VILNIAUS UNIVERSITETAS
KAUNO HUMANITARINIS FAKULTETAS

INFORMATIKOS KATEDRA

Verslo informacijos sistemų studijų programa
Kodas 62103S138

AKVILĖ ŽUKAUSKAITĖ

MAGISTRO BAIGIAMASIS DARBAS

DUOMENŲ GAVYBOS SISTEMA NAUDOJANT VEIKLOS MODELĮ

KAUNAS 2010

VILNIAUS UNIVERSITETAS
KAUNO HUMANITARINIS FAKULTETAS

INFORMATIKOS KATEDRA

AKVILĖ ŽUKAUSKAITĖ

MAGISTRO BAIGIAMASIS DARBAS

DUOMENŲ GAVYBOS SISTEMA NAUDOJANT VEIKLOS MODELĮ

Leidžiama ginti _____

Magistrantas _____
(parašas)

Darbo vadovas _____
(parašas)
Prof. doc. (HP) Saulius Gudas

Darbo įteikimo data _____
Registracijos Nr._____

KAUNAS 2010

TURINYS

PAVEIKSLŲ SĄRAŠAS

LENTELIŲ SĄRAŠAS

SANTRUMPŲ SĄRAŠAS

OLAP - Online Analytical Processing.

CRISP-DM- Cross-Industry Standard Process for Data Mining.

SQL - Structured Query Language.

BI – Business Intelligence.

IT – informacinės technologijos.

CRM - *C*ustomer Relationship Management.

BICC – Business Intelligence Competency Center.

DSD – duomenų srautų diagrama.

ASP - Active Server Pages.

API - Application Programming Interface.

DG – duomenų gavyba.

ŽUKAUSKAITĖ, Akvilė (2010) *Enterprise Model Based Data Mining System*. MBA Graduation Paper. Kaunas: Vilnius University, Kaunas Faculty of Humanities, Department of Informatics, 60 p.

SUMMARY

The main goal of the graduation paper is creation Enterprise Model Based Data Mining System. The main tasks of this work are:

1. Submit business intelligence systems analyses; select the best business intelligence system for small companies.
2. Submit data mining algorithms for analyses.
3. Create data mining process model using enterprise model.
4. Perform detailed analysis of data mining tools for small enterprises;
5. Use real data to verify the adaptation of tools for small businesses;
6. Use experimental data to realize the data mining process for real business.

For the graduation paper are used analyses of nonfiction literature, analogy, comparison, modelling and experimental methods.

The graduation paper composes of analysis part, proposed solution methodology part and experiment.

Results of work: designed data mining activity model for small logistics company.

The graduation paper includes 60 sheets, 9 tables and 36 pictures.

ĮVADAS

Šiuolaikinėse verslo organizacijose vis didesnis dėmesys skiriamas informacijai, jos valdymui ir efektyviam panaudojimui. Tai sąlygoja ne tik didėjantys informacijos kiekiai, bet ir poreikis analizuoti duomenis, ieškoti juose sąryšių, šablonų, priklausomybių ar modelių, kurie padėtų priimti sprendimus. Dauguma organizacijų aiškiai suvokia būtinybę geriau valdyti informaciją, daugelis įmonių yra įsitikinusios, jog informacija yra jos turtas.

Šiuolaikinėje visuomenėje apdorojamos informacijos kiekiai vis didėja, dėl to ją surinkti bei apdoroti darosi vis sunkiau. Verslo organizacijos suvokia būtinybę efektyviai valdyti informaciją, tačiau konkrečiai nežino kokią naudą jie gali gauti iš jau turimos, surinktos informacijos. Minimali duomenų analizė ir daromos išvados neskatina verslo vystymosi, o kartais jį net stabdo. Todėl reikalingas praktinis duomenų gavybos naudojimas įmonėse. Nors duomenų gavyba ir nepakeičia patyrusių verslo analitikų, bet duoda jiems galingą įrankį, kuris pagerina ir pagreitina jų atliekamą darbą. Įrankių pagalba galima greičiau apdoroti didelius informacijos kiekius ir rasti tarp duomenų įprastai nepastebimus ryšius.

Intelektinės sistemos (duomenų gavyba, OLAP sistemos bei žinių bazės) įmonėms yra naudingos, nes padeda didinti pelną, darbo našumą, veiklos rezultatus taupyti laiką, didinti konkurenciją, atrasti naujus klientų elgesio motyvus.

Temos aktualumas. Veiklos modelio panaudojimas duomenų gavybos procese leistų sujungti veiklos modelį su duomenų gavybos modeliu. Taip galima siekti geresnių rezultatų duomenų gavybos srityje bei efektyviau valdyti duomenų gavybos procesą. Tai ypač svarbu mažoms įmonėms kurios dar tik pradeda naudoti duomenų gavybos įrankius savo veikloje, nes norint gauti geriausius duomenų gavybos rezultatus duomenų gavyba reikia sklandžiai įtraukti į įmonės veiklos procesus.

Problemos esmė. Veiklos modelio pagalba norima pagerinti duomenų gavybos procesą (efektyviau valdyti) siekiant geresnių jos rezultatų.

Problemos ištyrimo lygis. Duomenų gavyba nėra nauja veiklos sritis, tačiau ji sparčiai vystosi ir tobulėja. Kiekvienas darbas duomenų gavybos srityje, padedantis didinti duomenų gavybos valdymo efektyvumą yra svarbus ir aktualus. Itin svarbus yra duomenų gavybos procesų modelio atsiradimas, kurio tikslas integruoti duomenų gavybą į verslo aplinką. Jį suformavo trys kompanijos (Daimler Chrysler, SSPS, Teradata). CRISP-DM (angl. *Cross-Industry Standard Process for Data Mining*) – standartinį duomenų gavybos proceso modelis (The CRISP-DM consorcium, 2003). Kaip galima sujungti duomenų gavybą ir sprendimų priėmimų sistemas siekiant optimalesnių, efektyvesnių bei

veiksmingesnių verslo operacijų išsamiai dėstoma „*Data Mining and Decision Support for Business Competitiveness: A European Virtual Enterprise*“ projekto medžiagoje (Mladenic D., 2003).

Darbo objektas. Duomenų gavybos sistema paremta veiklos modeliu.

Darbo tikslas. Sukurti duomenų gavybos sistemą paremtą veiklos modeliu.

Darbo uždaviniai.

1. Atlikti intelektinių verslo sistemų analizę, parinkti tinkamiausią mažų įmonių problemoms spręsti;
2. Atlikti duomenų gavybos metodų apžvalgą bei algoritmų analizę.
3. Sudaryti duomenų gavybos proceso modelį, panaudojant veiklos modelį.
4. Atlikti duomenų gavybos įrankių pritaikomų mažose įmonėse detalią analizę;
5. Panaudojant realius duomenis patikrinti įrankių pritaikymą mažose įmonėse.
6. Panaudojant eksperimentinius duomenis realizuoti duomenų gavybos proceso pritaikymą realiai įmonei.

Darbo struktūra. Darbą sudaro įvadas, išvados ir 3 pagrindiniai skyriai – teorinis/analitinis, siūlomo sprendimo metodikos, eksperimentinis.

Teoriniame skyriuje pateikiama intelektinių verslo sistemų ir duomenų gavybos algoritmų analizė. Taip pat CRISP-DM duomenų gavybos modelio bei duomenų gavybos programinės įrangos apžvalga.

Siūlomo sprendimo metodikos skyriuje pateikiamas sudarytas duomenų gavybos proceso modelis. Jis suformuotas remiantis pateikta veiklos modelių analize. Veiklos modelio pagalba sprendžiamos konkrečios organizacijos veiklos problemos. Šiame skyriuje taip pat pateikiama išsami duomenų gavybos įrankių pritaikomų mažose įmonėse analizė.

Eksperimentiniame skyriuje aprašomas XLMiner, Microsoft SQL Server 2008 Data Mining Add-Ins bei TreePlan pritaikymas konkrečioje įmonėje.

Darbe naudoti literatūros šaltiniai: informacija surinkta naudojant internete, duomenų bazėse esančias užsienio ir Lietuvos autorių publikacijas, straipsnius, empirinius tyrimus.

Tyrimo metodai.

1. Mokslinės literatūros analizė – problemos aktualumui pagrįsti, teorinės medžiagos rinkimui;
2. Analogija ir palyginimas – duomenų gavybos sistemų palyginimui;
3. Modeliavimas – duomenų gavybos procesų atvaizdavimui veiklos modeliais;
4. Eksperimentas – sistemos testavimui.

Darbo rezultatų teorinė ir praktinė reikšmė:

- Atlikta duomenų gavybos sistemų ir algoritmų analizė, aptarti jų privalumai, trūkumai, panaudojimo sritys, duomenų gavybos algoritmų tendencijos;
- Išanalizuota duomenų gavybos programinė įranga pritaikoma mažų įmonių problemoms spręsti;
- Sudarytas duomenų gavybos proceso modelis, paremtas veiklos modeliu.

Darbo struktūra ir apimtis. Darbą sudaro 67 puslapiai, pateikti 36 paveikslėliai ir 9 lentelės ir 3 priedai.

1. INTELEKTINIŲ SISTEMŲ ANALIZĖ

Šiame skyriuje pateikiama pagrindinių intelektinių sistemų: duomenų gavybos, žinių bazių ir OLAP sistemų apžvalga bei palyginimas. Taip pat duomenų gavybos algoritmų analizė, duomenų gavybos modelio bei programinės įrangos apžvalga.

1.1. Verslo intelektinės sistemos

Verslo intelekto sistemos (angl. *Business Intelligence* arba *BI*) sąvoka pirma kartą buvo paminėta 1958 m. Hans Peter Luhn IBM žurnale publikuotame straipsnyje „*A Business Intelligence System*“. 1989 m. Gartner Group (IT tyrimų ir konsultavimo įmonė, Konektikutas, JAV) mokslininko Howard Dresner straipsnyje buvo pateiktas terminas „*Business Intelligence*“ su rinkiniu metodų ir sąvokų. Jų tikslas buvo pagerinti verslo sprendimų priėmimo procesus naudojant įvairius duomenų išteklius. Nuo to laiko verslo intelekto sistemos plėtojasi siūlydamos vis naujesnius ir įvairiau pritaikomus įrankius verslui.

Verslo intelekto sistemos - tai taikomosios programos bei technologijos, naudojamos duomenų rinkimui, saugojimui, analizei bei priėjimui prie duomenų, kurie galėtų padėti organizacijoms greičiau priimti verslo sprendimus. BI apima veiklos sprendimų priėmimo sistemas, užklausų ir ataskaitų kūrimo internetu, analitinio apdorojimo (angl. *OLAP sistemos*), statistinės analizės, prognozavimo ir duomenų gavybos (angl. *data mining*) programas (Rossetti L., 2009).

UAB Alna Business Solutions tinklapyje ši sąvoka apibrėžiama taip: „tai sprendimus priimanti ar sprendimus padedanti priimti sistema, kuri sukaupia gausybę kaupiamų duomenų, juos išgrynina, agreguoja ir pateikia vartotojui labiausiai tinkama ir priimtina forma bei būdu - įvairios lentelės, grafikai, „prietaisų skydeliai“ ir pan.“ (UAB Alna Business Solutions, 2007). Tokia forma pateikta informacija dažnai naudojama, kai įmonė ketina priimti svarbų sprendimą ar tiesiog nori stebėti verslo pokyčius.

Visame pasaulyje yra firmų užsiimančių verslo intelekto sistemų kūrimu bei platinimu. Populiariausi ir geriausiai žinomi jų produktai „Microsoft“, „WebFOCUS“, „Cognos“, „Business Objects“, „Corporate Planner“, „Microstrategy“, „Siebel“ , „Oracle“.

Verslo intelekto sistemų populiarumą ir aktualumą labai gerai parodo *GARTNER* kompanijos atliktos apklausos rezultatai (1 pav). Buvo apklausta daugiau nei 1500 vyriausiųjų informacijos specialistų (angl. *CIO*). Siekta išsiaiškinti kokiems procesams ir kokiom technologijom verslas teikė

pirmenybę savo veikloje 2009 metais. (Pettey C., 2009). Kaip matome BI šiame reitinge užima pakankamai aukštą vietą. Todėl šiai sričiai skiriama vis daugiau dėmesio.

1 lentelė

10 geriausių verslo ir technologijų prioritetų 2009

Verslo prioritetų TOP 10		**Technologinių prioritetų TOP 10**	
Verslo procesų tobulinimas	1	Verslo intelektinės sistemos	1
Verslo išlaidų mažinimas	2	Verslo aplikacijos (ERP, CRM ir kt.)	2
Darbuotojų darbo efektyvumo gerinimas	3	Serverių ir saugyklų technologijos (vizualizacijos)	3
Naujų klientų pritraukimas ir išlaikymas	4	Pasenusių aplikacijų modernizavimas	4
Informacijos ir analitikos naudos didinimas	5	Bendravimo technologijos	5
Naujų produktų ir paslaugų kūrimas (inovacija)	6	Darbo tinkle, balso ir duomenų komunikacijos	6
Efektyvesnis klientų ir rinkų planavimas	7	Techninė infrastruktūra	7
Permainų valdymas	8	Saugumo technologijos	8
Ryšių su esamais klientais plėtimas	9	Į servisus orientuotos aplikacijos ir architektūros	9
Plėtra į naujas rinkas ir geografines vietoves	10	Dokumentų valdymas	10

Šaltinis: sudaryta autoriaus pagal Pettey, 2009.

Intelektinių sistemų svarbą galima pastebėti ieškant informacijos internete, knygose, duomenų bazėse. Daugybėje įvairių šaltinių galima rasti informacijos apie BI svarbą, teikiamus privalumus. R.M. Bogza ir Dorin Zaharie straipsnyje *„Business intelligence as a competitive differentiatos“* yra aiškinama, kokių konkurencinių pranašumų įmonei gali suteikti naudojamos BI technologijos. Straipsnio autoriai BI apibrėžia, kaip reikiamos informacijos pateikimą reikiamiems žmonėms reikiamu laiku. Teigiama, jog BICC (*Business Intelligence Competency Center*) atskleidžia daugybę konkurencinių privalumų verčiant neapdorotus duomenis į žinias, kurios teigiamai veikia veiklos strategijas. BICC keliami iššūkiai straipsnyje skirstomi į 6 sritis:

1. duomenų iššūkiai (greitas nesutvarkytų duomenų, pateiktų skirtingais formatais, apdorojimas);
2. technologiniai iššūkiai (BI sistema turi į vieną visumą susieti visą įmonės vertės grandinėje atsirandančią informaciją, o tam dažnai prireikia itin sudėtingų technologinių sprendimų);

3. procesų iššūkiai (BI yra procesas, o visi įmonėse vykstantys procesai yra kintantys arba pasikartojantys. Bet kuriuo atveju sėkmingos veiklos paslaptis yra žmonės. Kuo įmonėje vykstantys procesai bus artimesni žmonėms, aiškūs ir suprantami, tuo geresni bus veiklos rezultatai);

4. strateginiai iššūkiai (BI padeda pasirinkti tinkamą veiklos strategiją, tačiau dažnai atskiriems skyriams reikalingos skirtingos strategijos. Taigi pagrindinis BI strateginis iššūkis yra kaip tinkamai pristatyti reikiamą informaciją reikiamiems skyriams);

5. vartotojo iššūkiai (BI organizacijoje turi pateikti vartotojui tokią informaciją, kuri atitiktų jo pageidavimus, jo išprusimą bei jo užimamas pareigas);

6. kultūriniai iššūkiai (BI pateikiama informacija turi atitikti įmonės vidinę kultūra, jos veiklos standartus ir netrukdyti įprastinės jos veiklos).

Straipsnyje grafiškai atvaizduotos sritys, kurias apima bei kurioms naudą teikia BI. Teigiama, jog BICC įmonėje turėtų būti tam tikras centras apimantis skirtingos įgūdžių sritys (IT, analitinė veikla ir verslumas). Šio centro tikslas yra geriau suprasti veiklą, interpretuoti dabartinius veiklos rezultatus bei kuo tiksliau numatyti ateities perspektyvas (Bozga R.M, 2008).

BI sparčiai plinta įvairiuose veiklos srityse. Viena iš jų yra verslo kompiuterinė aplinka. Apie tai plačiau aprašyta Lida Xu, Li Zeng, Zhongzhi Shi, Qing He, Maoguang Wang straipsnyje „*Research on Business Intelligence in Enterprise Computing Environment*“. Straipsnyje teigiama, jog paskutiniu metu populiariausios tradicinės kompiuterinės sistemos yra e-komercijos sistemos, vadybos informacinės sistemos (*management information systems*), sandorių vykdymo sistemos bei verslo resursų planavimo sistemos. Visos šios sistemos skirtos palengvinti darbą, bet netinka apdoroti nuolat didėjančius duomenų kiekius bei išgauti sudėtingesnes ataskaitas. Tam prireikia kompiuterinių BI sprendimų – įvairių BI įrankių, technologijų bei pasiūlymų. Straipsnyje apžvelgiama techninė verslo sistemų struktūra, trumpai aptariami BI algoritmai (Xu L., 2007).

Nors BI sparčiai plinta visame pasaulyje daugeliui organizacijų žengti žingsnį ir pradėti naudotis BI yra gana sudėtinga. Larissa T. Moss savo straipsnyje *Organizational Barriers to Business Intelligence (Part 1)* išskyrė pagrindinius barjerus, kodėl joms yra sunku pradėti sklandžiai naudoti BI sistemas. Dažniausiai nesklandumų kyla ne dėl technologinių sprendimų, o būtent dėl to, jog organizacijos nenori keisti savo organizacinių, kultūrinių bei infrastruktūrinių įpročių. Visų pirma, jos vengia pakeisti valdymo struktūrą. Dažnas įvairiose veiklose, projektuose dalyvaujantis asmuo nori užimti aukštesnes pareigas, kuri reiškia didesnį atlygį, paaukštinimą ar pan. BI technologijos dažnai priverčia pasikeisti darbuotojus pareigomis, o tai veiklos subjektams sukelia papildomų problemų. Dar vienas svarbus aspektas yra įgūdžiai. Dažnai įmonės bei organizacijos neturi tinkamų specialistų, arba jiems trūksta įgūdžių, profesionalumo. Šiuo atveju svarbu nemaišyti įgūdžių su patirtimi ir išsiaiškinti

ar darbuotojas, kuriam patikima BI sistemos diegimas/priežiūra būtų išties kompetentingas šioje srityje. BI naudojimui, kaip ir visiems kitiems rimtesniems sprendimams organizacijose, turi pritarti veiklos subjekto valdytojai (t.y. direktoriai, aukščiausiųjų kategorijų vadybininkai ir pan.). Jei vykdomo projekto tiek finansiškai, tiek idėjiškai neremia vykdomoji valdžia arba ji nėra suinteresuota projekto vykdymu, didelė tikimybė, jog jo įgyvendinti nepavyks. Kuo BI technologijų diegimą mažiau riboja finansai, tuo didesnė tikimybė, jog bus parinktas optimaliausias variantas konkrečiam verslo subjektui. Idealiausia kai BI technologijų diegimą bei įgyvendinimą prižiūri lėšas tam skyrę vadovai, nes jie yra suinteresuoti lėšas panaudoti tikslingai bei gauti maksimalią naudą (Moss L.T., 2008).

1.2. Verslo intelektinių sistemų tipai ir jų palyginimas

Duomenų gavyba (angl. *Data Mining*) yra anksčiau nežinomos ir potencialiai naudingos informacijos ištraukimas iš sukauptų duomenų. Duomenų gavybos procesas susideda iš trijų pagrindinių etapų – tyrinėjimo, modelio sudarymo ir įvertinimo. Duomenų gavyba įdomi tuo, kad jos technologija sugeba faktiškus duomenis paversti naudinga informacija ir žiniomis, tinkamomis veiklos valdymui, rinkos analizei, sprendimų priėmimui (Sekliuckis V., 2006, p. 285).

Duomenų gavyba yra naudojama daugelyje veiklos sričių. Viena iš jų yra marketingas. A. Coskun Samli, Terrance L. Pohlen1 ir Nenad Bozovic straipsnyje „*A Review of Data Mining Techniques as they Apply to Marketing: Generating Strategic Information to Develop Market Segments*“ aprašoma, kuo duomenų gavyba yra svarbi apdorojant marketingui skirtus duomenis. Duomenų gavyba naudoja sudėtingas statistines analizes bei modeliavimo technikas. Pagrindinis jos tikslas - atrasti tokias struktūras, modelius bei informaciją, kurie būtų nepasiekiami naudojant tradicinius statistinius metodus. Straipsnyje pateiktoje lentelėje yra pateikti 7 žingsneliai, kurių kiekvienas atitinka tam tikrą marketingo funkciją. Kiekvienai tai funkcijai pateikta po kelis duomenų gavybos siūlomus sprendimo radimo būdus (Cozkun Samli A., 2002).

Apie duomenų gavybos svarbą siekiant išgauti žinias iš duomenų kalbama ir Guozheng Zhang, Faming Zhou, Fang Wang, Jian Luo straipsnyje „*Knowledge creation in marketing based on data Mining“*. Straipsnyje bandoma paaiškinti tiek pačią data mining sąvoką, tiek duomenų gavybos pranašumus lyginant su tradiciniais analizės būdais (Zhang G., 2008).

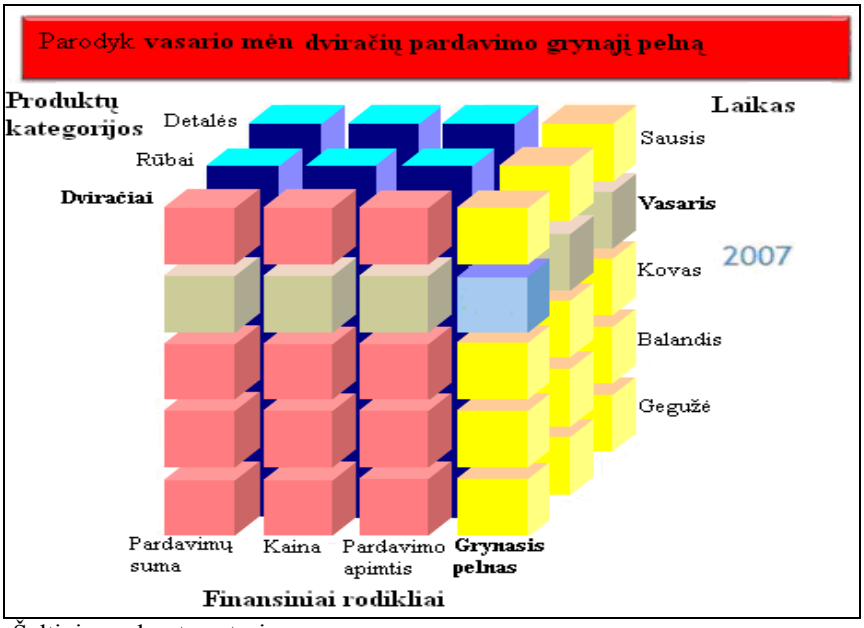
Duomenų gavybos panaudojimą e. mokymuisi naudojant CelGrid sistemą aprašo Paweł B. Myszkowski, Halina Kwaśnicka and Urszula Markowska-Kaczmar straipsnyje „*Data Mining techniques in e-learning CelGrid system*“. Straipsnyje aprašoma, kaip data mining technika gali padėti e. mokymuisi sugebėdama apdoroti daugybę duomenų ir padėdama studentams rasti daugiau nei

informaciją kurią pateikia dėstytojai. Studentams suteikiama galimybė rasti įvairias sąsajas tarp pateiktų duomenų, kurios leistų daug paprasčiau ir greičiau įsisavinti pateiktą informaciją siejant ją su tuo, kas jau žinoma (Myszkowski P.B., 2008).

Dongsong Zhang and Lina Zhou straipsnyje "*Discovering Golden Nuggets: Data Mining in Financial Application*" rašoma apie galimybę duomenų gavybą naudoti finansiniams duomenims apdoroti bei sprendimams priimti. Naudojant duomenų gavybą finansinių duomenų analizei siekiama išgauti įvairių rinkų finansinių rodiklių priklausomybes bei rinkų ateities prognozes (Zhang D., 2004).

Anglų kalboje trumpinys OLAP siejamas su *OnLine Analytical Processing* sąvoka. Terminas apibūdina programinius produktus skirtus įvairiapusiškai analizuoti verslo informaciją realiu laiku. Sąveika su tokiomis sistemomis vyksta interaktyviai, atsakymai net į daug skaičiavimų reikalaujančias užklausas gaunami per kelias sekundes. Galutinė informacija gali būti pateikta ne tik skaičiais, bet ir grafiniu pavidalu.

Dažniausiai OLAP sistemų duomenų bazėje sukaupta informacija vienu metu gali naudotis daug vartotojų, todėl tokiose programose dažniausiai numatyti ir įvairaus lygio saugumo apribojimai skirtingas priėjimo prie duomenų teises turintiems vartotojams (Naujoji komunikacija, 2003).



Šaltinis: sudaryta autoriaus

1 pav. OLAP kubo pavyzdys

OLAP - tai greita daugiamačių duomenų analizė. OLAP sistemų struktūra - tai daugiamačiai kubai, kuriuos galima nagrinėti įvairiais pjūviais (dimensijom). Duomenų gavybos principas kiek kitoks – ji ieško tendencijų, modelių, dėsningumų dideliuose duomenų kiekiuose. Tam pasitelkia matematinius, statistinius bei modelių atpažinimo metodus (IT Ekspertas, 2005). OLAP sistemos turi

kelis skirtingus sistemos fizinės realizacijos modelius: ROLAP (tai reliacinė OLAP sistemų architektūra), MOLAP (daugiamatė OLAP sistemų architektūra) ir HOLAP (hibridinė architektūra gaunama sujungus ROLAP ir MOLAP savybes) (Sekliuckis V., 2006, p. 293). Taip pat vartojami akronimai WOLAP (internetinis OLAP), DOLAP (darbalaukio OLAP), RTOLAP (realaus laiko OLAP).

Antrame paveiksle (1 pav.) pavaizduotas OLAP kubas galintis pateikti skirtingų produktų, skirtingų mėnesių skirtingus finansinius rodiklius. Paveiksle išskirtas tas kubo elementas, kuris atspindi užduotą užklausą, t.y. atrenkama prekė – dviratis, mėnuo – vasaris, finansinis rodiklis – grynasis pelnas.

OLAP sistemų programinė įranga pritaikoma ir dažniausiai naudojama pardavimų ir rinkodaros analizei, vartotojų ir produktų pelningumui, finansinių ataskaitų konsolidacijai, paslaugų ir prekių poreikiui numatyti, finansiniam modeliavimui, vartotojų grupėms nustatyti ir segmentuoti (Naujoji komunikacija, 2003).

„Žinių bazė (angl. *knowledgebase*; trumpinama KB, kb, arba Δ) yra speciali duomenų bazė, skirta žinioms tvarkyti. Ji suteikia priemones kompiuterizuotam žinių rinkimui, tvarkymui ir paieškai“ (Leonardo da Vinci Programa, 2009)

Žinių bazės dažniausiai būna geriausios praktiškos veiklos rezultatai, žmonijos sukaupti faktai, taisyklės, principai, kiti pažinimo subjektai. Todėl žinių bazėse, priešingai nei duomenų bazėse, pateikiami duomenys, esantys dokumentuose, o ne tokie objektai kaip knygos, straipsniai, dokumentai. Visi šie duomenys sudaro žinių visumą, charakterizuojančią pažinimo subjektus. Žinių bazėse saugoma nedaug žinių, bet jo užimama vieta yra didelė lyginant su paprastomis duomenų bazėmis. Žinių bazė, kurią sudaro faktai ir taisyklės, nėra pastovi, ji visą laiką papildoma nauja informacija ir naujomis taisyklėmis. Kai kurios iš jų taip pat gali būti panaikintos arba pakeistos. Taigi žinių bazė turi būti nuolat atnaujinama. (Broniukaitis R., 1997).

Ekspertinėms sistemoms dažniausiai naudojamos dvi žinių bazės – statistinė ir dinaminė. Statistinėje saugomos žinios, charakterizuojančios konkrečią pažinimo sritį, nekintančios uždavinio sprendimo eigoje, dinaminėje esantys duomenys nekinta sprendžiant konkretų uždavinį, tačiau gali keistis sprendimų procese (Broniukaitis R., 1997).

Yra dvi pagrindinės žinių bazių kategorijos:

- Žinių bazė skirta mašinoms. Tokia bazė naudojama siekiant automatizuoti dedukcinį protavimą. Jas sudaro duomenų rinkiniai, dažniausiai saugomi taisyklių pavidalu. Taisyklės aprašo žinias siekiant loginio suderinamumo. Gali būti naudojami loginiai operatoriai (konjunkcija,

disjunkcija, papildymas ir neigimas), kuriais žinių bazė sukuriama iš menkiausių detalių. Galiausiai naudojama tradicinė dedukcija atlikti išvadoms.

- Žinių bazė skirta žmogui. Tokios bazės sukurtos tam, kad žmonės galėtų gauti ir naudoti žinias, pirmiausiai apmokymo tikslais. Žinių bazėje duomenis turi būti nuolatos atnaujinami, ji turi turėti gerą išgavimo sistemą (paieškos mechanizmą), struktūrizuotą turinį ir klasifikuotą struktūrą.

Žinių bazės naudojamos ne tik priimant sprendimus, bet ir sprendžiant dirbtinio intelekto problemas ir uždavinius. Šios žinių bazių rūšys gali siūlyti sprendimus problemoms, jos kartais būna pagrįstos ir vartotojo grįžtamuoju ryšiu.

Visos intelektinės sistemos skirtos efektyvesniam įmonėse kaupiamų duomenų ir informacijos panaudojimui. Tačiau visos jos veikia skirtingai ir turi savų privalumų.

Duomenų gavybos sistemų unikalumas yra tas, jog jų pateikiamas rezultatas būna ne atsakymai į pateiktus klausimus, o naujų priklausomybių radimas. Duomenų gavybos sistemos sugeba atrasti tokias priklausomybes, apie kurias net nebuvo įtarta (atsakyti į neužduotus klausimus). Nagrinėjant tokius duomenis, gauti rezultatai net ir gerai veikiančioje įmonėje padeda atrasti nedidelius nesklandumus, kurie padeda pagerinti įmonės veiklą.

