VILNIAUS UNIVERSITETAS
KAUNO HUMANITARINIS FAKULTETAS

INFORMATIKOS KATEDRA

Verslo informatikos studijų programa
Kodas 62109P101

AUKSĖ STRAVINSKIENĖ

MAGISTRO BAIGIAMASIS DARBAS

DUOMENŲ GAVYBOS METODIKA

Kaunas, 2008

VILNIAUS UNIVERSITETAS
KAUNO HUMANITARINIS FAKULTETAS

INFORMATIKOS KATEDRA

AUKSĖ STRAVINSKIENĖ

MAGISTRO BAIGIAMASIS DARBAS

DUOMENŲ GAVYBOS METODIKA

Leidžiama ginti _____

Magistrantas_____
(parašas)

Darbo vadovas _____
(parašas)

prof. Saulius Gudas
(darbo vadovo mokslo laipsnis, mokslo pedagoginis vardas, vardas ir pavardė)

Darbo įteikimo data _____

Registracijos Nr. _____

Kaunas, 2008

TURINYS

LENTELIŲ SĄRAŠAS

PAVEIKSLŲ SĄRAŠAS

SANTRUMPŲ SĄRAŠAS

BI – Intelektinės sistemos verslui (angl. Business Intelligence);

CRISP-DM – ESPRIT fondo įkurtas konsorciumas, kurio paskirtis plėtotoji duomenų gavybos idėją, programinę įrangą ir t.t;

CROWS – duomenų gavybos procesų modelis;

DB - duomenų bazė;

DBVS – duomenų bazių valdymo sistemos;

DG- duomenų gavyba;

DS – duomenų saugykla;

OLAP - Online Analytical Processing;

MS – Microsoft;

SPSS - Statistical Package for the Social Sciences;

SQL – Structured Query Language.

SANTRAUKA

STRAVINSKIENĖ, Auksė. (2008) Methodology of data mining. MA Graduation Paper. Kaunas: Vilnius University, Kaunas Faculty of Humanities, Department of Informatics. 84 pages.

SUMMARY

Systems of data mining are defined as knowledge, which using various statistic, mathematical, models for patterns recognition methods, is generalized and properly processed. Applying technologies of data mining it is being searched for new subordinations, meaningful consistent patterns between an organization's data, tendencies among business processes, reflective model-based data, new and unfamiliar information about business processes.

Analysis of data mining systems is relevant not only for organizations. In order to adjust to alterations of market, it is important to provide information to students on time and properly, during sessions students should get acquainted with current information, studies programs should be approach towards alterations of market.

The main aim of the work is to explore and compare systems of data mining with reference to received results, to make methodology of data mining and to project laboratory works which could help to reveal the importance of data mining in a process of learning and in a modern enterprise.

These goals have been set in order to achieve the aim of master's work:

- Survey of accomplished research, analysis of literature;
- Analysis of data mining technologies and algorithms;
- Analysis of software;
- Projection of process model of data mining;
- Projection of laboratory works.

Technologies of data mining, working principles of algorithms, possibilities of software equipments, creating models of data mining, are demonstrated in prepared laboratory works. Models of data mining is created using means of MS SQL server 2005 programmable packet Analysis Services, applying representative Microsoft AdventureWorks DW data base.

ĮVADAS

Šiuolaikinėje verslo aplinkoje duomenų bazėse, duomenų saugyklose informacijos apimtys vis didėja, saugomi neriboti duomenų kiekiai, kurie reikalauja greito, intelektualaus duomenų apdorojimo realiame laike. Pagrindiniais verslo įmonių uždaviniais tampa savos organizacijos stebėjimas, verslo rezultatų, finansinės būklės ir duomenų analizavimas. Norint išlikti konkurencingais šiandieninėje rinkoje, reikia ne tik gerai pažinti savo organizaciją, bet ir įvertinti konkurentus, stebėti bendradarbiavimą su klientais bei atlikti rinkos analizę. Verslo įmonių sėkmė tampa tiesiogiai proporcinga organizacijos sugebėjimui tinkamai surinkti ir analizuoti reikalingus duomenis. Šiems uždaviniams spręsti gali būti naudojamos duomenų gavybos, duomenų sandėlių ir mugių, intelektinių sistemų verslui (angl. *Business Intelligence*) technologijos.

Duomenų gavybos sistemos apibūdinamos kaip žinios, kurios pasinaudojant įvairiais statistiniais, matematiniais, modelių atpažinimo metodais yra apibendrinamos ir tinkamai apdorojamos. Pasinaudojant duomenų gavybos technologijomis ieškoma naujų priklausomybių, prasmingų dėsningumų tarp organizacijos duomenų, tendencijų tarp verslo procesų, duomenų atvaizdavimo modelių, naujos ir nežinomos informacijos apie verslo procesus.

Duomenų gavybos sistemų analizavimas yra svarbus ne tik organizacijoms. Norint prisitaikyti prie rinkos pokyčių, svarbu laiku ir tinkamai pateikti informaciją studentams, paskaitų metu supažindinti su aktualiausia informacija, priartinti studijų programas prie rinkos pokyčių. Studijų metu studentai turi išmokti valdyti duomenų gavybos įrankius, suprasti duomenų gavybos procesą bei mokėti praktiškai pritaikyti įgytas žinias. Įvertinus tokios metodinės informacijos trūkumą buvo nuspręsta pasirinkti šią temą. Sukurta duomenų gavybos metodika bus panaudota magistratūros Verslo informacijos sistemų studijų programoje.

Šis **darbas aktualus** ne tik universitetui, bet ir studentams, nes jame analizuojami pagrindiniai duomenų gavybos proceso etapai bei suformuoti laboratoriniai darbai. Šiuose darbuose pateikiama teorinė medžiaga, darbo tikslas, darbo užduotis, kuri turi būti realizuojama duomenų gavybos priemone.

Atsižvelgiant į paminėtus faktus, svarbu išsiaiškinti duomenų gavybos proceso metodiką, sistemas bei jų naudingumą. Todėl darbe tiriamos duomenų gavybos technologijos, algoritmų panaudojimo savybės, pasinaudojant pavyzdinėmis duomenų bazėmis (Microsoft AdventureWorks DW) sudaryti duomenų gavybos modeliai. Tyrime taikoma mokslinės literatūros, statistinių duomenų bei lyginamoji analizė, sintezė ir apibendrinimas, naudojami tyrimo ir praktinio atlikimo metodai.

Darbo tyrimo **objektas** – duomenų gavybos procesas. Pagrindinis **tikslas** yra ištirti ir palyginti duomenų gavybos sistemas, remiantis gautais rezultatais, sudaryti duomenų gavybos

metodiką ir suprojektuoti laboratorinius darbus, kurie padėtų atskleisti duomenų gavybos reikšmę mokymosi procese bei šiuolaikinėje įmonėje. Šiam tikslui pasiekti iškelti duomenų gavybos technologijų ir algoritmų analizės, programinės įrangos analizės bei duomenų gavybos modelių kūrimo uždaviniai.

Magistrinio darbo tikslui pasiekti iškelti tokie **uždaviniai**:

- atliktų tyrimų apžvalga, literatūros analizė;
- duomenų gavybos technologijų ir algoritmų analizė;
- programinės įrangos analizė;
- duomenų gavybos proceso modelio projektavimas;
- laboratorinių darbų projektavimas.

Šie uždaviniai įgyvendinti remiantis mokslinės literatūros, palyginimo analizės, sintezės ir apibendrinimo, tyrimo ir praktinio atlikimo metodais.

Duomenų sandėlių sistemų analizei atlikti taikoma:

- mokslinės literatūros apžvalga;
- komercinių produktų tyrimas;
- eksperimentinių modelių sudarymas;
- sudarytų modelių testavimas ir analizė.

Sukurtuose laboratoriniuose darbuose pademonstruota duomenų gavybos technologijos, algoritmų veikimo principai, programinės įrangos įrankių galimybės, kuriant duomenų gavybos modelius.

Pirmoje dalyje išanalizuotos duomenų gavybos technologijos, jų paskirtis, funkcijos ir galimybės. Apžvelgti duomenų gavybos algoritmai ir jų panaudojimo savybės. Taip pat atlikta duomenų gavybos produktų analizė.

Antroje dalyje pateikiama duomenų gavybos proceso metodika bei pasiūlymai jos patobulinimui.

Trečioje dalyje pasinaudojant pavyzdine Microsoft AdventureWorks DW duomenų baze MS SQL server 2005 programinio paketo Analysis Services priemonėmis sudaryti eksperimentiniai duomenų gavybos proceso laboratoriniai darbai.

Darbo struktūra ir apimtis: darbą sudaro trys pagrindinės dalys (analizė, duomenų gavybos proceso metodika, eksperimentinis skyrius). Darbo apimtis yra 84 lapai, pateikiama 14 lentelių ir 80 paveikslėlių.

1. ANALIZĖS DALIS

Analizės dalyje pateikiama apibendrinta mokslinės literatūros apžvalga, informacija apie duomenų gavybos sistemas, susisteminta programinės įrangos analizė. Trumpai apžvelgiamos duomenų gavybos atsiradimo prielaidos, šių sistemų poreikis šiandieninėje verslo aplinkoje. Pateikiami pagrindiniai skirtumai tarp duomenų gavybos ir kitų duomenų analizės įrankių: OLAP, duomenų sandėliavimo sistemų (angl. *Data Warehousing*) ir statistikos įrankių.

1.1. Duomenų gavybos samprata

Šiuolaikinėje verslo aplinkoje vis didesnę reikšmę įgyja nematerialieji ištekliai (vidiniai organizacijos duomenys, sukauptos darbuotojų žinios ir t.t.), kuriuos labai svarbu tinkamai panaudoti norint išlikti konkurencingoje aplinkoje. Organizacijos viduje sukaupta informacija svarbi moksliniu, ekonominiu bei socialiniu požiūriu. Ekonominiu požiūriu, žinios tampa svarbiu konkurencingumo rodikliu ir, tinkamai panaudotos, gali padidinti organizacijos pelningumo rodiklius. Sukauptos žinios yra neatsiejamas organizacijos duomenų kiekio didėjimo procesas. Didėjant duomenų saugyklose saugomų duomenų kiekiams, daugėja juos apdorojančių procesų skaičius, o tam reikalingas ne tik kokybiškas duomenų apdorojimas, bet ir kokybiškesnis apdorotų duomenų panaudojimas.

