algoritmas nėra tinkamas.

Vėliau buvo apskaičiuotas klasifikatoriaus tikslumas naudojant kryžminį patikrinimą. Skaidinių preliminariai buvo sudaryta tiek, kad jų skaičius būtų lygus pusei blogų paskolų įrašų skaičiui kiekvienoje imtyje – 4 skaidiniai vartojamųjų paskolų duomenims, o būsto – 6. Pasinaudojus programa „Weka“ gauti tokie rezultatai:

- vartojamųjų paskolų imtyje – vidutiniškai teisingai klasifikuojami 84, o klaidingai 6 įrašai; apytiksliai gaunamas 93.3% tikslumas;
- būsto paskolų imtyje – vidutiniškai teisingai klasifikuojami 138, o klaidingai 10 įrašų; apytiksliai gaunamas 93.2% tikslumas;

Modelio kryžminio tikrinimo metu kiekvieniems duomenims taip pat buvo sudarytos nesutapimų matricos iš vidutinių reikšmių. Vartojamųjų ir būsto paskolų rezultatai atitinkamai pateikti 4 bei 5 lentelėje.

Klasifikuoti kaip			
Gera paskola	Bloga paskola		
78	4	Gera paskola	Tikroji reikšmė
2	6	Bloga paskola	

4 lentelė 1R algoritmo nesutapimų matrica vartojamųjų paskolų duomenims

Klasifikuoti kaip			
Gera paskola	Bloga paskola		
134	2	Gera paskola	Tikroji reikšmė
8	4	Bloga paskola	

5 lentelė 1R algoritmo nesutapimų matrica būsto paskolų duomenims

Kaip matome klasifikavimui naudojant 1R algoritmą gaunami pakankamai geri rezultatai. Bet jei atkreiptume dėmesį į būsto paskolų nesutapimų matricą, pamatytume, kad vidutiniškai algoritmais gana prastai nustato blogąsias paskolas(tik 4 teisingai klasifikuotos klasės) jas dažniau klasifikuodama, kaip gerąsias. Visi šie neatitikimai tikėtina yra dėl to, kad modelyje naudojamas tik vienas iš duomenų imties atributų, o mūsų turimose imtyse apsispręsti pagal vieną atributo reikšmę sudėtinga, todėl klasifikavimo metu reikėtų atsižvelgti į kelias atributo reikšmes.

3.1.1.2 RIPPER algoritmas

RIPPER – tai dar vienas klasifikavimo taisyklių sudarymo algoritmas. Šis metodas naudoja visus atributus bei išgauna geresnį tikslumą, kai duomenys yra triukšmingi, nes modelio konstravimo metu naudoja kontrolinę imtį.

Taigi pradžioje surūšiuojami visi klasių įrašai pagal dažnumą – A={Y1,Y2...Yk}, čia Y1 – rečiausiai mokymo imtyje sutinkama klasė, o Yk – dažniausiai. Visų pirma klasei Y1 priklausantys įrašai priskiriami teigiamiems, o visi likę – neigiamiems. Vėliau konstruojamos taisyklės teigiamiems įrašams klasifikuoti. Kitame žingsnyje pereinama prie klasės Y2 taisyklės formavimo ir taip, kol ateinama iki paskutinės klasės Yk. Kiekviena taisyklė konstruojama taip: pradžioje prijungiami nauji konjunktai tol, kol taisyklė pasidaro tiksli mokymo duomenims, t.y. jos tikslumas lygus 1. Po to taisyklė redukuojama priklausomai nuo to, kaip ji klasifikuoja kontrolinės imties įrašus pagal šį matą:

$g(r) = \frac{v^+ - v^-}{v^+ + v^-}$, čia taisyklės:

v^+ – teisingai klasifikuojamų klasės įrašų skaičius,

v^- – klaidingai klasifikuojamų klasės įrašų skaičius.

Taigi redukavimas vyksta pašalinant konjunktus iš taisyklės. Tuomet, jei matas g(r) padidėja, paliekama naujoji taisyklė. Taip redukuojamos visos taisyklės, kol gaunamas optimalus taisyklių sąrašas, geriausiai klasifikuojantis kontrolinės imties duomenis.

Mūsų turimų duomenų atveju buvo tik 2 taisyklės, kurios klasifikavo blogąsias ir geràsias paskolas. Kadangi blogųjų paskolų klasės duomenų buvo mažiau, tai RIPPER algoritmas pirmiausia bandė klasifikuoti būtent jas, o likusieji duomenys buvo priskirti gerosioms paskoloms. Taigi naudojant programą „Weka“ buvo gauti tokie rezultatai:

- vartojamųjų paskolų duomenų taisyklės:

  (algosDalis >= 23.08) and (klientoAlga <= 2000) → blogaPaskola(12.0/4.0)

  kiti duomenys → gera paskola(78/0.0) – skliaustuose esantys skaičiai nurodo, kiek taisyklė klasifikuoja įrašų, o po pasvirusio brūkšnelio nurodoma, kiek įrašų klasifikuojama klaidingai.

  Šios taisyklės teisingai klasifikuoja 86 iš 90 arba 95,6% įrašų;

- būsto paskolų taisyklės:

  (uzstatas >= 8200.41) and (imoka >= 191.67) and (klientoAlga <= 820) and (imoka <= 227.78) → blogaPaskola (10.0/0.0)

  kiti duomenys → gera paskola(138/2.0)

  Šios taisyklės teisingai klasifikuoja 146 iš 148 arba 98.6% įrašų.

Kaip matome šių taisyklių tikslumas panašus į 1R algoritmo sudaryto modelio, tačiau joms sudaryti naudojamos kelios atributų reikšmės, o tai leidžia daryti prielaidą, kad šie modeliai bus patikimesni.

Taip pat buvo atliktas algoritmo tikrinimas naudojant kryžminį patikrinimą su 4 bei 6 skaidiniais atitinkamai vartojamosioms bei būsto paskolų duomenims ir gauti tokie rezultatai:

- vartojamosios paskolos – vidutiniškai teisingai klasifikuojami 83, o klaidingai – 7 įrašai; gaunamas apytiksliai 93.3% tikslumas;
- būsto paskolos – vidutiniškai teisingai klasifikuojami 144, o klaidingai – 4 įrašai; gaunamas apytiksliai 97.3% tikslumas.

Abiejų duomenų imčių nesutapimų matrica pateikta 6 lentelėje.

Vartoj. paskolos klasifikuota kaip			Būsto paskolos klasifikuota kaip			
Gera	Bloga		Gera	Bloga		
77	5	Gera	134	2	Gera	Tikroji

2	6	Bloga	2	10	Bloga	reikšmė

6 lentelė RIPPER algoritmo nesutapimų matrica turimoms duomenų imtims

Taigi šis algoritmas bei juo remiantis sudarytas modelis duoda panašius, kartais net prastesnius, rezultatus (blogesni duomenys kryžminio patikrinimo metu būsto paskoloms) nei algoritmas 1R. Todėl, jei nauji duomenys bus labai panašūs į duotuosius, mums pakaks su 1R algoritmu suformuotos taisyklės, tačiau klasifikavimo modelis suformuotas su RIPPER algoritmu bus patikimesnis, nes atsižvelgia į didesnį duomenų požymių skaičių ir taip iškeliami didesni reikalavimai.