Tuo tarpu OLAP daugiamačio kubo paskirtis yra kitokia. Jo užduotis yra pavaizduoti sukauptus duomenis įvairiomis dimensijomis. Taip vartotojui pateikiamas realaus pasaulio vaizdas, kuris jam priimtinas ir suprantamas. OLAP technologija neieško jokių paslėptų ryšių, tiesiog sugeba sukauptus duomenis atvaizduoti dinamiškai, su galimybe detalizuoti duomenis (angl. *drill down*). OLAP kubas puikiai tinka dideliems duomenų kiekiams atvaizduoti.

Žinių bazės kitaip nei dvi pirmosios verslo sistemos dirba ne su duomenimis o su faktais bei taisyklėmis. Jos dažniausiai skirtos gan siaurai sričiai, kuriai projektuojama sistema. Žinios, kurias sudaro faktai ir taisyklės, nėra pastovios, jos visą laiką papildomos naujais faktais ir taisyklėmis. Kai kurios iš jų taip pat gali būti panaikintos arba pakeistos. Taigi žinių bazė turi būti nuolat atnaujinama.

Visas šias sistemas galima pritaikyti įvairioms įmonėms užsiimančiomis įvairia veikla. Tačiau mažoms įmonėms kasdienėms situacijoms spręsti aktualiausia turėtų būti duomenų gavyba. Ją galima pritaikyti labai plačiai įvairaus profilio įmonėse. Tolesniuose skyriuose detaliau išnagrinėti duomenų gavybos algoritmai bei galimybės ją taikyti mažose įmonėse.

1.3. Prognozės verslo intelektinių sistemų raidai

Valerie Logan, plačiai pasaulyje žinomos Technology Solutions Group, HP komandos narė. Jos pateiktame straipsnyje apie 2008 metų verslo intelektinių sistemų prognozes „*Top 10 Trends in*

Business Intelligence for 2008“ išskiriamos kelios pagrindinės verslo intelektinių sistemų veiklos kryptys:

- BI naudojimas kartu su debesies architektūra;
- BI taikymas dideliems nestruktūrizuotos informacijos kiekiams apdoroti;
- BI tampa ne tiek IT, kiek verslo dalimi;
- Vis daugiau dėmesio BI skiria papildomom funkcijom – resursų planavimui, CRM ir pan.

Svarbiu akcentu laikoma tai, jog specialūs BI sistemų įrankiai suteikia vis daugiau galimybių, galinčių padėti sumažinti atotrūkį tarp priimamų sprendimų ir įmonės veiklos naudojant pačią naujausią, laiku pateikiamą apdorotą informaciją sprendimus priimantiems asmenims. Vis dažniau BI technologijos naudojamos itin plačiai – nuo kelio užkirtimo sukčiavimams iki vertės grandinės efektyvumo gerinimo (Logan V. 2008).

Taip pat pastebimas vis didėjantis nestruktūrizuotos naudingos informacijos kiekis pateikiamos el. paštu, tiesioginėmis žinutėmis, video įrašais ir pan. Tad viena iš BI veiklos krypčių įvardijamas šios informacijos aptikimas ir efektyvus apdorojimas. Iš BI technologijų taip pat reikalaujama suderinamumo su kitais naudojamais įrankiais ir pritaikomumas konkrečiai kompanijai pagal jos poreikius (Logan V. 2008).

Dar vienu svarbiu akcentu yra laikomas BI technologijų suderinamumas su internetu. Kuo toliau tuo labiau siekiama, kad tokios interneto galimybės kaip blog‘ai, wiki puslapiai, tiesioginės žinutės taptų BI tiekimo mechanizmo dalis (Logan V. 2008).

Mycustomer.com tinklapyje yra pateikiamos 2009 metų tendencijos. Publikacijoje „*Business intelligence trends for 2009“ Gartner* grupės analitikai Bill Hostmann, Kurt Schlegel, Bill Gassman, Nigel Rayner, Neil McMurchy ir Matthew W. Cain išskyrė šiuos pagrindinius punktus (Hostman B., 2009):

- 2012 metais daugiau nei 35% kompanijų esančių top 5,000 pasaulinių kompanijų sąraše reguliariai klys spręsdamos apie reikšmingus pokyčius jų veikloje bei rinkose;
- Iki 2010 metų, 20% organizacijų turės tam tikrai pramonės šakai specifines analitines priemones kaip paslaugas pateikiamas per programinę įrangą (SAS) kaip standartinį jų verslo portfelio komponentą;
- Iki 2012 merų verslo vienetai skirs mažiausiai 40% bendrojo biudžeto verslo intelekto sistemoms;
- 2009 metais bus plėtojama sprendimų priėmimo produktų kategorija sugebanti apjungti socialinę programinę įrangą su BI platformos galimybėmis;

Kaip matome iš visų šių pastebėjimų, verslo intelekto sistemos yra nuolat tobulėjanti ir besiplečianti sritis. Vis daugiau didelių, vidutinių ir net smulkių įmonių pripažįsta BI sistemų naudą, vis dažniau jas diegia savo kompanijose siekdami efektyvesnės veiklos.

1.4. Duomenų gavybos algoritmai ir jų palyginimas

Duomenų gavybos sritis yra labai plati. Ji apima nemažą skaičių įvairių metodų, algoritmų, taikomųjų sistemų. Vieni duomenų gavybos algoritmai geriau klasifikuoja duomenis, kiti geriau prognozuoja. Visi jų turi savų privalumų bei trūkumų. Skirtingi algoritmai veikdami skirtingais metodais padeda gauti skirtingus rezultatus. Dažnai dėl to rezultatai tampa reikšmingesni, svarbesni ir lengviau panaudojami, nes leidžia į situaciją pažvelgti iš skirtingų pozicijų. Dažniausiai išskiriami šeši duomenų gavybos metodai (Fayyad U., 1996).

1. Klasifikavimas: remiantis duomenų savybėmis skirsto juos į iš anksto nustatytą kiekį tam tikrais parametrais aprašytų klasių;
2. Regresija: remiantis turimais duomenimis sudaromo modelį leidžiantis prognozuoti realias būsimas skaitines reikšmes.
3. Klasterizavimas: skirsto duomenis į iš anksto nenumatytą kiekį grupių ar klasterių pagal skiriamuosius bruožus.
4. Apibendrinimas: apima metodus padedančius sutrumpinti duomenis. Šie metodai dažniausiai naudojami interaktyviom žvalgomosioms analizėm, automatinėm ataskaitom ruošti;
5. Priklausomybių modeliavimas: randa modelį, kuris aprašo priklausomybes tarp kintamųjų.
6. Pakeitimų ir nukrypimų aptikimas: skirtas atskleisti svarbiausius pokyčius tarp anksčiau išmatuotų ir naujausių duomenų.

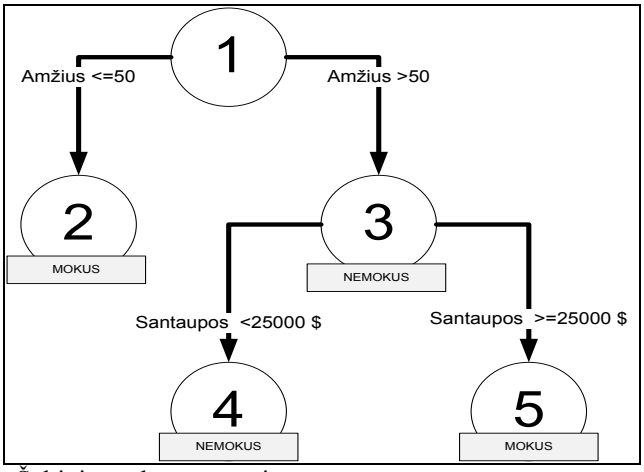
Kiekvienam iš šių metodų galima priskirti net po keletą įvairių algoritmų. Toliau pateikiami dažniausiai sutinkami ir populiariausi duomenų gavybos algoritmai (Fayyad U., 1996), (Wu X., 2008).

Sprendimų medis - tai struktūrinė schema, kuri yra panaši į medį. Išsišakojimai reiškia vieną ar kitą atsakymą į siekiamus išsiaiškinti klausimus. Tokiu būdu yra sudaromos taisyklės, kurios klasifikuoja nagrinėjamą duomenų aibę atsižvelgdamos į elementų savybes. Proceso pradžioje turėta duomenų aibė yra tol skaidoma į šakas, kol kiekviena iš jų tampa homogeniška. Pavyzdyje (2 pav.) vaizduojama situacija, kurioje sprendžiama, ar banko klientui suteikti paskolą ar ne.

Pagrindiniu sprendimų medžių privalumu laikomas jų aiškumas ir suprantamumas bei lankstumas.

Sprendimo medžio formavimo etapai (StatSoft, 2008):

- Prognozavimo tikslumo kriterijų nustatymas – norint minimizuoti sprendimų medį būtina nustatyti kriterijų eiliškumą (prioritetus), jų reikšmingumą.
- Šakojimųsi parinkimas – medis turi prasidėti nuo vieno „šakninio“ (root) išsišakojimo ir skaidytis į šakas pagal nustatytą hierarchiją.
- Šakojimųsi baigimo nustatymas – labai svarbu nustatyti kada baigti skaidymą. Nereiktų pasiekti tokio šakojimosi lygio kai kiekvieną galutinę reikšmę atitinka tik vienas iš rūšiuojamų objektų.



Šaltinis: sudaryta autoriaus

2 pav. Sprendimų medžio pavyzdys

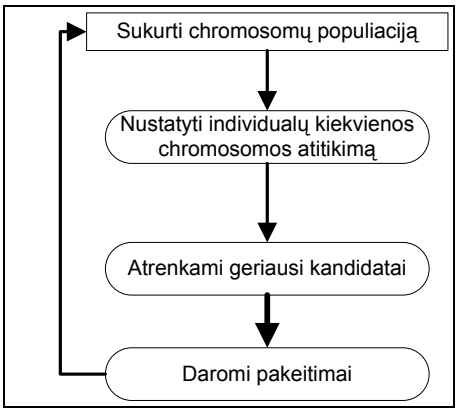
Yra du pagrindiniai būdai šakojimųsi baigimo nustatymui (StatSoft, 2008):

- minimalus n (minimum n) – skaidoma tol, kol kiekvienas mazgas turi daugiau ar nustatytą minimalų skaičių objektų. Nustatytas minimalus objektų skaičius vadinamas minimaliu n;
- objektų frakcijos (fraction of objects) – skirstymą leidžia tol, kol frakcijų (dalių) į kurias skirstomi objektai yra daugiau ar tiek kiek nustatyta.

Tinkamo dydžio medžio parinkimas – sprendimų medį geriau pasirinkti tokio dydžio, kuris gerai neklasifikuoja mokymosi duomenų, tačiau kiek įmanoma geriau prognozuoja testinius duomenis. Tinkamam sprendimų medžio dydžio parinkimui yra sukurta daug metodų – FACT-style direct stopping, Test sample cross-validation, V-fold cross-validation ir kt. (StatSoft, 2008).

Algoritmai, kurių veikimo principas yra pagrįstas selekcijos, mutacijos bei kryžminimo procesais ir kurie remiasi evoliucijos koncepcija, vadinami **genetiniais algoritmais** (3 pav.). Jie skirti neuroninių tinklų mokymui, funkcijų optimizavimui, klasifikavimo, prognozavimo ar regresijos uždavinių sprendimui. Algoritmo veikimo principas paremtas galimų sprendinių kodavimu simbolių eilutėmis (chromosomomis).

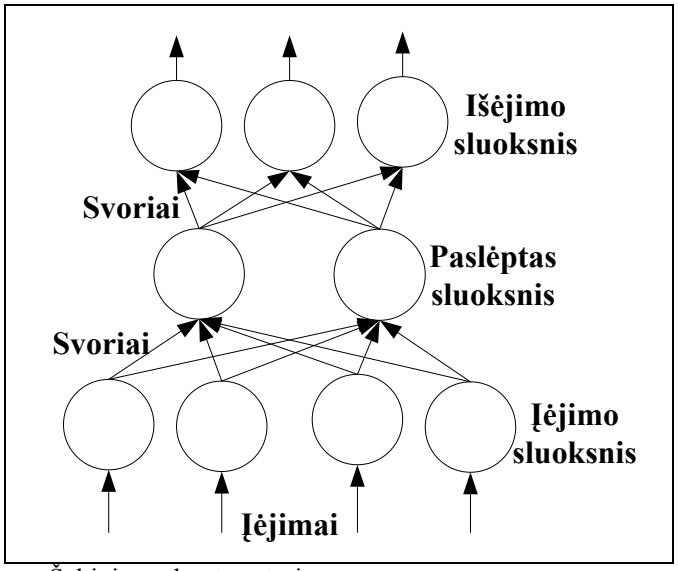
Išskirtinis šio algoritmo bruožas tas, kad ieškoma ne vieno galimo sprendinio, o tam tikros jų aibės, t.y. generuojama visa galimų chromosomų populiacija. Pranašesni už daugelį kitų algoritmų tuo, kad uždaviniui spręsti užtenka minimalios informacijos. Vienas iš didžiausių trūkumų – negarantuoja optimalaus problemų sprendimo. Genetiniai algoritmai apima neuroninius tinklus, ekspertines sistemas, fuzzy logiką ir pan. (Ciesiūnas A., 2007). Genetiniai algoritmai dažniausiai naudojami finansiniuose modeliavimuose, informacinėse sistemose, sprendimų priėmimo sistemose.



Šaltinis: sudaryta autoriaus pagal Court Merz, 2007.

3 pav. Genetinio algoritmo pavyzdys

Neuroninių tinklų algoritmai yra naudojami prognozavimo bei klasifikavimo uždaviniams spręsti. Tai netiesiniai modeliai savo struktūra primenantys biologinius neuroninius tinklus.



Šaltinis: sudaryta autoriaus

4 pav. Neuroninių tinklų algoritmo pavyzdys

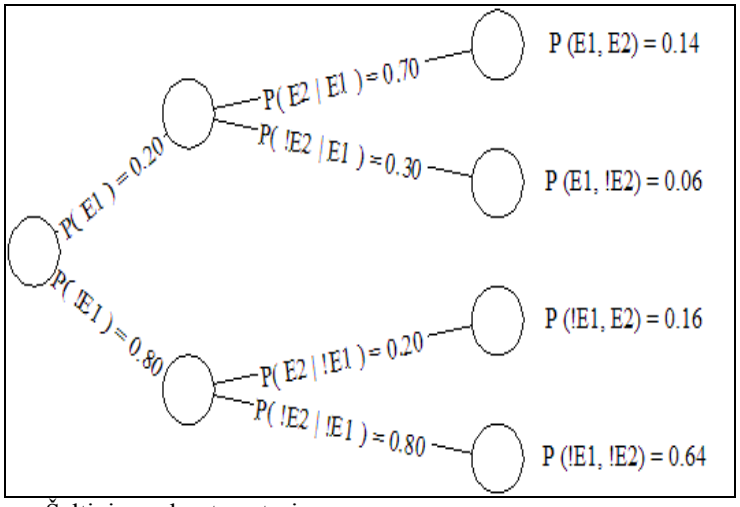
Neuroniniai tinklai suteikia galimybę mokytis iš duomenų ir praplėsti ateities numatymo ribas. Pats tinklas yra sudarytas iš kelių sluoksnių: įvesties, išvesties ir vieno ar daugiau paslėptų sluoksnių (4

pav.). Įvesties sluoksnyje yra elementą aprašanti informacija, išvesties sluoksnyje yra gaunamas rezultatas, o paslėptuose sluoksniuose realizuojama uždavinio logika (Ciesiūnas A., 2007).

Nagrinėjami elemento duomenys yra perverčiami į didesnį ar mažesnį svorį ir šių svorių parinkimas yra formuluojamas kaip pagrindinis šio algoritmo uždavinys. Pagrindinė neuroninių tinklų savybė yra galimybė mokytis. Neuroninių tinklų algoritmas yra pakankamai galingas ir produktyvus, tačiau reikalauja didelių išteklių. Todėl šis metodas rekomenduotinas tuomet, kai reikalinga tiksli analizė, o žadamas pelnas leidžia investuoti į brangius skaičiavimų išteklius (Ciesiūnas A., 2007).

Bajeso klasifikacijoje atsižvelgdama į pasiskirstymo tikimybes Bajeso klasifikacija gali pasiekti optimalų rezultatą. Bajeso metodas remiasi tikimybių teorija. Yra vienas apribojimas, kurio Bajeso taisyklė negali peržengti – tikimybių įvertinimas mokymo duomenyse. Yra pastebėta, kad kai kuriose situacijose, kuomet sprendimas yra paremtas tam tikrais kriterijais, ar kai duomenys yra visiškai atsitiktiniai, Bajeso klasifikacijos taikyti negalima. Yra išskiriami keli Bajeso klasifikacijos atvejai.

Naive Bajeso Klasifikacija – labiausiai tinka kai įėjimo duomenų pasiskirstymas didelis. Nepaisant savo paprastumo Naive Bajeso klasifikacija dažnai būna efektyvesnė nei kiti įmantresni klasifikavimo metodai (Xiao H., 2004).



Šaltinis: sudaryta autoriaus

5 pav. Bajeso tinklų pavyzdys

Gausinė Bajeso Klasifikacija – įvertina ir tikimybines tankio funkcijas. Naudodami Gauso teorijas mes galime naudoti tą patį Bajeso klasifikavimo modelį tik su tam tikru kovariacijos (priklausomybės) aptikimu. Paprastai tai duoda daug tikslesnių rezultatų (Xiao H., 2004).

Bajeso *tinklai* - naudoja sąlygines tikimybes tarp įvairių kintamųjų. Paprastai neįmanoma sugeneruoti visų sąlyginių tikimybių iš turimų duomenų. Mūsų užduotis yra išrinkti svarbiausiais iš jų

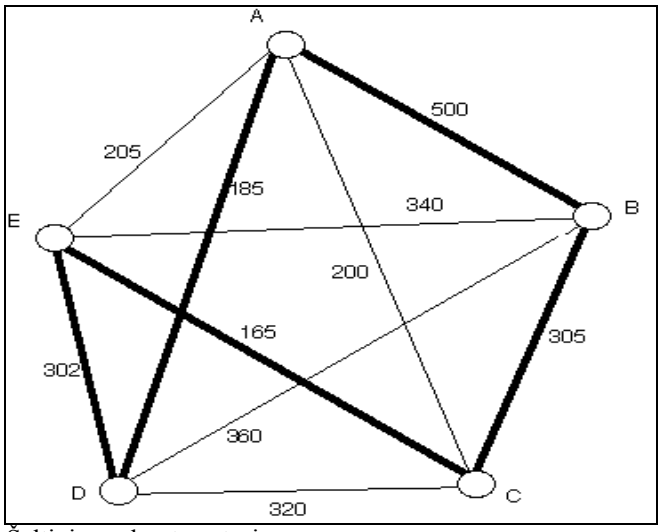
ir naudoti jas klasifikavimo procese. Taigi galima teigti, jog Bajeso tinklas yra apibendrintos tikimybinės tankio funkcijos (Xiao H., 2004).

Bajeso tinklų pavyzdyje (5 pav.) pavaizduota situacija kuomet norima nustatyti kokios oro sąlygos bus ateinančias dvi dienas. E1 – tai tikimybė, jog rytoj lis, E1! – tikimybė, kad nelis. E2 – tikimybė kad poryt lis, o E2! – kad nelis. Įvertinus tikimybes matome, jog didžiausia tikimybė, jog artimiausias dvi dienas nelis.

Artimiausio kaimyno metodu klasifikavimo uždaviniai yra sprendžiami ne grupuojant elementus pagal nustatytas ar analizės metu rastas taisykles, bet pagal elemento turimų savybių panašumą į jo kaimynus. Algoritmo tikslas yra klasifikuoti naują objektą pagal atributus. Šie algoritmai remiasi nagrinėjamo objekto lyginimu su prieš tai buvusiais. Klasifikatoriai nenaudoja jokio iš anksto sudaryto modelio.

Artimiausių kaimynų metodo veikimas yra labai paprastas (Teknomo K., 2006):

- nusistatome parametrą K – artimiausių kaimynų skaičių;
- apskaičiuojame atstumus tarp naujo elemento ir jau išskirstytų, ar įvestų apmokymo elementų;
- surūšiuojame elementus pagal atstumą ir atrenkame K arčiausiai esančių elementų;
- nustatome, kokiai sričiai (parametrui) priskirti K arčiausiai esantys kaimynai;
- naujam elementui parenkame sritį (parametrą), kuri turi daugiausiai atrinktų kaimynų.



Šaltinis: sudaryta autoriaus

6 pav. Artimiausių kaimynų algoritmo pavyzdys

Paveiksle (6 pav.) pavaizduota situacija, kuomet pasirinktas parametras k=4, nes iš kiekvienos viršūnės galima pasiekti keturis artimiausius kaimynus. Pradžios taškas A. Reikia apeiti visus taškus naudojant artimiausių kaimynų algoritmą. Kaip matome iš taško A trumpiausias kelias iki taško D, iš D

į E ir t.t., kol grįžtama iki pradinio taško A. Didžiausias šio algoritmo trūkumas - sudaryto modelio dydis.

Naujoji duomenų gavybos karta yra įvairūs hibridiniai algoritmai. Jie sudaromi apjungiant kelis paprastus algoritmus į vieną, bandant padidinti jų efektyvumą, sumažinti atskirų algoritmų trūkumus. Hibridinės sistemos ir hibridiniai modeliavimo bei valdymo metodai suteikia naujų galimybių efektyviau valdyti gamybos ir verslo procesus. Hibridinių algoritmų yra labai daug, tad visus juos apžvelgti būtų sudėtinga, tad pasirinkau keletą.

Neuro-fuzzy algoritmas yra hibridinis algoritmas apimantis neuroninių tinklų bei fuzzy sistemas. Taisyklėmis paremta fuzzy sistema, neuro-fuzzy modeliuose, suprantama kaip neuroninis tinklas. Neuro-fuzzy algoritmas yra naudojamas įvairioms sritims – kontrolei, duomenų analizei, sprendimų priėmimams, ir pan.

NBTree yra sprendimų medžio bei Naive Bajeso hibridas. Jis sukuria sprendimų medžius, kurių „lapai“ yra Naive Bajeso klasifikatoriai tiems atvejams, kurie pasiekia būtent tą sprendimų medžio šaką. Konstruojant medį naudojamas kryžminis patvirtinimas tam, kad nustatyti ar sprendimų medžio mazgas turi šakotis toliau, ar čia jau reikia naudoti Naive Bajeso modelį (Witten I.H., 2005, p. 408)

Neuro-ekspertinė sistema apjungia klasikinės ekspertinės sistemos bei neuroninių tinklų savybes. Kadangi klasikinės ekspertinės sistemos neturi savybės apsimokyti, tam puikiai tinka neuroniniai tinklai. O neuroninių tinklų trūkumą – negalėjimą paaiškinti savo pateikiamų sprendimų padeda spręsti klasikinės ekspertinės sistemos. Sistemoje žinių bazė formuojama neuroninio tinklo pagalba. Taisyklių formavimo blokas išanalizavęs neuroninio tinklo įėjimus ir išėjimus formuoja „jei... tai...“ (angl. *If.... then*) taisykles (Simutis R., 2008).

Sprendimų medžio/genetinis algoritmo pagrindinė idėja yra dviejų fazių mokymas. Sprendimų medžiai naudojami klasifikuoti pavyzdžius į didelius skyrius. Genetiniai algoritmai priešingai – yra stipresni, lankstesni, tad sugeba susidoroti su duomenimis skirstydamas juos į mažesnius skyrelius. Pirmojoje fazėje algoritmas tradiciškai skirsto duomenis į skyrius, kurie turi vieną parametrą – didelis skyrius arba mažas skyrius, priklausomai nuo to, kiek elementų tam skyriui priskiriama (lyginama su iš anksto nustatytų „slenksčiu“). Antroje fazėje algoritmas nustato taisykles aprašančias pavyzdžius priklausančius mažiems skyriams (Carvalho D., 2000).

Toliau pateiktoje lentelėje (2 lentelė) atliktas pagrindinių, aukščiau aptartų, duomenų gavybos algoritmų palyginimas. Įvardinti pagrindiniai jų privalumai, trūkumai, ir dažniausios bei aktualiausios panaudojimo sritys. Nors hibridinių algoritmų yra labai daug, šioje lentelėje jie pateikti viename

punkte. Taip yra todėl, kad labai sunku išskirti, kurie iš jų yra populiariausi ar geriausi. Be to, lyginant su kitais duomenų gavybos algoritmais, hibridiniai algoritmai yra palyginti nauja algoritmų evoliucijos pasekmė.

2 lentelė

Pagrindinių duomenų gavybos algoritmų palyginimas

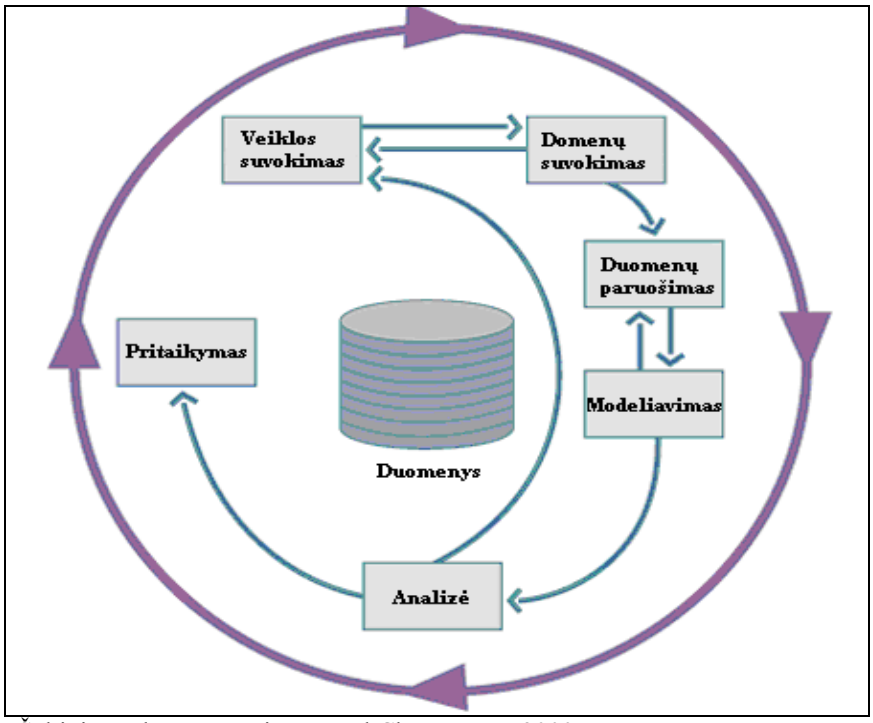
Algoritmas	**Privalumai**	**Trūkumai**	**Panaudojimas**
Sprendimų medžio	- Nustato netikėtus ryšius; - Identifikuoja pogrupių skirtumus; - lengvai suprantamas; - nesunkiai apdoroja duomenis su daug požymių; - pagrįstai išnaudoja laiką.	- galimi netikri ryšiai; - sunkiai randa sudėtingus ryšius tarp elementų;	- ligoms nustatinėti; - įrangos veikimo problemoms spręsti; - skiriamų paskolų dydžiui nustatyti; - gali būti naudojamas prognozavimui; - kai reikia pasirinkti vieną iš kelių galimų veikimo planų; - norint grafiškai pavaizduoti skirtingus problemos sprendimo būdus, iliustruoti turimas galimybes, apribojimus, skirtingų variantų sąsajas.
Bajeso klasifikacija	- greitai apmokomas; - plačiai panaudojamas.	- kai kuriose situacijose, kuomet sprendimas yra paremtas tam tikrais kriterijais, ar kai duomenys yra visiškai atsitiktiniai, Bajeso klasifikacijos taikyti negalima.	- Plačiai naudojamas tekstų klasifikacijai; - dokumentų skirstymui pagal turinį (spam‘o filtravimui) - Bajeso tinklai naudojami žinių modeliavimui bioinformatikoje, medicinoje; - informacijos atitaisymui; - vaizdų apdorojimui; - sprendimų priėmimo sistemose.
Artimiausių kaimynų	- gerai apdoroja triukšmingus duomenis - efektyvus dirbant su dideliais duomenų kiekiais	- sudaryto modelio dydis - reikalauja apibrėžti k parametrą; - reiklus kompiuterio resursam	- vaizdų apdorojimui; - klasifikavimo uždavinių sprendimui.
Neuroniniai tinklai	- gali mokytis; - galingas; - produktyvus; - gali apdoroti daug duomenų; - dirba su įvairiais duomenų tipais.	- reikalauja didelių išteklių - ilgai trunka mokymasis - sunkiai interpretuojami rezultatai.	- medicininių fenomenų aptikimui; - akcijų rinkų prognozavimui; - kreditų paskyrimams; - įrengimų būklės stebėjimui.
Genetiniai	- ieškoma ne vieno galimo sprendimo, o jų aibės; - uždaviniui spręsti nereikia daug informacijos.	- negarantuoja optimalaus problemos sprendimo;	- finansiniuose modeliuose; - informacinėse sistemose - darbų tvarkaraščio sudarymui; - sprendimų priėmimo sistemose.
Hibridiniai	- apjungiami keli algoritmai; - algoritmai papildo vienas kitą	- kai kurie algoritmai sunkiai suderinami.	- prognozavimui, klasifikavimui ir pan., priklausomai nuo naudotų algoritmų savybių.

Šaltinis: sudaryta autoriaus.

Kiekvienai sprendžiamai situacijai galima pritaikyti net po keletą duomenų gavybos algoritmų. Dažnai jų pateikiami rezultatai kiek skiriasi, tačiau lyginant skirtingų algoritmų rezultatus galima gauti efektyvesnį galutinį rezultatą.

1.5. CRISP-DM duomenų gavybos modelis

CRISP-DM (angl. *Cross-Industry Standard Process for Data Mining*) – standartinis duomenų gavybos proceso modelis (7 pav.), kuris buvo sudarytas 1996 metais. Šio modelio sudarytojai, duomenų gavybos pradininkai – Daimler Chrysler, SSPS, Teradata. Pagrindinis šių organizacijų tikslas integruoti duomenų gavybą į verslo aplinką, plėtojant duomenų gavybos principus. Modelis buvo kuriamas remiantis ne tiek teorija, kiek praktika, todėl jis yra toks paplitęs ir dažnai naudojamas (The CRISP-DM consorcium, 2003).