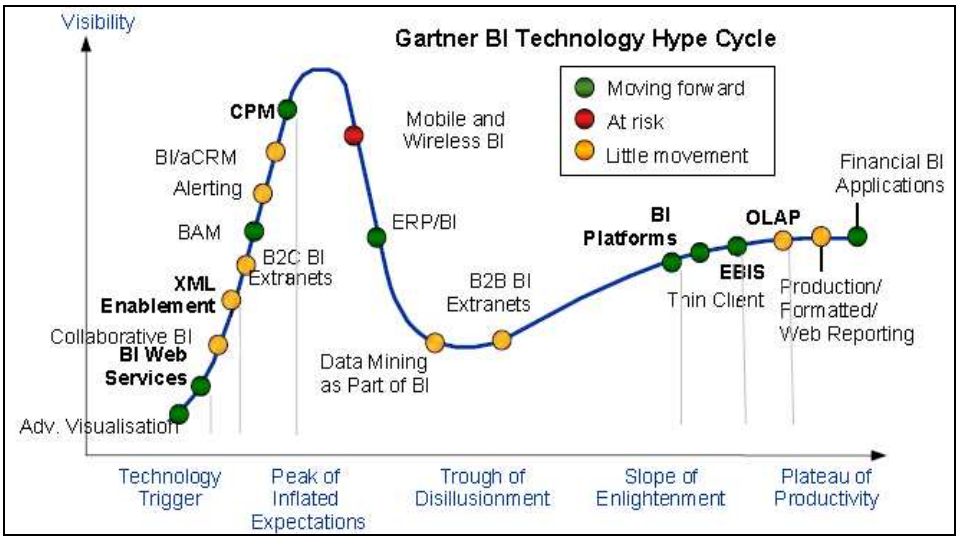
Taip atsiranda sąvoka intelektinės sistemos verslui (angl. *Business Intelligence*, toliau BI). Šių sistemų esmė - surinkti iš įvairių šaltinių duomenis, juos susisteminti ir galutiniam vartotojui pateikti naudingą ir kokybišką informaciją. Ši informacija dažniausiai naudojama svarbių verslo sprendimų priėmimui, verslo analizei ar naujų verslo pokyčių nustatymui. BI technologijų pagrindinis tikslas – iš organizacijoje sukauptų duomenų išgauti maksimalią naudą, reikiamu momentu aprūpinant specialistus reikalinga informacija. BI atsiradimą sąlygojo poreikis padidinti organizacijos pajamas, tuo pačiu metu mažinant kaštus ir didinant konkurencingumą. Taip pat svarbu efektyviai modeliuoti organizaciją supančią aplinką, suteikiant darbuotojams galimybę greitai ir ekonomiškai pasiekti reikalingą informaciją [1].

Intelektines sistemas verslui sudaro:

- OLAP;
- duomenų gavyba;
- ataskaitų rengimo programinė įranga;
- elektroninės lentelės (Spreedsheet);
- kitos sistemos.

Informacinių technologijų mokslininkų grupė Gartner, kiekvienais metais atlieka intelektinių sistemų verslui rinkos tyrimus, siekdama išsiaiškinti kiekvienos technologijos pažangą

ir užimamą padėtį informacinių technologijų rinkoje [2]. 1 paveiksle Pateikiama duomenų gavybos pozicija 2003 metų informacinių technologijų rinkoje.



Šaltinis: Microsoft Gartner. (2003) Business Intelligence in Europe: Relevant and Valuable Information for better desicions in a Real-time Enterprise context , p. 14.

1 pav. Duomenų gavybos pozicija BI rinkoje

Duomenų gavyba yra prasmingų dėsningumų, modelių ir tendencijų radimo procesas dideliuose informacijos kiekiuose, naudojant modelių atpažinimo, statistinius bei matematinius metodus [1]. Šį procesą galima apibrėžti ir kaip analitinį procesą, kuris analizuoja didelius duomenų kiekius, ieškodamas sisteminių ryšių tarp kintamųjų, naujų priklausomybių. Įvertinus gautus rezultatus suformuojamas modelis. Pagrindinis duomenų gavybos principas, kad ieškoma dar neatrastų dėsningumų, sąryšių tarp kintamųjų. Tai gali būti naujų ir naudingų modelių, šablonų, struktūrų ar net taisyklių paieška, panaudojant automatinius ir pusiau automatinius įrankius. Duomenų gavybos pagalba atrandama tokia informacija, kuri nėra lengvai pastebima ar apie jos egzistavimą net neįtariama.

Duomenų gavybos tikslas - didelėse duomenų saugyklose (kaip racionalios duomenų bazės, duomenų sandėliai, duomenų centrai ar kitos saugyklos), esant dideliam duomenų kiekiui, atrasti naują požiūrį ir priklausomybes tarp kintamųjų, sudarant prognozavimo, taisyklių, atvaizdavimo ar kitus modelius. Šiam tikslui pasiekti naudojami įvairūs duomenų gavybos metodai, algoritmai ir taikomosios programos [3].

Duomenų gavyba nėra paprastas procesas. Jis reikalauja kruopštaus ir tikslingo pasiruošimo. Sėkmingai duomenų gavybai svarbu suprasti verslo aplinką, įvertinti jos poreikius, išskirti probleminę sritį, kuriai reikalinga ši technologija. Ne mažiau svarbus kriterijus sėkmingai duomenų gavybai – kokybiški ir tinkamai paruošti duomenys. Šis kriterijus dažniausiai sunkiausiai įgyvendinamas, nes organizacijose duomenys yra saugomi įvairiose duomenų saugyklose, dažnai trūksta tam tikrų reikšmių, jos neteisingos ar klaidingos [4].

Šiandieninėje verslo aplinkoje yra labai svarbus faktinių duomenų pavertimas naudinga ir kokybiška informacija. Tam įgyvendinti panaudojamos duomenų išgavimo technikos, algoritmai ir speciali programinė įranga. Algoritmai, kurie yra pagrindas duomenų gavybos procese, gali būti kilę iš statistikos, biologinių struktūrų, evoliucijos teorijos, matematinių teorijų ir t.t. Duomenų gavybos priemonių pasirinkimą lemia daug faktorių, tai gali būti duomenų struktūra, kilmė, problema, kurią bandoma išspręsti. Didelę įtaką priemonių pasirinkimui gali turėti ir veiklos sritis, kurioje reikalinga išsami duomenų analizė. Šių priemonių pagalba gali atrasti naudingas žinias, reikalingas organizacijos veiklai valdyti, rinkos ar pardavimų analizei, sprendimų priėmimui, klientų elgesio modelių kūrimui, veiklos efektyvumui didinti bei prognozavimui. Duomenų gavybos sistema pagrįsta taisyklėmis, kai duomenų gavybos priemonėmis iš organizacijos duomenų bandoma išgauti eksperto žinias ir jas išreikšti taisyklėmis.

Pažangios organizacijos, visame pasaulyje, jau naudoja duomenų gavybą ir analizę. Pagrindiniai organizacijų tikslai - surasti ir patraukti vertingesnius klientus, performuoti produktų pasiūlą, padidinti pardavimus ir kiek įmanoma sumažinti nuostolius dėl klaidų ar sukčiavimų. Duomenų analizė taikoma ne tik versle. Pirmieji duomenų gavybos modeliai buvo pritaikyti statistikoje. Vėliau panaudoti analizuojant medicininių operacijų ir tyrimų efektyvumą, genetinės informacijos apdorojimui, chemijos inžinerijoje, gaminių kokybės kontroliavimui bei finansinei analizei atlikti [4].

1.2. Duomenų gavybos istorija

Duomenų gavybos užuomazgos pastebėtos prieš 40 metų ir turi ilgą vystymosi istoriją. Duomenų analizė nebuvo apibrėžiama kaip duomenų gavyba. 1960 m. buvo pradėti naudoti duomenų gavybos modeliai, tačiau jie buvo pritaikomi statistikoje, o duomenų gavyba buvo vadinama statistine analize. Statistika yra daugelio technologijų pagrindas, iš kurios sukurta duomenų gavyba bei analizė. Duomenų ir duomenų ryšių tyrinėjimams yra naudojami klasikinės statistikos metodai: regresijos analizė, standartinis pasiskirstymas, standartinis nuokrypis, klasterinė analizė ir patikimumo intervalai. Statistinės programinės įrangos pradininkai buvo šios organizacijos: SAS (angl. *Statistical analysis software*) ir SPSS (angl. *Statistical Package for the Social Sciences*). Šie gamintojai iki šiol užima lyderių pozicijas duomenų gavybos rinkoje.

Nuo 1980 metų sparčiai pažengus informacinėms technologijoms, naudojant galingus kompiuterius, duomenų gavyba buvo siejama su dirbtiniu intelektu (angl. *Artificial intelligence*). Didelis dėmesys buvo skirimas į euristinius metodus, kai stengtasi problemų sprendimui panaudoti duomenų apdorojimo modelius, panašius į žmogaus mąstymą. Tradicinės duomenų analizės technikos buvo papildytos neaiškių logikų (angl. *Fuzzy logic*), neuroninių tinklų (angl. *Neutral networks*) metodais [5].