3.1.2 Sprendimų medžiai

Vienas iš paprasčiausių ir dažniausiai naudojamų klasifikavimo modelių yra sprendimų medžiai. Modelį sudaro medžio struktūros klasifikatorius, kurio kiekviena viršūnė atitinką vieną iš duomenų įrašų atributų, o šakos nurodo atributo reikšmes.

Viršūnė, kuri turi būti aukštesniame hierarchijos lygyje, dažnai nustatinėjama naudojant atributų požymių priklausomybę nuo ieškomos klasės. Kuo šie du požymiai labiau priklauso vienas nuo kito, tuo pasirinkto požymio viršūnė yra aukščiau hierarchijoje. Pvz., [CPH+08] straipsnyje aprašomas rizikos vertinimas, ar banko klientai sugebės grąžinti paimtą būsto paskolą. Atributas, kuris turi didžiausią įtaką rizikos nustatymui, yra konstruojamo medžio viršuje. Taigi straipsnyje tyrimais nustatoma, kad pajamų ir įmokos santykis turi didžiausią reikšmę paskolos grąžinimo galimybei. Taigi ši viršūnė pati aukščiausia medžio struktūroje. Taip pagal informatyvumą sudarinėjamos ir kitos medžio viršūnės.

Vėliau, sudarius pilną sprendimų medį, duomenų klasifikavimas vykdomas iš aukščiausios višūnės einant žemėjančia tvarka, kol pasiekiamas lapas, kuris atitinka vieną iš turimų klasių. Be to, kiekvienas iš šių kelių atitinka klasifikavimo taisyklę.

3.1.2.1 C4.5 algoritmas medžio konstravimui

C4.5 vienas iš populiariausių sprendimų medžio sudarymo algoritmų. Algoritmas pagal mokymo duomenų požymius bei norimas gauti klases suformuoja sprendimų medį. Kaip anksčiau buvo minėta naujai sudarytas medis turi pasižymėti tokiomis savybėmis:

- kievienas medžio lapas atitinka klasę;
- viršūnės atitinka turimų duomenų požymius, o remiantis jų reikšmėmis, viršūnė yra sujungta su žemesniu medžiu;

Medžio viršūnių formavimas vykdomas rekursiškai naudojant mokymo duomenų imtį [KOQ09]:

1. jei turimos imties T įrašai priklauso vienai klasei Yj, tuomet sprendimų medžio viršūnė yra lapas atitinkantis klasę Yj;
2. jei imties T įrašai priklauso daugiau nei vienai klasei, paimamas informatyviausias turimų duomenų požymis ir pagal jo reikšmes viršūnė skaidoma į du ar daugiau vaikus. Jiems priskiriami imties T poaibiai, sudaryti pagal pasirinkto požymio reikšmes. Šis algoritmas naudojamas kiekvienam naujai gautam medžio vaikui;

Jei antrajame žingsnyje, patikrinus visus atributus, negaunama viršūnė, kuri turėtų tik vienos klasės reikšmes – klasė priskiriama lapui pagal tai, kokių jos įrašų yra daugiausia nagrinėjamoje viršūnėje.

Vėliau, sudarius medį, vyksta jo tobulinimas, t.y. viršūnių skaičiaus mažinimas. Šis žingsnis padeda supaprastinti medžio struktūrą, kai duomenys susideda iš daug klasių bei atributų.

3.1.2.2 C 4.5 kredito vertinimo modelių apžvalga

Šis algoritmas jau yra gana populiarus bei dažnai naudojamas bankinių duomenų vertinimui. [CGP04] naudoja šį algoritmą banko pelno galimam pelno (nuostoliams) įvertinti. Trumpai apžvelgsime jo gautus rezultatus, vėliau bandysime gauti sprendimų medį pagal mūsų turimus požymius skirtą kredito rizikai įvertinti. Taigi straipsnyje [CGP04] naudojami požymiai yra šie:

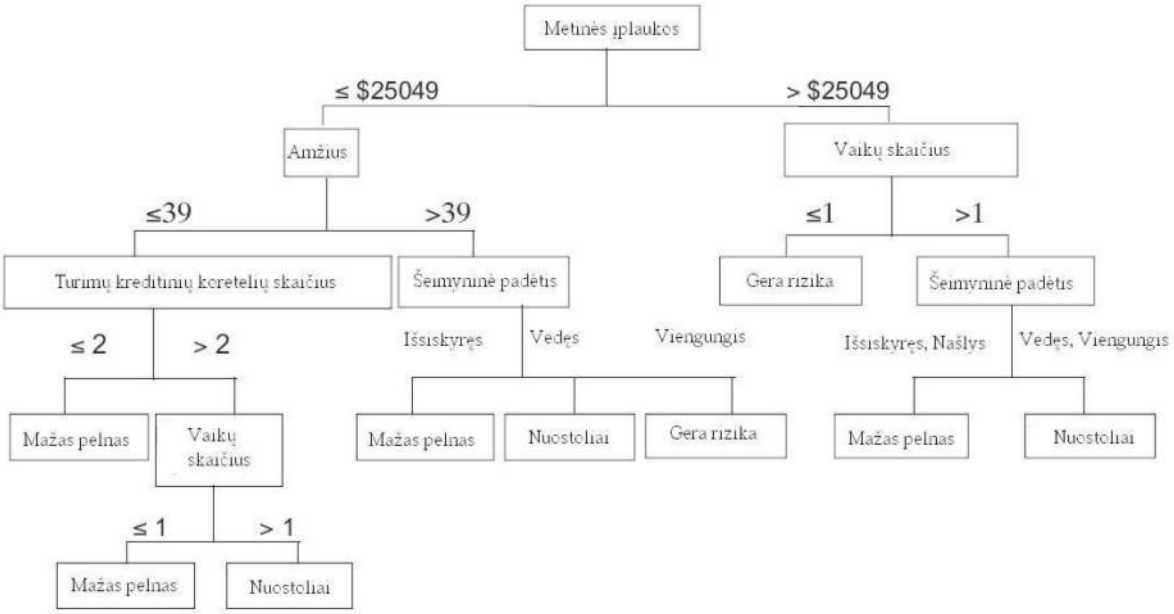
- amžius;
- metinės įplaukos;
- šeimyninė padėtis;
- vaikų skaičius;
- turimų kredito kortelių skaičius;
- klientas turi pasiėmęs būsto paskolą;

Galimos trys išeities klasės:

- gera rizika – klientas bus pelningas, jam pakolą išduoti galima;
- mažo pelno – klientas neduos didelio pelno, gali neatiduoti paskolos.
- nuostoliai – klientas tikriausiai bus nuostolingas, paskolos išduoti nerekomenduojama;

Pagal autoriaus duomenis sugeneruotas klasifikavimo medis, pateiktas 13 pav. Taigi, pradedat stebėti įrašus nuo medžio viršunės iki lapo ir tikrinant kiekvieną iš jų, galima nesunkiai klasifikuoti turimus ar naujai ateinančius duomenis. Taip pat kiekvieną kelią nuo aukščiausios

viršūnės iki lapo galime sudaryti kaip atskiras taisykles, pvz., geros rizikos klientai gali būti tie, kurių metinės pajamos viršija 25049$ ir kurie neturi vaikų arba turi vieną vaiką. Iš viso šis klasifikavimo medis teisingai klasifikuoja 76% testavimo imties duomenų.