Šaltinis: sudaryta autoriaus pagal Chapman P., 2000.

7 pav. Standartinis duomenų gavybos procesų modelis

Duomenų gavybos procesas - tai uždaras ciklas, kai įvykdžius vieną žingsnį neišvengiamai reikia grįžti į prieš tai padarytus žingsnius. Tai svarbu norint užtikrinti proceso vientisumą, kokybišką eigą bei norint išvengti galimų klaidų. Norint duomenų gavybos procesą atlikti kokybiškai reikia tuos pačius veiksmus kartoti po keletą ar net keliasdešimt kartų. Tai sunkus ir didelio pasiruošimo reikalaujantis procesas.

Duomenų gavybos sistemą sudaro 6 pagrindiniai duomenų gavybos kūrimo žingsniai (Chapman P., 2000):

1. Veiklos suvokimas - išanalizuoti verslo aplinką, apibrėžiant verslo problemas bei sukuriant veiklos modelį;
2. Duomenų suvokimas – jo metu suformuojama duomenų gavybos duomenų bazę (t.y. pasirenkamas duomenų gavybai naudojamas duomenų šaltinis);
3. Duomenų paruošimas – organizacijos duomenų ištyrimas ir parengimas duomenų gavybai;
4. Modeliavimas – duomenų gavybos modelio sukūrimas;
5. Analizė – duomenų gavybos modelio įvertinimas, išskleidimas ir rezultatų apžvalga;
6. Pritaikymas – gautų rezultatų pateikimas vartotojui ir pritaikymas įmonės veikloje.

Galima teigti, kad šis modelis yra hierarchinės struktūros, nes kiekvienas kūrimo žingsnis gali būti skaidomas į smulkesnius. Kiekvienas konkretus uždavinys išplėtojamas į specializuotas užduotis, kurios galutiniame etape pateikia laukiamus rezultatus (Chapman P., 2000). Tačiau toks sudėtingas modelis geriausiai tinka didelėms įmonėms. Mažoms įmonėms užtenka naudoti supaprastintą šio modelio variantą, kai dalis procesų yra supaprastinami ar visai panaikinami.

1.6. Duomenų gavybos programinės įrangos apžvalga

Pasaulyje vis daugiau įmonių savo verslo analizei ir prognozavimui naudoja įvairius duomenų gavybos įrankius. Visų šių įrankių gamintojai sudaro duomenų gavybos rinką. Šiuo metu rinkoje esantys duomenų gavybos įrankiai atlieka duomenų analizę, verslo ar vartotojų elgesio prognozavimą, galimybę naudoti įvairius statistinius modelius ir vizualines priemones.

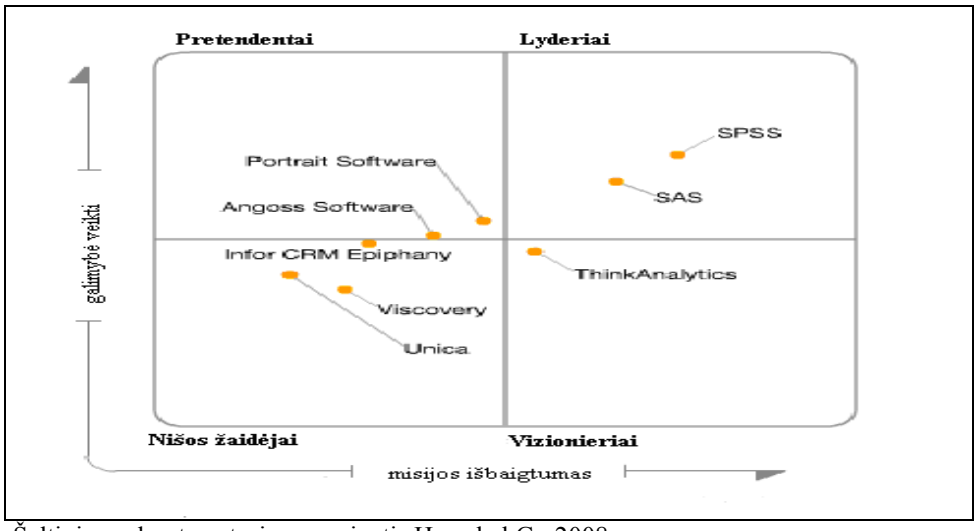
Klasikiniai įrankiai (komerciniai)	SAS Enterprise Miner 5.3
	SPSS Clementine 12
Klasikiniai įrankiai (atviro kodo)	Rapid-I: Rapidminer 4.2
	Universitat Konstanz: KNIME 1.3.5
	Universitat Waikato:Weka 3.4.13
Savarankiškai veikiantys įrankiai	KXEN Analytic Framework 4.04
Sunažinto funkcionaluno įrankiai (skirti tam tikrai taikymo sričiai)	Viscovery SOMine 5.0
	Prudsys Discoverer 5.5/Basket Analyzer 5.2
	Bissantz Delta Master 5.3.6
Padidinto funkcionaluno įrankiai	SAP NetWeaver 7.0 Data Mining Workbench
	ORACLE 11g Data Mining
	SQL Server 2005 Analysis Services

Šaltinis: sudaryta autoriaus remiantis Mayato, 2008

8 pav. Duomenų gavybos programinės įrangos skirstymas

Pastaraisiais metais duomenų gavybos įrankių rinka yra labai dinamiška, todėl kompanijos, nusprendusios pasinaudoti duomenų gavybos nauda, gali rinktis iš daugybės įrankių. Daugiau nei 50 mokamų ir daugiau nei 15 nemokamų duomenų gavybos įrankių sąrašą pateikia KDnuggets (angl. *The Association for Knowledge Discovery and Data Mining*) (KDnuggets, 2009a). Business Application Research Center (BARC) vien Vokietijos BI programinės įrangos pardavimui prognozuoja 10 – 12 % augimą iki 2012m. (Bange A., 2008). Nagrinėjant duomenų gavybos programinę įrangą svarbu atkreipti dėmesį į jos klasifikaciją. Pagrindiniai jos tipai pateikti 8 paveikslėlyje.

Kaip matome, pirmosios dvi grupės yra klasikiniai duomenų gavybos įrankiai, tačiau pastaruoju metu labai padaugėjo nemokamų įrankių, tad jie išskiriami į atskirą klasikinių duomenų gavybos įrankių grupę. Trečioji grupė yra savarankiškai veikiantys duomenų gavybos įrankiai. Šie automatizuoti įrankiai veikia be rankinio duomenų paruošimo ir parametrų nustatymo. Ketvirtoji - sumažinto funkcionalumo įrankiai, jie skirti tam tikrai pritaikymo sričiai (pvz.: kontrolei), analizės atvejui (pvz.: užduoties kontrolei ir klasifikacijai) ar jų kombinacijai. Paskutinioji grupė yra įrankiai kuriuose vis dažniau atsiranda platesnės funkcijos duomenų analizei (Mayato, 2008).



Šaltinis: sudaryta autoriaus remiantis Herschel G., 2008

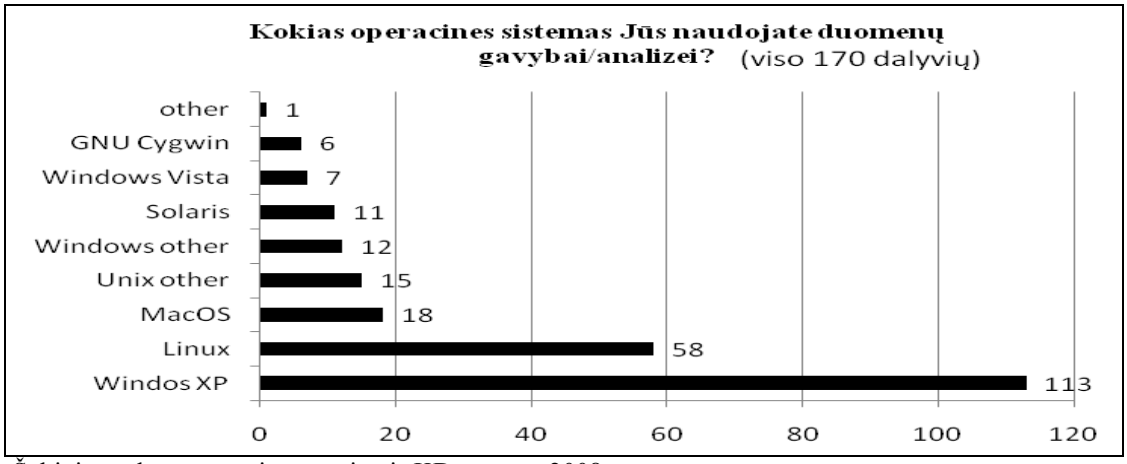
9 pav. 2008 metų duomenų gavybos įrankių "Magic Quadrant"

Pasaulyje pirmaujanti informacinių technologijų mokslinių tyrinėjimų bei konsultantų kompanija Gartner 2008m. birželį pateikė 9 paveiksle pavaizduotą Magic Quadrant išskirstydama duomenų gavybos įrankius (Herschel G., 2008).

Kaip matome iš Gartner kompanijos modelio rinkos lyderiai (geriausiai tenkinantys klientų poreikius bei turi didelį poveikį rinkos kryptims ir augimui) yra SPSS ir SAS. Rinkos pretendentai (*challengers*) – Portrait Software, Angoss Sotware.

Apžvelgiant duomenų gavybos programinę įrangą svarbu išsiaiškinti kokią operacinę sistemą analitikai dažniausiai naudoja, kadangi skirtingoms operacinėms sistemoms yra siūlomi skirtingų firmų produktai.. KDnuggets pateikia 2008 m. duomenis. Jie atvaizduoti 10 paveikslėlyje.

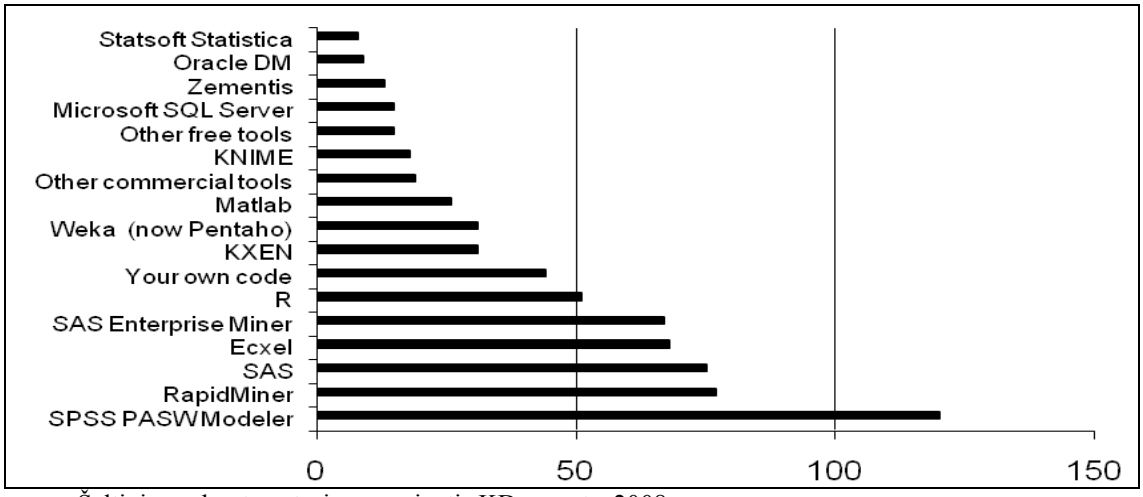
Kaip matome populiariausia duomenų gavybai naudojama operacinė sistema 2008 m. buvo Windows XP (net 66.5%) ir Linux (34.1%). Jos smarkiai aplenkė trečioje vietoje likusią Mac sistemą. Peržvelgus bei palyginus keleto metų duomenis pateikiamus KDnuggets galima pastebėti tas pačias tendencijas ir tas pačias populiariausias operacines sistemas.



Šaltinis: sudaryta autoriaus remiantis KDnuggets, 2008.

10 pav. Operacinės sistemos dažniausiai naudojamos duomenų gavybai/analizei

KDnuggets tinklapyje taip pat pateikiama informacija apie 2009 metų pirmojo pusmečio dažniausiai naudojamą programinę įrangą duomenų gavybai (KDnuggets, 2009b). Apklausoje dalyvavo 364 analitikai įvardiję dažniausiai jų naudojamus duomenų gavybos įrankius. Apklausos rezultatai pateikiami 11 paveikslėlyje.



Šaltinis: sudaryta autoriaus remiantis KDnuggets, 2008.

11 pav. Duomenų gavybos įrankių populiarumo grafikas

Kaip matome populiariausi yra SPSS PASW Modeler įrankiai. Beveik vienodai balsų surinko RapidMiner bei SAS sistemos. Kiek mažiau jų teko MS skaičiuoklei Excel bei SAS Enterprise Miner įrankiams.

1.7. Intelektinių sistemų analizės skyriaus išvados

- Remiantis atlikta verslo intelektinių sistemų analize, dėl savo funkcionalumo ir įvairiapusiško taikymo, mažų įmonių verslo problemoms spręsti pasirinkta duomenų gavybą.
- Atlikus duomenų gavybos algoritmų analizę pastebėta, jog tą pačią verslo problemą galima spręsti naudojant skirtingus algoritmus ir gauti skirtingus rezultatus interpretavimui.
- Apžvelgus CRISP-DM standartinį duomenų gavybos modelį nustatyta, jog mažoms įmonėms galima naudoti paprastesnį jo variantą, neįtraukiant visų, jame aprašomų, duomenų gavybos etapų.
- Apžvelgus duomenų gavybos įrankius nustatyta, jog geriausiai duomenų gavybai mažose įmonėse tinka įrankiai veikiantys kaip MS Excel skaičiuoklės papildomos funkcijos ir apdorojantys MS Excel formatu pateiktus duomenis, kadangi daug įmonių naudojasi Windows XP operacine sistema bei MS Excel skaičiuokle.

2. DUOMENŲ GAVYBA IR VEIKLOS MODELIAVIMAS

Siūlomo sprendimo metodikos skyriuje pateikiami veiklos modeliai vaizduojantys duomenų gavybos procesą bei duomenų gavybos proceso modelis, pritaikytas mažoms įmonėms.

Mažos įmonės vengia naudoti duomenų gavybos įrankius dėl daugelio priežasčių, tačiau galima išskirti penkias pagrindines:

- Informacijos stoka apie duomenų gavybos įrankius;
- Manymas, jog duomenų gavybos įrankiai brangūs ir reikalauja didelių eksploatacinių išlaidų;
- Teikiamos naudos nežinojimas;
- Specialistų reikalingumas, kurių mažos įmonės nepajėgia pasisamdyti;
- Manymas, jog duomenų gavybos įrankių pateikiami rezultatai sudėtingi ir sunkiai pasiekiami.

Duomenų gavyba mažose įmonėse mažai nagrinėjama ir dauguma siūlomų produktų yra orientuoti į dideles, ar bent jau vidutinio dydžio įmones. Jau prieš kelis metus Melissa Solomon tyrinėjo mažų įmonių norą naudotis aukštosiomis technologijomis. Ji išskyrė dešimtuką IT sričių į kurias mažos įmonės planavo investuoti pinigus. Nors pirmąsias vietas užėmė saugumo užtikrinimo produktai bei bendravimo programos, į sąrašą įtraukta ir duomenų gavyba. Šiuo atveju ji pateikta kaip įrankiai skirti klientų veiksmų analizei. Surinkus informaciją apie tai, ką jie veikia bei kuo domisi, ją galima naudoti analizuojant įmonės darbą bei atsižvelgiant į paklausą planuojant naujus pasiūlymus (Solomon M., 2005).

Tačiau kokia duomenų gavybą yra svarbi mažoms įmonėms puikiai atskleidžiama Barny Ritholz straipsnyje „Intro to Data Mining for Small Businesses“. Pagrindinė straipsnio idėja yra ta, jog duomenų gavyba gali padėti įmonėms anksčiau sureaguoti į rinkos pokyčius ir lengviau išgyventi net per finansines krizes. Straipsnyje teigiama, kad jei mažos įmonės pačios naudotų duomenų gavybos įrankius, jos galėtų pačios pastebėti rinkos pokyčius, jų tendencijas ir įvairias anomalijas. Taip būtų greičiau sureaguota į pokyčius nei tuo atveju, kai apie rinkos krizes praneša ekonomikos biurai surinkę ir apdoroję tūkstančių panašių įmonių informaciją. Šiuo atveju tas laiko skirtumas gali būti labai reikšmingas ir padėti įmonei išvengti nuostolingų investicijų (Ritholz., 2009)

Duomenų gavybą mažose įmonėse panaudoti galima labai įvairiai. Daug kas priklauso nuo įmonės veiklos bei pasirinktų duomenų gavybos įrankių. Veiklos modeliai, atvaizduojantys duomenų gavybos procesą įmonėje sudaryti tarptautiniais pervežimais užsiimančiai įmonei.

2.1. Įmonės veiklos ir problemos aprašymas

Duomenų gavyba skirta anksčiau nežinomos ir potencialiai naudingos informacijos ištraukimui iš sukauptų duomenų. Ji naudinga ne tik didelėms, bet ir mažoms įmonėms. Šiame darbe veiklos modelis pritaikomas nedidelei tarptautiniais pervežimais užsiimančiai įmonei. Tai nedidelė Kauno mieste įsikūrusi įmonė sėkmingai gyvuojanti jau daugiau nei 10 metų. Pagrindinė jos veikla - ieškoti vežėjų įvairių rūšių produkcijai gabenti iš vienos valstybės į kitą. Įmonėje dirba nedaug darbuotojų, tad ji stengiasi apsieiti be brangios programinės įrangos. Naudojama Windows XP operacinė sistema, 2007 MS Office paketas, buhalterinė programa. Vykdydama veiklą įmonė MS Excel pagalba kaupia duomenis apie vykdomus užsakymus registruodama užsakovo, vežėjo duomenis bei informaciją apie krovinius.

Pagrindinė sprendžiama problema yra veiklos optimizavimas. Siekiama atsisakyti užsakymų, kurie įmonei nuostolingi ar teikia mažai naudos. Tokiu būdu bandoma stiprinant santykius su geriausias pervežimo sąlygas siūlančiomis įmonėmis bei tapti nuolatiniu jų klientu. Tai leistu sudaryti palankesnius sandorius bei gauti įvairių nuolaidų. Taip būtų siekimą pagrindinio įmonės tikslo – tapti Lietuvos kompanija, siūlančia pigiausius krovinių pervežimus Europos ribose. Tapus nuolatiniu įmonės, pervežančios krovinius klientu, galima dar labiau sumažinti pervežimų kainas.

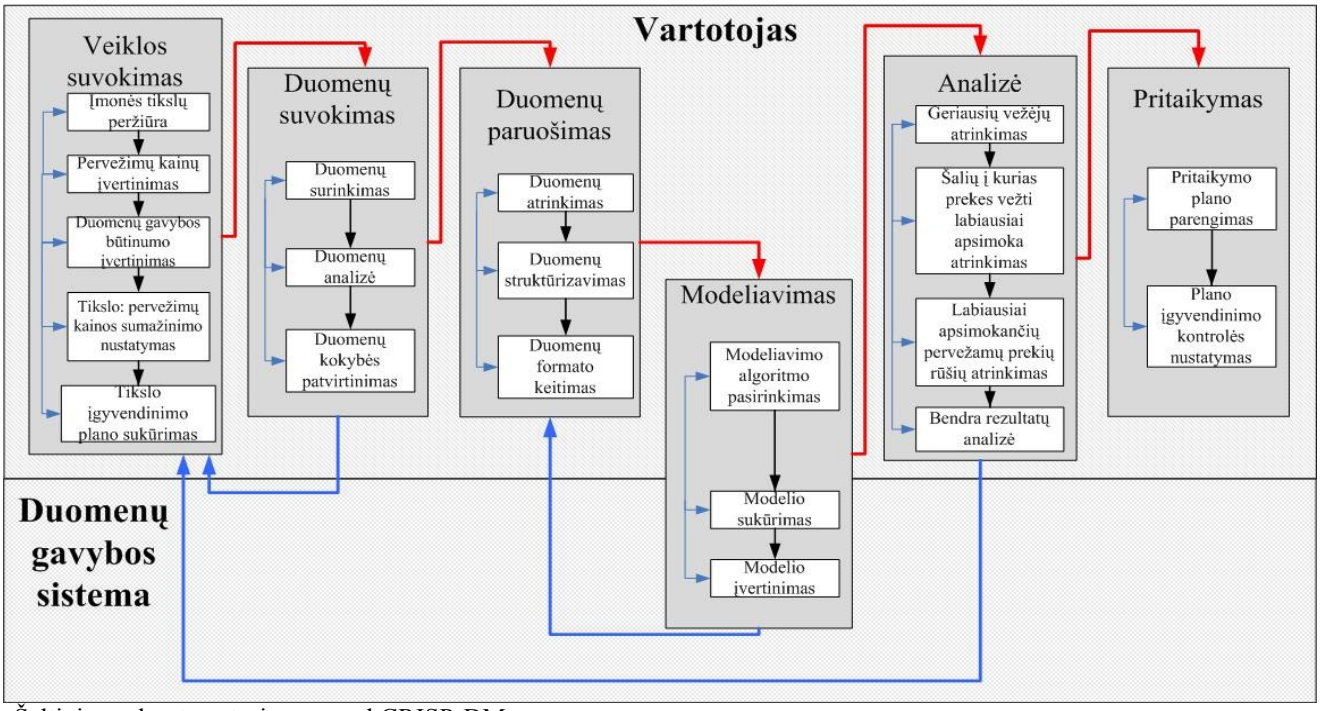
2.2. Problemos sprendimo modeliai

Veiklos modelio paskirtis – aprašyti veiklos sritį, jos svarbiausius procesus, duomenų transformacijas sistemose. Analizuojant veiklos modelį tikslinamos darbų sekos, jų turinys, sąryšiai. Nors veiklos modelių yra įvairių, tačiau jie visi padeda suprasti organizacijos ar modeliuojamos srities struktūrą bei elgseną. Naudojant tokius modelius lengviau suprasti ar galutiniai vartotojai ir kūrėjai vienodai suprantą tam tikrą sritį, jos elgseną, struktūrą. Duomenų gavybos standartinis modelis atvaizduotas trim skirtingais būdais siekiant išanalizuoti darbų sekas, jų turinį bei sąryšius. Tolesniuose skyriuose pateikti veiklos modeliai skirti spręsti konkrečiai problemai – užsakymų atrinkimui. Jie visi braižyti naudojant standartinį duomenų gavybos procesų modelį.

2.2.1. Darbų sekų modelis (Workflow)

Darbų sekų modelis (workflow) atvaizduoja biznio eigos komponentes - procesus ir darbų perdavimo eigą iš vieno proceso kitiems. Be to, darbų sekų modelis parodo, kuris organizacijos

padalinys atlieka ar atsako už konkretų procesą. Tai leidžia analizuoti darbų rezultatų perdavimą tarp organizacijos padalinių (darbuotojų), analizuoti organizacijos veiklos eiliškumą, ieškoti neefektyvumo priežasčių. Sudarius darbų sekų modelį, kiekvienas procesas ir procesų seka gali būti tiriami, ieškant trūkumų, įvertinant laiko ir kainos parametrus. Esant reikalui, kiekvienas darbų sekų modelio atvaizduotas procesas gali būti detalizuojamas, sudarant atskirą jo darbų sekų modelį (GUDAS S., 2002, p.12).



Šaltinis: sudaryta autoriaus pagal CRISP-DM

12 pav. Duomenų gavybos modelis, taikomas konkrečiai problemai spręsti (Workflow)

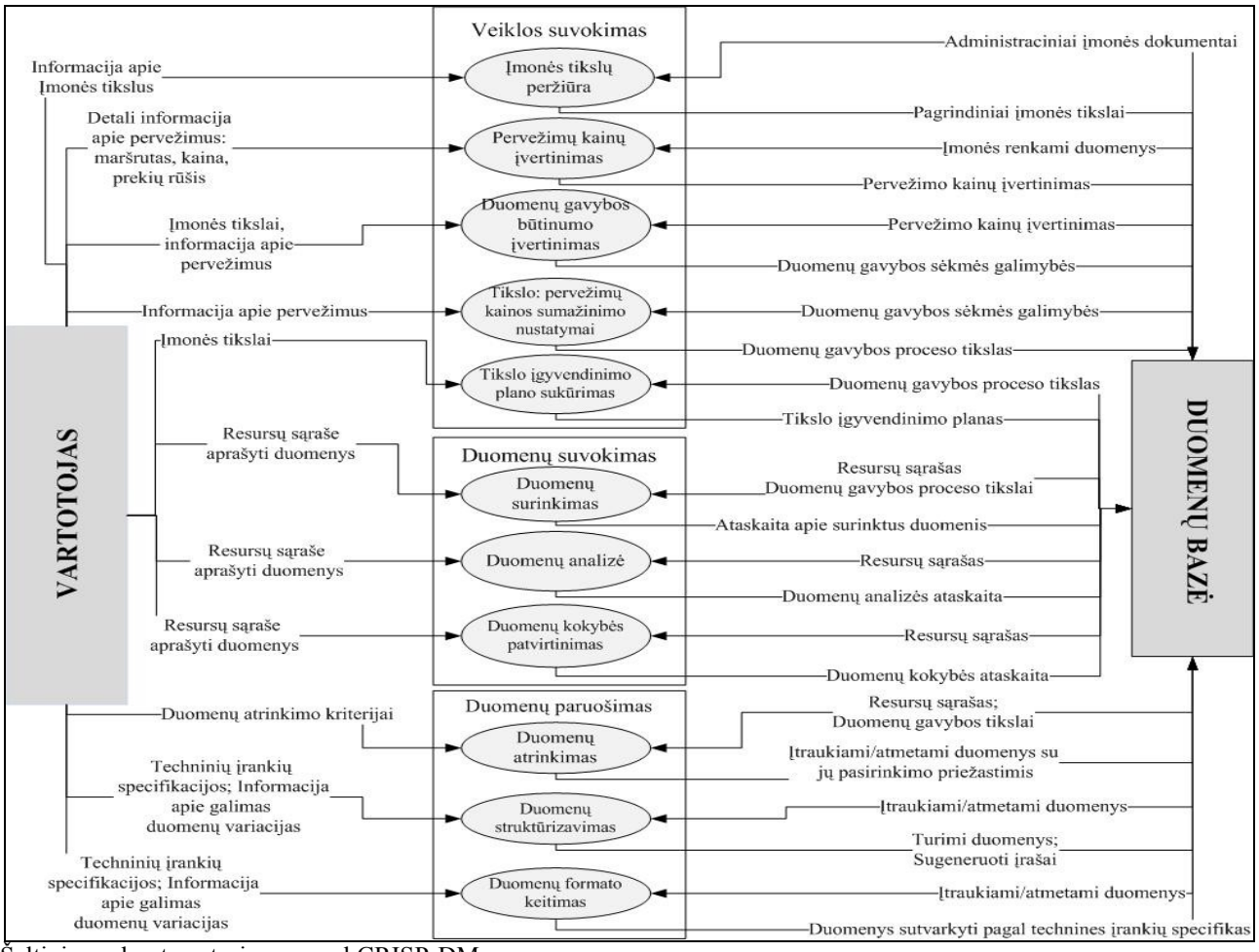
Darbų sekų modelyje atvaizduotos veiklos eigos komponentės – procesai ir darbų perdavimo eiga iš vieno proceso kitiems. Modelyje aiškiai matomas duomenų gavybos procesų eiliškumas. Kadangi kiekvieną iš pagrindinių procesų galima suskaidyti į kelis smulkesnius procesus, jie atvaizduoti pagrindiniuose procesuose. Kaip matome iš grafiko, visi pagrindiniai procesai turi būti atliekami eilės tvarka, tačiau bet kada galima grįžti prie ankstesnio proceso (modelyje pavaizduoti pagrindiniai ir dažniausiai naudojami grįžimo atvejai). Tą patį galima daryti ir žvelgiant į smulkiuosius procesus – visuomet galima grįžti prie ankstesnio proceso.

Kaip matome modelis sudarytas remiantis standartiniu duomenų gavybos modeliu, palikti visi pagrindiniai jo procesai, tačiau dauguma mažų procesų yra pritaikyti modeliuojamai situacijai.

2.2.2. DSD

Duomenų srautų diagramos (Data Flow Diagrams) skirtos veiklos sričiai apibrėžti, t.y. sistemos funkcijoms (procesams) ir jų sąveikoms vaizduoti. Grafinis duomenų vaizdavimas leidžia vartotojams gauti aiškiai suprantamą bendrą sistemos vaizdą, matyti kaip dera atskiros sistemos dalys. DSD diagramos rodo apdorojimo žingsnius kaip duomenų srautus. Joje srautai ir procesai turi turėti savo pavadinimus (Sekliuckis V., 2006, p. 47). Kiekvienas procesas turi turėti informacinius ar materialius įėjimo ir išėjimo srautus diagrama. Duomenų srautų diagramos dalis yra būdinga daugeliui analizės metodų, kadangi jos naudojimas nesudėtingas, žymėjimas intuityvus.

Kadangi duomenų gavybos modelio DSD diagrama labai didelė, ji pateikta devintame ir dešimtame paveiksluose (13 pav. ir 14 pav.).