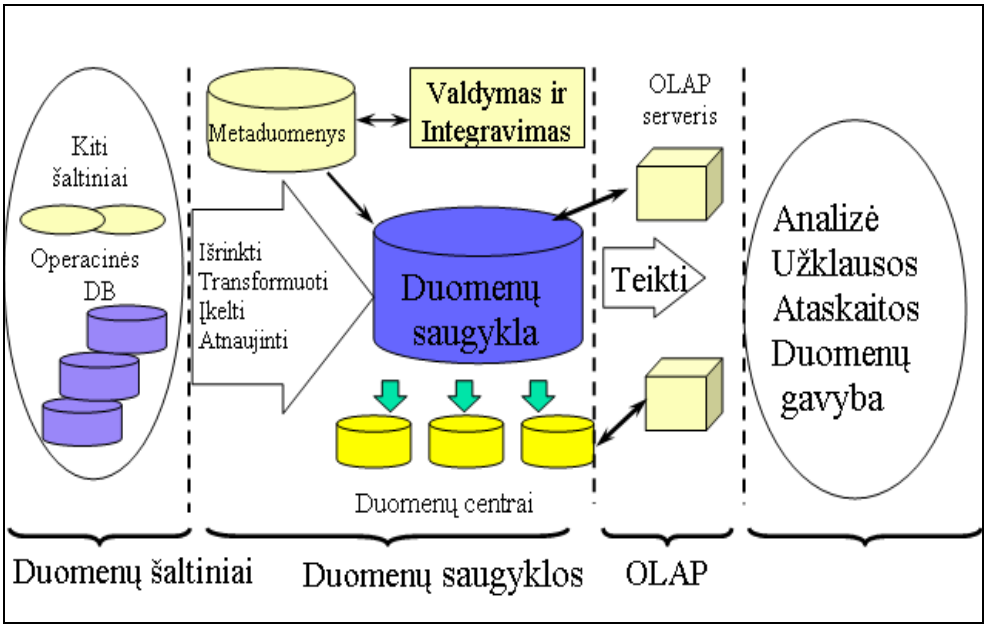
1990 metais pradėta kalbėti apie duomenų gavybos sąvoką. Šio termino atsiradimą sąlygojo didžiulių duomenų bazių atsiradimas, kurios talpina terabaitinius duomenų kiekius. O 1994 m. pasirodė pirmasis duomenų gavybos programinis produktas, kurį pristatė SSPS.

1.3. Duomenų gavybos veikimo principas

Duomenų gavyba padeda verslo analitikams surasti dėsningumus ir ryšius duomenyse. Ji nenurodo šių dėsningumų reikšmės organizacijai. Duomenų gavybos ir analizės metu atrasti prognozuojami ryšiai nebūtinai turi būti suprantami kaip vartotojų elgsenos motyvas, tai gali būti tam tikros taisyklės, dėsningumai. Nepakanka naudotis algoritmų suformuotais rezultatais. Norint priimti gerus sprendimus, svarbu ne tik tinkamai įvertinti duomenų gavybos rezultatus, bet ir patikrinti juos logiškai, įvertinant situaciją, patikrinant sukaupta patirtimi.

Duomenų gavybos procese svarbu valdyti visus duomenų gavybos proceso etapus. Tinkamai įvertinti verslo problemos sudėtingumą, duomenų gavybos tikslus, gerai suprasti organizacijos duomenų sudėtingumą bei tinkamai pasirinkti algoritmus. Pasirinkus duomenų gavybos įrankį ir suprantant, kaip veikia duomenų gavybos modelis, galima pasiekti optimalius rezultatus. Laikantis šių principų, galima greitai sudaryti tikslų duomenų gavybos modelį, kuris pavaizduos laukiamus rezultatus [7].

Duomenų gavyba yra vienas iš duomenų analizės tipų. Duomenų analizės architektūra (2 pav.) pavaizduoja visą duomenų srauto kelią iki galutinio rezultato.



Šaltinis: sukurta autoriaus pagal K. Thearling (2003).

2 pav. Duomenų analizės architektūra

Duomenų gavyba nepakeičia patyrusių verslo analitikų ar vadybininkų, bet duoda jiems galingą naują įrankį, kuris pagerina jų atliekamą darbą. Suderinant du svarbus aspektus (duomenų

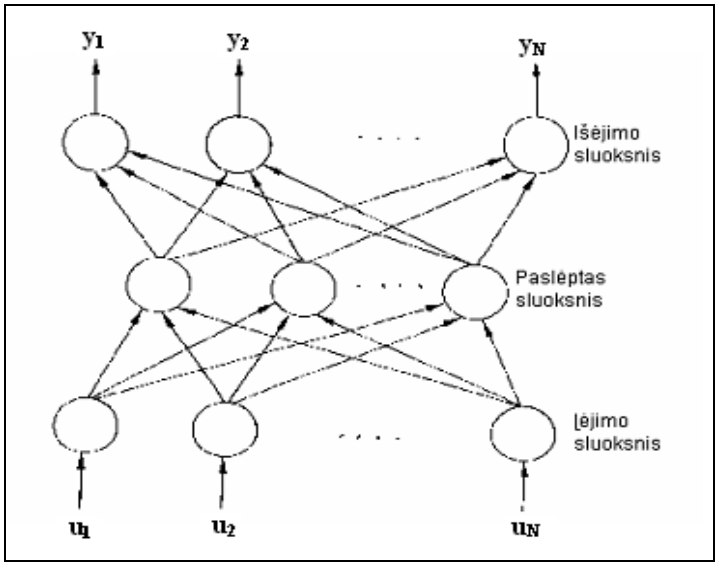
gavybos sistemą ir verslo analizės specialistą) įmonė gali pasiekti puikių rezultatų: padidinti konkurencingumą, pagreitinti sprendimų priėmimo laiką bei atrasti naujus klientų elgesio motyvus.

1.4. Duomenų gavybos būdai

1.4.1. Neuroniniai tinklai

Neuroniniai tinklai kaip ir Sprendimų medžio bei Naïve Bayes modeliai yra naudojami siekiant atrasti visus galimus duomenų ryšius, duomenų tyrinėjimui, klasifikacijai ir prognozavimui. Šis metodas yra vienas sudėtingiausių, jis išsamiai ieško paslėptos informacijos, todėl duomenų gavybos procesas užtrunka ilgiausiai. Neuroniniai tinklai - netiesinis modelis, kuris naudojamas spręsti klasifikavimo ir prognozavimo uždavinius.

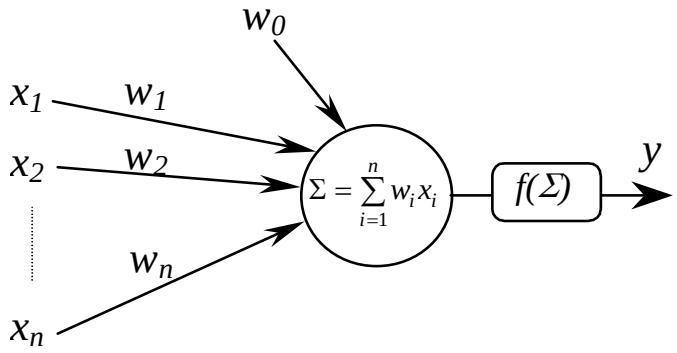
Neuroninių tinklų struktūra analogiška biologinėms neuroninėms sistemoms. Klasikinis neuroninis tinklas sudarytas iš kelių sluoksnių: įvesties, išvesties mazgų bei vieno ar daugiau paslėptų sluoksnių (3 pav.). Įėjimai į mazgus paslėptame sluoksnyje ir į išėjimo mazgus turi savo svorinius koeficientus. Įėjimo mazgai – tai elementus aprašanti informacija. Išvesties mazgai — gaunamas duomenų gavybos rezultatas (klasifikavimo uždaviniuose - priklausymas klasėms, o prognozavimo uždaviniuose – numatomas rezultatas) [8].



Šaltinis: prof. R. Simučio paskaitų medžiaga, 2008 m.

3 pav. Neuroninio tinklo sluoksniai

Kiekvienas neuroninio tinklo sluoksnio mazgas yra susijęs su šalia esančiais sluoksnio mazgais. Atliekant istorinių duomenų modeliavimą, pasirenkama tam tikra neuroninio tinklo struktūra ir tinklo apmokymo metu nustatomi tinklo svoriai (1 formulė). Tinklo struktūra gali daug kartų kisti, o tai priklauso nuo modeliavimo rezultatų.



(1)

Čia $x_1 \ldots x_n$ – įėjimo mazgai;

$w_1 \ldots w_n$ – tinklo svoriai;

w_0 – slenksčio reikšmė;

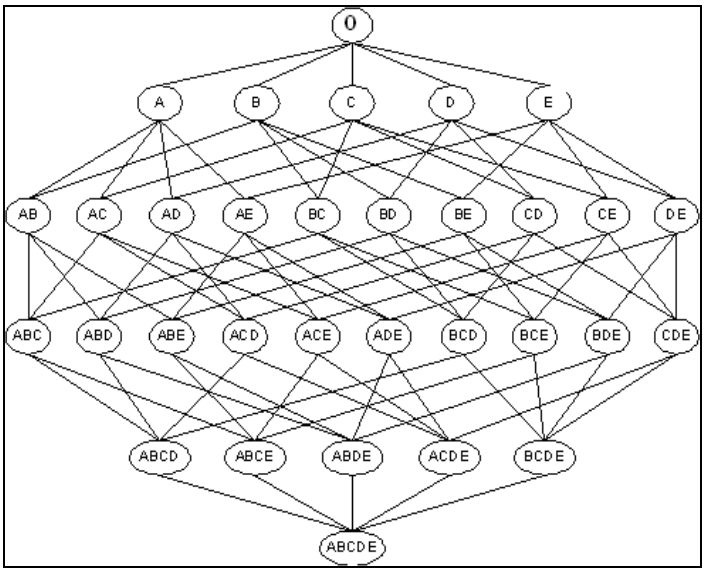
$f()$ – perdavimo funkcija;

y – neurono išėjimas.

Modeliavimo procese svarbiausia tinkamai parinkti tinklo svorius – tai pagrindinis algoritmo uždavinys.

1.4.2. Asociacijų analizė

Asociacijos algoritmas naudojamas norint tarp didelių duomenų kiekių, surasti dėsningumus objektų ar reiškinių grupėse, bei suformuoti statistines taisykles. Šio algoritmo pagrindinis tikslas yra grupuoti didelės apimties duomenų rinkinių elementus, kurie tam tikru vartotojo duomenų apdorojimo metu yra dažniausiai panaudojami. Tie elementai grupuojami į elementų rinkinius ir generuojamos taisyklės, kurios naudojamos prognozavimui (4 pav.).



Šaltinis: sudaryta autoriaus.