13 pav. Sprendimų medis mokumo rizikai įvertinti.

Tačiau matome, kad šis medis nėra tinkamas mūsų duomenims, nes dalies požymių apie klientus, mes neturime todėl negalime klasifikuoti turimų įrašų, be to šis medis sudarytas naudojant kitos šalies bankinius duomenis.

Šis algoritmas taip pat yra naudojamas autoriaus [IG01] analizuojant banko kliento paskolos grąžinimo riziką t.y. ar klientas sugebės atiduoti paskolą ar ne. Autorius pateikia geriausias taisykles gautas suformavus medį naudojant C 4.5 algoritmą.

Nagrinėjant duomenis buvo naudojam 13 duomenų požymių 7 iš jų buvo skaitiniai, o 6 kategoriniai:

- skaitiniai kintamieji – amžius, kiek metų žmogus gyvena toje pačioje vietoje, kiek metų žmogus dirba tame pačiame darbe, įplaukos, paskolos dydis, Kreditas/(laikotarpis*įplaukos), paskolos laikotarpis metais
- kategoriniai kintamieji – ar klientas vedęs, ar turi vaikų, ar turi automobilį, ar turi nusipirkęs gyvenamąją būstą, jo profesija, gyvenamasis miestas.

Pagal turimus požymius įrašai suskirstomi į dvi klases: pirmoji klasė reiškia, kad klientas grąžino pasiimtą paskolą, antroji, kad paskola nebuvo grąžinta. Apmokymo duomenis sudaro 1300 įrašų iš kurių 909 yra pirmosios klasės ir 391 antrosios klasės įrašai.

Taigi naudojant C 4.5 klasifikavimo algoritmą buvo gautas sudėtingas medis, autorius jo nepateikia, tačiau yra pateikiamos iš jo sudarytos 187 klasifikavimo taisyklės. 7 lentelėje surūšiuotos taisyklės mažėjimo tvarka atsižvelgiant į aktualumą, apibūdinantį taisyklės naudingumą, ar ji yra lengvai suprantama, netikėta, novatoriška. Aktualumą siūloma skaičiuoti pagal formulę:

RI = P(A,B) – P(A)P(B),

kur A→B

A – yra kairėje taisyklės pusėje esantys konjunktai

B – prognozuojama klasė

Taisyklė	Klasė	Apimtis	Tikslumas (%)	Aktualumas
Taisyklė 5: kreditas/(laikotarpis*įplaukos)<=0.964258, paskolos laikotarpis<=5	1	263	97.2	0.125879
Taisyklė 108: amžius<=25, turi vaikų = 0, ar turi nusipirkęs gyvenamąją būstą = 0, paskolos dydis > 1073, kreditas/(laikotarpis*įplaukos)<=0.256667, paskolos laikotarpis > 8	2	28	90.9	0.107183
Taisyklė 86: paskolos dydis > 1073, paskolos laikotarpis > 8	2	451	54.6	0.104145
Taisyklė 2: kiek metų žmogus gyvena toje pačioje vietoje<=24, paskolos laikotarpis <= 5	1	288	96.3	0.097437
Taisyklė 7: kiek metų žmogus gyvena toje pačioje vietoje<=24, paskolos laikotarpis > 2, paskolos laikotarpis <= 5	1	288	97.7	0.097437
Taisyklė 37: turi automobilį = 0, paskolos dydis > 619, gyvenamasis miestas=BOLU	2	18	92.6	0.096817
Taisyklė 101: turi vaikų = 1, įplaukos<= 1604, paskolos laikotarpis > 8, gyvenamasis miestas=BOLU	2	18	92.6	0.096817
Taisyklė 61: vedęs=1, turi automobilį = 0, gyvenamasis miestas=BOLU	2	17	92.2	0.086866
Taisyklė 182:	2	86	76.5	0.086866

kreditas/(laikotarpis*įplaukos)<=0.18638, paskolos laikotarpis >= 8, profesija = dirbantis savarankiškai				
Taisyklė 139: įplaukos > 612, paskolos dydis > 1073, paskolos laikotarpis > 8	2	331	54.8	0.078665

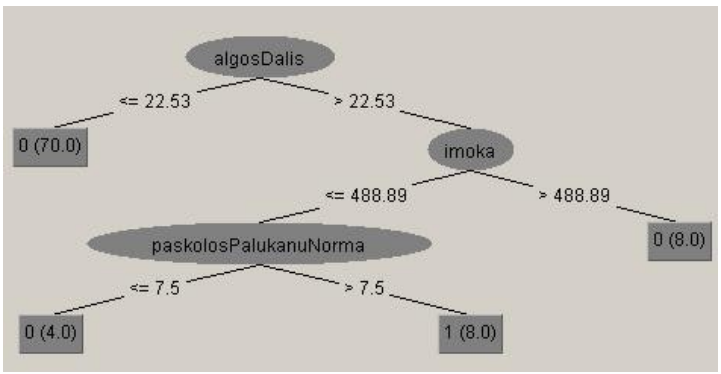
7 lentelė – 10 aktualiausiu taisyklių banko klientų paskolų duomenims

Taigi, iš lentelėje pateiktų duomenų matome, kad dalis taisyklių galėtų klasifikuoti kai kuriuos mūsų turimus duomenis pvz., taisyklė 5. Tačiau taip būtų klasifikuojama tik nedidelė turimų įrašų dalis bei, kaip minėjome anksčiau, šios taisyklės yra tinkamos užsienio rinkos duomenų imčiai, kuria naudojantis jos buvo sudarytos, todėl turimiems įrašams jos nėra tinkamos.