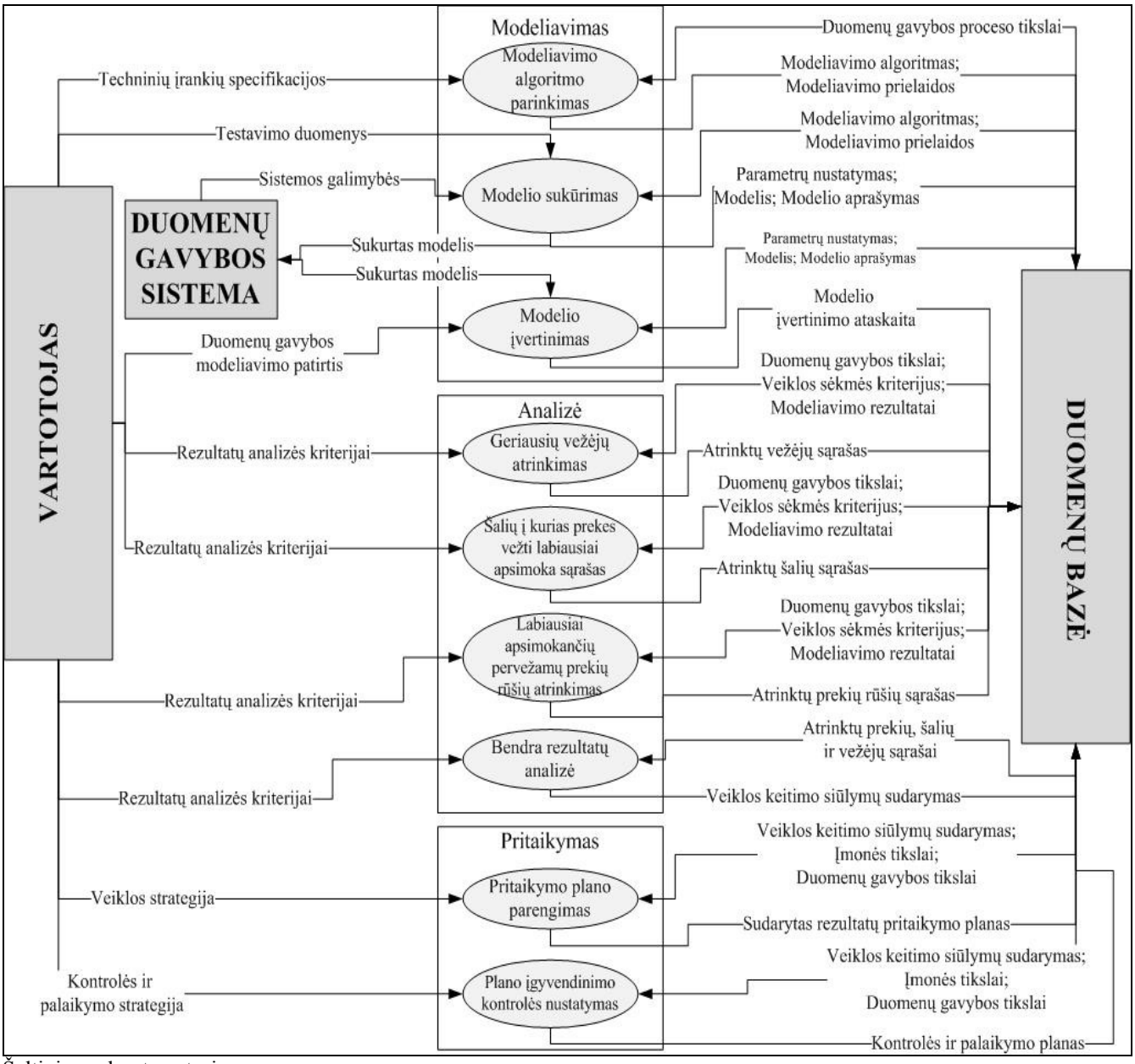


Šaltinis: sudaryta autoriaus pagal CRISP-DM

13 pav. Duomenų gavybos modelis taikomas konkrečiai problemai spręsti (DSD) I dalis

Kaip matome vartotojas ir duomenų bazė informacijai saugoti reikalinga visuose duomenų gavybos etapuose.

14 paveiksle, be funkcijų, vartotojo bei duomenų bazės taip pat pavaizduota duomenų gavybos sistema, reikalinga duomenų gavybos modeliui kurti. Pateiktame DSD modelyje pavaizduoti duomenų gavybos procesai suskirstyti į 6 blokus. Nurodomi duomenų srautai - procesų įeiga ir koks yra kiekvieno proceso rezultatas, kuris talpinamas duomenų bazėje.



Šaltinis: sudaryta autoriaus

14 pav. Duomenų gavybos modelis taikomas konkrečiai problemai spręsti (DSD) II dalis

Duomenų srautų diagramoje aiškiai matosi visi duomenų srautai į veiklos procesus ir iš jų. Tai leidžia aiškiai įsivaizduoti veiklos dalyviams, kaip ir ką reikia įgyvendinti.

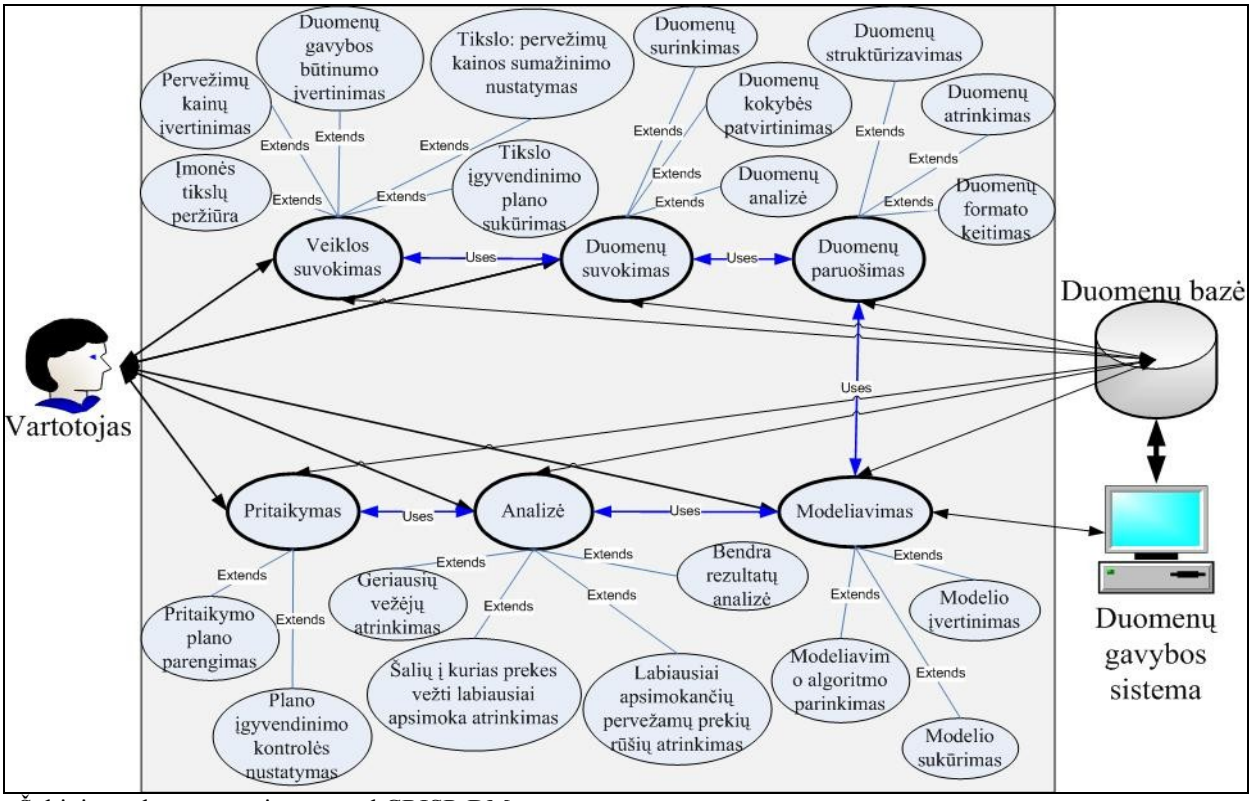
2.2.3. Use Case

Lietuvių kalboje šis terminas formuluojamas skirtingai: „ panaudojimo atvejų modelis“, „veiklos uždavinių modelis“, „vartotojo reikalavimų modelis“. Use Case modelio sąveikos apima informacijos srautų ir (arba) materialinių srautų perdavimą veiklos procesui arba gavimą iš veiklos proceso.

Use Case modelis gali būti taikomas dviem tarpusavyje susijusiems tikslams (Sekliuckis V., 2006, p. 56):

- analizuojamos veiklos srities modeliui aprašyti – nurodyti svarbiausias veiklos dalyvių sąveikas su veiklos uždaviniais;
- kompiuterizuojamos veiklos srities informaciniams reikalavimams specifikuoti.

Tarp funkcijų šiame modelyje naudojamos dviejų tipų sąsajos – *uses* (kai vienas veiklos procesas naudoja kito veiklos proceso rezultatus) ir *extends* (kai vienas veiklos procesas yra kito proceso sudėtyje) (Sekliuckis V., 2006, p. 55).



Šaltinis: sudaryta autoriaus pagal CRISP-DM

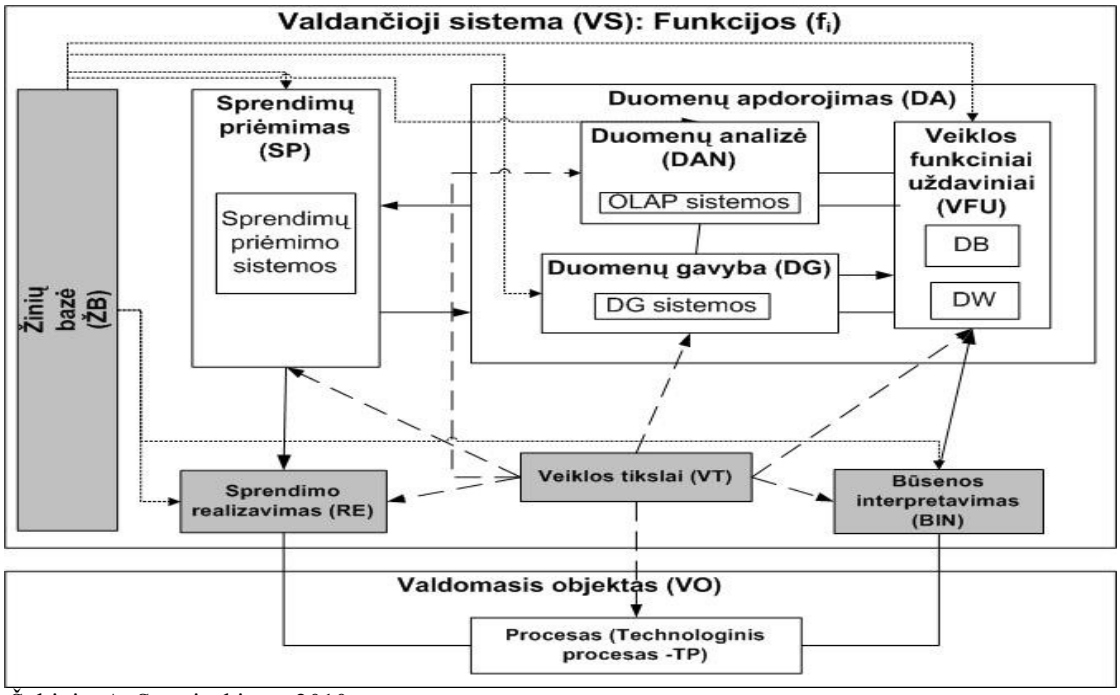
15 pav. Duomenų gavybos modelis taikomas konkrečiai problemai spręsti (Use Case)

Svarbiausios standartinio duomenų gavybos modelio veiklos dalyvių sąveikos su veiklos uždaviniais nurodyti Use Case diagramoje 15 paveiksle.

Kaip matome iš modelio duomenų gavybos modelyje yra 2 pagrindiniai veiklos dalyviai: vartotojas ir duomenų gavybos sistema. Šiame modelyje neatsiejama veiklos dalis yra duomenų bazė, kurioje saugomi duomenų analizei reikalingi duomenys, įvairios ataskaitos. Duomenų gavybos modelyje išskiriami 6 pagrindiniai etapai: veiklos suvokimas, duomenų suvokimas, duomenų paruošimas, modeliavimas, analizė, pritaikymas. Visi jų sudaryti iš smulkesnių procesų.

2.3. Duomenų gavybos proceso veiklos modelis

Duomenų gavyba versle nėra naujas reiškinys, tačiau dažnai duomenų gavybos įrankius taiko didelės įmonės. Didelių įmonių vadovai labiau vertina duomenų analizės svarbą, bei gali skirti tam daugiau lėšų. Mažos įmonės dažnai susiduria su problema, nes dauguma duomenų gavybos įrankių kainuoja daug, lyginant su įmonės gaunamomis pajamomis. Mažos įmonės, prireikus analizuoti duomenis, naudoja elementarias skaičiuokles kaip MS Excel. Tačiau gautas rezultatas dažnai neduoda tiek naudos, kiek duotų duomenų gavybos įrankių panaudojimas.



Šaltinis: A. Stravinskienė, 2010.

16 pav. Duomenų gavybos proceso vieta įmonės valdyme

Verslo intelektinių sistemų, paremtų veiklos modeliu, principinė schema pateikta 16 pav. Joje esanti žinių bazė formuojama metamodelio pagrindu. Ten gali būti kaupiamos taisyklės, šablonai, dokumentų specifikacijos ar kitos struktūros. Ši bazė įtakoja duomenų apdorojimo, sprendimo priėmimo, sprendimo realizavimo bei būsenos interpretavimo etapus. Tokią schemą galima taikyti bet

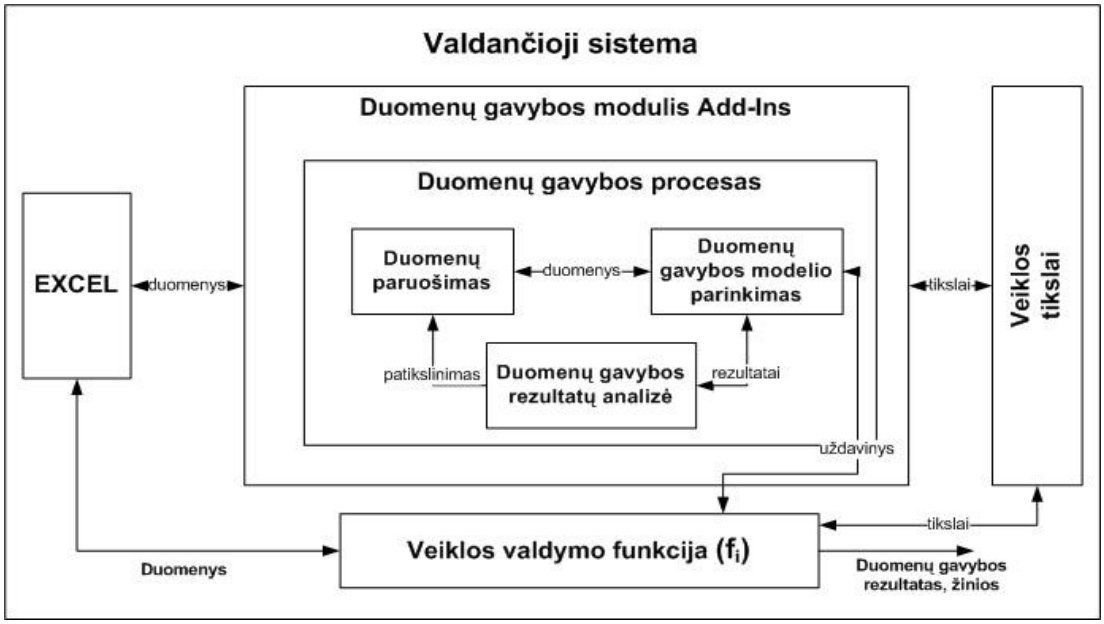
kokio dydžio įmonėms, tačiau mažoms įmonėms ji yra kiek per sudėtinga. Mažose įmonėse duomenų gavybos procesas paprastesnis, nes ten dažnai sprendžiamos standartinės problemos, naudojami paprastesni įrankiai. Dauguma mažų įmonių savo veikloje nenaudoja nei žinių nei duomenų bazės. Kaupiami duomenys dažniausiai nėra didelės apimties ir įvairovės, jie kaupiami MS Excel failuose. Esminiai duomenų gavybos proceso skirtumai didelėje ir mažoje įmonėse pateikti 3 lentelėje.

3 lentelė

Duomenų gavybos proceso didelėje ir mažoje įmonėje skirtumai

DG procesas didelėje įmonėje	DG procesas mažoje įmonėje
Naudoja duomenis iš įvairių duomenų šaltinių, kelių duomenų bazių.	Naudoja duomenis iš vieno MS Excel failo.
Duomenų gavybą atlieka analitikai.	Duomenų gavybą atlieka įmonės vadovas ar kitas darbuotojas neturintis spec. žinių.
Modeliavimui gali naudoti bet kokį duomenų gavybos metodą algoritmą.	Modeliavimui gali naudoti ribotą skaičių metodų ir algoritmų.
Duomenų gavybos sistemą galima praplėsti.	Duomenų gavybos modulis yra riboto didžio.
Galima koreguoti užklausas, algoritmus.	Algoritmų taikymo galimybes riboja duomenų gavybos modulio savybės.
Reikalingos specifinės žinios.	Gali naudotis ir pradedantys vartotojai.
Apdoroja didelius duomenų kiekius.	Ribotas failo dydis
Duomenų gavybą galima taikyti įvairios specifikos ir dydžio įmonėms.	Duomenų gavybos moduliai gali tikti ne visų specifikų įmonėms.

Duomenų gavybos proceso modelis, į kurį integruotas CRISP-DM standartinis duomenų gavybos procesų modelis pritaikytas mažai įmonei pateiktas 17 paveiksle.



Šaltinis: sudaryta autoriaus

17 pav. Duomenų gavybos proceso modelis mažai įmonei

Duomenų gavybos proceso modelis mažai įmonei nuo standartinio skiriasi elementų skaičiumi. Jis daug paprastesnis. Čia patį duomenų gavybos procesą sudaro tik trys etapai – duomenų paruošimas,

duomenų gavybos modelio parinkimas, duomenų gavybos rezultatų analizė. Jame nėra duomenų suvokimo bei pritaikymo funkcijų, o modeliavimo funkcija pakeičiama modelio parinkimu. Esminis šių modelių skirtumas yra tas, jog duomenys mažos įmonės duomenų gavybos modelyje yra imami ne iš duomenų bazės, bet iš MS Excel skaičiuoklės failo, kuriame ir kaupiami.

Modelių elementai paaiškinti 4 lentelėje. Dalis elementų modeliuose sutampa, tačiau daugumos elementų duomenų gavybos proceso modelyje nėra.

4 lentelė

Duomenų gavybos proceso modelių elementų apibrėžimas

Modelio elementas	Duomenų gavybos proceso modelis įmonei	Duomenų gavybos proceso modelis mažai įmonei
1. Valdančioji sistema	Įmonės valdymo sistema, kurios padaliniai naudoja BI priemones.	
2. Žinių bazė	Duomenų bazė, kurioje saugoma kaupiamos veiklos srities žinios. Žinių bazė formuojama metamodelio pagrindu.	------
3. Duomenų gavybos modulis	------	Duomenų gavybos įrankis veikiantis kaip MS Excel papildoma funkcija (modulis).
4. Būsenos interpretavimas	Veiklos srities problemos identifikavimas. Nustatoma kokie duomenys reikalingi problemos sprendimui	------
5. Veiklos tikslai	Veiklos tikslas kurio siekiama panaudojant duomenų gavybą. Pagrindinis sistemos funkcionavimo veiksnys. Jiems kintant, keičiasi proceso eiga, parametrai, užduotys. Kiekvienas duomenų gavybos etapas turi atsižvelgti į keliamą tikslą.	
6. Valdomas objektas	Probleminė veiklos sritis, kurioje įžvelgiama problema.	------
7. Veiklos valdymo funkcija	------	Uždavinys, kurio įgyvendinimui pasitelkiama duomenų gavyba.
8. Duomenų apdorojimas	Etapas kuriame vykdomi duomenų analizės, gavybos uždaviniai.	------
9. Duomenų gavybos procesas	------	Integruotas CRISP-DM modelis
10. Duomenų paruošimas	------	Įmonės duomenų ištyrimas ir parengimas duomenų gavybai.
11. Duomenų gavybos modelio parinkimas	------	DG modelio parinkimas iš tų, kuriuos siūlo duomenų gavybos modulis.
12. Duomenų gavybos rezultatų analizė	------	DG modelio įvertinimas, rezultatų apžvalga.
13. Sprendimo priėmimas	Apibendrinami ir įvertinami duomenų apdorojimo etapo rezultatai, priimamas konkretus sprendimas.	------
14. Sprendimo realizavimas	Intelektinės verslo sistemos pateikiamo rezultato apdorojimas. Juo remiantis priimami sprendimai. Parenkamos priemonės realizavimui.	------
15. MS Excel failas	------	MS Excel skaičiuoklės failas kuriame įmonė kaupia duomenis.

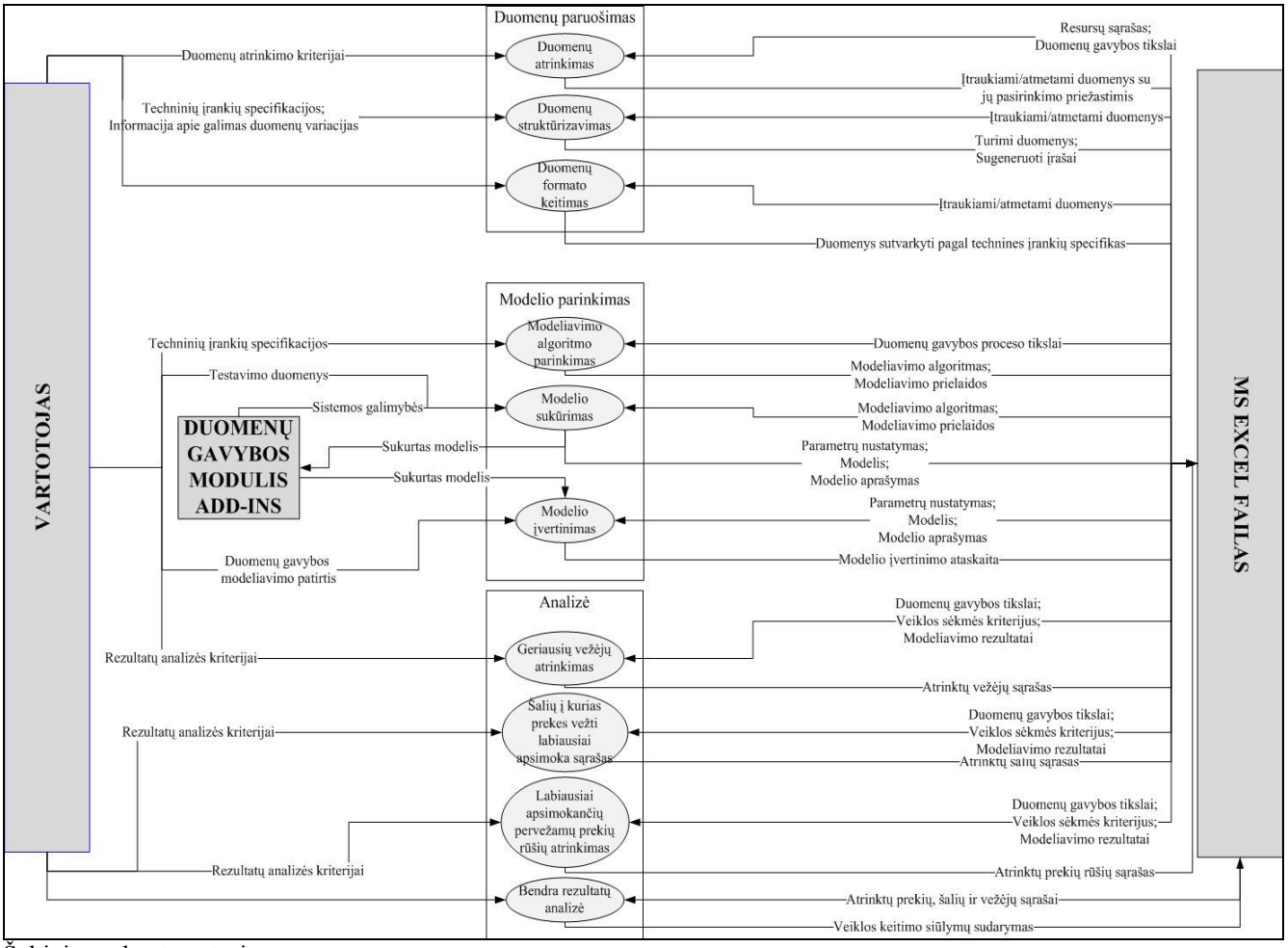
Šaltinis: sudaryta autoriaus

Standartinį duomenų gavybos procesą sudaro šeši nuoseklūs etapai. Mažai įmonės poreikiams tenkinti užtenka trijų iš jų. Pateikti trys etapai yra iteratyvus, tad bet kuriuo laiko momentu lengvai

galima grįžti į ankstesnį etapą. Trūkstant papildomų duomenų iš analizes galima grįžti į duomenų paruošimo etapą.

Visi duomenų gavybos procesai yra įtakojami pagrindinio veiklos tikslo – veiklos optimizavimo. Veiklos valdymo funkcija (f_i) šiuo atveju yra užsakymų atrinkimas stebint kas įtakoja tam tikrą užsakymo kainos intervalą. Jų atrinkimas ir yra pagrindinė duomenų gavybos sistemos užduotis.

Duomenų gavybos procesų modelis mažai įmonei įgyvendinamas remiantis toliau pateiktais veiklos modeliais. Modeliuose vaizduojami duomenų gavybos proceso etapai naudojami mažose įmonėse. Informaciniai ir materialūs duomenų srautai einantys iš/į proceso iki vartotojo, duomenų gavybos modelio ar MS Excel failo pateikti duomenų srautų diagramoje.

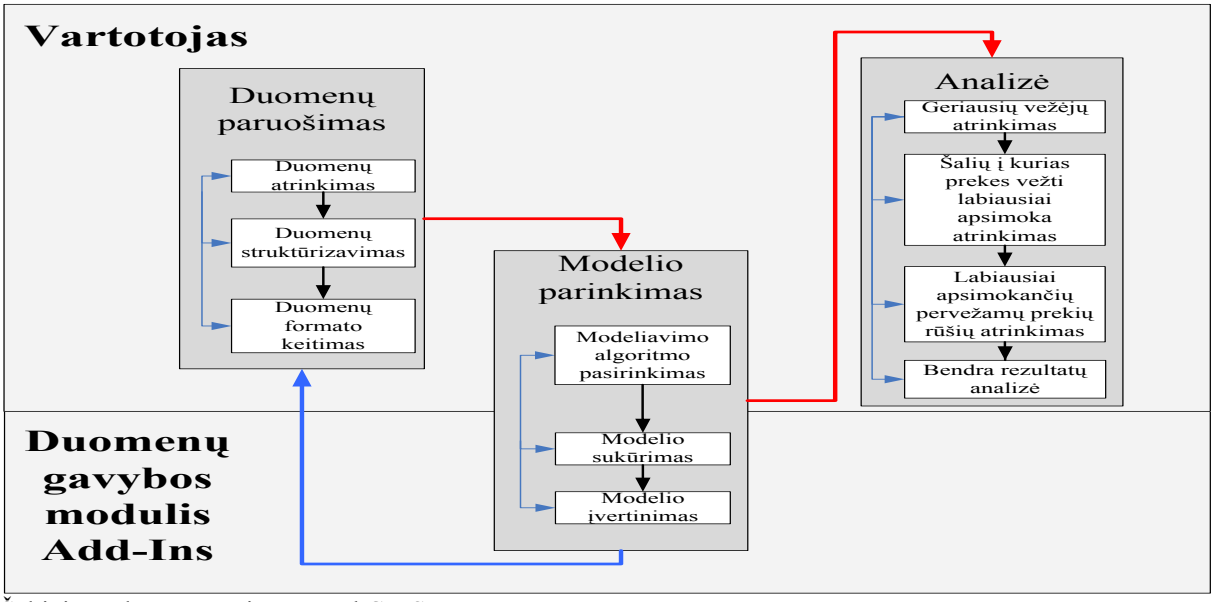


Šaltinis: sudaryta autoriaus

18 pav. DG modelis taikomas konkrečiai mažos įmonės problemai spręsti (DSD)

Kaip matome mažoje įmonėje vietoje duomenų bazė naudojamas MS Excel failas, duomenų gavybos sistemą atstoja paprastesnis duomenų gavybos įrankis – modulis instaliuojamas į MS Excel

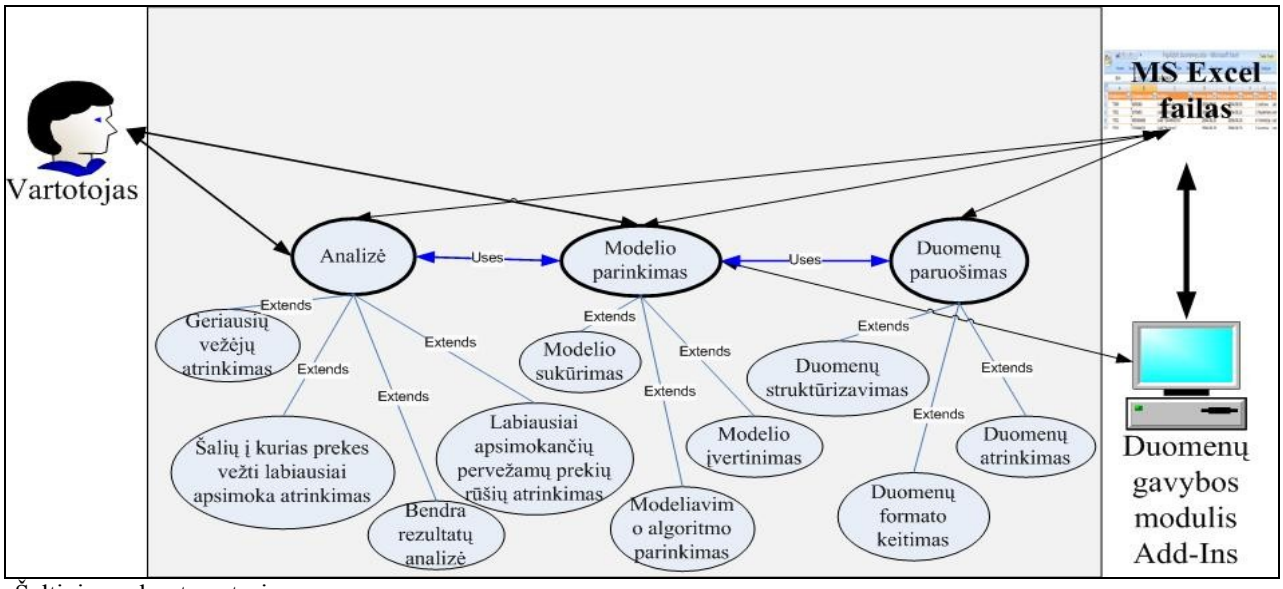
skaičiuoklę. Darbų sekų modelyje vaizduojami trys iteratyvūs duomenų gavybos etapai – duomenų paruošimas, modelio parinkimas ir analizė.