4 pav. Asociacijų taisyklių tinklas

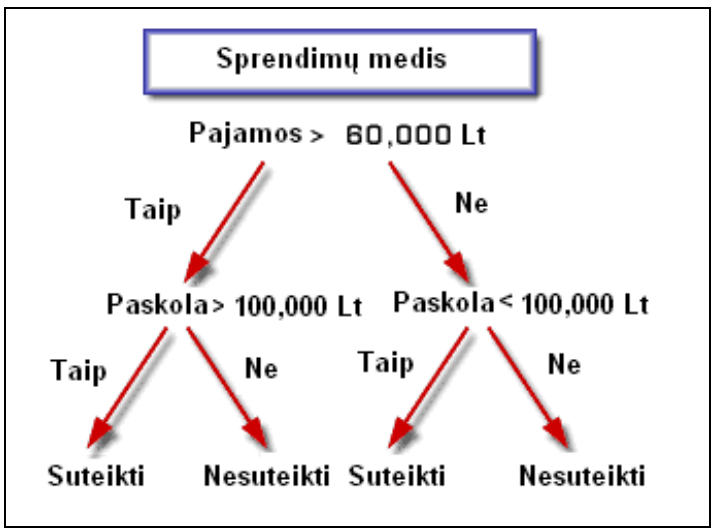
Duomenų gavyboje asociacijų analizė dažniausiai naudojama pardavimų, pirkimų ar paslaugų rinkos analizei. Dažniausiai pritaikoma norint išanalizuoti, kokias prekių grupes perka pirkėjai. Asociacijos algoritmo pagalba išsiaiškinami produktų deriniai, kuriuos dažniausiai tam tikros kategorijos vartotojai. Organizacijai asociacijos metodo rezultatas labai svarbus vertinant ir kuriant reklamos kampaniją. Taip pat analitikai gali įvertinti rezultatus, formuojant nuolaidų sistemas, nupigintų prekių grupes ir t.t.

Šiam algoritmui svarbus yra parametrų parinkimas, todėl jis labiau tinkamas didesniems ir sudėtingesniems duomenų gavybos uždaviniams spręsti. Kitu atveju, esant paprastesniem uždaviniams, galima naudoti sprendimų medžio algoritmą.

1.4.3. Sprendimų medžiai

Sprendimų medžio algoritmas yra vienas iš paprasčiausių ir dažniausiai naudojamų duomenų gavybos algoritmų. Galima teigti, kad tai pradinis duomenų gavybos taškas duomenų analizės procese. Sprendimų medis vienas iš klasifikacijos algoritmų, kuris klasifikuoja pavienius ir tolydžius požymius į duomenų gavybos modelį. Modelis formuojamas taip, kad nagrinėtų, kurie įvedamų elementų požymiai labiausiai veikia prognozuojamo požymio rezultatus. Pagrindinis Sprendimų medžio modelio tikslas – surasti visas galimas įvesties duomenų požymių kombinacijas ar būsenas, kurios turi didžiausią įtaką galutiniam rezultatui.

Šio duomenų gavybos algoritmo principas – medžio struktūra (5 pav.). Požymiai suklasifikuojami ir išdėstomi hierarchiškai. Kiekvienas svarbesnis požymis turi nuo jo priklausančius mažesnės reikšmės požymius. Šiuo sprendimų medžio principu yra sudaromos taisyklės, analizuojančios suklasifikuotų duomenų aibės elementų savybes. Vienas pagrindinių Sprendimų medžio privalumų yra tai, kad jis lengvai suprantamas įvairių sričių specialistams [9].



Šaltinis: sukurta autoriaus.

5 pav. Sprendimų medis

Sprendimų medžio pagalba gali būti sprendžiami tokie uždaviniai, kaip pirkėjų tikslinių grupių, perkančių konkrečią prekę, nustatymas ir t.t.

1.4.4. Genetiniai algoritmai

Pagrindinis genetinių algoritmų principas – tai veikimas pagal biologinės evoliucijos šabloną bei, remiantis prognozavimo ir klasifikavimo taisyklių išgavimu, surasti apytikslį užduoties sprendimą. Šie algoritmai nėra tinkami duomenų gavybos šablonų suradimui, dažniausiai jie naudojami kitų duomenų gavybos algoritmų apmokymo procese (kaip neuroniniai tinklai). Jie vadinami genetiniais algoritmais, nes ieško elementų, kurie svarbūs tolimesniam duomenų gavybos procesui ir jie ieškomi tol, kol surandamas geriausias modelis. Pasirinkus pradinį duomenų rinkinį įvertinamas jo tinkamumas, atrenkami pagal tam tikrus atrankos kriterijus nauji elementai. Atrinktieji elementai pakeičiami naudojant tam tikras kombinacijas ir sukuriamas naujas duomenų rinkinys. Vėliau viskas kartojama, atrenkant naujus, tinkamiausius elementus ir sudarant naują duomenų rinkinį. Ciklas kartojamas, kol gaunamas tinkamas rezultatas. Naudojami istoriniai duomenys, turintys parametrus, skirtus modelio kūrimui. Genetiniai algoritmai yra puikus būdas modelių optimizavimui, tačiau tai reikalauja atlikti papildomus skaičiavimus [10].

Genetiniai modeliai naudojami ateities procesų prognozavimui, surandama „gimininga“ informacija duomenų rinkiniuose.

1.4.5. Artimiausio kaimyno algoritmas

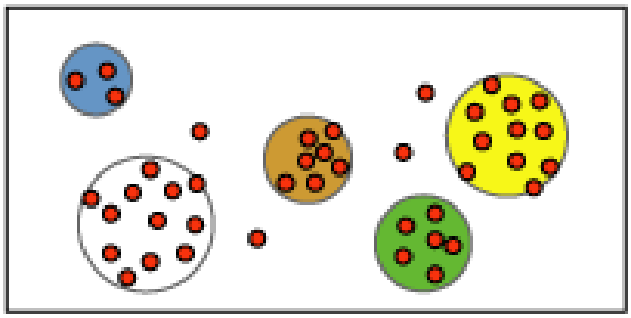
Artimiausio kaimyno algoritmas – vienas iš klasifikavimo uždavinių sprendimo būdų. Klasifikavimo uždavinius galima spręsti ne tik panaudojant grupavimo ar taisyklių metodus, bet ir ieškant elementų panašumo į artimiausius elementus (kaimynus). Pagrindinis šio algoritmo principas yra tai, kad išskiriami kriterijai, pagal kuriuos bus ieškomos elementų grupės duomenų aibėje. Ieškojimo proceso metu gretimi elementai lyginami tarpusavyje ir ieškoma dėsningų panašumų. Šio algoritmo pagrindinis trūkumas yra tas, kad yra ribotas modelio dydis, apdorojamas nedidelis kiekis duomenų [10].

1.4.6. Naïve Bayes algoritmas

Šis algoritmas taip pat sprendžia klasifikavimo ir prognozavimo uždavinius. Pagrindinis šio algoritmo veikimo principas – įvertinti elementų būsenas ir joms priskirti tam tikras atributų reikšmes. Tai vienas iš algoritmų, greičiausiai formuojančių duomenų gavybos modelius. Jo pagalba gali būti sprendžiami tokie uždaviniai, kaip elektroninės reklamos įtaka pirkimams, pelnui. Dažnai algoritmai naudojamas nustatyti tikslines pirkėjų grupes reklamuojamam produktui [10].

1.4.7. Grupavimas

Grupavimo algoritmas taikomas (6 pav.), norint išsiaiškinti grupes, turinčias panašius požymius, savybes ar charakteristikas. Panaudojant pasikartojimo principą, tarp elementų ieškoma panašius požymius turinčių elementų ir jie grupuojami į atskiras grupes. Grupavimui pasirenkami kriterijai, pagal kuriuos bus ieškoma panašių savybių, pasikartojančių tarp elementų. Sudaromos panašius požymius turinčios grupės ir duomenų analizė atliekama ieškant tarpusavio ryšių tarp grupių. Šio algoritmo pagalba gali būti sprendžiami rinkos ir klientų susiskirstymo uždaviniai, nekilnojamojo turto ar finansinių rodiklių prognozavimas. Šis algoritmas naudojamas ir kaip prognozavimo metodas, nes pagal suskirstytas grupes galima prognozuoti dėsningumus ateičiai [11].



Šaltinis: sukurta autoriaus

6 pav. Grupavimas

1.4.8. Asociacijų taisyklės

Asociacijos algoritmas – tai taisyklių rinkinys, kuris pavaizduoja tam tikrus dėsningumus duomenų aibėje. Šio algoritmo pagalba nustatomi ryšiai tarp įvykių, reiškinių ar daiktų grupėse. Didelėse duomenų bazėse ieškoma sąryšių tarp vienu momentu įvykusių reiškinių. Atrasti panašūs elementai ciklo metu grupuojami į elementų rinkinius. Atradus sąryšius, formuojamos elgsenos taisyklės, kurios naudojamos tolimesnei duomenų analizei. Asociacijos algoritmas jautrus algoritmo parametrų parinkimui, todėl mažoms problemoms spręsti geriau yra naudoti Sprendimų medžio algoritmą. Dažniausiai šis metodas naudojamas pardavimų ar pirkimų analizei. Nagrinėjama, kokios prekės perkamos kartu, kokia tikimybė, kad bus būtent toks derinys ir t.t. Tokio tipo uždaviniai gali būti sprendžiami rengiant reklamos kampanijas, kuriant nuolaidų sistemas [12].

1.4.9. Sekos nustatymas

Sekų nustatymo algoritmas naudojamas nustatyti tam tikru laiko momentu vykstančius procesus. Ieškoma sąryšių proceso sekoje. Svarbus kriterijus – įvykiai turi būti nuoseklūs, nes algoritmas jautrus atsitiktinei tvarkai. Dažniausiai šio algoritmo pagalba sprendžiami šie uždaviniai: internetinės svetainės lankomumas, dažniausiai peržiūrimos svetainėje produktų grupės, dažniausiai

internetu užsakomi produktai. Panaudojus šį metodą ir išanalizavus gautus rezultatus, galima prognozuoti klientų elgseną ar interneto svetainės modelį ateityje [10].

1.4.10. Regresija

Regresijos algoritmas naudojamas prognozavimo uždaviniams spręsti. Algoritmas esamų duomenų pagalba prognozuoja, kas bus ateityje. Pirmiausiai įvertinama esamos reikšmės, nustatant galimas reikšmių pasikeitimo tendencijas. Tai vienas iš klasikinės statistikos metodų. Šis metodas naudojamas sprendžiant sudėtingas ir aktualias pardavimų, gamybos, produkcijos problemas [10].