3.1.2.3 C 4.5 sprendimų medžiai turimoms duomenų imtims

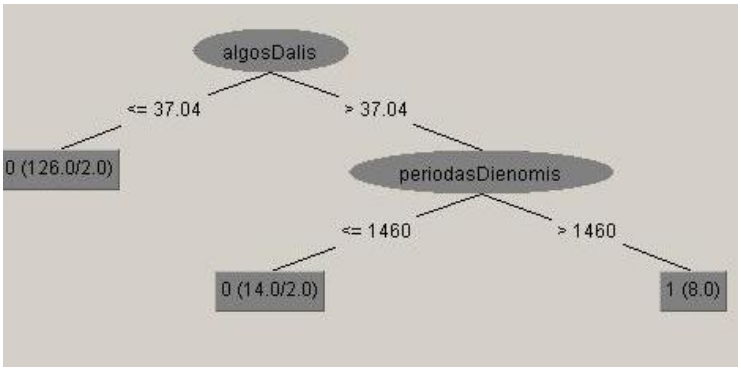
Taigi, nustačius, kad sudarytieji modeliai netinka mūsų turimiems duomenims, naudojantis programa „Weka“ buvo sudaryti sprendimų medžiai, kurie pagal turimus požymius geriausiai klasifikuoja turimas duomenų imtis, gauti tokie rezultatai:

- vartojamųjų paskolų informacija – 14 pav. vizualiai vaizduojamas sprendimų medis klasifikuojantis šią duomenų imtį. Lapuose gerą paskolą simbolizuoja 0, o blogą - 1. Šis medis teisingai klasifikuoja visus turimus įrašus, tačiau, žinoma, mes šiais rezultatais negalime pasitikėti visu 100%. Kaip buvo minėta kiekvienas kelias nuo medžio viršūnės iki šaknies atitinka klasifikavimo taisyklę, taigi iš šio medžio gaunamos keturios taisyklės:
  - algosDalis<=22.53 → gera paskola(70);
  - algosDalis>22.53 ir imoka > 488.89 → gera paskola (8);
  - algosDalis>22.53 ir imoka <= 488.89 ir paskolosPalukanuNorma <=7.5 → gera paskola (4).
  - algosDalis>22.53 ir imoka <= 488.89 ir paskolosPalukanuNorma > 7.5 → bloga paskola (8).



14 pav. sprendimų medis vartojamosioms paskoloms

- būsto paskolų informacija – 15 pav. vizualiai vaizduojamas sprendimų medis, klasifikuojantis šią duomenų imtį. Kaip ir vartojamųjų paskolų atveju gerąją paskolą atitinka 0, blogą – 1. Taip sudarytas medis teisingai suskirsto 144 iš 148 arba 97.3% duomenų. Taip pat šis medis sudaro tris klasifikavimo taisykles:
  - algosDalis<=22.53 → gera paskola(126/2);
  - algosDalis<=22.53 ir periodasDienomis <=1460 → gera paskola(14/2);
  - algosDalis<=22.53 ir periodasDienomis >1460 → bloga paskola(8);



15 pav. sprendimų medis būsto paskoloms

Vėliau buvo vykdomas algoritmo tikrinimas naudojant kryžminį patikrinimą, vartojamosioms paskoloms naudojant 4 skirstinius, o būsto – 6. Gauta nesutapimų matrica pateikta 8 lentelėje. Kaip matome nesutapimų matricoje apie vartojamąsias paskolas vidutinis tikslumas yra 92.2% (83 teisingai klasifikuotų įrašų iš 90), jis nebėra lygus 100%, tačiau yra pakankamai geras įvertis.

Vartoj. paskolos klasifikuota kaip			Būsto paskolos klasifikuota kaip			
Gera	Bloga		Gera	Bloga		
78	4	Gera	132	4	Gera	Tikroji reikšmė
3	5	Bloga	4	8	Bloga	

8 lentelė klasifikavimo medžio nesutapimų matrica duomenų imtims

Taigi mūsų nagrinėti sprendimų medžiai sudaryti algoritmo C 4.5 pagalba gerai klasifikuoja turimas duomenų imtis. Be to, šis metodas vizualiai pateikia suformuotas taisykles, todėl turint tokį modelį klasifikavimas tampa paprastesnis. Kaip matome ir šiame metode mūsų suformuotas atributas „algosDalis“ yra aukščiausioje medžio struktūroje, todėl galima teigti, kad jis yra vienas svarbiausių klasifikavimo atributų. Taip pat iš nesutapimų matricos matome, kad blogųjų paskolų priskirtų gerosioms skaičius yra nedidelis, todėl modelis tūrėtų būti gana patikimas.

3.1.3 Klasifikavimo modelių apibendrinimas

Ištirtų algoritmų bei juos naudojant sudarytų modelių tikslumai pateikti 9 lentelėje. Kaip matome iš lentelėje pateiktų duomenų visi sudaryti modeliai gana tiksliai klasifikuoja turimus duomenis, tačiau, kaip buvo minėta anksčiau 1R algoritmas nėra tinkamas. O sprendimas panaudoti klasifikavimo taisyklių algoritmą RIPPER buvo sėkmingas – tiek modeliai parodo gana gerą rezultatą, tiek kryžminio tikrinimo metu gaunami gana patikimi duomenys.

Blogųjų paskolų klasifikavimui taip pat buvo ištirtas sprendimų medžių sudarymo metodas C4.5. Jis buvo pasirinktas, nes yra vienas populiariausių ir dažniausiai naudojamų klasifikavimo medžių sudarymo algoritmų. Naudojant jį yra sukurta klasifikavimo modelių skirtų kredito rizikai įvertinti, tačiau atlikus jų analizę paaiškėjo, kad jie nėra tinkami turimai informacijai, nes duomenų požymiai skiriasi nuo naudojamų sukurtuose modeliuose. Todėl naudojant šį algoritmą sudarėme modelius, geriausiai nustatančius blogąsias paskolas turimose duomenų imtyse, ir gavome gana gerus rezultatus.

Algoritmas	**Modelio tikslumas**		**Algoritmo tikslumas (kryžminio patikrinimo būdu)**	
	Vartojamosios paskolos	**Būsto paskolos**	**Vartojamosios paskolos**	**Būsto paskolos**
Klasifikavimo taisyklės – 1R	95.6%	93.2%	93.3%	93.2%
Klasifikavimo taisyklės – RIPPER	95.6%	98.6%	92.2%	97.3%
Sprendimų medžiai – C4.5	100%	97.3%	92.2%	94.6%

9 lentelė klasifikatorių bei jų modelių tikslumai

3.2 Skaitinė prognozė

Kaip ir klasifikavimo uždavinyje prognozėje naudojami tokie patys vartojamųjų ir būsto paskolų duomenys. Prognozės tikslas sudaryti kuo tikslesnę funkciją Y=f(x), kuri identifikuoja turimus duomenis, kur:

Y – priklausomas klasės kintamasis (vėluojamų mokėti paskolos įmoką dienų skaičius),

X – nepriklausomi kintamieji (kiti duomenų atributai).

Sudarinėjant prognozavimo funkciją siekiama sumažinti modelio nuostolių funkciją, kuri nurodo, kaip tiksliai modelis įvertina turimus įrašus.