Šaltinis: sudaryta autoriaus pagal CRISP-DM

19 pav. DG modelis, taikomas konkrečiai mažos įmonės problemai spręsti (Workflow)

Svarbiausios veiklos dalyvių sąveikos su veiklos uždaviniais pateiktos Use Case diagramoje.



Šaltinis: sudaryta autoriaus

20 pav. DG modelis taikomas konkrečiai mažos įmonės problemai spręsti (Use Case)

Kaip matome iš modelių, duomenų gavybai mažose įmonėse užtenka paprastesnio duomenų gavybos proceso nei dideliai įmonei.

2.4. Duomenų gavybos įrankių parinkimo įmonei analizė

Nors duomenų gavybos įrankių yra labai daug ir įvairių, analizei parinkti trys. Juos renkantis dėmesys buvo kreipiamas į kelis kriterijus. Pirmiausia buvo žiūrima, jog įrankiai veiktų MS Windows aplinkoje, kadangi ją naudoja didelė dalis įmonių. Taip pat svarbus buvo įrankių suderinamumas su MS Excel skaičiuokle, nes ją naudoja daugelis nedidelių įmonių. Taip pat buvo kreipiamas dėmesys į produkto įsigijimo bei eksploatavimo kaštus, kadangi mažos įmonės negali skirti daug lėšų nei produkto įsigijimui, nei darbuotojų mokymui ar specialistų samdymui.

Taigi įrankių analizei parinkti trys įrankiai: Microsoft SQL Server 2008 Data Mining Add-Ins, XLMiner bei Tree plan. Visi jie veikia MS Windows aplinkoje kaip MS Excel skaičiuoklės elementai. Sekančiuose skyreliuose pateikiama išsami jų analizė.

2.4.1. Microsoft SQL Server 2008 Data Mining Add-Ins

Tai Microsoft kompanijos siūlomas priedas MS Excel skaičiuoklei skirtas duomenų gavybai. Šis priedas turi nemažai įvairių įrankių duomenų tvarkymui ir jų modeliavimui.

MS SQL Server 2008 Data Mining Add-Ins turi dvi imties sudarymo funkcijas:

- Atsitiktinis duomenų padalijimas: duomenis atrenka nurodžius imties dydį procentais ar eilutėmis.
- Padalijimas su perviršiu: imtis sudaroma atrenkant duomenis, o kartu subalansuojant kurio nors stulpelio reikšmes.

Taip pat yra funkcija skirta duomenų padalijimui – dalija duomenis į mokymosi ir testavimo, nurodžius kiek procentų duomenų paimti mokymuisi. Ši funkcija pateikiama ir vedliuose, norint klasifikuoti, prognozuoti duomenis.

Taip pat labai naudingos trūkstamų/netinkamų duomenų redagavimo bei duomenų keitimo funkcijos. Jos ypač praverčia jei įmonės sukaupti duomenys yra netvarkingi, juose pasitaiko nereikšminių duomenų ir pan. Surikiavus nurodyto stulpelio duomenis galima panaudoto duomenų apkarpymo funkcija. Ji leidžia nereikšmines (retai pasitaikančias) reikšmes ištrinti, pakeisti į nulines ar kitas galimas reikšmes. Pasirinkus skaitinių reikšmių ar datų stulpelį iš duomenų galima atrinkti reikiamą intervalą. Laukelių pavadinimų keitimo funkcija parinkus stulpelį jame suskaičiuoja vienodų įrašų skaičių bei pateikia jų pavadinimą, kurį galima keisti. Klasterizavimo (išskirtinumų išryškinimo) funkcija padeda išskirti duomenų eilutes, kurios neatitinka „standarto“. Toje eilutėje taip pat išskiriamas konkretus elementas, kuris nulemia tą skirtumą.

Tačiau renkantis duomenų gavybos įrankį įmonei svarbiausia atkreipti dėmesį į jų siūlomas duomenų modeliavimo galimybes. Microsoft SQL Server 2008 Data Mining Add-Ins turi įrankių duomenų prognozavimui, klasifikavimui bei bendrų bruožų radimui. **Prognozavimas** pritaikomas kiekvienoje mažoje įmonėje. Nagrinėjamas Excel priedas siūlo net kelis metodus jam atlikti.

- Laiko eilutės (prognozavimas): nurodžius prognozuojamą reikšmę ir kiek periodų į priekį prognozuojame gauname suprognozuotas reikšmes ir visų reikšmių grafiką.
- Logistinė regresija (prognozės modeliavimas): nurodžius priklausomojo kintamojo pageidaujamas reikšmes, programa sukuria modelį, kuriame keičiant įtaką darančių kintamųjų reikšmes, galima įvertinti jų įtaką priklausomam kintamajam. Taip pat pateikiamos visų galimų elementų įtaka galutiniam rezultatui. Galima modeliuoti ir visus kiekvieno kintamojo elementus siekiant nurodyto galutinio rezultato.
- Sprendimų medžio algoritmas (apytikris skaičiavimas): nurodžius analizuojamą (skaitinį) kintamąjį sudaromas sprendimų medis. Kiekviename mazge ant nubrėžtos tiesės pateikiamas rombas, kuris apima kintamojo reikšmes būdingas tam mazgui (vietoje histogramos).

Bendrų bruožų radimo funkcija padeda duomenis suskirstyti į grupes pagal panašius požymius. Microsoft SQL Server 2008 Data Mining Add-Ins turi tris variantus šiam metodui.

- Naive Bayes‘o algoritmas (įtaka priklausomam kintamajam): nurodžius priklausomą kintamąjį, jį sugrupuoja ir nurodo kiek kiekvieną grupę įtakoja kiti elementai.
- Asociacijų taisyklės (pirkinių krepšelio analizė): nurodžius kintamųjų stulpelius, pateikiamas rezultatas: dažniausiai kartu pasirenkami elementai, jų pasitaikymo dažnis, bei bendra suma, kurią uždirba įmonė. Taip pat pateikiamos pirkinių krepšelių rekomendacijos ir jų duodama nauda.
- Asociacijų taisyklės: algoritmas ieško elementų kurie eina kartu. Susiję elementai sudaro rinkinius, kurių kiekį ir reikšmingumą bendram duomenų kiekiui nustato algoritmas. Rezultate pateikiami elementų rinkiniai bei taisyklės.

Daugiausiai įvairių galimybių pateikiama **klasifikavimo metodui**. Tai leidžia klasifikavimą atlikti naudojant skirtingus algoritmus, o vėliau lyginti gautus rezultatus bei daryti išsamesnes išvadas.

- Klasterizavimas (kategorijų nustatymas): pasirinkus kintamuosius ir kategorijų skaičių duomenys suskirstomi naudojant klasterizavo metodą. Pateikiami būdingiausi bruožai.
- Sprendimo medžio algoritmas: sudaromas klasifikavimo modelis kuris leidžia prognozuoti nuo kitų elementų priklausančias vieno stulpelio reikšmes.

- Logistinė regresija (užpildymas pagal pavyzdį): nurodžius stulpelį su pavyzdiniais įrašais apie eilutės priklausymą kuriai nors kategorijai, funkcija jas įvertina ir suskirsto likusius duomenis į kategorijas. Taip pat pateikia informaciją apie kategorijai būdingas savybes.
- Logistinė regresija (scenarijų analizė): ši analizė gali būti dvejopa: 1) Tikslo siekimas: pasirinkus eilutę ir nurodžius vieno eilutės elemento reikšmę (tikslią, apytikslę ir pan.) bei elementą, kurį galime keisti, randamas geriausias keičiamo elemento variantas. 2)Kas jeigu: matome kas atsitiktų stebimam (priklausomam) eilutės elementui, jei pakeistume kurį nors kitą.
- Sprendimų medžiai (Klasifikavimas): nurodžius stulpelį kurio duomenis analizuosime remiantis sprendimų medžio algoritmu juos suklasifikuoja ir pateikia interaktyvių sprendimų medžio pavidalu. Sudarytą medį galima išskleisti ar sumažinti, peržiūrėti kiekvieno mazgo analizuojamo kintamojo reikšmes.
- Klasterizavimas: nurodžius stulpelius kuriuos analizuoti sudaroma klasteriu diagrama, kurioje matosi kaip klasteriai susiję tarpusavyje. Taip pat histogramomis pateikiamos klasterių savybės, labiausiai klasterį įtakojantys elementai, taip pat suteikiama galimybė lyginti skirtingų klasterių savybes.

MS SQL Server 2008 Data Mining Add-Ins yra lengvai naudojamas, specialių žinių nereikalaujantis produktas. Jis turi puikius, paprastus ir aiškius vedlius, kurių pagalba galima lengvai perprasti naudojamų metodų ir algoritmų rezultatus. Rezultatai pateikiami grafiškai. Dažniausiai histogramų ar sprendimų medžių pavidalu.

2.4.2. XLMiner

Tai Cytel Inc. produktas skirtas duomenų gavybai ir diegiasi į MS Excel kaip priedas. Jis priešingi nei MS SQL Server 2008 Data Mining Add-Ins skirtas daugiau statistinei analizei. Tačiau duomenų analizei naudoja ir duomenų gavybos algoritmus skirtus duomenų analizei.

XLMiner kaip ir anksčiau nagrinėtas MS Excel priedas turi galimybę apdoroti duomenis. Jis leidžia sudaryti paprastas imtis (kai jos dydis nurodomas tiesiog eilučių skaičiumi), dalinti duomenis su perviršiu (imtis sudaroma atrenkant duomenis, kartu subalansuojant kurio nors stulpelio reikšmes), bei sudaryti imtis naudojant startas (populiacija suskirstoma į panašių elementų grupes (stratas). Duomenys iš kiekvienos stratos parenkami atsitiktiniu būdu). XLMiner taip pat turi ir duomenų padalijimo funkciją, tik duomenis dalija ne į dvi dalis, bet į tris – mokymosi, patvirtinimo ir testavimo.

Jei duomenys nėra visiškai pilni arba jų formatas yra ne toks, kokio norime, duomenis galime paredaguoti. Trūkstami duomenys (ar kitokios reikšmės, kurias nurodėme kaip trūkstamas) pakeičiami

stulpelio moda, mediana, vidurkiu, vartotojo nurodyta reikšme. Jei reikia eilutes su trūkstamai duomenimis galima tiesiog ištrinti.

XLMiner kai kuriems algoritmams reikalauja vienokio ar kitokio duomenų formato. Programa jį leidžia keisti trimis būdais:

- Laikinų duomenų sukūrimas: sukuriama tiek stulpelių su binarinėmis reikšmėmis, kad identifikuotų visas skirtingas žodines reikšmes.
- Kategorinių duomenų kūrimas: visos žodinės reikšmė pakeičiamos atitinkamomis skaitinėmis (1,2,3 ir pan.)
- Tolydžiųjų duomenų rūšiavimas: dažnai naudinga tolydžiuosius duomenis suskirstyti į klases, priskiriant joms tam tikrą reikšmę: klasės vidurkį, klasės medianą, ar rangą.

XLMiner nors ir skirtas daugiau statistinei analizei, kiekvienam metodui leidžia naudoti po keletą duomenų gavybos algoritmų. **Prognozuojant** duomenis bet kuriuo iš keturių algoritmų rezultate pateikiamos realios reikšmės, sumodeliuotos prognozuojamos reikšmės bei jų skirtumai. XLMiner leidžia naudoti šiuos algoritmus:

- Daugybinė tiesinė regresija (Multiple Linear Regression): algoritmas parenka regresijos koeficientus taip, kad skirtumas tarp prognozuojamos ir realios reikšmės būtų mažiausias. Turint didesnį kiekį duomenų rekomenduojama apriboti kuriamų modelių skaičių nurodant maksimų jų skaičių. Baigus modeliavimą tinkamiausią modelį galima pasirinkti iš sąrašo įvertinus pateiktas klaidų tikimybes.
- Artimiausių kaimynų metodas (K –Nearest Neighbors): nurodžius artimiausių kaimynų skaičių k, prognozavimą atlieka su nurodytu skaičiumi ar geriausia reikšme iš intervale 1- k.
- Regresiniai medžiai: nurodžius maksimalų duomenų padalijimų skaičių ir minimalų įrašų skaičių sprendimo rezultate. Rezultate duomenys pateikiami tiek lentelės tiek grafinio regresinio medžio pavidalu.
- Neuroniniai tinklai: nurodžius tinklo architektūrą(paslėptų sluoksnių skaičių), mazgų skaičių kiekviename sluoksnyje pateikiamas rezultatas su nurodytais mazgų svoriais.

Bendrų bruožų radimui naudojamas tik vienas algoritmas – asociacijų taisyklės (association rules). Jis naudojamas atrenkant susijusius duomenis (pvz.: nagrinėjant pirkinių krepšelį – t.y. kas su kuo perkama kartu). Sudaromos taisyklės ir nurodomos jų tikimybės procentais. Taip pat sudaroma taisyklių lentelė.

Klasifikavimui kaip ir prognozavimui naudojami keli algoritmai:

- Diskriminantinė analizė: k klasių sukuriama k diskriminantinių funkcijų. Galima nustatyti kad duomenis pateiktų kanonine forma.
- Logistinė regresija: priklausomas kintamasis gali turėti tik 2 reikšmes (pvz.: 0 ir 1). Nurodžius apytikslį patikimumo lygį ir maksimalų iteracijų skaičių sukuriamas nurodytas skaičius modelių iš kurių atsižvelgus į klaidos tikimybes galima išsirinkti tinkamiausią. Taip pat pateikiama elemento tikimybė priklausyti kiekvienai klasei.
- Klasifikavimo medžiai: nurodžius minimalų įrašų skaičių rezultate bei pasirinkus medžio pateikimo formą pateikiamas grafinis klasifikavimo medis bei jo sudarymo taisyklės.
- Naive Bayes‘o algoritmas: pateikiamos sėkmės tikimybės, jog elementas priklauso kuriai nors klasei. Taip pat pateikiama sąlyginių tikimybių matrica nurodanti kiekvieno kintamojo reikšmės tikimybę priklausyti kuriai nors klasei.
- Neuroniniai tinklai: nurodžius tinklo architektūrą – paslėptų sluoksnių skaičių, mazgų skaičių kiekviename sluoksnyje pateikiamas rezultatas su nurodytais mazgų svoriais bei elemento priklausymo kiekvienai klasei tikimybės.
- Artimiausių kaimynų metodas: nurodžius artimiausių kaimynų skaičių k parenkamas geriausias jų skaičius iš intervalo nuo 1 iki k. Taip pat pateikiama tikimybės jog elementas priklauso vienai ar kitai klasei.

XLMiner siūlo **duomenų sumažinimo** metodo algoritmus, kurių neturi MS SQL Server 2008 Add-Ins:

- Principinių komponentų analizė: naudojamas kai duomenys tarpusavyje labai koreliuoti – sumažina duomenų skaičių ir padaro juos mažiau koreliuotais. Nurodžius fiksuotą komponentų skaičių ar mažiausią komponentų procentą ir parinkus metodą (kovariacijos ar koreliacijos matricas) pateikiamas rezultatas – matrica su atitinkamomis elemento priklausymo principiniam komponentui reikšmėmis.
- K – means klasterizavimas: nurodžius klasterių bei iteracijų skaičių pateikiama: i) klasterio savybės; ii) matrica su atstumais tarp klasterių centrų; iii) įrašų patenkančių į kiekvieną klasterį skaičius; iv) atstumai tarp kiekvieno elemento ir kiekvieno klasterio centro; v) elemento priklausymas kuriam nors klasteriui.
- Hierarchiniai klasteriai (segmentavimas): nurodžius klasterių skaičių, pasirinkus panašumo matavimo būdą (Euklido atstumus, Jaccard koeficientus ar sutapimo koeficientus) bei klasteriamo metodą (iš galimų penkių) pateikiami klasteriai su atstumais, dendogramos, prognozuojami klasteriai, atstumų matrica.

Lyginant su MS SQL Server 2008 Data Mining Add-Ins XLMiner yra sudėtingesnis produktas. Jis labiau tiktų daugiau žinių turintiems ar norintiems įgyti asmenims. Programos pateikiamuose modeliuose galima rasti daug daugiau nustatymų, naudojami vedliai ne tokie intuityvūs, tačiau leidžia daug įvairiau analizuoti informaciją pasirinktu metodu. Rezultatai dažniau pateikiami ne grafiškai bet lentelėse, tačiau yra daug išsamesni nei MS SQL Server 2008 Data Mining Add-Ins. Pateikiamas ne tik galutinis, bet ir tarpiniai rezultatai, kurie leidžia peržiūrėti įvairius koeficientus, taisykles ir kintamuosius sukurtus modeliavimo eigoje.

2.4.3. TreePlan

Tai Mike Middleton sukurtas MS Excel priedas skirtas sprendimų medžiams braižyti. Jis, priešingai nei abu anksčiau nagrinėti produktai, skirtas tik **prognozavimui**. Ir šiam metodui taiko tik vieną, sprendimų medžio, algoritmą. TreePlan leidžia braižyti sprendimo medžius juose žymint sprendimų, įvykių mazgus, atskirų įvykių tikimybes bei uždirbamas ar išleidžiamas pinigų sumas. Šis įrankis puikiai tinka sprendžiant nuoseklias problemas, kai galimos kelios alternatyvos. Tokiomis problemomis gali būti naujo produkto įvedimas į rinką ar naujos technologijos išbandymas. Dažniausiai tokiose situacijose yra kelios veiksmų alternatyvos. Pavyzdžiui, galima naują produktą į rinką įvesti iš karto arba tam tikrais etapais.

TreePlan braižomas sprendimų medis suskaičiuoja kiekvienos alternatyvos tarpines bei galutines vertes. O tai leidžia pasirinkti optimaliausias alternatyvas. Nors šis produktas neturi vedlių kaip Microsoft SQL Server 2008 Data Mining Add-Ins ar XLMiner, tačiau yra gan paprastas ir lengvai naudojamas.

Lyginant su ankstesniais produktais šis nėra labai patogus vartotojui, nes reikia iš anksto žinoti visus duomenis bei juos į modelį suvedinėti ranka. Tai nėra patogu, kai norima analizuoti daug duomenų.

2.4.4. Duomenų gavybos įrankių palyginimas

Nors visi trys įrankiai veikiantys kaip MS Excel priedai yra skirti duomenų gavybai, jų savybės gerokai skiriasi. MS SQL Server 2008 Data Mining Add-Ins skirtas kasdieniam naudojimui didelės patirties neturintiems asmenims. Jis labai patogus vartotojams pripratusiems prie MS Excel skaičiuoklės naudojimo, kadangi jo funkcijos išdėstytos panašiai. Pateikiami labai intuityvūs vedliai padedantys naudotis įvairiomis funkcijomis. XLMiner reikalauja statistinių žinių, todėl labiau tinkamas

patyrusiems vartotojams, kurie galėtų daryti išsamesnes išvadas pagal gautus rezultatus. Jo naudojimas nėra toks paprastas, nes vedliai turi daug įvairių nustatymų, be to pateikiami rezultatai reikalauja papildomų žinių, nes pateikiama daug statistinių rodiklių, koeficientų. TreePlan nėra patogus, nes reikalauja duomenis suvesti ranka. Tačiau jis yra pakankamai paprastas ir lengvai naudojamas. Didžiausias jo trūkumas, jog jis nėra toks universalus ir tinka tik vieno tipo problemoms. Išsamesnis šių įrankių palyginimas iš vartotojo pusės pateiktas 5 lentelėje.

5 lentelė

Duomenų gavybos įrankių palyginimas

	XLMiner	**TreePlan**	**Microsoft SQL Server 2008 Data Mining Add-Ins**
Grafinis vaizdavimas	Duomenis atvaizduoja histogramomis, matricomis, sklaidos diagramomis.	Duomenis leidžia vaizduoti sprendimų medžiais	Duomenis atvaizduoja histogramomis bei grafiniais sprendimų medžiais.
Kitos galimybės			Leidžia testuoti ir patvirtinti sukurtus modelius.
Programinės įrangos reikalavimai	Operacinė sistema: MS Windows NT 4.0/ 2000/ XP ar vėlesnė; Microsoft Excel 2000/2003/2007 (neveikia su Microsoft Excel 97).	Operacinė sistema: MS Windows NT 4.0 / 2000/ XP ar vėlesnė; MS Excel 1997/2000/2003/2007. Taip pat veikia su Macinosh.	Operacinė sistema: Windows Server 2003/2008; MS Windows XP/Vista/7. Reikalinga prieiga prie SQL Server 2008 Analysis Services: Enterprise, Standard.
Naudojimo patogumas	• Tinklapyje pateikiama išsami naudojimo instrukcija su pavyzdžiais. • Funkcijos turi vedlius, tačiau juose daug nustatymų, kas sudėtinga nepatyrusiam vartotojui. • Pateikiami tarpiniai skaičiavimai.	• Pateikiama išsami naudojimo instrukcija su pavyzdžiais. • Duomenis į sprendimų medį reikia suvedinėti ranka. • Netinka apdoroti dideliems duomenų srautams.	• Žinyne pateikiama informacija apie funkcijų panaudojimo galimybes, rezultatų paaiškinimas. • Funkcijos turi labai supaprastintus vedlius. • Pateikiamas tik galutinis rezultatas.

Šaltinis: sudaryta autoriaus.

Be šių savybių, svarbu atsižvelgti ir į įrankių naudojamus algoritmus duomenų gavybai. Įrankių naudojami algoritmai pateikiami 6 lentelėje.

Kaip matome iš įrankių apžvalgos bei palyginimo visi šie įrankiai yra skirtingi, nors skirti tai pačiai veiklos sričiai – duomenų gavybai. Jie leidžia išnaudoti įvairias duomenų gavybos galimybes bei gauti rezultatus teikiančius naudą įmonei. Norint įmonei parinkti geriausią iš jų, reikia atsižvelgti į kelis kriterijus:

- įmonės veiklos specifiką;
- darbuotojų, dirbsiančių su įrankiu, žinias;

- produkto įsigijimo bei išlaikymo kaštus.

Paskutinysis kriterijus ypač svarbus mažoms įmonėms, nes jos neturi galimybės skirti didelės sumos produkto įsigijimui bei palaikymui. Taip pat jos negali samdyti specialistų, kurie dirbtų su duomenų gavybos įrankiais ir analizuotų jų pateikiamus rezultatus. Joms reikia įrankių, kurie būtų patogūs ir paprasti naudotis, nes dažniausiai duomenų gavyba mažose įmonėse rūpinasi vienas ar du žmonės.

6 lentelė

Įrankių naudojamų algoritmų palyginimas

Algoritmas	**MS SQL Server 2008 Data Mining Add-Ins**	**XLMiner**	**TreePlan**
Daugybinė tiesinė regresija		√	
Artimiausių kaimynų metodas		√	
Sprendimų medis	√	√	√
Laiko eilutės	√		
Logistinė regresija	√	√	
Regresiniai medžiai		√	
Neuroniniai tinklai		√	
Diskriminantinė analizė		√	
Naive Bayes	√	√	
Klasterizavimas	√		
Asociacijų taisyklės	√	√	
Principinių komponenčių analizė		√	
K-means klasterizavimas		√	
Hierarchiniai klasteriai		√	

Šaltinis: sudaryta autoriaus.

Nagrinėtų įrankių įsigijimo kaštai yra prieinami mažoms įmonėms. Microsoft SQL Server 2008 Data Mining Add-Ins yra nemokamas, tačiau reikia turėti įsidiegus ne tik Microsoft Excel 2007, bet ir SQL Server 2005 ar 2008 Analysis Services (enterprise/standard). Norint įsidiegti XLMiner užtenka turėti Microsoft Excel 2000/2003/2007 skaičiuoklę, tačiau reikia mokėti už XLMiner (XLMiner 3.0 professional edition licencija 999$) programos licenziją. TreePlan veikia su Microsoft Excel 1997/2000/2003/2007 skaičiuoklės versijom, licencija kainuoja 49$.

2.5. Siūlomo sprendimo metodikos skyriaus išvados

- Naudojant standartinį CRISP-DM duomenų gavybos modelį sudaryti darbų sekų, duomenų srautų, use case modeliai problemai spręsti.
- Remiantis sudarytais duomenų gavybos modeliais problemai spręsti, suformuotas duomenų gavybos proceso modelis, kuris gali būti pritaikomas mažos įmonės veikloje.

- Remiantis atlikta duomenų gavybos įrankių, pritaikomų mažose įmonėse analize, galima daryti išvadą, jog jie tinkami ir pasirinktos verslo problemos sprendimui eksperimentinėje darbo dalyje.

3. DUOMENŲ GAVYBOS PROCESO MODELIO PRITAIKYMAS

Eksperimentiniame skyriuje bandoma anksčiau išnagrinėtus įrankius pritaikyti tarptautiniais pervežimais užsiimančios įmonės veiklos optimizavimui. Parinkti įrankiai veikia kaip MS Excel programos dalys, tad puikiai tinka įmonei iki šiol duomenis apie užsakymus kaupusiai MS Excel pagalba. Pritaikymas remiasi antrajame darbo skyriuje pateiktu duomenų gavybos proceso modeliu mažai įmonei.

Naudojant antrajame skyriuje sudarytus veiklos modelius konkrečiai problemai, buvo atlikta tarptautiniais pervežimais užsiimančios įmonės veiklos analizė. Visas eksperimentas yra paremtas duomenų gavybos proceso modeliu ir atliktas atsižvelgiant į visus jo etapus. Pradžioje identifikuotos įmonės problemos, ištirti įmonės kaupiami duomenys, kurie gali būti tinkami toms problemoms spręsti. Juos sutvarkius, parinkti modeliai duomenų gavybos procesui. Gauti rezultatai įvertinti analizės etape. Modelyje pateiktas realizavimo etapas neįgyvendintas, tačiau suformuoti siūlymai nagrinėtai įmonei.

3.1. Eksperimentui naudojami duomenys

Eksperimentiniame skyriuje naudojami realūs tarptautiniais pervežimais užsiimančios įmonės duomenys kaupiami MS Excel formatu. Taip duomenys įmonėje kaupiami nuo pat jos įkūrimo. Įmonė neturi tikslo pakeisti šį, jai įprastą duomenų kaupimo metodą, kadangi registruojamų duomenų nėra daug, o darbuotojai yra įgudę duomenis kaupti MS Excel skaičiuoklės pagalba. Užsakymų analizei naudojami duomenys pateikiami 7 lentelėje.

Kaip matome iš lentelės, nagrinėjama užsakymo kaina priklauso nuo šešių kintamųjų – vežėjo, krovinio pristatymo trukmės, krovinio rūšies, šalies iš ir į kuria gabenamas krovinys bei krovinio svorio.

Eksperimentui naudojami 2004 m. rugpjūčio mėn. – 2009 m. birželio mėn. duomenys. Duomenys yra išsamūs – pateikiami visi užsakymo kainai įtaką darantys veiksniai. Prieš analizuojant sukauptus duomenis teko išspręsti duomenų netikslumo problemą – užpildyti nulines reikšmes, įrašyti trūkstamus duomenis. Tai darant buvo atsižvelgiama į kitas to paties stulpelio reikšmes ir trūkstamas vietas užpildyti dažniausiai pasitaikančiomis.

Siekiant kuo tiksliau įvertinti duomenis, bei gauti tikslesnius rezultatus, užsakymų kainos pateikiamos be nuolatiniams klientams taikomų nuolaidų. Tai leidžia objektyviau vertinti ir analizuoti realias užsakymų kainas.

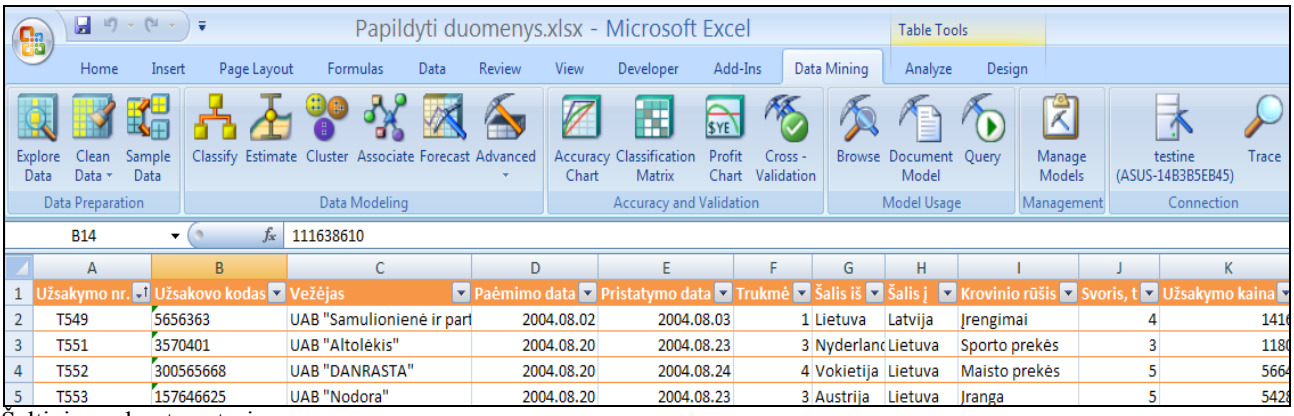
7 lentelė

Eksperimentui naudojami duomenys

Kintamasis	**Kintamojo apibūdinimas**
Užsakymo numeris	Nurodo užsakymo numerį. Analizei jis nėra reikšminis.
Užsakovo kodas	Įmonės užsakiusios pervežimą kodas. Analizei jis nėra reikšminis
Vežėjas	Užsakymą pervežusios įmonės pavadinimas. Užsakymo kainai įtaką darantis kintamasis.
Paėmimo data	Krovinio paėmimo data. Reikalinga užsakymo trukmei skaičiuoti.
Pristatymo data	Krovinio pristatymo data. Reikalinga užsakymo trukmei skaičiuoti.
Trukmė	Krovinio pristatymo trukmė (užsakymo pristatymo ir priėmimo datų skirtumas). Užsakymo kainai įtaką darantis kintamasis.
Šalis iš	Šalis iš kurios gabenamas krovinys. Užsakymo kainai įtaką darantis kintamasis.
Šalis į	Šalis į kurią gabenamas krovinys. Užsakymo kainai įtaką darantis kintamasis.
Krovinio rūšis	Užsakymo kainai įtaką darantis kintamasis.
Svoris	Krovinio svoris. Užsakymo kainai įtaką darantis kintamasis.
Užsakymo kaina	Nuo šešių elementų priklausantis kintamasis.