1.4.11. Laiko eilutės

Prognozuojant laiko eilučių algoritmo pagalba svarbus kriterijus yra laikas. Duomenų gavybos metu analizuojamas reikšmių kitimas per laikotarpį. Modelis atsižvelgia į laiko savybes. Tai gali būti hierarchinis periodų išsidėstymas, kalendorinės datos, tam tikri laiko momentai. Laiko eilučių pagalba sprendžiami akcijų ar vertybinių popierių kainos pokyčiai metų ketvirčiais. Algoritmas numato ateities reikšmes, kurios yra nežinomos ir prognozuojamos pagal per laikotarpį kintančias reikšmes [10].

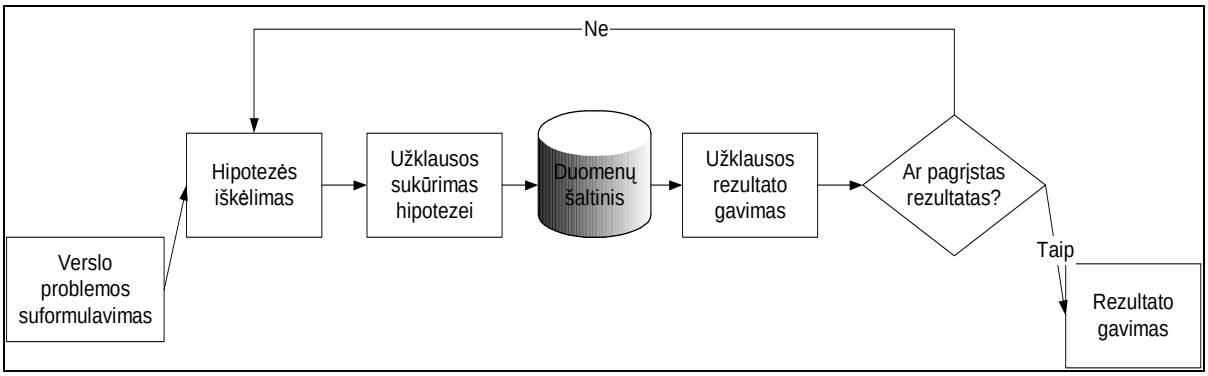
1.5. Duomenų gavybos ir kitų duomenų analizės įrankių palyginimas

Vienas iš dažniausiai kylančių klausimų apie duomenų profesionalų apdorojimą yra skirtumas tarp duomenų gavybos, statistikos ir OLAP (angl. *On-Line Analytical Processing*). Šie duomenų analizės įrankiai naudojami duomenų analizės procese. Tai yra skirtingi įrankiai, tačiau jie gali puikiai vienas kitą papildyti.

1.5.1. Duomenų gavyba ir OLAP

OLAP yra sprendimų palaikymo įrankių dalis. Įprastos užklausos pavaizduoja, kas yra duomenų bazėje. Tuo tarpu OLAP įrankis yra naudojamas, norint nustatyti, kodėl vyksta tam tikri dalykai. Vartotojas suformuoja hipotezę, tikrina jos teisingumą užklausų pagalba ir lygina su duomenimis. Taip galima nustatyti faktorius, kurie lemia paskolų nevykdymą. Analitikas gali spėti, kad žmonės, turintys mažas pajamas, yra nepatikimi ir OLAP pagalba patvirtinti (ar paneigti) šią prielaidą. Gauto rezultato pagalba analitikas gali nuspręsti tolimesnius veiksmus – skolinti pinigų ar ne. OLAP analizė generuoja hipotezinių šablonų, ryšių eilę ir naudoja užklausas duomenų bazėms, kad patvirtintų ar paneigtų juos. OLAP analizę galima apibūdinti kaip deduktyvinį (darantis

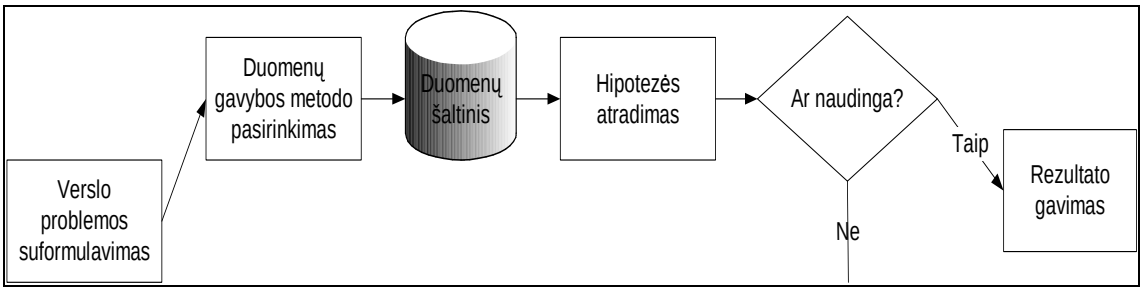
išvadas) procesą (7 pav.). Jis tampa daug sudėtingesnis ir atimantis nemažai laiko ieškant geros hipotezės ir analizuojant duomenų bazę su OLAP, kad patvirtintų ar paneigtų ją [13].



Šaltinis: sudaryta autoriaus.

7 pav. OLAP proceso modelis

Duomenų gavyba skiriasi nuo OLAP, nes ji vietoj hipotezinių šablonų tikrinimo naudoja pačius duomenis, kad atskleistų šiuos modelius. Tai yra induktyvinis procesas (8 pav.). Analitikas, kuris norėjo nustatyti rizikos faktorius paskolos nevykdymui, naudoja duomenų gavybos įrankį. Duomenų gavybos įrankis nurodo, kad žmonės, turintys didelę skolą ir mažas pajamas yra nepatikimi, bet šis įrankis gali plėtoti duomenų analizės procesą toliau ir taip pat nustatyti šabloną, apie kurį analitikas net nepagalvojo. Taip gali būti atrastos naujos priežastys, padėsiančios nustatyti naujus rizikos faktorius. Duomenų gavybos proceso metu nustatoma, kad paskolų išdavimo metu atsiranda nenumatyti veiksniai, kaip amžius, skolos ir pajamų santykis, kurie taip pat yra rizikingi ir turi būti vertinami suteikiant paskolą [14].



Šaltinis: sudaryta autoriaus.

8 pav. Duomenų gavybos proceso modelis

Pagrindinis skirtumas tarp OLAP ir duomenų gavybos yra tai, kaip yra apdorojami duomenys analizei. OLAP paremtas dedukcine analize, tuo tarpu duomenų gavyba remiasi indukcine duomenų analize. Šie du duomenų analizės įrankiai gali puikiai papildyti vienas kitą. Duomenų gavybos pagalba galima atrasti svarbius paslėptus sąryšius tarp duomenų šaltinio atributų ar lentelių. Tuomet OLAP gali smulkiau paaiškinti atrastus sąryšius bei suformuoti detalesnę ataskaitą [15].

Pagrindiniai OLAP ir duomenų gavybos skirtumai pateikiami 1 lentelėje.

1 lentelė

OLAP ir duomenų gavybos esminiai skirtumai

	OLAP	**Duomenų gavyba**
Analizės metodas	Dedukcinė duomenų analizė	Induktyvi duomenų analizė
Duomenų apdorojimo technikos	Hierarchinis duomenų skaidymas	Hierarchinis, dimencinis ir t.t. duomenų skaidymas
Analizės įrankiai	Užklausos	Algoritmai
Duomenų šaltinis	Duomenų bazės, duomenų centrai, duomenų sandėliai	Duomenų bazės, duomenų sandėliai, duomenų centrai, OLAP
Metodo principas	Analitikas suformuoja hipotezę ir panaudodamas OLAP įrankius, ją patvirtina.	Duomenys naudojami generuoti hipotezei.

Šaltinis: sudaryta autoriaus.

1.5.2. Duomenų gavyba ir statistika

Duomenų gavyba turi daug panašumų su statistika. Galima teigti, kad duomenų gavybos atsiradimą sąlygojo klasikinės statistikos metodų netobulumas. Iki duomenų gavybos atsiradimo statistiniai metodai puikiai apdorojo nedidelius duomenų kiekius. Didėjantys duomenų kiekiai duomenų bazėse sąlygojo duomenų gavybos atsiradimą ir sudarė sąlygas naujų metodų, paremtų efektyvių sprendimų tyrinėjimo, raidai.

Nauji metodai apima palyginti tokius naujus algoritmus, kaip neuroniniai tinklai ir Sprendimų medžiai, ir naują požiūrį į tokius senus algoritmus, kaip diskriminantinė analizė. Tradiciniai statistiniai metodai priklauso nuo modeliuotojo, apibrėžiančio funkcines būsenas ir sąveikas [16].

Esminis dalykas yra tas, kad duomenų gavyba yra taikoma įvairioms dirbtinio intelekto formoms ir statistiniams modeliams, pritaikytiems įprastoms verslo problemoms spręsti tokioje formoje, kurioje šie modeliai tampa naudingi tiek patyrusiam žinių darbuotojui, tiek kvalifikuotam statistikos profesionalui.

Panašiai, kaip statistika, duomenų gavyba nėra tik modeliavimas ir prognozė, bet tai – ištisas problemų sprendimo procesas. Sėkmingam duomenų išgavimui svarbiausias yra supratimas, ko verslui reikia iš tikrųjų, kadangi to negali įvertinti netgi patys naujausi ir sudėtingiausi algoritmai. Dar vienas svarbus aspektas yra duomenų kokybė, kadangi tik iš kokybiškų duomenų galima išgauti kokybiškus duomenis ir kokybiškai atlikti patį duomenų išgavimą. Tikrovėje ši sąlyga sunkiai įvykdoma, kadangi realūs duomenys beveik nebūna paruošti duomenų gavybai ir jie turi būti integruojami iš skirtingų duomenų šaltinių, turi klaidų arba neteisingų ar trūkstamų reikšmių [17].