3.2.1 Daugialypė tiesinė regresija

Daugialypę tiesinę regresiją sudaro tiesinė prognozavimo funkcija:

$$f(X) = B_0 + \sum_{j=1}^{k} B_j X_j ,$$

čia X – nepriklausomi kintamieji, B – koeficientai, k – atributų skaičius. Norėdami gauti lygties koeficientus B_k turėsime minimizuoti kvadratinę nuostolių funkciją:

$$L(Y, f(X)) = \quad y - \quad (X))^2 ,$$

Ši funkcija parodo, kiek išprogramuotos reikšmės neatitinka realių įverčių. Funkciją galima užrašyti matricų pavidalu. Nepriklausomus kintamuosius žymėsime vektoriumi X, jį papildysime vienetine koordinate $X=\{1,X_1 \ldots X_k\}$, o koeficientus žymėsime matrica $B=\{B_0,B_1 ... B_k\}$, klasės kintamųjų matrica $Y=\{y_1,...,y_k\}$. visų kintamųjų matricą sudarytą iš vektorių $x_1,x_2,...,x_n$ žymėsime Q:

$$Q = \begin{pmatrix} 1 & X_{11} & \ldots & X_{1k} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & X_{n1} & \ldots & X_{nk} \end{pmatrix}$$

Tada koeficientus B randame minimizuodami anksčiau aprašytą kvadratinę nuostolių funkciją:

$$B = \sum_{i=1}^{n} (y_i - f(X_i))^2 \quad = \sum_{i=1}^{n} (y_i - BX_i^T)^2 \ .$$

X^T žymi transponuotą matricą, tada B išreiškiame naudojami matricų sandaugą:

$$B = \quad y - \quad B^T)^T (y - \quad B^T) .$$

Vėliau išdiferencijavę šią funkciją pagal B ir prilyginę nuliui gauname sprendinį lygų:

$$B = \quad ^T Q(Q^T X)$$

Ir galiausiai gauname tiesinės regresijos lygtį:

$$y(x) = \quad X^T .$$

3.2.1.1 Regresijos tiesės charakteristikos

Taigi sudarius prognozavimo lygtį jos patikimumui įvertinti yra naudojami keli parametrai. Apibrėšime keletą populiariausių ir aktualiausių prognozavimo modelio charakteristikų, apibūdinančių regresijos lygties patikimumą.

Pirmasis parametras – tai absoliutinė vidutinė paklaida, kuri nurodo, kiek vidutiniškai skiriasi realios reikšmės nuo prognozuojamųjų su regresijos tiese. Šis parametras lygus:

$$AVP = \frac{\sum_{i=1} s(y(x) - \cdot)}{n}$$

Kita charakteristika, parodanti, kaip stipriai tikrosios klasės reikšmės susijusios su išprognozuotomis, yra koreliacijos koeficientas:

$$\rho = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_{ti} - \bar{y}_{ti})(y_{pi} - \bar{y}_{pi})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ti} - \bar{y}_{ti})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{pi} - \bar{y}_{pi})^2}}$$

čia y_t – tikrosios reikšmės, o y_p – išprognozuotosios, atitinkamai $\bar{y}_t$, $\bar{y}_p$ – tikrųjų ir išprognozuotųjų klasių vidurkiai. Koreliacijos reikšmė kinta tarp -1 ir 1, nuo reikšmės priklauso priklausomybės stiprumas, taigi reikšmių intervalai bei priklausomybių stiprumas pateiktas 10 lentelėje.

Labai stipri	Stipri	Silpna	Nėra ryšio	Silpna	Stipri	Labai stipri
-1	<-0.9	>-0.9 ir <-0.4	>-0.4 ir <0.4	>0.4 ir <0.9	>0.9	1

10 lentelė Koreliacijos koeficiento reikšmių skalė

3.2.1.2 Modelio sudarymas

Kaip minėjome anksčiau duomenys naudojami tiesinėje regresijoje turi būti skaitiniai, todėl iš turimų duomenų imčių buvo pašalinti visi simboliniai bei datų atributai. Be to buvo palikti mūsų sukurti atributai, nes jie buvo naudoti klasifikavimo modelių sudarymo metu, todėl taip pat tūrėtų suteikti papildomos informacijos kuriant prognozavimo modelį.

Taigi pasiruošę informaciją bei pasinaudoję programa „Weka“ sudarėme tiesinės regresijos modelius turimiems duomenims. Taip pat reikėjo modifikuoti programos kodą, kad galėtume gauti modelių vidutinę paklaidą bei koreliaciją. Taigi pasinaudojus „Weka“ buvo gauti tokie rezultatai:

- vartojamosioms paskoloms gauta regresijos tiesė:

  400.9703 * klientoAlga + 0.0002 * paskolosDydis - 0.5637 * paskolosPalukanuNorma - 98.3523 * paskolosPeriodasMen + 3.2253 *

paskolosPeriodasDienomis - 401.0776 * imoka + 3.0751 * algosDalis - 400.9505 * liekaKitomsIslaidoms - 33.1413

Šio modelio vidutinė paklaida yra lygi 6.99 dienos bei koreliacija yra lygi 0.4904.

- būsto paskoloms gauta regresijos tiesė:

  blogaPaskola = 0.0068 * paskolosSuma - 0.321 * periodasDienomis + 9.0489 * periodasMenesiais + 293.067 * klientoAlga - 293.2105 * imoka + 0.5542 * algosDalis - 293.0731 * liekaKitomsIslaidoms - 0.0001 * uzstatas + 20.1975 * baudaUzPaskola + 628.2785 * baudaUzPalukanas - 3.4396

  Šio modelio vidutė paklaida yra lygi 10.4259 dienos bei koreliacijos koeficientas yra 0.5182.

Taigi pateiktuose modeliuose prie kiekvieno nepriklausomo kintamojo matome konstantas. Teigiama konstanta rodo teigiamą įtaką prognozuojamai klasei, neigiama – klasės reikšmės sumažėjimą, pvz., regresijos modelio vartojamosioms paskolos atributo „paskola Mėnesiais“ neigiama konstanta (-98.3523) rodo, kad augant paskolos terminui mėnesiais vėluojamų dienų skaičius sparčiai mažėja, tuo tarpu didėjant algos daliai, kuri skiriama paskolai – vėluojamų dienų skaičius tūrėtų augti.

Be to, kaip ir klasifikavimo metu, šį metodą patikrinome pasitelkdami kryžminį patikrinimą: vartojamosioms paskoloms naudojome 4 skirstinius, o būsto – 6. Metodo tikrinimą atlikus su programa „Weka“ buvo gauti tokie rezultatai:

- vartojamosioms paskoloms – koreliacijos koeficientas lygus 0.4116, vidutinė paklaida lygi 7.5507 dienos;
- būsto paskoloms – koreliacijos koeficientas lygus 0.4037, vidutinė paklaida lygi 11.1078 dienos.

Taigi pateikti modelių koreliacijos koeficientai apibūdina jų patikimumą. Regresijos tiesės, prognozuojančios vėluojamų mokėti įmokas dienų skaičių, apibūdina silpną priklausomybę tarp realių bei išprognozuotų reikšmių. Į šiuos modelius būtų galima iš dalies atsižvelgti prognozuojant blogąsias paskolas, tačiau visiškai pasitikėti nevertėtų, tai patvirtina kryžminio patikrinimo metu gauti rezultatai, kurių koreliacijos koeficientai atskleidė silpną priklausomybę tarp tikrų ir išprognozuotų vėluojamų mokėti paskolą dienų skaičiaus.