Šaltinis: sudaryta autoriaus.

Eksperimentui naudojami trys siūlomo sprendimo skyriuje aptarti įrankiai: MS SQL Server 2008 Data Mining Add-Ins, XLMiner, TreePlan. MS Excel programos lange jie matomi kaip atskiri meniu punktai (21 pav.). Nors juos į skaičiuoklę reikia instaliuoti atskirai jų išdėstymas ir naudojimas yra labai panašus į pagrindines MS Excel skaičiuoklės funkcijas.



	A	B	C	D	E	F	G	H	I	J	K
1	Užsakymo nr.	Užsakovo kodas	Vežėjas	Paėmimo data	Pristatymo data	Trukmė	Šalis iš	Šalis į	Krovinio rūšis	Svoris, t	Užsakymo kaina
2	T549	5656363	UAB "Samulionienė ir part	2004.08.02	2004.08.03	1	Lietuva	Latvija	Įrengimai	4	1416
3	T551	3570401	UAB "Altolėkis"	2004.08.20	2004.08.23	3	Nyderlanc	Lietuva	Sporto prekės	3	1180
4	T552	300565668	UAB "DANRASTA"	2004.08.20	2004.08.24	4	Vokietija	Lietuva	Maisto prekės	5	5664
5	T553	157646625	UAB "Nodora"	2004.08.20	2004.08.23	3	Austrija	Lietuva	Įranga	5	5428

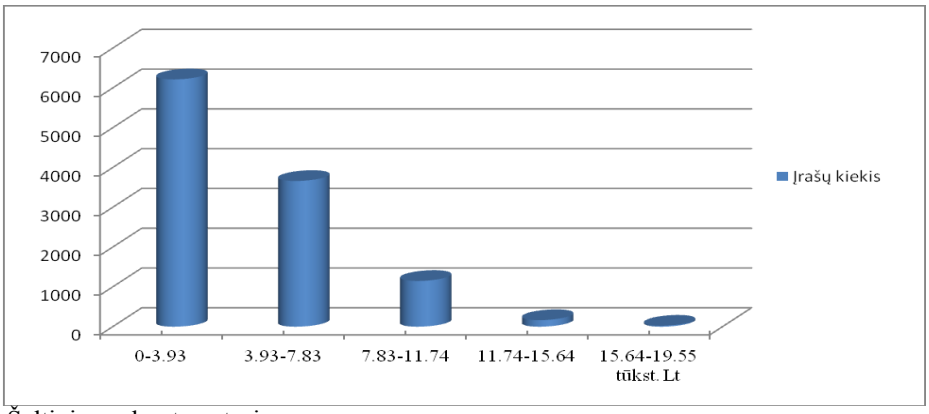
Šaltinis: sudaryta autoriaus.

21 pav. MS Excel skaičiuoklės langas su add-in

Tolesniuose skyreliuose pateiktas šių įrankių pritaikymas konkrečioms situacijos su realiais duomenimis.

3.2. Microsoft SQL Server 2008 Data Mining Add-Ins pritaikymas įmonei

Siekiant optimizuoti įmonės veiklą tikslinga atrinkti užsakymus kurie pasitaiko retai bei kurių kaina yra labai didelė. Tokiu būdu galima daugiau dėmesio skirti tiems užsakymams, kurie įmonei neša pagrindinį pelną. Pradžioje naudojant MS SQL Server 2008 Data Mining Add-Ins funkciją Explore Data užsakymų kainos suskirstytos į penkis intervalus ir suskaičiuotas užsakymų dažnis kiekvienam intervalui.



Šaltinis: sudaryta autoriaus

22 pav. Užsakymo kainos pasiskirstymas

Naudojant duomenų gavybos klasifikavimo funkciją galima net keturiais skirtingais algoritmais nustatyti užsakymo kainą lemiančius kintamuosius. Pirmiausią šie kintamieji nustatyti naudojant sprendimų medžio algoritmą. Pateikiamas sprendimų medis yra interaktyvus ir labai didelis, tad rezultatai pateikti fragmentais.

Total Cases: 7854

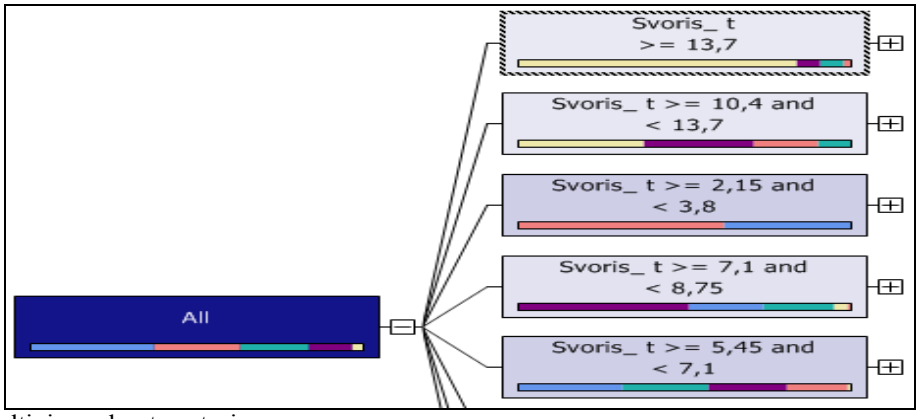
Value	Cases	Probab...	Histogram
< 2419.20290816	2748	34,96%	
>= 9443.8055542784	306	3,93%	
2419.20290816 - 471...	2080	26,47%	
4713.2597075968 - 7...	1663	21,17%	
7231.4886774784 - 9...	1057	13,47%	
Missing	0	0,00%	

Šaltinis: sudaryta autoriaus

23 pav. Sprendimų medžio žymėjimas

Užsakymo kainas į intervalus suskirstė pati programa, tad intervalai šiek tiek skiriasi nuo prieš tai pateiktų. Tačiau tai nedaro didelės įtakos rezultatams, kadangi į aukščiausios kainos intervalą patenka itin reti užsakymai.

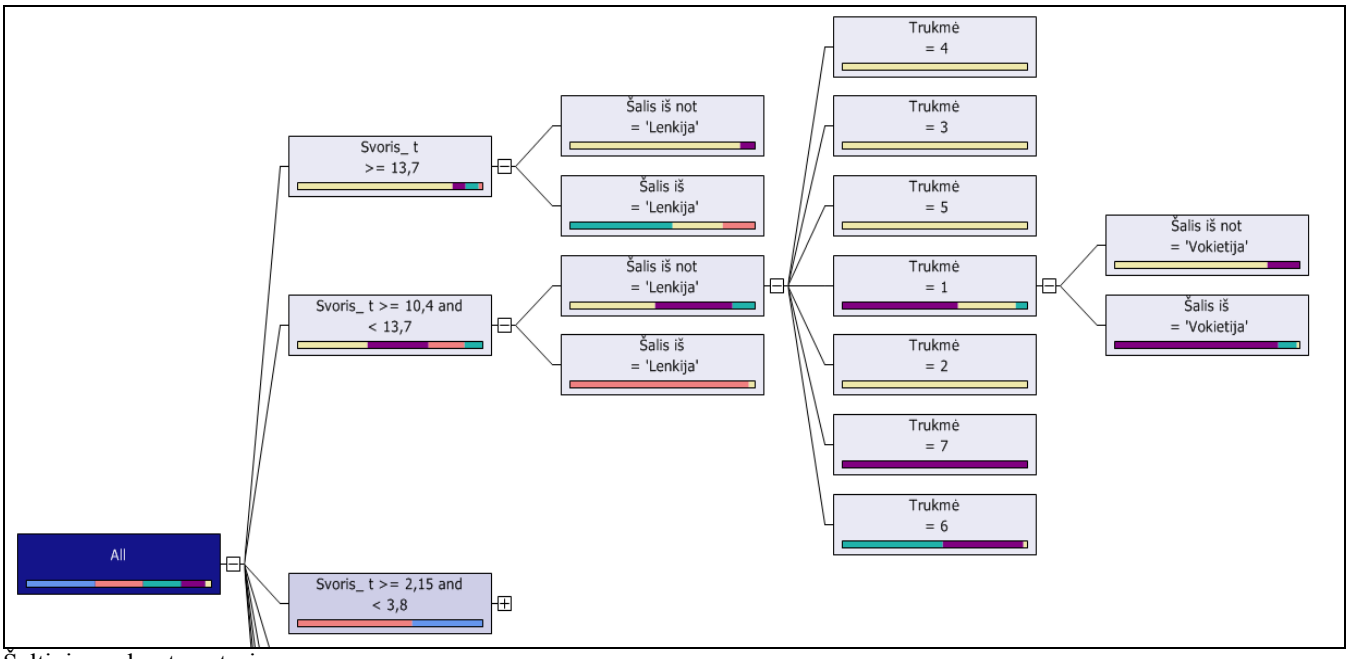
Kaip matome 24 pav. pradžioje duomenys skirstomi pagal jų svorį. Didžiausią kainą lemia krovinio svoris viršijantis 13,7 t., taip pat kai krovinio svoris yra nuo 10,4 iki 13,7 t.



Šaltinis: sudaryta autoriaus

24 pav. Sprendimų medis (skirstymas pagal svorį)

Detaliau skleidžiant gautą sprendimų medį gaunamas tolesniuose paveiksluose pateiktas vaizdas. Išskleidus visą medį matome ne tik aukštą kainą įtakojančiu veiksnius, bet ir kitiems kainos intervalams būdingas savybes. I sprendimo medžio dalyje (25 pav.) matome jog aukštą užsakymo kainą lemia ne tik krovinio svoris, bet ir kiti veiksniai.

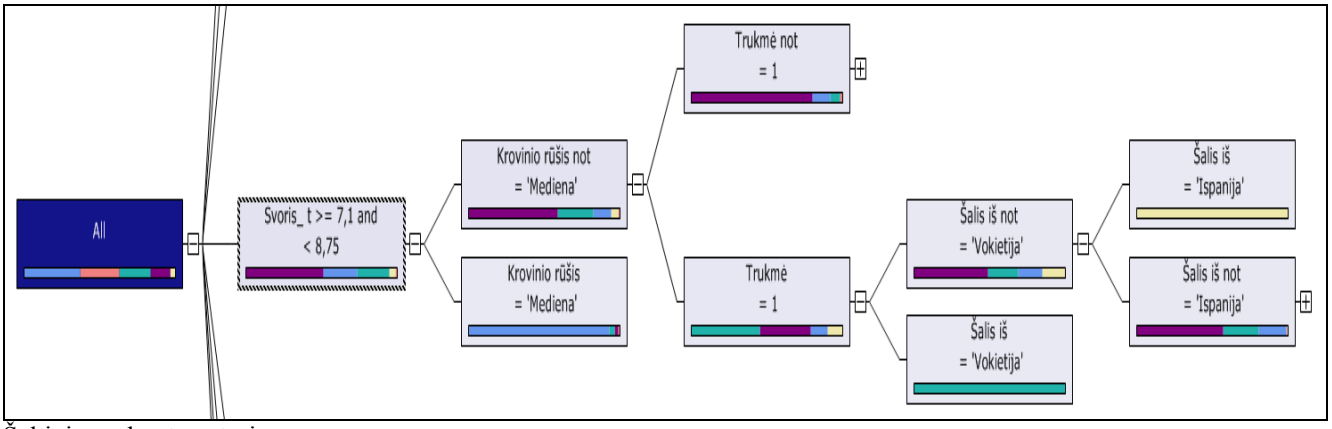


Šaltinis: sudaryta autoriaus

25 pav. Aukštą užsakymo kainą įtakojantys veiksniai (sprendimų medis) I dalis

Didžiausia užsakymo kaina būna tuomet kai užsakymo trukmė yra dvi, trys, keturios ar penkios dienos, o krovinys gabenamas ne iš Lenkijos (25 pav.). Taip pat jei trukmė yra viena diena, tačiau krovinys vežamas ne iš Lenkijos ir ne iš Vokietijos.

Jei krovinio svoris yra nuo 7,1 iki 8,75 t. aukšta kaina būna tuomet, kai per vieną dieną iš Ispanijos pristatomas krovinys, kurio rūšis yra ne mediena (26 pav.).



Šaltinis: sudaryta autoriaus

26 pav. Aukštą užsakymo kainą įtakojantys veiksniai (sprendimų medis) II dalis

Visas sprendimų medis pateikiamas pirmame priede (1 PRIEDAS). Tačiau kitose atšakose pateikiami elementai neturi didelės įtakos aukštai užsakymo kainai.

Classify Užsakymo kaina
Logistic Regression
Užsakymo kaina

Atributas	Vertė	Įtaka, kai kaina >= 9304.102641664
Šalis iš	Graikija	
Šalis iš	Rusija, Maskva	
Šalis iš	Rusija, Kaliningradas	
Šalis iš	Rusija, St.Peterburgas	
Šalis į	Liuksemburgas	
Šalis į	Olandija	
Vežėjas	UAB "Rima Gražienė ir partneriai "	
Krovinio rūšis	Sporto prekės	
Šalis į	Šveicarija	
Krovinio rūšis	Maistas	
Šalis iš	Rumunija	

Šaltinis: sudaryta autoriaus

27 pav. Aukštą užsakymo kainą įtakojantys veiksniai (logistinė regresija)

Tą patį atlikus kitais metodais gaunami kiek kitokie rezultatai. Atlikus klasifikavimą logistinės regresijos būdu (27 pav.) gauname, kad didžiausią įtaką užsakymams, kurių kaina daugiau nei 9304 Lt. daro šalis iš kurios vežamas krovinys bei krovinio rūšis. Labiausiai tikėtina, kad tokia aukšta užsakymo kaina bus jei krovinys gabenamas iš Graikijos, Rusijos (Kaliningrado srities bei Maskvos). Taip pat jei krovinys gabenamas į Liuksemburgą, Olandiją ar Šveicariją arba jei gabenamos sporto prekės.

Atlikus klasifikavimą naudojant Naive Bayes algoritmą gaunamas dar kitoks rezultatas. Kaip matome 28 pav. labiausiai aukštą užsakymo kainą įtakoja šalis į kurią gabenamas krovinys – Lietuva (85%), krovinio rūšis – maisto prekės (67%), šalis iš kurios gabenamas krovinys – Ispanija (45%),

Italija (16%)ar Lietuva (12%). Taip pat krovinio svoris 5,8 – 8,5 t. (12%), 8,5-11.7 t. (42%), ar daugiau nei 11.7 t. (42%).

Classify Užsakymo kaina

Attribute Characteristics

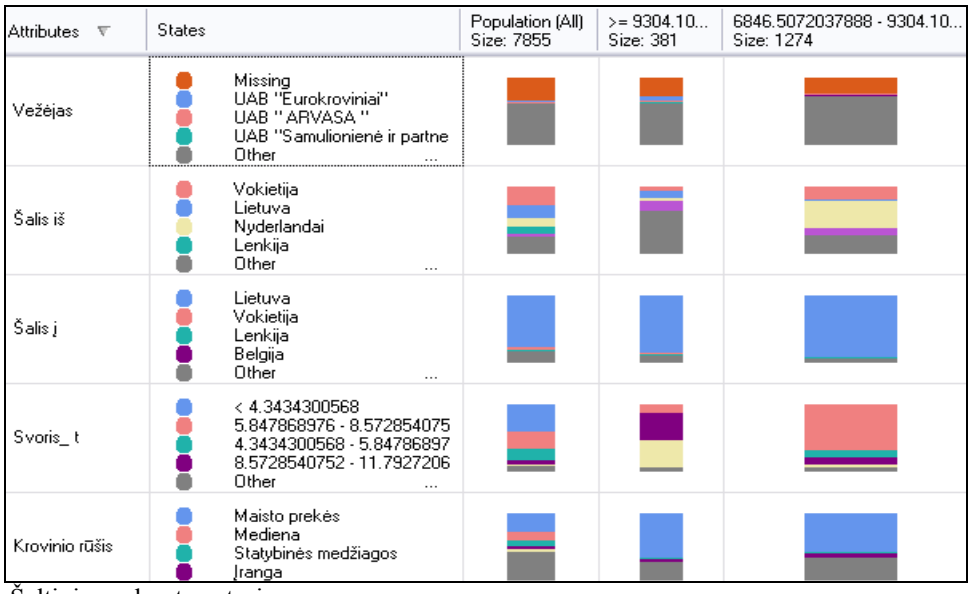
Užsakymo kaina=>= 9304.102641664

Atributas	Vertė	Tikimybė
Šalis į	Lietuva	85 %
Krovinio rūšis	Maisto prekės	67 %
Šalis iš	Ispanija	45 %
Svoris_ t	8.5728540752 - 11.7927206512	42 %
Svoris_ t	>= 11.7927206512	41 %
Vežėjas	Missing	29 %
Šalis iš	Italija	16 %
Svoris_ t	5.847868976 - 8.5728540752	14 %
Šalis iš	Lietuva	12 %
Šalis iš	Vokietija	7 %

Šaltinis: sudaryta autoriaus

28 pav. Aukštą užsakymo kainą įtakojantys veiksniai (Naive Bayes)

Klasifikavimas naudojant Naive Bayes algoritmą leidžia peržiūrėti ir kiekvieno kintamojo įtaką priklausomam kintamajam. Ji atvaizduojama spalvomis.

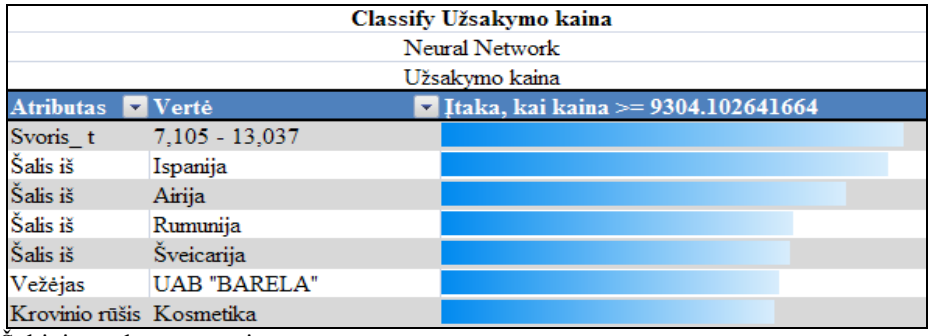


Šaltinis: sudaryta autoriaus

29 pav. Aukštą užsakymo kainą įtakojantys veiksniai (Naive Bayes pagal elementus)

Kaip matome (29 pav.) neišsiskiria nė vienas vežėjas, kuris įtakotų aukštą užsakymo kainą. Tačiau žvelgiant į šalių iš kurių gabenami kroviniai grafiką puikiai išsiskiria Vokietija, Lietuva, Italija. Iš šalių į kurias gabenami kroviniai išsiskiria Lietuva. Svorių stulpelyje labiausiai išsiskiria sunkūs kroviniai. Pagal krovinio rūšis – maisto prekės.

Klasifikuojant duomenis naudojant neuroninį tinklą gauti rezultatai pateikiami 30 pav. Didžiausią įtaką aukštai užsakymo kainai daro krovinio svoris (7105-13037t.), šalis iš kurios gabenamas krovinys (Ispanija, Airija, Rumunija, Šveicarija), krovinio rūšis (kosmetika)



Classify Užsakymo kaina

Neural Network

Užsakymo kaina

Atributas	Vertė	Įtaka, kai kaina >= 9304.102641664
Svoris_ t	7,105 - 13,037	
Šalis iš	Ispanija	
Šalis iš	Airija	
Šalis iš	Rumunija	
Šalis iš	Šveicarija	
Vežėjas	UAB "BARELA"	
Krovinio rūšis	Kosmetika	

Šaltinis: sudaryta autoriaus

30 pav. Aukštą užsakymo kainą įtakojantys veiksniai (neuroninis tinklas)

Kaip matome visi klasifikavimo algoritmai pateikia skirtingus rezultatus. Juos apibendrinti galima lentelėje (8 lentelė).

8 lentelė

Klasifikavimo algoritmų pateikiamų rezultatų palyginimas

Įtaką darantis kintamasis	Sprendimų medžio algoritmas	Logistinės regresijos algoritmas	Naive Bayes algoritmas	Neuroninių tinklų algoritmas
Krovinio svoris (t)				
8.5-10			√	√
≥ 10	√		√	
Užsakymo trukmė (dienos)				
3 - 5	√			
Šalis iš				
Ispanija	√		√	√
Graikija		√		
Airija				√
Italija			√	√
Rusija (Kaliningrado sritis)		√		
Vokietija			√	
Lietuva			√	
Rusija (Maskva)		√		
Šalis į				
Liuksemburgas		√		
Olandija		√		
Lietuva			√	
Krovinio rūšis				
Sporto prekės		√		
Maisto prekės			√	
Kosmetika				√

Šaltinis: sudaryta autoriaus

Palyginimas atliktas išskiriant pagrindinius aukštą užsakymo kainą įtakojančius veiksnius suskirstytus pagal sritis: krovinio svoris, užsakymo trukmė, šalis iš kurios gabenamas krovinys, šalis į

kurią gabenamas krovinys, krovinio rūšis. Iš palyginimo matyti, jog labiausia aukštą kainą įtakoja krovinio svoris (daugiau nei 8,5 t.),bei šalis iš kurios gabenami kroviniai (Ispanija arba Italija). Galima daryti išvadą, jog reiktų atsisakyti tokių krovinių.

3.3. XLMiner pritaikymas įmonei

Šioje darbo dalyje tiems patiems duomenims pritaikyti XLMiner klasifikavimo algoritmai. XLMiner testavimui naudota programos demonstracinė versija turinti tam tikrų apribojimų. Vienas jų – apdorojamas įrašų kiekis. Demonstracinė versija leidžia nagrinėti tik 200 įrašų eilučių. Tad eksperimentui naudojama tik 200 naujausių užsakymų įrašų. Be to XLMiner labiau skirtas statistinei analizei, todėl daugumoje algoritmų leidžia naudoti tik skaitines reikšmes. XLMiner klasifikavimą leidžia atlikti naudojant šešis algoritmus, tačiau demonstracinė versija pilnai neleidžia jų visų panaudoti.

Klasifikavimui naudojant diskriminantinę analizę kaip nepriklausomi kintamieji imami krovinio pristatymo laikas ir krovinio svoris (duomenys su skaitinėmis reikšmėmis). Kiti kintamieji nebuvo įtraukti į analizę, nes jų reikšmės yra tekstinės. Priklausomas kintamasis, užsakymo kaina, prieš analizę suskirstyti į penkis intervalus (į juos duomenis buvo suskirstyti SQL Server 2008 Data Mining Add-Ins įrankio pritaikymo pradžioje), nes klasifikavimo algoritmai leidžia apdoroti tik tokius duomenis, kur priklausomas kintamasis turi ne daugiau nei 30 reikšmių. Tokie apribojimai neleidžia gauti tikslių rezultatų.

Pirminės klasių tikimybės

Equal prior probabilities

Klasė	Tikimybė
11741 - 15647	0,25
24.45 - 3930	0,25
3930 - 7836	0,25
7836 - 11741	0,25

Šaltinis: sudaryta autoriaus

31 pav. Klasių tikimybės (diskriminantinės analizės algoritmas)

Iš algoritmo pateikiamų rezultatų matome, jog pradžioje visoms klasėms priskiriamos vienodos tikimybės (31 pav.).

XLMiner pateikiama kintamųjų įtaka užsakymo kainai nurodo tik kuris kintamasis (krovinio pristatymo trukmė ar svoris) daro didesnę įtaką užsakymo kainai. Kaip matome iš 32 paveikslo aukštai

užsakymo kainai (aukštesnei nei 11741 Lt.) didesnę įtaką daro krovinio svoris, o ne krovinio pristatymo trukmė.

Modelis

Klasifikavimo funkcija

Kintamieji	Klasifikavimo funkcija			
	11741 -	24.45 - 3930	3930 - 7836	7836 - 11741
Konstanta	-17,5658474	-6,0394578	-9,60890865	-23,7135048
Trukmė	0,68211371	0,86605036	1,07946491	1,71192718
Svoris, t	3,72670484	1,82797909	2,47057343	4,1056385

Šaltinis: sudaryta autoriaus

32 pav. Kintamųjų įtaka (diskriminantinės analizės algoritmas)

XLMiner taip pat pateikia kiekvienos duomenų eilutės priklausomumo vienai ar kitai klasei tikimybės (33 pav.). Jei modeliavimo metu nustatyta eilutės klasė skiriasi nuo tikrosios, tokia eilutė išskiriama žaliai.

Klasės su aukščiausiomis tikimybėmis išskirtos geltonai
Neatitikimai taro tikros ir numatytos klasių išskirti žaliai

Eilutės ID	Numatyta klasė	Tikra klasė	11741 - 15647 klasės tikimybė	24.45 - 3930 klasės tikimybė	3930 - 7836 klasės tikimybė	7836 - 11741 klasės tikimybė	Trukmė	Svoris, t
1	24.45 - 3930	24.45 - 3930	0,011081908	0,679078762	0,309537169	0,000302161	1	4
2	24.45 - 3930	24.45 - 3930	0,001235452	0,730302687	0,268280992	0,00018087	3	3
3	3930 - 7836	3930 - 7836	0,022723273	0,362091559	0,595306355	0,019878812	4	5
4	3930 - 7836	3930 - 7836	0,031077466	0,412012542	0,547202147	0,009707845	3	5
5	24.45 - 3930	24.45 - 3930	0,001235452	0,730302687	0,268280992	0,00018087	3	3
6	24.45 - 3930	24.45 - 3930	0,055034356	0,505048848	0,437724866	0,002191929	1	5
7	24.45 - 3930	3930 - 7836	0,011081908	0,679078762	0,309537169	0,000302161	1	4
8	24.45 - 3930	3930 - 7836	0,006597407	0,584042222	0,407949518	0,001410853	3	4
9	24.45 - 3930	24.45 - 3930	0,006597407	0,584042222	0,407949518	0,001410853	3	4
10	24.45 - 3930	24.45 - 3930	0,001565419	0,769883701	0,228469046	8,18329E-05	2	3

Šaltinis: sudaryta autoriaus

33 pav. Eilutės elementų priklausymas kuriai nors klasei

Klasifikuojant duomenis panaudojant Naive Bayes algoritmą klasių tikimybės jau pradžioje skiriasi (34pav.).

Pirminės klasių tikimybės

According to relative occurrences in training data

Klasė	Tikimybė
11741 - 15647	0,01
24.45 - 3930	0,62
3930 - 7836	0,34
7836 - 11741	0,03

Šaltinis: sudaryta autoriaus

34 pav. Klasių tikimybės (Naive Bayes algortimas)

Kaip matome didžiausia tikimybė, jog užsakymo kaina yra nuo 24.45 iki 3930 Lt. Mažiausia tikimybė jog užsakymo kaina bus nuo 11741 iki 15647 Lt.

Naive Bayes algoritmas parodo ne tik kiekvieno kintamojo daromą įtaką, bet ir kiekvienos kintamojo verčių tikimybes.

	Klasės			Klasės	
Įėjimo kintamieji	11741 - 15647		Įėjimo kintamieji	11741 - 15647	
	Vertė	Tikimybė		Vertė	Tikimybė
Trukmė	1	1	Šalis į	Austrija	0
	2	0		Baltarusija	0
	3	0		Belgija	0
	4	0		Danija	0
	5	0		Graikija	0
	6	0		Italija	0
	7	0		Latvija	0
	8	0		Lenkija	0
	9	0		Lietuva	1
Šalis iš	Airija	0		Nyderlandai	0
	Austrija	0		Vokietija	0
	Baltarusija	0	Svoris, t	1	0
	Belgija	0		2	0
	Čekija	0		3	0
	Jungtinė Britanija	0		4	0
	Estija	0		5	0
	Ispanija	1		6	0
	Italija	0		7	0
	Latvija	0		8	0,5
	Lenkija	0		9	0,5
	Lietuva	0		11	0
	Liuksemburgas	0		13	0
	Nyderlandai	0			

Šaltinis: sudaryta autoriaus

35 pav. Kiekvienos vertės tikimybės (Naive Bayes)

Kaip matome jei užsakymo kaina yra aukšta, krovinio pristatymo trukmė yra viena diena, krovinys gabenamas iš Ispanijos į Lietuvą. Krovinio svoris dažniausiai būna 8 - 9 t. Išsami visų intervalų rezultatų lentelė pateikiama antrame priede (2 PRIEDAS).

Iš rezultatų matome, jog aukštą užsakymo kainą lemia įvairūs veiksniai. Iš Naive Bayes klasifikavimo matome, jog reikšminiai veiksniai yra:

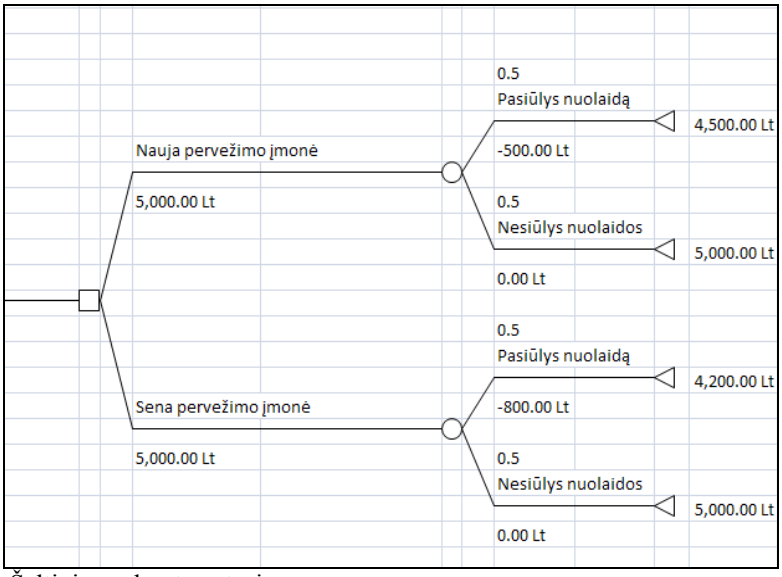
- Šalis iš kurios vežamas krovinys – Ispanija;
- Šalis į kurią vežamas krovinys – Lietuva;
- Krovinio svoris 8 arba 9 t.;
- Krovinio pristatymo trukmė.