1.6. Programinės įrangos apžvalga

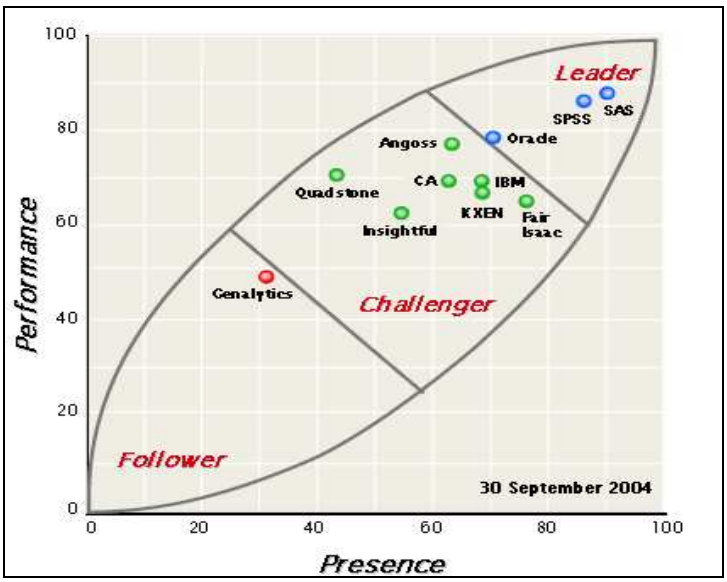
Duomenų gavybos rinką sudaro programinės įrangos gamintojai, kurie siūlo įrankius atrenkančius prognozuojančią informaciją iš didelės apimties duomenų šaltinių. Šie įrankiai atlieka duomenų analizę, kuri pagerina organizacijos vidinius resursus ir sukuria verslo krypties ir vartotojų elgsenos prognozes. Duomenų gavybos programinė įranga suteikia galimybes pasinaudoti statistiniais duomenų modeliais ir vizualinėmis priemonėmis [18].

Prognozuojama, kad duomenų gavybos rinka ateityje kasmet padidės 10% - 20%. 2004 metais duomenų gavybos programinių produktų rinkoje buvo didelis atotrūkis tarp gamintojų lyderių ir pasekėjų, tačiau pamažu rinka stabilizuojasi ir stiprėja. Šiuo metu rinką sudaro didieji programinės įrangos gamintojai (Oracle, SAS, IBM ir kt.) bei smulkieji duomenų analizėje specializuojantys programinės įrangos gamintojai (SPSS, KXEN, Angoss ir kt.) (9pav.).

Duomenų gavybos rinką galima suskirstyti į du segmentus:

- pažangi ir tiksli duomenų gavyba;
- duomenų gavyba platesnei analitikų auditorijai, su mažesnėmis technologinėmis galimybėmis.

Duomenų gavybos rinka šiuo metu suskirstyta į tris segmentus: lyderiai, pretendentai į lyderius ir pasekėjai. Programinės įrangos lyderiai siūlo stabilius, “subrendusius” produktus, kurie likusius programinės įrangos gamintojus visais aspektais pralenkia duomenų gavybos funkcijomis. Taip pat lyderiai užima stiprias pozicijas rinkoje. Pagal 2004 metų duomenis rinkos lyderiais buvo tokie programinės įrangos gamintojai kaip SAS, SPSS ir Oracle. Jie siūlo pažangias duomenų gavybos technologijas bei yra pirmieji šios programinės įrangos pradininkai. SPSS ir SAS siūlo daugiau statistinius duomenų gavybos įrankius [19].



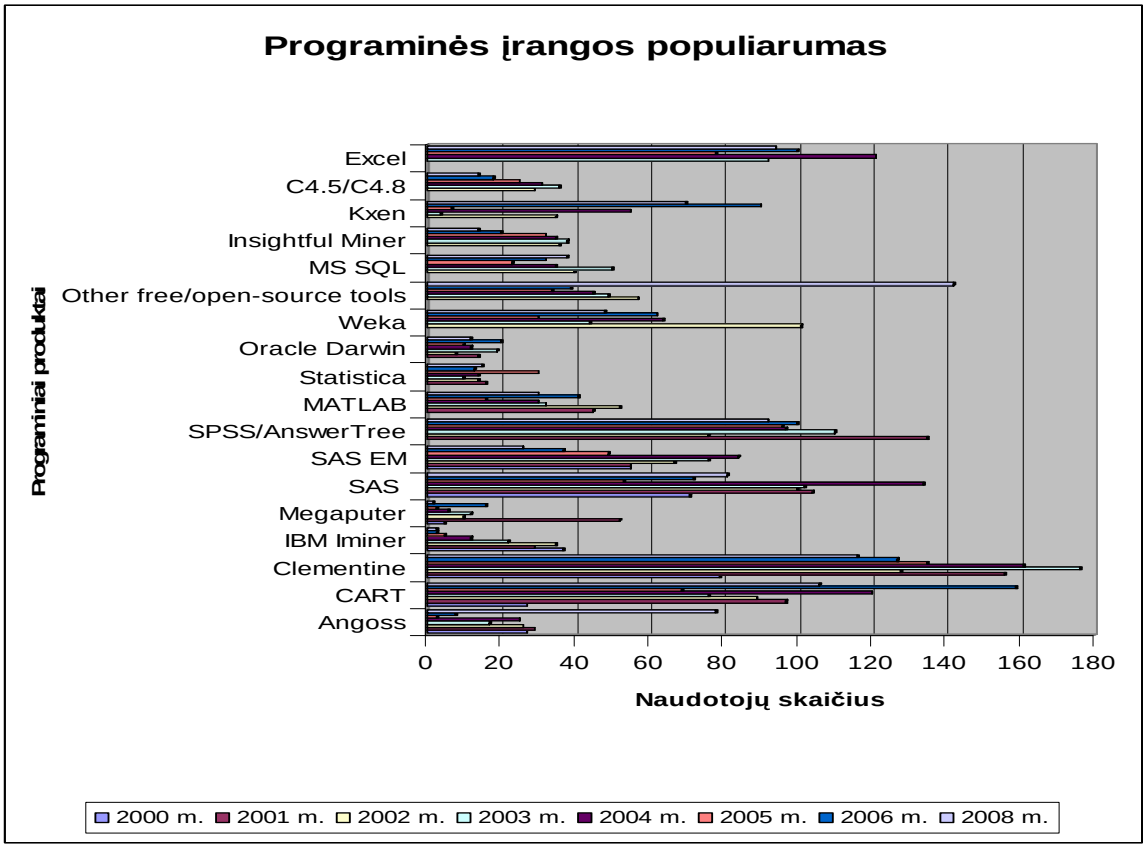
Šaltinis: Microsoft Gartner. (2003) Business Intelligence in Europe: Relevant and Valuable Information for better desicions in a Real-time Enterprise contex.

9 pav. Duomenų gavybos programinės įrangos lyderiai

KDnuggest (angl. The Association for Knowledge Discovery and Data Mining) – asociacija specializuojasi duomenų gavyboje. Ši asociacija vienija 10 000 duomenų gavybos specialistų. Kiekvienais metais vykdo įvairias apklausas, siekdama išsiaiškinti duomenų gavybos įrankių, metodikų ir kitų aspektų populiarumą šiandieninėje duomenų gavybos rinkoje [20].

Pasinaudojus minėtos organizacijos duomenimis, juos apibendrinus nustatyta, kad tarp analitikų populiari ši programinė įranga (10 pav.):

- SPSS gamintojo produktas Clementine;
- Salford gamintojo produktas CART;
- SAS produktas Enterprise Miner;
- Microsoft gamintojo MS SQL ir Excel produktai.

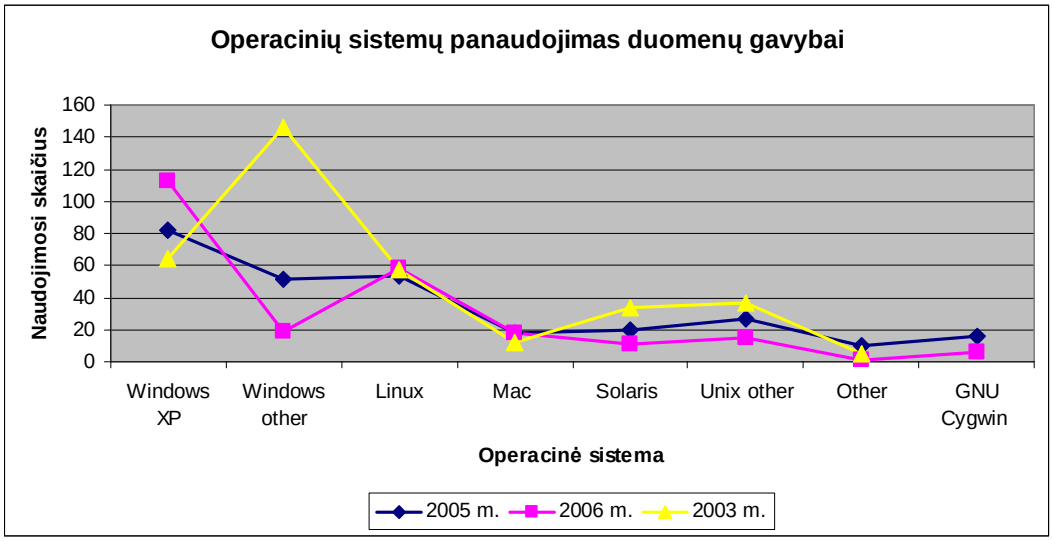


Šaltinis: sudaryta autoriaus pagal KDnugest (2008).

10 pav. Duomenų gavybos programinės įrangos populiarumas

Pagal duomenų gavybos analitikų naudojimosi operacinėmis sistemomis duomenis, galima teigti, kad dauguma specialistų naudoja Windows operacines sistemas (11 pav.). 2003 metais šias sistemas naudojo daugiau nei pusė duomenų gavybos specialistų. Ypatingai išaugo Windows XP operacinės sistemos naudojimas 2006 metais. Linux OS naudojasi žymiai mažesnis vartotojų skaičius, tačiau jis nekinta nuo 2003 metų,. Lyginant 2003 – 2006 metų laikotarpį, didžiausi

pokyčiai vyko tarp Windows sistemų. Kitos OS yra mažiau populiarios ir jomis naudojasi nedidelis skaičius analitikų.