3.2.1.3 Regresijos modelio gerinimas

Nors modeliai ir turi silpną priklausomybę, tačiau ji yra gana nežymi, todėl mes bandėme ją padidinti imdami ne visus atributus. Buvo atisakyta tarpusavyje tiesinių atributų (koreliacija artima vienetui), pvz., periodas dienomis bei mėnesiais. Jų buvo bandyta atsisakyti, nes jie nesuteikia papildomos informacijos bei dėl jų regresijos lygties koeficientų reikšmės gali žymiai

išaugti – tai mūsų modeliuose ir galime pastebėti, pvz., būsto paskolų tiesės koeficientas -293.0731 prie atributo „lieka kitoms išlaidoms“.

Išanalizavus programos „Weka“ galimybes buvo nustatyta, kad ji suteikia galimybę nustatyti tarpusavyje tiesinius atributus, todėl programą reikėjo tik minimaliai pritaikyti mūsų reikmėms. Taigi po tarpusavyje tiesinių atributų pašalinimo buvo gautos tokios regresijos tiesės:

- vartojamosioms paskoloms gauta regresijos tiesė:

  0.0004 * paskolosDydis - 0.4206 * paskolosPalukanuNorma - 0.0039 * paskolosPeriodaisDienomis - 0.1074 * imoka + 3.1189 * algosDalis + 0.0188 * liekaKitomsIslaidoms - 40.4973

  Šio modelio vidutinė paklaida yra lygi 6.9645 dienos bei koreliacija yra lygi 0.4778.

- būsto paskoloms gauta regresijos tiesė:

  0.0061 * paskolosSuma - 0.6487 * periodasMenesiais - 0.1153 * imoka + 0.4366 * algosDalis - 0.0076 * liekaKitomsIslaidoms + 9.7996 * baudaUzPaskola +698.6988 * baudaUzPalukanas - 4.1277

  Šio modelio vidutinė paklaida yra lygi 10.4909 dienos bei koreliacijos koeficientas yra 0.4996.

Pašalinus tarpusavyje tiesinius atributus, pasikeitė ir metodo kryžminio tikrinimo rezultatai:

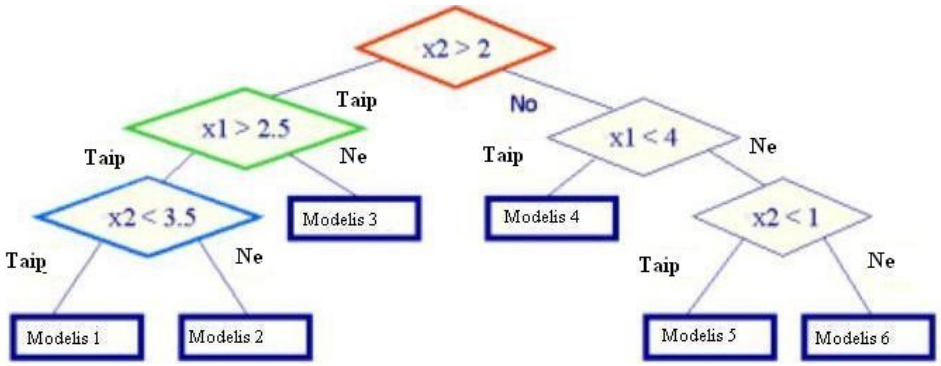
- vartojamosioms paskoloms – naudojant 4 skirstinius buvo gautas koreliacijos koeficientas lygus 0.4251, vidutinė paklaida lygi 7.273 dienos;
- būsto paskoloms – naudojant 6 skirstinius buvo gautas koreliacijos koeficientas lygus 0.3887, vidutinė paklaida lygi 11.1036 dienos;

Taigi, pašalinus atributus buvo gauti nežymiai prastesni koreliacijos bei vidutinės klaidos rezultatai (išskyrus vartojamųjų paskolų vidutinę klaidą), tačiau pačios regresijos lygtys tapo geresnės – buvo atsisakyta perteklinių duomenų bei žymiai sumažintos kai kurių atributų koeficientų reikšmės. Todėl galima teigti, kad šios tiesės bus vertingesnės nei gautos naudojant visus nepriklausomus atributus.

3.2.2 **M5 modelių medžiai**

Kaip pastebėjome, tiesinės regresijos modeliai apibūdina silpną priklausomybę tarp realių bei išprognozuotų mokėti paskolą dienų skaičiaus. Todėl nusprendėme pabandyti patikslinti prognozavimo įverčius, sukuriant modelių medžius dienų skaičiui prognozuoti. Taigi buvo pasirinktas vienas populiariausių jų sudarymo algoritmų „M5 modelių medžiai“ [Qui92]. Jis remiasi idėja, kad tiesinės regresijos modeliai sukuriami ne visiems duomenims, o tam tikroms

grupėms t.y. pirmiausia sudaromas sprendimų medis, o jo lapuose patalpinamos regresijos tiesės, kurios naudojamos duomenų prognozei, pvz., pateiktas 16 pav.



16 pav. M5 modelis

Algoritmo sprendimų medžio atributų parinkimas bei jų padalinimas į intervalus nustatomas naudojant standartinio nuokrypio įvertį. Skaičiuojamas šio dydžio sumažėjimas SNS(standartinio nuokrypio sumažėjimas) kiekvienai viršūnei[dat04]:

$$SNS = sd(T) - \sum_i \frac{T_i}{T} sd(T_i),$$

T – įrašų skaičius, kuris pasiekia tiriamą viršūnę,

T_i – įrašų skaičius po i-ojo padailinimo,

sn – standartinis nuokrypis, kuris lygus:

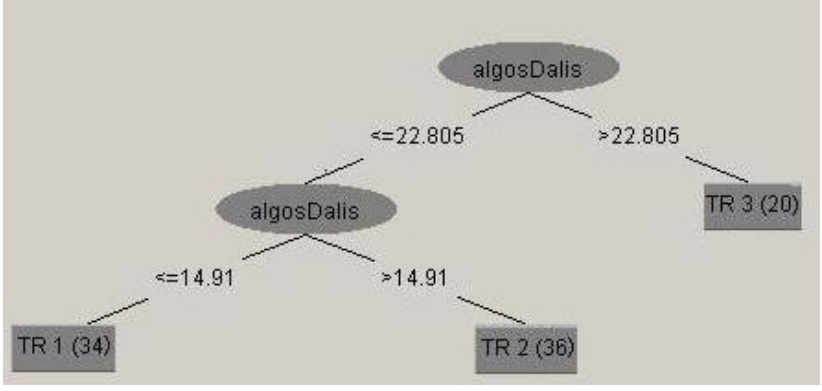
$$sn = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2},$$

μ – reikšmių vidurkis.

Po padalijimo paimami mažiausią SNS reikšmę įgyjantys atributo intervalai. Taip patikrinami visi galimi atributų skaidiniai, kol randama mažiausia standartinio nuokrypio sumažėjimo reikšmė.