Tačiau iš diskriminantinės analizės algoritmų rezultatų matome, jog krovinio svoris daro didesnę įtaką užsakymo kainai nei krovinio pristatymo trukmė. Tad didžiausi kriterijai lemiantys aukštą užsakymo kainą yra trys.

3.4. TreePlan pritaikymas įmonei

TreePlan įrankis, kaip jau minėta anksčiau, netinka tokiai duomenų analizei kaip MS SQL Server 2008 Data Mining Add-Ins ar XLMiner. Tačiau šis įrankis puikiai tinka modeliuoti situacijoms.

TreePlan įrankis pritaikytas mažai tarptautinių pervežimų įmonei siekiant pasirinkti pervežimo įmonę tolesniam bendradarbiavimui (36 pav.). Siekiama išsiaiškinti kas labiau apsimoka – tą patį pasiūlymą pateikti naujam vežėjui, ar tam, su kuriuo bendraujama jau seniai.



Šaltinis: sudaryta autoriaus

36 pav. TreePlan sprendimų medis

Sumodeliuotai situacijai paimti konkretūs duomenys. Tariama, jog pagal užsakymo kainą lemiančius kriterijus nustatyta 5000 Lt užsakymo kaina. Įmonė neturi duomenų apie tai, kaip dažnai pervežimo įmonės siūlo nuolaidas užsakymams. Todėl tikimybė, jog nuolaida pasiūlys nauja pervežimo įmonė ar ta, su kuria bendradarbiaujama jau seniai, palikta vienoda – 50%.

Įmonės darbuotojų teigimu ilgiau bendradarbiaujančios pervežimo įmonės siūlo 15-16 % nuolaidą užsakymams. Naujos pervežimo įmonės dažniausiai siūloma nuolaida tik 10 ar mažiau procentų. Tai sudaro atitinkamai 800 ir 500 Lt. Kaip matome iš grafiko geriau rinktis vežėją, su kuriuo bendraujama seniai. Nes pasiūlymas gali būti arba toks pats kaip ir naujosios pervežimo įmonės arba palankesnis.

Tokiu būdu galima modeliuoti ir sudėtingesnes situacijas, tačiau tam reikia papildomų duomenų.

3.5. Įrankių pritaikymo įmonei apibendrinimas

Nagrinėti įrankiai pritaikyti veiklos optimizavimui remiantis veiklos modeliu. Pradžioje ištirta įmonės veikla, apdoroti eksperimentinei daliai parinkti duomenys – užpildytos trūkstamos bei nulinės reikšmės. Remiantis šia informacija nustatytas pagrindinis jos tikslas – veiklos optimizavimas.

Kadangi skirtingiems duomenų gavybos metodams galima naudoti įvairius algoritmus, problemai spręsti buvo parinkti ne tik įrankiai, bet ir metodai bei algoritmai. Parinktam metodui buvo panaudoti skirtingi algoritmai. Tokiu būdu gaunami įvairiapusiški, išsamūs rezultatai leidžiantys į tą pačią situaciją pažvelgti iš skirtingų pusių.

Šiame darbe įrankiai pritaikomi įmonės veiklos optimizavimui, t.y. nepelningų, retai pasitaikančių užsakymų atrinkimui. Naudojant MS SQL Server 2008 Data Mining Add-Ins ir XLMiner įrankius siekta nustatyti, kas įtakoja aukštą užsakymo kainą. Rezultatai, gauti naudojant abu įrankius pateikti šeštoje lentelėje (9 lentelė).

9 lentelė

Įrankių atlikto klasifikavimo rezultatų palyginimas

Įtaką darantis kriterijus	**MS SQL Server 2008 Data Mining Add-Ins**				**XLMiner**	
	Sprendimų medis	*Logistinė regresija*	*Naive Bayes*	*Neuroniniai tinklai*	*Naive Bayes*	*Diskriminantinė analizė*
Krovinio svoris: 8.5-10 t			√	√	√	√
Trukmė: 1 diena					√	
Šalis iš: Ispanija	√		√	√	√	neanalizuota
Šalis iš: Italija			√	√		neanalizuota
Šalis iš: Lietuva					√	neanalizuota

Šaltinis: sudaryta autoriaus

Kaip matome labiausiai aukštą užsakymų kainą lemia krovinio svoris bei šalis iš kurios krovinys gabenamas. Iš šių rezultatų galima daryti išvadą, jog įmonei naudinga atsisakyti vežti krovinius iš Ispanijos. Tokiu būdu būtų galima daugiau dėmesio skirti kitiems užsakymams ir didinti jų skaičių, tuo pačiu mažinant krovinio vežimo kainą. Didesnis pervežimų skaičius panašiu maršrutu ir naudojantis tų pačių vežėjų paslaugomis didina lojalumą. Tai užtikrina galimybę gauti didesnę nuolaidą ir mažinti pervežimų kainas. Koks skirtumas tarp bendradarbiavimo su naujais ir senesniais parteriais matome iš TreePlan pritaikymo šiam klausimui spręsti.

3.6. Eksperimentinio skyriaus išvados

- Pritaikius MS SQL Server 2008 Data Mining Add-Ins, XLMiner bei TreePlan duomenų gavybos įrankius krovinių pervežimo įmonei, nustatyta jog įmonei būtų tikslinga atsisakyti vežti krovinius iš Ispanijos.
- Remiantis atlikta analize bei tyrimu, galima teigti, jog pasirinkti įrankiai puikiai tinka naudoti mažose įmonėse siekiančiose optimizuoti savo veiklą bei efektyviai panaudoti sukauptus duomenis.

IŠVADOS

1. Atlikus verslo intelektinių sistemų analizę nustatyta, jog mažų įmonių problemoms spręsti galima taikyti duomenų gavybą.
2. Atlikus duomenų gavybos algoritmų analizę pastebėta, jog tą pačią verslo problemą galima spręsti naudojant skirtingus algoritmus ir gauti skirtingus rezultatus interpretavimui.
3. Remiantis pateiktais problemos sprendimo modeliais, sudarytas ir realizuotas duomenų gavybos proceso modelis konkrečiai tarptautiniais pervežimais užsiimančios įmonės veiklai;
4. Atlikus duomenų gavybos programinės įrangos apžvalgą parinkti trys įrankiai (MS SQL Server 2008 Data Mining Add-Ins, XLMiner, TreePlan) veikiantys kaip MS Excel skaičiuoklės priedai. Įvertinus jų detalią analizę nustatyta jog visi jie puikiai tinka mažų įmonių veiklai ir yra tinkami mažų įmonių problemoms spręsti.
5. Pritaikius parinktus duomenų gavybos įrankius, nustatyta jog nagrinėtai pervežimo įmonei būtų tikslinga atsisakyti vežti krovinius iš Ispanijos. Taip būtų išvengta retų ir brangių užsakymų, tuo pačiu stiprinant ryšius (lojalumą) su tomis pervežimo kompanijomis, kurios gabena krovinius dažniau pasitaikančiais ir mažiau kainuojančiais maršrutais.
6. Panaudojus eksperimentinius duomenis realizuotas duomenų gavybos proceso modelio pritaikymas realiai įmonei.

LITERATŪRA

Mokslinės literatūros sąrašas

1. STRAVINSKIENĖ Auksė, ŽUKAUSKAITĖ Akvilė, GUDAS Saulius (2010) Duomenų gavybos įrankių pritaikymas mažose įmonėse. 16th International Conference on Information and Software Technologies IT 2010. ISSN 2029-0055

2. STRAVINSKIENĖ Auksė, GUDAS Saulius (2010) Duomenų gavyba paremta veiklos modeliu. 16th International Conference on Information and Software Technologies IT 2010. ISSN 2029-0055

3. BORGZA R.M., ZAHARI D. (2008) Business intelligence as a competitive differentiator. *IEEE Xplore.*

4. XU Lida, ZENG Li, SHI Zhongzhi HE Qing, WANG Maoguang (2007) Research on Business Intelligence in Enterprise Computing Environment. *IEEE Xplore.*

5. MOSS Larissa T. (2008), Organizational Barriers to Business Intelligence (Part 1) [interaktyvus] ITNetwork365 [žiūrėta 2009m. sausio 12d.]. Prieiga per internetą: <http://www.businessintelligence.com/article.asp?id=120&pagenum=1 >

6. COZKUN SAMLI A., POHLEN T. L., BOZOVIC N. (2002) A Review of Data Mining Techniques as they Apply to Marketing: Generating Strategic Information to Develop Market Segments. *IEEE Xplore.*

7. ZHANG Guozheng, ZHOU Faming, WANG Fang, LUO Jian (2008) Knowledge creation in marketing based on data Mining. *IEEE Xplore.*

8. MYSZKOWSKI Paweł B., KWAŚNICKA Halina and MARKOWSKA-KACZMAR Urszula (2008) Data Mining techniques in e-learning CelGrid system. *IEEE Xplore.*

9. ZHANG D., ZHOU L. (2004) Discovering Golden Nuggets: Data Mining in Financial Application . *IEEE Xplore.*

10. FAYYAD Usama, PIATETSKY-SHAPIRO Gregory, SMYTH Padhraic. (1996) From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence.*

11. WU Xindong, KUMAR Vipin, QUINLAN J. Ross ir kt. (2008) Top 10 algorithms in data mining. *Springer.*

12. CIESIŪNAS Arūnas, JUDŽENTIS Edvardas.(2007) Duomenų gavybos metodai. *10-osios Lietuvos jaunųjų mokslininkų konferencijos „Mokslas – Lietuvos ateitis“ medžiaga.*

13. CARVALHO D.R., FREITAS A.A. (2000) A hybrid decision tree/genetis algorithm for coping with the problem of small disjuncts in data mining. *The Pennsylvania State University*

14. BRONIUKAITIS, Romualdas. (1997) *Ekspertinės sistemos ir žinių bazės: mokomoji priemonė*. Vilnius: Vilniaus universiteto leidykla. 25 psl. UDK 681.518.2(075.8).

15. SEKLIUCKIS Vitolis, GUDAS Saulius, GARŠVA Gintautas (2006) Informacijos sistemos ir duomenų bazės : informacijos sistemų ir reliacinių duomenų bazių kūrimo pagrindai : vadovėlis. Kaunas: Technologija. 349 psl. ISBN 9955094869.

16. WITTEN I.H., FRANK E. (2005) *Data mining. Practical Machine Learning Tools and Techniques, Second Edition.* Elsevier Inc.522 psl. ISBN: 0-12-088407-0

17. GUDAS Saulius (2002) *Veiklos analizė ir informacinių poreikių specifikavimas : mokomoji knyga.* Kaunas: Naujasis lankas. 93 psl. ISBN 9955031263.

18. RITHOLZ B**.** (2009) Intro to Data Mining for Small Businesses [interaktyvus] *2010 American Express Company* [žiūrėta 2010m. sausio 11d.]. Prieiga per internetą: < http://www.openforum.com/idea-hub/topics/money/article/intro-to-data-mining-for-small-businesses-barry-ritholtz >

Informacijos šaltinių sąrašas

19. MLADENIC Dunja, LAVRAC N., ir kt., (2003) *Data Mining and Decision Support for Business Competitiveness: A European Virtual Enterprise* [interaktyvus]. SolEUNet [žiūrėta 2009m. balandžio 4d.]. Prieiga per internetą: < http://soleunet.ijs.si/website/html/euproject.html >

20. ROSSETTI Luca. (2009) *SearchDataManagement.com Definitions* [interaktyvus]. TechTarget [žiūrėta 2009m. balandžio 4d.]. Prieiga per internetą: < http://searchdatamanagement.techtarget.com/sDefinition/0,,sid91_gci213571,00.html >

21. UAB Alna Business Solutions (2007) *Kas yra verslo analizės sistema* [interaktyvus]. Vilnius: Alna Business Solutions UAB. [žiūrėta 2009m. sausio 13d.]. Prieiga per internetą: <http://www.alna.lt/abs/sprendimai/verslo_analizes_sistema_business_intelligence/Kas_yra_verslo_analizes_sistema/ >

22. PETTEY Christy (2009) *Gartner EXP Worldwide Survey of More than 1,500 CIOs Shows IT Spending to Be Flat in 2009* [interaktyvus]. Gartner [žiūrėta 2009m. sausio 13d.]. Prieiga per internetą: < http://www.gartner.com/it/page.jsp?id=855612 >

23. LOGAN Valerie (2008) *Top 10 Trends in Business Intelligence for 2008* [interaktyvus]. Technology Solutions Group, Hewlett-Packard Company [žiūrėta 2009m. sausio 13d.]. Prieiga per internetą: < http://www.dashboardinsight.com/articles/new-concepts-in-business-intelligence/top-10-trends-business-intelligence-for-2008.aspx >

24. HOSTMAN Bill, SCHLEGEL Kurt, GASSAMAN Bill, RAYNER Nigel ir kt. (2009) *Business intelligence trends for 2009* [interaktyvus]. MyCustomer [žiūrėta 2009m. sausio 13d.]. Prieiga per internetą: < http://www.mycustomer.com/item/134178 >

25. Naujoji komunikacija (2003) *Interaktyvios duomenų analizės įrankiai šiuolaikinėje įmonėje. OLAP duomenų bazės* [interaktyvus]. Naujoji komunikacija [žiūrėta 2009m. sausio 13d.].

Prieiga per internetą: < http://www.nk.lt/programine-iranga/interaktyvios-duomenu-analizes-irankiai-siuolaikineje-imoneje-olap-duomenu-bazes/ >

26. IT Ekspertas (2005) *OLAP duomenų bazės* [interaktyvus]. Vilnius: IT Ekspertas. [žiūrėta 2009m. sausio 13d.]. Prieiga per internetą: < http://itekspertas.projektas.lt/index.php?option =com_content&task=view&id=94&Itemid=53 >

27. Leonardo da Vinci Programa (2009) *Kas yra žinių bazė?*[interaktyvus] ETutors [žiūrėta 2009m. sausio 13d.]. Prieiga per internetą: < http://www.etutors-portal.net/portal-contents-lt-lt/folder.2007-12-08.8382965941-lt/folder.2007-12-08.3222969154-lt/kas-yra-ziniu-baze >

28. StatSoft (2008) *Classification Trees* [interaktyvus] StatSoft [žiūrėta 2009m. sausio 13d.]. Prieiga per internetą: < http://www.statsoft.nl/uk/textbook/stclatre.html >

29. XIAO Henry (2004) *Data Mining - Bayesian Approaches* [interaktyvus] School of Computing, Queen's University [žiūrėta 2009m. sausio 13d.]. Prieiga per internetą: < http://www.cs. queensu. ca/home/xiao/dm.html >

30. TEKNOMO Kardy (2006) *What is K Nearest Neighbors Algorithm?* [interaktyvus] Kardi Teknomo [žiūrėta 2009m. sausio 13d.]. Prieiga per internetą: < http://people.revoledu.com /kardi/tutorial/KNN/What-is-K-Nearest-Neighbor-Algorithm.html >

31. COURT Merz (2007) *The basic algorithm for a GA* [interaktyvus] Newcastle Engineering Design Centre, Newcastle University. [žiūrėta 2009m. sausio 13d.]. Prieiga per internetą: <www.edc.ncl.ac.uk/highlight/rhjanuary2007g01.php/ >

32. SIMUTIS Rimvydas (2008) Skaitinio intelekto taikymai. *Leidinys paruošas remiant Lietuvos valstybiniam mokslo fondui*, *pirma versija*.

33. The CRISP-DM consortium (2003) *Process Model* [interaktyvus] The CRISP-DM project. [žiūrėta 2009m. sausio 13d.].Prieiga per internetą: < http://www.crispm.org/Process/index.htm>

34. CHAPMAN Pete, CLINTON Julian, KERBER Randy ir kt. (2000) CRISP-DM 1.0 Step-by-step data mining guide *The CRISP-DM project.*

35. KDnuggets (2009a) *Software Suites for Data Mining, Analytics, and Knowledge Discovery* [interaktyvus] KDnuggets. [žiūrėta 2009m. sausio 13d.]. Prieiga per internetą: <http://www. kdnuggets.com/software/suites.html >

36. BANGE Axel (2008) *Market shares 2007* [interaktyvus] BARC Gmbh [žiūrėta 2009m. sausio 13d.]. Prieiga per internetą: < http://www.barc.de/en/marktforschung/bi-market-in-germany-2007.html >

37. MAYATO (2008) *mayato Study: Data Mining Software 2009* [interaktyvus] Mayato [žiūrėta 2009m. sausio 13d.]. Prieiga per internetą: < http://www.mayato.com/downloads/Summary_ mayato_Data-Mining-Study_2009.pdf >

38. HERSCHEL Gareth (2008) *Magic Quadrant for Customer Data-Mining Applications* [interaktyvus] Gartner [žiūrėta 2009m. sausio 13d.]. Prieiga per internetą: < http://mediaproducts.gartner.com/reprints/sas/vol5/article3/article3.htm l>

39. KDnuggets (2009b) *Data Mining Tools Used Poll* [interaktyvus] KDnuggets. [žiūrėta 2009m. sausio 13d.]. Prieiga per internetą: < http://www.kdnuggets.com/polls/2009/data-mining-tools-used.htm >

40. Statistics.com (2006) *XLMiner Data Mining Add-Ins for Excel* [interaktyvus] Cytel Software Corporation [žiūrėta 2010 m. sausio 13 d.]. Prieiga per internetą: < http://www.resample.com /xlminer/index.shtml >

41. Microsoft Corporation (2009) *Data Mining Add-Ins for Office 2007* [interaktyvus] Microsoft Corporation [žiūrėta 2010 m. sausio 20 d.]. Prieiga per internetą: < http://www.microsoft.com /sqlserver/2008/en/us/data-mining-addins.aspx >

42. MIDDLETON Mike(2009) *TreePlan Decision Tree Add-Ins* [interaktyvus] Decision Toolworks [žiūrėta 2010 m. sausio 20 d.]. Prieiga per internetą: < http://www.microsoft.com /sqlserver/2008/en/us/data-mining-addins.aspx >

VILNIAUS UNIVERSITETAS
KAUNO HUMANITARINIS FAKULTETAS
INFORMATIKOS KATEDRA

VERSLO INFORMACIJOS SISTEMŲ MAGISTRATŪROS PROGRAMOS
MOKSLO TIRIAMOJO DARBO PLANAS

Magistrantė Akvilė Žukauskaitė Tel. 8 600 22768

Magistratūros trukmė nuo 2008 09 01 iki 2010 06 30

TEMA Duomenų gavybos sistema naudojant veiklos modelį (Enterprise Model based Data Mining System)

Vadovas prof. Saulius Gudas

Darbo anotacija:

Tikslas: Sukurti duomenų gavybos sistemą paremtą veiklos modeliu.

Uždaviniai:

1. Atlikti intelektinių verslo sistemų analizę, parinkti tinkamiausią mažų įmonių problemoms spręsti;
2. Atlikti duomenų gavybos metodų apžvalgą bei algoritmų analizę.
3. Sudaryti duomenų gavybos proceso modelį, panaudojant veiklos modelį.
4. Atlikti duomenų gavybos įrankių pritaikomų mažose įmonėse detalią analizę;
5. Panaudojant realius duomenis patikrinti įrankių pritaikymą mažose įmonėse.
6. Panaudojant eksperimentinius duomenis realizuoti duomenų gavybos proceso pritaikymą realiai įmonei.

Metodai:

- Mokslinės literatūros analizė – problemos aktualumui pagrįsti, teorinės medžiagos rinkimui;
- Analogija ir palyginimas – duomenų gavybos sistemų palyginimui;
- Modeliavimas – duomenų gavybos procesų atvaizdavimui veiklos modeliais;
- Eksperimentas – sistemos testavimui.

Laukiami rezultatai:

- Atlikta duomenų gavybos sistemų ir algoritmų analizė, aptarti jų privalumai, trūkumai, panaudojimo sritys, duomenų gavybos algoritmų tendencijos;
- Išanalizuota duomenų gavybos programinė įranga pritaikoma mažų įmonių problemoms spręsti;
- Sudarytas duomenų gavybos proceso modelis mažai įmonei.

Mokslo – tiriamojo darbo planas

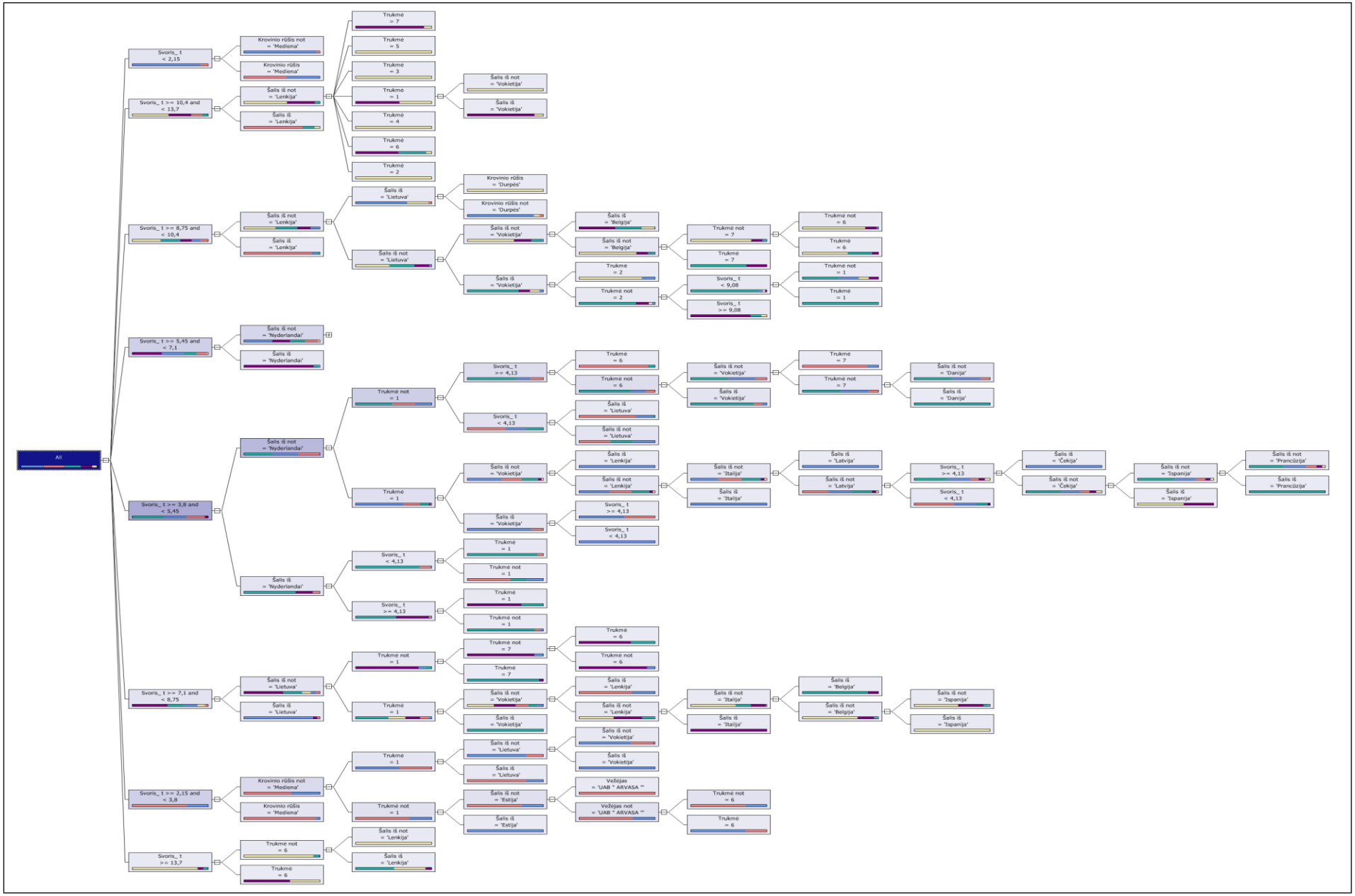
Semestras	(data)	Užduotys
S1	2008-10-25	Magistrinio darbo temos formulavimas.
	2008-11-01	Kalendorinio plano parengimas
	2008-12-01	Literatūros studijavimas ir esamos padėties apžvalga.
	2008-12-27	Darbo problemos formulavimas, galimų sprendimų ir hipotezių iškėlimas, darbo objekto, tikslo ir uždavinių nustatymas.
	2009-01-16	Pirmojo etapo ataskaitos ruošimas
S2	2009-03-25	Teorinės darbo problemos sprendimo medžiagos ruošimas.
	2009-04-15	Esančių modelių, metodų, algoritmų, sprendimų analizė, jų trūkumų ir privalumų nustatymas, naujų pasiūlymų pateikimas.
	2009-04-30	Išsamus jų aprašymas, preliminaraus sprendimo skyriaus paruošimas
	2009-05-15	Teorinio skyriaus parengimas.Antrojo etapo ataskaitos ruošimas
S3	2009-09-20	Eksperimentinės tyrimo metodikos ruošimas.
	2009-11-10	Duomenų ir kitos eksperimentiniam tyrimui reikalingos medžiagos rinkimas, apdorojimas, vertinimas
	2009-12-27	Eksperimentinių tyrimų atlikimas.
	2010-01-20	Preliminarių išvadų formulavimas. Trečiojo etapo ataskaitos ruošimas.
S4	2010-03-02	Teorijos skyriaus papildymas
	2010-04-01	Eksperimentinio skyriaus papildymas
	2010-04-25	Išvados apie gautų rezultatų praktinį pritaikymą, pasiūlymų ir rekomendacijų rengimas.
	2010-05-01	Galutinis magistrinio darbo, pagal metodinius nurodymus, sutvarkymas ir parengimas.

Magistrantas: ..... Vadovas: .....

PRIEDAI

1 PRIEDAS

SPRENDIMŲ MEDŽIO ALGORITMO REZULTATAI



2 PRIEDAS

NAIVE BAYES ALGORITMO REZULTATAI

	Klasės							
Įėjimo kintamieji	11741 - 15647		24.45 - 3930		3930 - 7836		7836 - 11741	
	Vertė	Tikimybė	Vertė	Tikimybė	Vertė	Tikimybė	Vertė	Tikimybė
Trukmė	1	0,999	1	0,516	1	0,338	1	0,166
	2	0	2	0,120	2	0,117	2	0,333
	3	0	3	0,120	3	0,191	3	0
	4	0	4	0,104	4	0,191	4	0
	5	0	5	0,056	5	0,073	5	0
	6	0	6	0,040	6	0,029	6	0,166
	7	0	7	0,024	7	0,044	7	0,166
	8	0	8	0,008	8	0,014	8	0
Šalis iš	Airija	0	Airija	0	Airija	0	Airija	0,166
	Baltarusija	0	Baltarusija	0,016	Baltarusija	0	Baltarusija	0
	Belgija	0	Belgija	0,016	Belgija	0,117	Belgija	0
	Čekija	0	Čekija	0,064	Čekija	0	Čekija	0
	D.Britanija	0	D. Britanija	0,024	D. Britanija	0,014	D. Britanija	0
	Estija	0	Estija	0,024	Estija	0	Estija	0
	Ispanija	0,999	Ispanija	0,024	Ispanija	0	Ispanija	0
	Italija	0	Italija	0,008	Italija	0,014	Italija	0
	Latvija	0	Latvija	0,032	Latvija	0	Latvija	0
	Lenkija	0	Lenkija	0,064	Lenkija	0	Lenkija	0
	Lietuva	0	Lietuva	0,137	Lietuva	0,044	Lietuva	0
	Liuksem.	0	Liuksem.	0,024	Liuksem.	0,058	Liuksem.	0
	Nyderlan.	0	Nyderlan.	0,104	Nyderlan.	0,147	Nyderlan.	0,499
	Norvegija	0	Norvegija	0,008	Norvegija	0	Norvegija	0
	Prancūzija	0	Prancūzija	0,016	Prancūzija	0,014	Prancūzija	0
	Rusija	0	Rusija	0,00	Rusija	0	Rusija	0
	Slovakija	0	Slovakija	0,024	Slovakija	0	Slovakija	0
	Suomija	0	Suomija	0,008	Suomija	0,014	Suomija	0
	Švedija	0	Švedija	0,008	Švedija	0,029	Švedija	0
	Šveicarija	0	Šveicarija	0,016	Šveicarija	0	Šveicarija	0
	Vokietija	0	Vokietija	0,370	Vokietija	0,514	Vokietija	0,333
Šalis į	Austrija	0	Austrija	0,008	Austrija	0	Austrija	0
	Baltarusija	0	Baltarusija	0,008	Baltarusija	0	Baltarusija	0
	Belgija	0	Belgija	0,008	Belgija	0,014	Belgija	0
	Danija	0	Danija	0,016	Danija	0	Danija	0
	Graikija	0	Graikija	0	Graikija	0,014	Graikija	0
	Italija	0	Italija	0,008	Italija	0,014	Italija	0
	Latvija	0	Latvija	0,008	Latvija	0	Latvija	0
	Lenkija	0	Lenkija	0,024	Lenkija	0	Lenkija	0
	Lietuva	0,999	Lietuva	0,862	Lietuva	0,941	Lietuva	0,999
	Nyderlan.	0	Nyderlan.	0,008	Nyderlan.	0	Nyderlan.	0
	Vokietija	0	Vokietija	0,048	Vokietija	0,014	Vokietija	0
Svoris, t	1	0	1	0,048	1	0	1	0
	2	0	2	0,048	2	0	2	0
	3	0	3	0,338	3	0	3	0
	4	0	4	0,266	4	0,367	4	0
	5	0	5	0,120	5	0,308	5	0
	6	0	6	0,112	6	0,058	6	0,166
	7	0	7	0,056	7	0,147	7	0
	8	0,499	8	0,008	8	0,088	8	0,166
	9	0,499	9	0	9	0,014	9	0,499
	11	0	11	0	11	0,014	11	0
	13	0	13	0	13	0	13	0,166

3 PRIEDAS

DUOMENŲ GAVYBOS ĮRANKIŲ PRITAIKYMAS MAŽOSE ĮMONĖSE

Auksė Stravinskienė[1], Akvilė Žukauskaitė[2], Saulius Gudas[3]

[1] Vilniaus universiteto Kauno humanitarinio fakulteto doktorantė
el. paštas aukse.stravinskiene@vukhf.lt
[2] Vilniaus universiteto Kauno humanitarinio fakulteto magistrantė
el.paštas akvile.zukauskaites@vukhf.lt
3 Vilniaus universiteto Kauno humanitarinio fakulteto profesorius
el.paštas gudas@vukhf.lt

Santrauka. *Straipsnyje analizuojama duomenų gavybos technologijų pritaikymas mažoms įmonėms, negalinčioms investuoti į informacines technologijas, tačiau siekiančioms optimizuoti veiklos procesus, efektyviai panaudoti sukauptus duomenis.*

Raktiniai žodžiai: duomenų gavyba, veiklos modeliavimas, veiklos procesų optimizavimas.