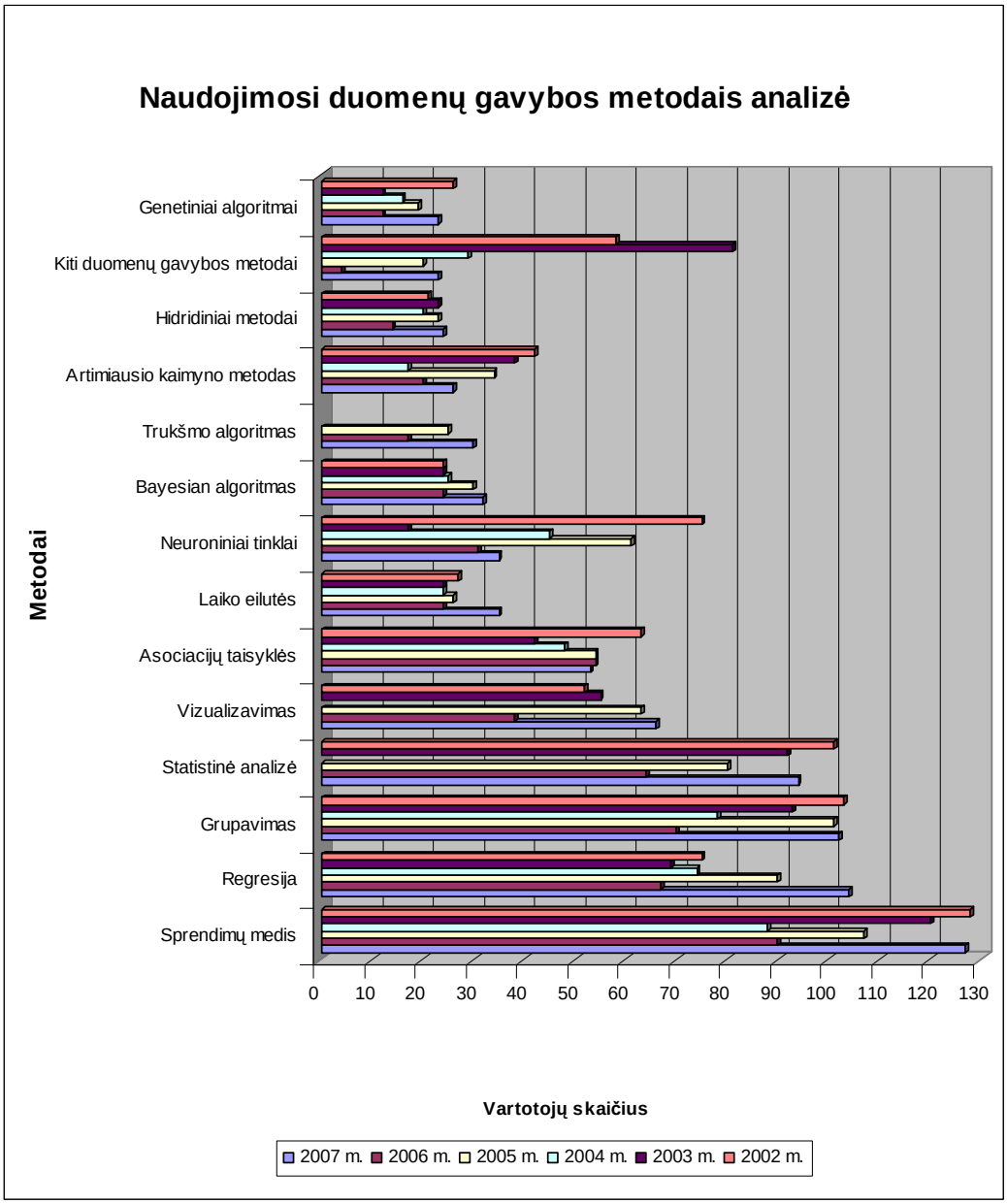
Šaltinis: sudaryta autoriaus pagal KDnuggest (2008).

11 pav. Operacinių sistemų naudojimas

Pagal duomenų gavybos algoritmų naudojimą duomenų analizei matome, kad analitikai daugiausia naudojasi šiais duomenų gavybos metodais (12 pav.):

- Sprendimų medžio algoritmas;
- Regresijos algoritmas;
- Grupavimo algoritmas.

Daugiausia naudojamasi Sprendimų medžio algoritmais. Galima teigti, kad šis algoritmas populiarus, nes yra vienas paprasčiausių duomenų gavybos metodų, lengvai suprantamas ir pritaikomas. Mažiausiai naudojasi Genetiniu ir Baysian algoritmu.



Šaltinis: sudaryta autoriaus pagal KDnuggest (2008).

12 pav. Duomenų gavybos algoritmų naudojimas

Toliau bus nagrinėjami šie duomenų gavybos produktai:

- Oracle – Darwin;
- MS SQL server 2005 Analysis Services;
- SPSS – Clementine;
- Megaputer – PolyAnalyst;
- IBM – DB2 Intelligent Miner.

2 lentelėje pateikiamas šių duomenų gavybos programinių produktų palyginimas. Išskiriami stipriausios ir silpniausios produkto pusės, funkcionalumo savybės.

2 lentelė

Programinių produktų palyginimas

Gamintojas	Produktas	Operacinė sistema	Pliusai	Minusai	Produkto funkcionalumas
SPSS	Clementine	Windows Server 2003, Windows 2000 Server, Windows 2000 Professional ar Windows NT 4.0 su Service Pack 6 ; Solaris 8 ar 9; HP-UX 11i; IBM AIX 5.2 ar OS/400; Microsoft Windows XP Home Edition, Windows XP Professional, Windows 2000 Professional, Windows Server 2003, Windows 2000 Server ar Windows NT 4.0.	Vizualizavimo palaikymo galimybė parodanti ryšius tarp duomenų rinkinių, turinčių didelius kiekius kintamųjų. Teksto gavybos galimybė leidžia ištraukti duomenis iš nestruktūrizuotų duomenų. Galimybė įtraukti Web duomenis į modeliuojamą aplinką ir procesus. Paprasta naudotis duomenų gavybos priemonėmis.	Nėra galimybės panaudoti genetinį algoritmą. Duomenų gavyboje nesiūlo ASP ir hosted serviso. Nesuteikia nuosavo portalo sąsajos, nors klientai gali naudotis visais modeliais. Painoka terminologija.	Produktas leidžia nustatyti ryšius tarp klientų, nustatant produktų ar paslaugų svarbumą jiems. Taip pat išanalizavus klientų poreikius, kurti reklamines kompanijas efektyviai ir tinkamai panaudojant reklamai skirtas lėšas. Daugiau statistinis produktas.
IBM corporation	DB2 Intelligent Miner	AIX, Solaris, Windows 2000, Windows NT, Windows Server 2003, Windows XP	Su šiuo programiniu produktu galime formuoti geresnius duomenų gavybos modelius nei su kitais įrankiais. Todėl, kad šis įrankis siūlo platų algoritmų pasirinkimą. Produktas turi labai vertingą API biblioteką, kuri suteikta galimybės sukurti klientų pritaikytą modelį. Produktas yra paremtas standartu ir kintamas. Galimybė apdirbti didelius kiekius duomenų.	Reikalinga draugiškesnė vartotojo sąsaja. Ieškoma būdų kaip panaudoti keletą gavybos operacijų vienam gavybos objektui.	Programinis produktas siūlo platų algoritmų pasirinkimą: klasterizavimas, klasifikavimas, vertės prognozavimas, ryšių ieškojimas, sekų modelių atradimas ir prognozuojantis modeliavimas.
MEGAPUTER Intelligence Inc	PolyAnalyst	Windows XP.	Galimybė ištraukti vertingas žinias iš teksto, papildant struktūriškai apibrėžtus duomenis, jų našumas ir grafinis vizualizavimas. Pirminei analizei ir preliminariai statistikai svarbu, kad duomenų rinkinius, sudarančius maždaug 20000 įrašų apdorojama per trumpesnį nei viena valanda laiko tarpą. Sekų analizė vyksta greitai ir tiksliai. Vizualinės priemonės puikiai atskleidžia reikšmingus sąryšius ir leidžia tolimesnę analizę. Dauguma diagramų gali būti eksportuojamos į kitas taikomąsias programas kaip MS Word ar MS PowerPoint.	Dauguma programinės įrangos savybių yra pritaikytos tiems analitikams, kurie supranta bendrus konceptus.	PolyAnalyst's Text Analysis apdirba teksto įrašus panaudodamas lingvistines, semantines ir statistines teksto analizės technologijų kombinacijas. PolyAnalyst Link Analysis leidžia parodyti ir vizualizuoti sąryšius tarp atrastų objektų ir demografinių grupių.

2 lentelės tęsinys

Gamintojas	Produktas	Operacinė sistema	Pliusai	Minusai	Produkto funkcionalumas
Oracle	Darwin	Windows NT/95 client/server UNIX , Sun Solaris ir HP-UX aplinka.	Galimybė apdoroti didelius duomenų kiekius, nes duomenų gavybos algoritmai analogiškai realizuojami. Modeliai gali būti lengvai išdėstyti kaip taikomoji dalis. Tai svarbu siekiant išsiaiškinti kuri duomenų gavybos dalis yra svarbiausia galutinio vartotojo vertei. Darwin labai patobulino vartotojo sąsają, dabar vartotojui leidžiančią vedlių pagalba vykdyti modelio kūrimo procesą. Darwin vedlys automatiškai kuria daugybinius modelius ir parenka geriausią peržiūrai. Darbų sekos ir rašymo galimybės leidžia duomenų gavybos žingsnių vaizdavimą ir automatizuoja duomenų gavybos procesą. Darwin leidžia išlaikyti aukšto lygio kontrolę.	Duomenų vizualizavimo įrankiai, kurie padeda vartotojui geriau suprasti duomenis formuojant duomenų gavybos modelius, yra pagrindinis Darwin trūkumas.	Darwin gali išgauti informaciją iš ASCII ir RDBMS. Darwin vartotojo sąsaja panaši į Windows NT/95, įskaitant išplėstus vedlius skirtus vartotojui kurti modelius. Lengvai transformuojami duomenys. Realizuoja tris duomenų gavybos algoritmus (neuroninių tinklų, klasifikavimo, regresijos medžius bei artimiausio kaimyno). Naudojama MS Excel grafinių duomenų gavybos rezultatų atvaizdavimui ir MS Internet Explorer tinklinei pagalbai. Planuojama naujesnes Oracle Darwin versijas papildyti klasterizavimo ir Naive Bayes algoritmais.
Microsoft	SQL server 2005 Analysis Services	Microsoft Windows platformos	Integruoti verslo tyrimų (business intelligence – BI) įrankiai, padėsiantys valdyti duomenis. Programos vedlys leidžia lengvai sukurti duomenų gavybos modulius, modeliai gali būti peržiūrimi keletu pjūvių. Draugiška vartotojo aplinka.	Tai integruotas duomenų gavybos įrankis, kuris teikia kiek ribotą įrankių pasirinkimą. Dirba tik Windows aplinkoje.	Integruota duomenų valdymo ir analizės platforma, skirta svarbioms įmonės verslui programoms. Produktas leidžia panaudoti dažniausiai naudojamus algoritmus. Galimybė duomenis atvaizduoti MS Excel pagalba. Duomenų analizei galima panaudoti ir OLAP rezultatus.