Taigi pamodifikavus programos „Weka“ M5 algoritmo įgyvendinimą, sudarėme M5 modelių medžius turimiems duomenims bei gavome tokius rezultatus.

Vartojamosios paskolos gautas medis pateiktas 17 pav.



17 pav. M5 modelio medis būsto paskoloms

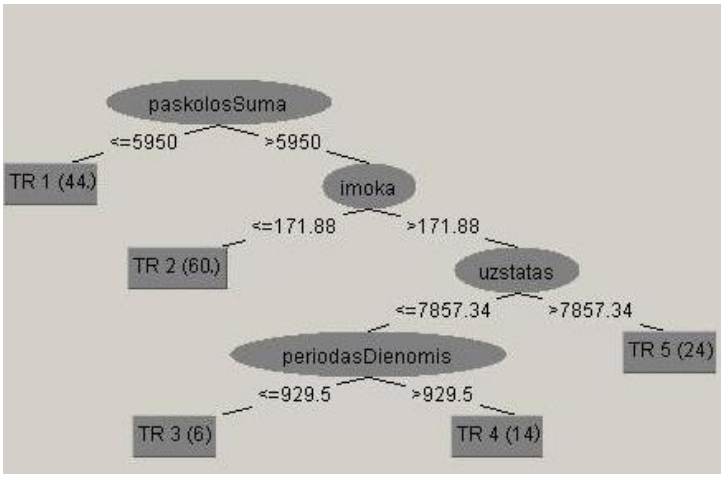
Medžio lapuose esantys tiesinės regresijos(TR) modeliai pateikti 11 lentelėje.

Numeris	Tiesė
TR1	0.0002 * klientoAlga + 0.1888 * algosDalis - 2.6358
TR2	0.0016 * klientoAlga + 0.1876 * algosDalis - 4.0208
TR3	-2.5159 * algosDalis + 83.4842

11 lentelė vartojamųjų paskolų M5 modelio regresijos tiesės

Šio modelio koreliacijos koeficientas yra 0.715, o vidutinė paklaida 4.8188 dienos.

Būsto paskoloms gautas modelio medis pateiktas 18 pav.



18 pav. M5 modelio medis būsto paskoloms

Būsto paskolų modelio taisyklės pateiktos 12 lentelėje.

Numeris	Regresijos tiesė
TR1	0.0012 * paskolosSuma - 0.0036 * periodasDienomis - 0.0064 * klientoAlga + 1.5478
TR2	0.0015 * paskolosSuma - 0.0048 * periodasDienomis - 0.0088 * klientoAlga - 0.0004 * uzstatas + 8.0372
TR3	0.0017 * paskolosSuma - 0.0162 * periodasDienomis - 0.0318 * klientoAlga + 0.001 * uzstatas + 30.5003
TR4	0.0017 * paskolosSuma - 0.0133 * periodasDienomis - 0.0318 * klientoAlga + 0.001 * uzstatas + 24.7403
TR5	0.0017 * paskolosSuma - 0.0057 * periodasDienomis - 0.0696 * klientoAlga - 0.1255 * imoka + 0.0053 * uzstatas + 36.7974

12 lentelė būsto paskolų M5 modelio regresijos tiesės

Šio modelio koreliacijos koeficientas yra lygus 0.8194, o vidutinė paklaida 6.3029 dienos.

Šis metodas taip pat buvo patikrintas naudojant kryžminį patikrinimą: vartojamosioms paskoloms naudojant 4 skirstinius, o būsto – 6, ir buvo gauti tokie rezultatai:

- vartojamosioms paskoloms – koreliacijos koeficientas lygus 0.5714, vidutinė paklaida lygi 5.5763 dienos;
- būsto paskoloms – koreliacijos koeficientas lygus 0.6649, vidutinė paklaida lygi 7.3828 dienos.

Taigi, vėluojamų dienų prognozei pasitelkus M5 modelių medžių sudarymo algoritmą, buvo gauti pastebimai geresni koreliacijos koeficiento bei vidutinės paklaidos įverčiai turimoms duomenų imtims. Be to, kryžminio patikrinimo rezultatai taip pat pagerėjo. Todėl, įvertinę visus rezultatus galime teigti kad šie modeliai patikimiau įvertins vėluojamų mokėti paskolą dienų skaičių nei anksčiau nagrinėti paprastos daugialypės tiesinės regresijos modeliai. Tačiau ir naujieji modeliai suteikė tik silpną priklausomybę tarp realių ir išprognozuotų reikšmių(9 lentelė), todėl prognozės metu jais besąlygiškai pasitikėti nevertėtų.

3.2.3 Prgnozės modelių apibendrinimas

Ištirtų algoritmų bei juos naudojant sudarytų modelių tikslumai pateikti 13 lentelėje. Pagal pateiktus duomenis matome, kad tiek M5 algoritmas, tiek juo suformuoti modeliai patikimiau prognozuoja vėluojamų mokėti paskolą dienų skaičių, nei paprastas daugialypės tisinės regresijos modelis. Tai yra nes M5 modelyje regresijos tiesės sudarinėjamos ne visiems duomenims, o tam tikroms jų grupėms.

Algoritmas	**Modelio duomenys koreliacija (vidutinė dienų paklaida)**		**Algoritmo duomenys (kryžminio patikrinimo būdu) koreliacija(vidutinė dienų paklaida)**	
	Vartojamosios paskolos	**Būsto paskolos**	**Vartojamosios paskolos**	**Būsto paskolos**
Tiesinė regresija	0.4778 (6.9645)	0.4996 (10.4909)	0.4251 (7.273)	0.3887 (11.1036)
M5 modelio medis	0.715 (4.8188)	0.8194 (6.3029)	0.5714 (5.5763)	0.6649 (7.3828)

13 lentelė prognozės algoritmų bei jų modelių koreliacijos

Rezultatai ir išvados

Atlikus darbą buvo gauti šie rezultatai:

- Ištirti duomenų vitrinų sudarymo metodai bei sukurti duomenų vitrinų modeliai turimoms duomenų struktūroms;
- Sukurta programinė įrangą leidžianti konvertuoti gautus „Microsoft Excel“ rinkmenas į mums reikalingą arff formatą;
- Ištirti klasifikavimo algoritmai, bei sukurti modeliai geriausiai nustatantys blogąsias paskolas;
- Ištirti prognozavimo algoritmai, bei sukurti modeliai geriausiai įvertinantys vėluojamų mokėti paskolą dienų skaičių;
- Nustatyti svarbiausi bankinių duomenų atributai.

Taigi, pirmiausia, gavus banko duomenis buvo išnagrinėtos duomenų struktūros, sukurti žvaigždės bei plokščio tipo duomenų vitrinų modeliai, kurie apibendrina ir aiškiau pateikia turimus duomenis. Be to, banko informacijos perdavimui buvo naudojami plokščio tipo duomenų vitrinų modeliai. Jie buvo pasirinkti, nes supaprastina duomenų paruošimą – visi duomenys perduodami pagal iš anksto susitartą struktūrą, tai leidžia supaprastinti informacijos konvertavimą (viena programa, kuri konvertuoja gaunamus duomenis į mums tinkamą formatą).