1. Įvadas

Duomenų gavyba versle nėra naujas reiškinys, tačiau dažnai duomenų gavybos įrankius taiko didelės įmonės. Didelių įmonių vadovai labiau vertina duomenų analizės svarbą, bei gali skirti tam daugiau lėšų. Mažos įmonės dažnai susiduria su problema, nes dauguma duomenų gavybos įrankių kainuoja daug lyginant su įmonės gaunamomis pajamomis, todėl mažos įmonės, prireikus analizuoti duomenis, naudoja elementarias skaičiuokles kaip MS Excel. Tačiau gautas rezultatas dažnai neduoda tiek naudos, kiek duotų duomenų gavybos įrankių panaudojimas. Straipsnyje siūloma šią problemą spręsti naudojant įrankius kurie veikia ne kaip atskiros programos, bet kaip MS Excel skaičiuoklės dalys.

Tema aktuali tuo, kad vis daugiau mažų įmonių vadovų siekia optimizuoti savo įmonių veiklą, efektyviai panaudoti kaupiamus duomenis, sutaupyti lėšų ir aiškiai įvertinus įmonės perspektyvas planuoti tolesnę veiklą. Darbo tikslas yra pasiūlyti duomenų gavybos įrankius, kurie tiktų naudoti mažoms įmonėms, įvertinant jų naudojimo patogumą bei galimybes.

Straipsnyje pateikiamas trijų įrankių, kurie gali būti efektyviai naudojami mažose įmonėse palyginimas. Pagrindinis įrankių privalumas yra tai, jog visi jie integruojami į Microsoft Office paketo skaičiuoklę Excel.

2. Duomenų gavyba ir jos samprata

Duomenų gavyba (angl. Data Mining) yra anksčiau nežinomos ir potencialiai naudingos informacijos ištraukimas iš sukauptų duomenų. Ji įdomi tuo, kad jos technologija sugeba faktiškus duomenis paversti naudinga informacija ir žiniomis, tinkamomis veiklos valdymui, rinkos analizei, sprendimų priėmimui. (Sekliuckis V., 2006).

Standartinį duomenų gavybos proceso modelį CRISP-DM (angl. Cross-Industry Standard Process for Data Mining) sudaro šeši, nuosekliai išsidėstę etapai: veiklos suvokimas, duomenų suvokimas, duomenų paruošimas, modeliavimas, analizė ir pritaikymas. Šį modelį 1996 metais sudarė duomenų gavybos pradininkai – Daimler Chrysler, SSPS, Teradata. Modelis buvo kuriamas remiantis ne tiek teorija, kiek praktika, todėl jis yra paplitęs ir dažnai naudojamas (The CRISP-DM consorcium, 2003).

Duomenų gavybos procese visi šie šeši etapai vykdomi nuosekliai. Kadangi tai uždaras ciklas, tai įvykdžius vieną žingsnį neišvengiamai reikia grįžti į prieš tai padarytus žingsnius. Iteratyvumas svarbus siekiant užtikrinti proceso vientisumą, kokybišką eigą bei norint išvengti galimų klaidų. Norint duomenų gavybos procesą atlikti kokybiškai reikia tuos pačius veiksmus kartoti po keletą ar net keliasdešimt kartų.

Duomenų gavybos taikymo sritis yra labai plati. Ji apima nemažą skaičių įvairių metodų, algoritmų, taikomųjų sistemų. Kiekvienai problemai spręsti reikia stengtis parinkti tinkamą duomenų gavybos metodą (Fayyad U., 1996).

- Klasifikavimas: remiantis duomenų savybėmis skirsto juos į iš anksto nustatytą kiekį tam tikrais parametrais aprašytų klasių;
- Regresija: remiantis turimais duomenimis sudaromo modelį leidžiantis prognozuoti realias būsimas skaitines reikšmes.
- Klasterizavimas: skirsto duomenis į iš anksto nenumatytą kiekį grupių ar klasterių pagal skiriamuosius bruožus.
- Apibendrinimas: apima metodus padedančius sutrumpinti duomenis. Šie metodai dažniausiai naudojami interaktyviom žvalgomosioms analizėm, automatinėm ataskaitom ruošti;

- Priklausomybių modeliavimas: randa modelį, kuris aprašo priklausomybes (struktūrinio bei kiekybinio lygio) tarp kintamųjų.
- Pakeitimų ir nukrypimų aptikimas: skirtas atskleisti svarbiausius pokyčius tarp anksčiau išmatuotų ir naujausių duomenų.

Kiekvienam metodui galima pritaikyti net po keletą algoritmų, kurie veikdami skirtingais metodais, padeda gauti skirtingus rezultatus. Dažnai dėl to rezultatai tampa reikšmingesni, svarbesni ir lengviau panaudojami, nes leidžia į situaciją pažvelgti iš skirtingų pozicijų. Vieni duomenų gavybos algoritmai geriau klasifikuoja duomenis, kiti geriau prognozuoja. Visi jų turi savų privalumų bei trūkumų. Dažniausiai naudojami bei labiausiai išnagrinėti algoritmai: sprendimų medžio, Bajeso klasifikacija, klasterizavimo, artimiausių kaimynų, neuroniniai tinklai, genetiniai ir hibridiniai. Pastaruosius sudaro kelių algoritmų kombinacijos.

Įvairūs duomenų gavybos įrankiai skiriasi savo savybėmis bei leidžiamais naudoti metodais bei algoritmais. Mažų įmonių vadovai turėtų į tai atsižvelgti siekdami optimizuoti savo įmonių veiklą, efektyviai panaudoti kaupiamus duomenis, sutaupyti lėšų ir aiškiai įvertinus įmonės perspektyvas planuoti tolesnę. Gerai, kai tas pats įrankis tą pačią analizę leidžia atlikti naudojant skirtingus algoritmus. Tokių rezultatų nagrinėjimas suteikia daug naudingos informacijos.

3. Duomenų gavybos įrankių panaudojimo reikšmė mažose įmonėse

Duomenų gavybos įrankių panaudojimo didelėse įmonėse tema yra plačiai aptarinėjama ir analizuojama. Pateikiama daugybė programinių produktų, skirtų apdoroti ir analizuoti duomenis. Tačiau tik neseniai pradėta atkreipti dėmesį į mažas įmonės, kuriose būtų galima pritaikyti duomenų gavybos įrankius.

Pagrindinės priežastys, kodėl mažose įmonėse nenaudojamos duomenų gavybos sistemos:

- Informacijos stoka apie duomenų gavybos sistemas;
- Manymas, jog duomenų gavybos įrankiai brangūs ir reikalauja didelių eksploatacinių išlaidų;
- Teikiamos naudos įmonei nežinojimas;
- Specialistų reikalingumas, kurių mažos įmonės nepajėgia pasisamdyti;
- Papildomos išlaidos darbuotojų apmokymui.

Taip pat dažnai manoma jog duomenų gavybos įrankių pateikti rezultatai yra sudėtingi ir sunkiai suprantami. Dauguma mažų įmonių vengia naudoti duomenų gavybos sistemas, nes nežino tikrosios duomenų gavybos įrankių naudos įmonei. Tačiau taip elgiasi ne visos įmonės. Jau prieš kelis metus Melissa Solomon tyrinėjo mažų įmonių norą naudotis aukštosiomis technologijomis. Jis išskyrė dešimtuką IT sričių į kurias mažos įmonės planavo investuoti pinigus. Nors pirmąsias vietas užėmė saugumo užtikrinimo produktai bei bendravimo programos, į sąrašą įtraukta ir duomenų gavyba. Šiuo atveju ji pateikta kaip įrankiai skirti klientų veiksmų analizei. Surinkus informaciją apie tai, ką jie veikia bei kuo domisi, ją galima naudoti analizuojant įmonės darbą bei atsižvelgiant į paklausą planuojant naujus pasiūlymus (Solomon M., 2005).

Tačiau kokia duomenų gavybą yra svarbi mažoms įmonėms puikiai atskleidžiama Barny Ritholz straipsnyje „Intro to Data Mining for Small Businesses“. Pagrindinė straipsnio idėja yra ta, jog duomenų gavyba gali padėti įmonėms anksčiau sureaguoti į rinkos pokyčius ir lengviau išgyventi net finansines krizes. Straipsnyje teigiama, kad jei mažos įmonės pačios naudotų duomenų gavybos įrankius, jos greičiau galėtų pastebėti rinkos pokyčius, jų tendencijas ir įvairias anomalijas. Taip būtų greičiau sureaguota į pokyčius nei tuo atveju, kai apie rinkos krizes praneša ekonomikos biurai surinkę ir apdoroję tūkstančių panašių įmonių informaciją. Šiuo atveju tas laiko skirtumas gali būti labai reikšmingas ir padėti įmonei išvengti nuostolingų investicijų (Ritholz., 2009)

Mažos įmonės turėtų daugiau dėmesio kreipti į duomenų rinkimą bei jų analizę. Viena pagrindinių ir svarbiausių duomenų gavybos funkcijų mažoms įmonėms būtų prognozavimas. Įvertinus pakankamai rodiklių įmonėms būtų daug lengviau planuoti savo veiklą, būsimas pajamas ir išlaidas. Priklausomai nuo įmonės veiklos galima prognozuoti jos pajamas, pardavimus, užsakymų skaičių ir pan.

Įmonėms pardavinėjančioms prekes didelės naudos galėtų duoti ir pirkinių krepšelio analizė. Įvertinus, kokius prekių rinkinius dažniausiai renkasi pirkėjai, būtų galima vykdyti įvairias akcijas, kai perkant vieną rinkinio prekę pasiūloma kita, dažniausiai perkama kartu. Žinant vartotojų įpročius daug lengviau įtikinti jį pirkti.

Duomenų gavybą mažose įmonėse panaudoti galima labai įvairiai, tačiau teisingai naudojant duomenų gavybos įrankius įmonė visada dirbs optimaliau bei efektyviau. Daug kas priklauso nuo įmonės veiklos bei pasirinktų duomenų gavybos įrankių.

4. Duomenų gavybos įrankių palyginimas

Kaip jau minėta anksčiau, mažos įmonės dažnai neturi pakankamai lėšų duomenų gavybos įrankiams įsigyti, tačiau siekia optimizuoti veiklos procesus bei efektyviai panaudoti sukauptus duomenis. Viena iš išeičių tokioms įmonėms būtų naudoti įrankius, kurie veikia ne kaip atskiros programos, o kaip tam tikrų, jau naudojamų, programų papildymas. Šiuo atveju toliau išnagrinėti trys produktai, kurie integruojami į Microsoft Office paketo skaičiuoklę Excel ir veikia kaip papildomos skaičiuoklės funkcijos. MS Excel skaičiuoklę naudoja daugelis įmonių, todėl duomenų

gavybos funkcijos tiesiog praplėstų kasdien naudojamų skaičiuoklės funkcijų rinkinį. Toliau esančioje lentelėje (žr. 1 lentelė) pateikta pagrindinių duomenų gavybos funkcijų apžvalga trijuose nagrinėjamuose įrankiuose – XLMiner, TreePlan bei Microsoft SQL Server 2008 Data Mining Add-Ins.

XLMiner bei Microsoft SQL Server 2008 Data Mining Add-Ins turi funkcijų tiek duomenų tvarkymui tiek modeliavimui. Viena iš svarbesnių duomenų tvarkymo funkcijų yra **duomenų padalijimas** – jis leidžia duomenis skirstyti į grupes, kurių viena naudojama modelio mokymui, o kita testavimui. Jei duomenys nepilni – trūksta įrašų ar norima dalį pakeisti įrankiuose tam galima rasti ne vieną funkciją. Įmonėms reiktų įvertinti tai, kad kuo tikslesni bei išsamesni duomenys nagrinėjami, tuo geresni rezultatai gaunami.

Abu įrankiai siūlo gan panašių **klasifikavimo įrankių**. Platus šių funkcijų pasirinkimas leidžia vartotojams tuos pačius duomenis apdoroti skirtingais būdais, stebėti kuo panašūs ar kuo skiriasi rezultatai ir daryti dar nuoseklesnes išvadas. Microsoft SQL Server 2008 Data Mining Add-Ins suteikia ir kitokias galimybes panaudoti klasifikavimą – ji leidžia modeliuoti tam tikras situacijas. Pavyzdžiui siūloma funkcija „tikslo siekimas“ pasirinkus eilutę ir nurodžius vieno eilutės elemento reikšmę (tikslią, apytikslę ir pan.) bei elementą, kurį galime keisti, randa geriausias keičiamo elemento variantą, o funkcija “kas jeigu“ leidžia pamatyti kas atsitiktų stebimam (priklausomam) eilutės elementui, jei pakeistume kurį nors kitą. Tokios modeliavimo funkcijos leidžia įvertinti įvairias galimybes nepatiriant nuostolių. **Bendrų bruožų radimo** funkcijas įrankiuose pateikia pirkinių krepšelio analizė. Tai labai naudinga tiek prekybos, tiek paslaugų įmonėms nagrinėjant vartotojų elgseną ir siekiant pagerinti, pagreitinti jų aptarnavimą tuo pačiu pagerinant santykius su vartotoju – didinant jo lojalumą.

Viena iš dažniausiai įmonėse naudojamų funkcijų – prognozavimas, kuris padeda lengviau įvertinti ateities perspektyvas bei patikimiau planuoti įmonės veiksmus. Tai vienintelė funkcija, kuria teikia TreePlan įrankis, Jis leidžia braižyti sprendimo medžius juose žymint sprendimų, įvykių mazgus, atskirų įvykių tikimybes bei uždirbamas ar išleidžiamas pinigų sumas. Šis įrankis puikiai tinka sprendžiant nuoseklias problemas, kai galimos kelios alternatyvos (pvz,: į rinką įvesti naują produktą, išbandyti naują technologiją).

1 lentelė. Duomenų gavybos įrankių funkcijų palyginimas

	XLMiner	**TreePlan**	**Microsoft SQL Server 2008 Data Mining Add-Ins**
Duomenų tvarkymas			
Duomenų padalijimas	Dalina duomenis į 3 grupes: mokymosi, patvirtinimo ir testavimo.	Nėra	Dalina duomenis į mokymosi ir testavimo.
Imties sudarymas	• Padalijimas su perviršiu. • Paprasta imti. • Imtis naudojant stratas.	Nėra	• Atsitiktinis duomenų padalijimas. • Padalijimas su perviršiu.
Trūkstamų/netinkamų duomenų redagavimas	Trūkstami duomenų papildymas.	Nėra	• Duomenų apkarpymas. • Laukelių pavadinimų keitimas. • Klasterizavimas.
Duomenų formato keitimas	• Laikinų duomenų sukūrimas. • Kategorinių duomenų kūrimas. • Tolydžiųjų duomenų rūšiavimas.	Nėra	Nėra
Duomenų modeliavimas			
Prognozavimas	• Daugybinė tiesinė regresija • Artimiausių kaimynų metodas. • Regresiniai medžiai. • Neuroniniai tinklai.	Sprendimų medis (visų galimų variantų įvertinimas)	• Laiko eilutės (prognozavimas). • Logistinė regresija (prognozės modeliavimas). • Sprendimų medžio algoritmas (apytikris skaičiavimas).
Klasifikavimas	• Diskriminantinė analizė. • Logistinė regresija. • Klasifikavimo medžiai. • Naive Bayes‘o algoritmas. • Neuroniniai tinklai.	Nėra	• Klasterizavimas (kategorijų nustatymas). • Sprendimo medžio algoritmas. • Logistinė regresija (užpildymas pagal pavyzdį). • Logistinė regresija (scenarijų

	• Artimiausių kaimynų metodas.		analizė) • Sprendimų medžiai (Klasifikavimas). • Klasterizavimas.
Bendrų bruožų radimas	Asociacijų taisyklės (association rules).	Nėra	• Naive Bayes'o algoritmas (Įtaka priklausomam kintamajam). • Asociacijų taisyklės (pirkinių krepšelio analizė). • Asociacijų taisyklės.
Duomenų sumažinimas	• Principinių komponentų analizė. • K – means klasterizavimas. • Hierarchiniai klasteriai (segmentavimas).	Nėra	Nėra

Kaip matome XLMiner bei Microsoft SQL Server 2008 Data Mining Add-Ins įrankiai turi labai panašias funkcijas, tuo tarpu TreePlan yra skirtas kiek kitokioms verslo užduotims spręsti. Tačiau visų jų aplinkos bei specifika šiek tiek skiriasi. Tolesnėje lentelėje (žr. 2 lentelė) apžvelgta būtent ši sritis.

2 lentelė. Duomenų gavybos įrankių palyginimas

	XLMiner	**TreePlan**	**Microsoft SQL Server 2008 Data Mining Add-Ins**
Grafinis vaizdavimas	Duomenis atvaizduoja histogramomis, matricomis, sklaidos diagramomis.	Duomenis leidžia vaizduoti sprendimų medžiais	Duomenis atvaizduoja histogramomis bei grafiniais sprendimų medžiais.
Kitos galimybės			Leidžia testuoti ir patvirtinti sukurtus modelius.
Programinės įrangos reikalavimai	Operacinė sistema: Microsoft Windows NT 4.0/ 2000/ XP ar vėlesnė; Microsoft Excel 2000/2003/2007 (neveikia su Microsoft Excel 97).	Operacinė sistema: Microsoft Windows NT 4.0 / 2000/ XP ar vėlesnė; Microsoft Excel 1997/2000/2003/2007. Taip pat veikia su Macinosh.	Operacinė sistema: Windows Server 2003/2008; Windows XP/Vista/7. Reikalinga prieiga prie SQL Server 2008 Analysis Services: Enterprise, Standard.
Naudojimo patogumas	• Tinklapyje pateikiama išsami naudojimo instrukcija su pavyzdžiais. • Funkcijos turi vedlius, tačiau juose daug nustatymų, kas sudėtinga nepatyrusiam vartotojui. • Pateikiami tarpiniai skaičiavimai.	• Pateikiama išsami naudojimo instrukcija su pavyzdžiais. • Duomenis į sprendimų medį reikia suvedinėti ranka. • Netinka apdoroti dideliems duomenų srautams.	• Žinyne pateikiama informacija apie funkcijų panaudojimo galimybes, rezultatų paaiškinimas. • Funkcijos turi labai supaprastintus vedlius. • Pateikiamas tik galutinis rezultatas.

Microsoft SQL Server 2008 Data Mining Add-Ins patiks pradedantiems vartotojams dėl savo vedlių, kurie veikia gan intuityviai, vartotojui reikia pasirinkti tik pagrindinius nustatymus, be to dauguma rezultatų pateikiami vienokia ar kitokia grafine forma. Vaizdinis rezultatų pateikimas leidžia daug greičiau susigaudyti bei daryti išvadas. TreePlan spendimų medžio pavidalu vaizduoja galimus sprendimus bei leidžia pasirinkti optimaliausius. Tam naudojami atskirų sprendimo medžio šakų duodamos naudos skaičiavimai. Jis nors ir neturi vedlių kaip Microsoft SQL Server 2008 Data Mining Add-Ins, tačiau yra gan paprastas ir lengvai naudojamas. Tuo tarpu XLMiner labiau tiktų daugiau žinių turintiems ar norintiems įgyti asmenims. Programos pateikiamuose modeliuose galima rasti daug daugiau nustatymų, naudojami vedliai ne tokie intuityvūs, tačiau leidžia daug įvairiau analizuoti informaciją pasirinktu metodu. Rezultatai pateikiami ne tiek vaizdžiai grafiškai kiek lentelėse, bet yra daug išsamesni. Pateikiamas ne tik galutinis rezultatas bet ir tarpiniai, kurie leidžia peržiūrėti įvairius koeficientus, taisykles ir kintamuosius sukurtus modeliavimo eigoje.

Be visų šitų savybių įmonėms reiktų įvertinti ir produkto įsigijimo kaštus. Microsoft SQL Server 2008 Data Mining Add-Ins yra nemokamas, tačiau reikia turėti įsidiegus ne tik Microsoft Excel 2007, bet ir SQL Server 2005 ar

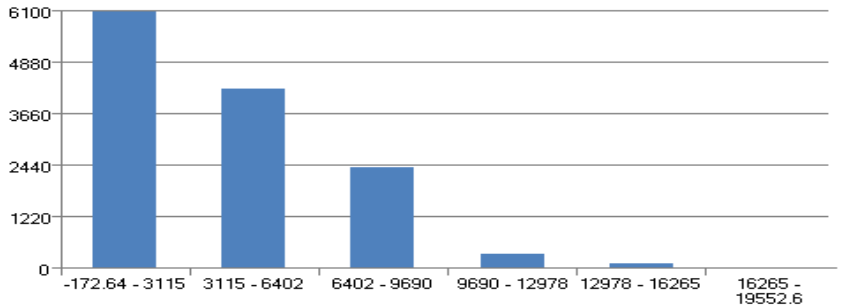
2008 Analysis Services (enterprise/standard). Norint įsidiegti XLMiner užtenka turėti Microsoft Excel 2000/2003/2007 skaičiuoklę, tačiau reikia mokėti už XLMiner (XLMiner 3.0 professional edition licencija 999$) programos licenziją. TreePlan veikia su Microsoft Excel 1997/2000/2003/2007 skaičiuoklės versijom, licencija kainuoja 49$.

5. Eksperimentas

Įmonė, kurios duomenys naudojami šiame skyriuje yra nedidelė tarptautinių pervežimų įmonė, kuri ieško vežėjų įvairių rūšių produkcijai gabenti iš vienos valstybės į kitą. Pagrindinė įmonės sprendžiama problema yra veiklos optimizavimas. Siekiama atsisakyti užsakymų, kurie nuostolingi įmonei, tuo pačiu stiprinant santykius su geriausias pervežimo sąlygas siūlančiomis įmonėmis. Tai leistų pasiekti pagrindinio įmonės tikslo – tapti kompanija, siūlančia pigiausius krovinių pervežimus Europos ribose. Tapus nuolatiniu įmonės, pervežančios krovinius klientu, galima dar labiau sumažinti pervežimų kainas.

Eksperimentui naudojami duomenys: užsakovas, vežėjas, užsakymo data, pristatymo data, šalis į kurią vežamas krovinys, šalis iš kurios vežamas krovinys, krovinio rūšis, užsakymo kaina.

Pradžioje sudarytas grafikas rodantis užsakymo kainų pasiskirstymą. Kaip matome dažniausiai pasitaikantys užsakymai yra tie, kurių vertė iki 9690.00 Lt. Didesnės vertės užsakymai yra retesni, todėl galima daryti pirminę išvadą, jog juos vykdyti neapsimoka.



Paveikslas 1. Histograma vaizduojanti užsakymų kainų pasiskirstymą

Reikia išsiaiškinti kokie pagrindiniai veiksniai lemia tokia aukštą užsakymo kainą. Naudojant Microsoft SQL Server 2008 Data Mining Add-Ins įrankį duomenų gavybai klasifikavimas atliktas remiantis neuroninių tinklų algoritmu.

Atributas	Vertė	Užsakymo kaina >= 9715,66
Šalis iš	Ispanija	
Vežėjas	FTU Arkadiusz Nowakowski	
Vežėjas	UAB "BARELA"	
Vežėjas	UAB "Artemsta"	
Vežėjas	UAB "Rima Gražienė ir partneriai "	
Šalis iš	Rusija, Maskva	
Krovinio rūšis	Medikamentai	
Šalis į	Rusija, St.Peterburgas	
Vežėjas	UAB "Septyni kalnai"	
Šalis iš	Turkija	
Šalis į	Liuksemburgas	
Vežėjas	UAB "Rasvita"	
Šalis iš	Airija	
Vežėjas	UAB "Nėgė"	
Šalis iš	Rumunija	

Paveikslas 2. Įtaką aukštai užsakymo kainai darantys veiksniai

Kaip matome iš rezultatų labiausiai tokią aukštą užsakymo kainą lemia šalis iš kurios gabenamas krovinys: Ispanija, Rusija (Maskva, Sankt Peterburgas), Turkija. Taip pat vežėjai gabenantys krovinius: FTU Arkadiusz Nowakowski, UAB „BARELA“, UAB „Artemsta“ ir kiti.

Analogiškos analizės rezultatų su XLMiner neįmanoma gauti dėl to, jog užsakymo kaina yra tolydusis kintamasis. Šį įrankį galima panaudoti tolesnei duomenų analizei tyrinėjant kiekvieno elemento pasitaikymo tikimybes. Atlikus klasifikavimą, naudojant neuroninių tinklų algoritmą, pagal šalį iš kurios gabenami kroviniai tolesnei analizei galima panaudoti klasių (kurių yra tiek kiek yra šalių) tikimybes.

Klasė	Klasės tikimybė
Airija	0.005
Belgija	0.05
Bulgarija	0.005
Čekija	0.015
Danija	0.03
Didžioji Britanija	0.015
Ispanija	0.05
Italija	0.05
Lenkija	0.06

Lietuva	0.315
Nyderlandai	0.16
Rusija (Maskva)	0.005
Slovakija	0.005
Švedija	0.01
Ukraina	0.005
Vokietija	0.225

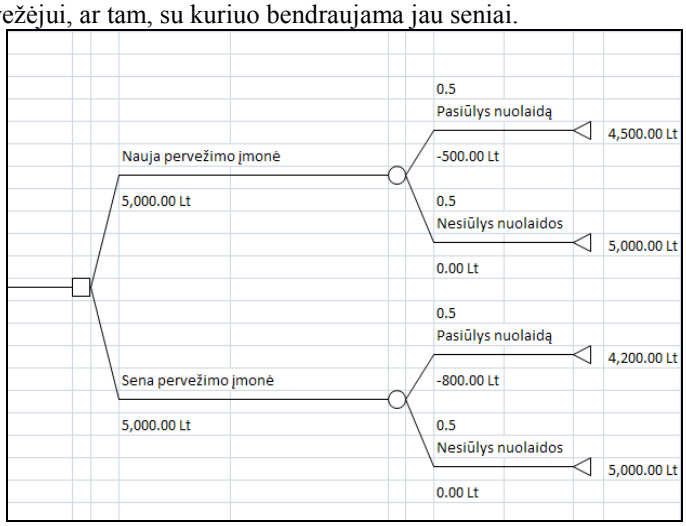
Paveikslas 3. Šalių pasitaikymo tikimybės naudojant neuroninius tinklus

Išnagrinėjus šalių pasitaikymo tikimybes galima nuspręsti, ar verta atsisakyti prekes gabenti į šią šalį. Jei tikimybė, jog maža, galbūt ir verta atsisakyti. Jei didelė reikia žiūrėti ar atsisakę nepatirsime didesnių nuostolių.

Nagrinėjamu atvejų galime apžvelgti šalis, kurias kaip didžiausią kainą lemiančius veiksnius atrinkome tyrimo pradžioje. Tai būtų Ispanija, Rusija (Maskva), Italija, Lietuva. Kaip matome tikimybė kad krovinys gabenamas iš Ispanijos, Rusijos (Maskva), Italijos yra tik 0.05 (5%), kai tuo tarpu tikimybė jog krovinys bus gabenamas iš Lietuvos yra net 0.315 (31.5%).

Taigi, įmonei vertėtų pamąstyti apie atsisakymą prekes gabenti iš Ispanijos, Rusijos, Italijos bei daugiau investuoti į krovinių gabenamų iš Lietuvos kainų mažinimą.

TreePlan įrankis tokiai analizei netinka. Tačiau juo galima apskaičiuoti kas labiau apsimoka – tą patį pasiūlymą pateikti naujam vežėjui, ar tam, su kuriuo bendraujama jau seniai.



Paveikslas 4. Skirtingų vežėjų pasiūlymų palyginimas

Kaip matome iš grafiko geriau rinktis vežėją, su kuriuo bendraujama seniai. Nes pasiūlymas gali būti arba toks pats kaip ir naujosios pervežimo įmonės arba palankesnis.

6. Išvados

Kaip matome duomenų gavyba nėra skirta tik didelėms įmonėms. Naudojant atitinkamus įrankius ją galima taikyti ir nedidelėse, tačiau apie ateities perspektyvas galvojančiose įmonėse. Joms puikiai tinka įrankiai, kurie veikia ne kaip atskiros programos, o kaip MS Excel skaičiuoklės papildomos funkcijos.

Išnagrinėti trys duomenų gavybos įrankiai – XLMiner, TreePlan, Microsoft SQL Server 2008 Data Mining Add-Ins. Šie įrankiai puikiai tinka naudoti mažose įmonėse, nes turi pakankamai daug įvairių funkcijų. Įvairios įrankių funkcijos gali padėti analizuoti duomenis, įvertinus gautus rezultatus mažosios įmonės gali optimizuoti veiklos procesus, efektyviai panaudoti sukauptus duomenis.

Literatūra

[1] **Sekliuckis V., Gudas S., Garšva G..** (2006) Informacijos sistemos ir duomenų bazės: informacijos sistemų ir reliacinių duomenų bazių kūrimo pagrindai: vadovėlis. Kaunas: Technologija. 349 ps. ISBN 9955094869.

[2] **The CRISP-DM consortium** (2003) Process Model [interaktyvus] The CRISP-DM project. [žiūrėta 2010m. sausio 13d.]. Prieiga per internetą: <http://www.crisp-dm.org/Process/index.htm>

[3] **Usama F., Piatetsky-Shapiro G.** (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*.

[4] **Solomon M**.(2005). Ten High-End Technologies You Can Afford [interaktyvus] *2010 CDW Corporation* [žiūrėta 2010m. sausio 13d.]. Prieiga per internetą: < http://www.biztechmagazine.com/article.asp ?item_id=87 >

[5] **Ritholz B.** (2009) Intro to Data Mining for Small Businesses [interaktyvus] *2010 American Express Company* [žiūrėta 2010m. sausio 11d.]. Prieiga per internetą: <http://www.openforum.com/idea-hub/topics/money/article/intro-to-data-mining-for-small-businesses-barry-ritholtz >

[6] **Ciesiūnas A., Judžentis E.** (2007) Duomenų gavybos metodai. *10-osios Lietuvos jaunųjų mokslininkų konferencijos „Mokslas – Lietuvos ateitis“ medžiaga.*