Vėliau atlikus klasifikavimo algoritmų tinkamumo tyrimus bei išanalizavus rezultatus nustatėme, kad RIPPER algoritmu sudarytos taisyklės, suteikia labai panašius turimų duomenų klasifikavimo rezultatus, kaip ir sprendimų medžiai sukurti naudojant C4.5 metodą. Taigi galime teigti, kad vienas iš šių klasifikatorių tūrėtų gana tiksliai įvertinti naujai ateinančius duomenis, jei jų informacija bus panaši į turimą, tačiau visiškai remtis gautais modeliais nederėtų. Taip pat atlikus algoritmų vertinimą naudojant kryžminį patikrinimą nustatėme, kad geriausiu tikslumu pasižymi klasifikavimo taisyklių sudarymo metodas RIPPER. Todėl galima teigti, kad būtent šis algoritmas tinkamiausias turimoms duomenų imtims.

Duomenų tyrimo metu be klasifikavimo taip pat buvo sudaryti prognozės modeliai, skirti vėluojamų mokėti paskolos įmokas dienų skaičiui prognozuoti. Išanalizavus tyrimo rezultatus buvo nustatyta, kad M5 algoritmu sudaryti modeliai bei pats algoritmas (kryžminio patikrinimo metu) suteikia geresnius rezultatus nei daugialypės tiesinės regresijos modeliai, todėl galime teigti, jis yra patikimesnis bei tinkamesnis turimoms duomenų imtims. Tačiau nors ir buvo gauti gana geri koreliacijos koeficientai, jie pasižymėjo tik silpna priklausomybe tarp realių ir išprognozuotų vėluojamų mokėti paskolą dienų skaičiaus. Todėl prognozės metu šiais modeliais visiškai pasitikėti nevertėtų – jie galėtų būti naudojami, kaip papildoma informacija.

Taip pat tyrimo metu siekiant didesnio sudaromų modelių tikslumo prie esamų duomenų buvo pridėti nauji, išvestiniai atributai. Būtent vienas iš jų – algos dalis, skiriama paskolos įmokai, buvo vienas iš svarbiausių požymių, naudojamų klasifikacijos bei prognozės modelių sudarymo metu. Taigi duomenų papildymas naujais atributais pasiteisino, nes šis ir kiti naujai sukurti atributai suteikė papildomos informacijos apie turimus duomenis ir buvo vertingi duomenų tyrimo metu.

Apibendrinat, galima teigti, kad sukurti vertinimo modeliai pakankamai tiksliai klasifikuoja blogąsias paskolas bei prognozuoja vėluojamų dienų skaičių. Tačiau jie yra tinkamiausi turimo laikotarpio duomenims, nes skirtingais laikotarpiais informacija gali kisti. Todėl pakartotinai, kas kelis mėnesius, reikėtų paimti naujus duomenis ir sugeneruoti modelius naudojant mūsų ištirtus klasifikavimo bei prognozavimo algoritmų nustatymus. Taip būtume užtikrinti, kad modeliai bus patikimesni bei tikslesni, nes jų sudarymui bus naudojama naujausia informacija.

Literatūros sąrašas

[IG01] Nazli Ikizler, H.Altay Guvenir "Mining Interesting Rules in Bank Loans Data", Ankara Bilken universitetas, 10 „Turkijos Dirbtinio intelekto simpoziumas", 2001

[Das98] Rajanish Dass „Data mining in banking and finance: a note for bankers", 3p. 1998.

[BAB+01] Corinne Baragoin, Christian M. Andersen, Stephan Bayerl, Graham Bent, Jieun Lee, Christoph Schommer „Mining Your Own Business in Banking Using DB2 Intelligent Miner for Data", IBM 8-21p. rugpjūtis 2001.

[CCK+99]Peter Cabena, Hyun Hee Choi, Il Soo Kim, Shuichi Otsuka, Joerg Reinschmidt, Gary Saarenvirta „Intelligent Miner for Data Applications Guide", IBM 6p. kovas 1999.

[Inm02] W.H. Inmon „Building the data warehouse", WILEY 2002

[Tag] Dr. Taghiyare „Customer Relationship Management Instructor"

[Mac07] Algirdas Mačiulis „Duomenų tyrimas",Vilnius VU 2007

[Hol93] R.C. Holte "Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. Machine Learning", „Machine Learning" 11:63-91 p., 1993

[MHW03] Craig Nevill-Manning, Geoffrey Holmes and Ian H. Witten Department of Computer Science, Vaikato univertisetas, Hamiltonas, Naujoji Zelandija „The Development of Holte's 1R Classifier" Naujoji Zelandija tarptautinė duomenų tyrimo konferenicja 1995

[Tho00] Lyn C. Thomas "A Survey of Credit and Behavioural Scoring; Forecasting financial risk of lending to consumers." „ International Journal of Forecasting", 16 numeris, 149-172p. gegužė 2000

[CGP04] Koh Hian Chye, Tan Wei Chin Goh, Chwee Peng "Credit Scoring Using Data Mining Techniques", Singapūras duomenų valdymo institutas, 26 numeris, 25-27p. liepos 1 2004

[MKo00] Daniel L. Moody. Mark A.R. Kortink "From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design", Stokholmas „Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'2000) " birželio 5 2000

[KQu99] Ron Kohavi, Ross Quinlan "Decision Tree Discovery", Oksfordo universitetas, 267 – 276p. 2002

[ZZY03] Shichao Zhang, Chengqi Zhang, Qiang Yang "Data Preparation For Data Mining", „Dirbtinio Intelekto Taikymas", 17 numeris, 375-381p., 2003

[JDR99] Paul Jermyn, Maurice Dixon, Brian J Read "Preparing Clean Views of Data for Data Mining", Londonas Guildhal universitetas, 1999

[CPH+08] Jingping Chen , Haiwei Pan, Qilong Han, Linghu Chen, Jun Ni “Credit Risk Assessment Model Based On Domain Knowledge Constraint”, „2008 metų tarptautinis kompiuterių mokslų symposiumas“, 144-149p., 2008

[Qui92] Ross J. Quinlan “Learning with Continuous Classes“, „5 Autralijos dirbtinio intelekto konferencija“, 267 – 276p., 1992.

[dat04] http://datamining.ihe.nl/research/model-trees.htm, 2004

Santrumpos

ARFF – angl. Attribute-Relation File Format;

SNS – standartinio nuokrypio sumažėjimas;

Xls – angl. Excel Spreadsheet;

1R – angl. One Rule;

RIPPER – angl. Repeated Incremental Pruning to Produce Error Reduction;

TR – teisinė regresija;

AVP – absoliutinė vidutinė paklaida;

SN – standartinis nuokrypis;

M5 - Model trees, version 5;