Šaltinis: sudaryta autoriaus.

Kiekvienas programinis produktas vartotojui pateikia tam tikrą sąrašą duomenų gavybos algoritmų, su kuriais galima atlikti duomenų analizę (3 lentelė). Ne visus algoritmus galima panaudoti konkrečiame programiniame produkte.

3 lentelė

Duomenų gavybos algoritmai

Produktas/ algoritmas	Sprendimo medžių algoritmas	Statistiniai algoritmai	Neuroniniai tinklai	Artimiausių kaimynų algoritmas	Bayes algoritmas	Grupavimo algoritmas	Taisyklių algoritmas	Laiko eilučių algoritmas	Sekų nustatymo algoritmas	K vidurkio algoritmas	Asociacijų taisyklių algortimas	Kohonen algoritmas
Clementine	X	X	X				X			X	X	X
Intelligent Miner	X	X	X					X	X	X	X	
PolyAnalyst	X											
Darwin	X		X	X	X	X						
SQL server 2005	X			X		X		X			X	

Šaltinis: sudaryta autoriaus.

Pasirenkant duomenų gavybos programinius produktus svarbu, kad vartotojui būtų lengva juo naudotis, duomenų gavybos procesą palengvintų vartotojo žinynai, vedliai. Duomenų gavybos proceso metu, vartotojas turi turėti galimybę pasirinkti duomenų gavybos algoritmus ar galimybę įvesti naujus algoritmus. Svarbios ir duomenų atvaizdavimo galimybės. 4 lentelėje pateikiamos svarbios duomenų gavybos produktų savybės:

- ODBC jungtis, kurios pagalba galima duomenų gavyba su įvairios paskirties duomenų bazėmis;
- Galutinio duomenų gavybos rezultato valdymas;
- Ataskaitų generavimas ir automatiškas antraščių sukūrimas;
- Native Database Drivers.

4 lentelė

Duomenų gavybos produktų savybės

Produktas	Automatiškai generuojama antraštė	Duomenų formato išsaugojimas	ODBC jungtis	Native Database Drivers	Ataskaitų formavimas	Produktas
Clementine	X		X			X
Darwin		X	X		X	X
Intelligent Miner				X		X
MS SQL server			X	X	X	X
PolyAnalyst	X					

Šaltinis: sudaryta autoriaus

Analizės dalies išvados

1. Atlikta duomenų gavybos sistemų analizė ir nustatyta, kad duomenų gavyba yra perspektyvi informacinių technologijų šaka, kurią galima pritaikyti optimizuojant verslo rezultatus, vertinant konkurentus, planuojant ateities perspektyvas.
2. Duomenų gavybos algoritmai yra puiki priemonė duomenų analizei ir juos panaudojant galima spręsti įvairaus pobūdžio uždavinius.
3. Išanalizavus pagrindinius duomenų gavybos ir kitų duomenų analizės sistemų privalumus ir skirtumus, nustatyta, kad duomenų gavybos, OLAP ir statistikos įrankiai papildo vienas kitą.
4. Atlikta duomenų gavybos programinės įrangos analizė. Nustatyta, kad duomenų gavybos įrankiai siūlo įvairius duomenų analizės būdus ir jų pasirinkimas priklauso nuo problemos pobūdžio ir analitiko patirties.

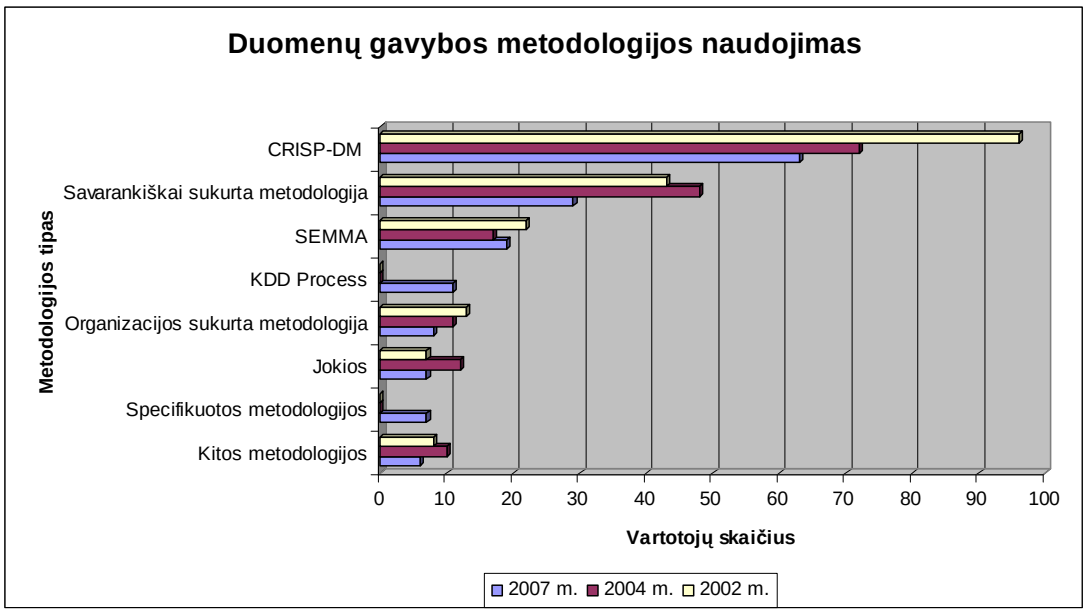
2. DUOMENŲ GAVYBOS PROCESO METODIKA

Duomenų gavyba ir analizė tai procesas, kuris naudoja daugybę duomenų analizės įrankių, kad surastų modelius ir ryšius duomenyse, kurie gali būti panaudoti efektyvioms prognozėms. Pagrindinis duomenų gavybos principas, tai naujų priklausomybių tarp kintamųjų radimas. Pirmas ir paprasčiausias analitinis žingsnis duomenų gavyboje ir analizėje yra duomenų supratimas – apibendrinti jų statistinius atributus, vizualiai apžvelgti juos (naudojant lenteles ir diagramas) bei surasti galimus ryšius tarp kintamųjų.

Bet vienas duomenų supratimo etapas negali sukurti veiksmų plano. Reikia suformuoti prognozuojantį modelį, paremtą šablonais, nustatytais iš žinomų rezultatų, tuomet patikrinti tą modelį su pirminiu duomenų rinkiniu. Geras modelis neturėtų kada nors būti supainiotas su tikrove, bet jis gali būti naudingas vadovas, norint suprasti savo verslą. Galutinis žingsnis yra empiriškai (patirties būdu) patvirtinti modelį.

Vartotojui iškeliamas sudėtingas uždavinys – pasirinkti tinkamas duomenų gavybos priemones. Svarbu pasirenkant metodą vadovautis aplinkos logika, sukaupta patirtimi, tinkamai įvertinti tikslus ir apribojimus. Nėra vienintelio ir paties geriausio modelio, tik atsižvelgus į anksčiau paminėtus aspektus bei įvertinus metodo savybes, galima pasirinkti tinkamiausią metodą.

Prieš pradedant realizuoti duomenų gavybos uždavinius, svarbu pasirinkti tinkamą duomenų gavybos metodologiją. Šiuo metu analitikai naudoja dvi pagrindines duomenų gavybos metodologijas CRISP - DM ir SEMMA (13 pav.). Darbe naudosime CRISP - DM metodologiją, nes ji populiariausia ir ją naudoja didieji programinės įrangos gamintojai kaip SSPS, Oracle.

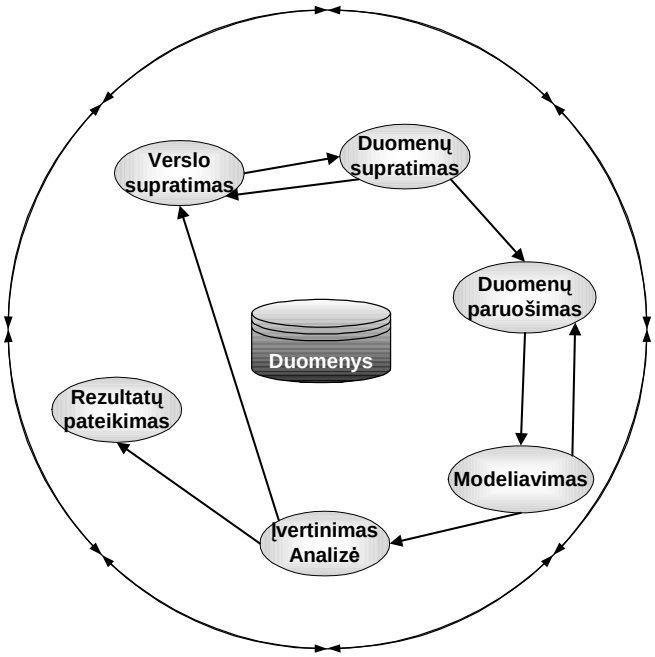


Šaltinis: sukurta autoriaus pagal KDnuggest (2008).

13 pav. Duomenų gavybos metodologijos

2.1. Duomenų gavybos proceso modelis

CRISP-DM (angl. *Cross-Industry Standard Process for Data Mining*) – standartinis duomenų gavybos proceso modelis, 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. Jų sudarytas duomenų gavybos procesų modelis (14 pav.), toliau minimas kaip CRISP – DM, atvaizduoja pagrindinius gavybos proceso principus [21].



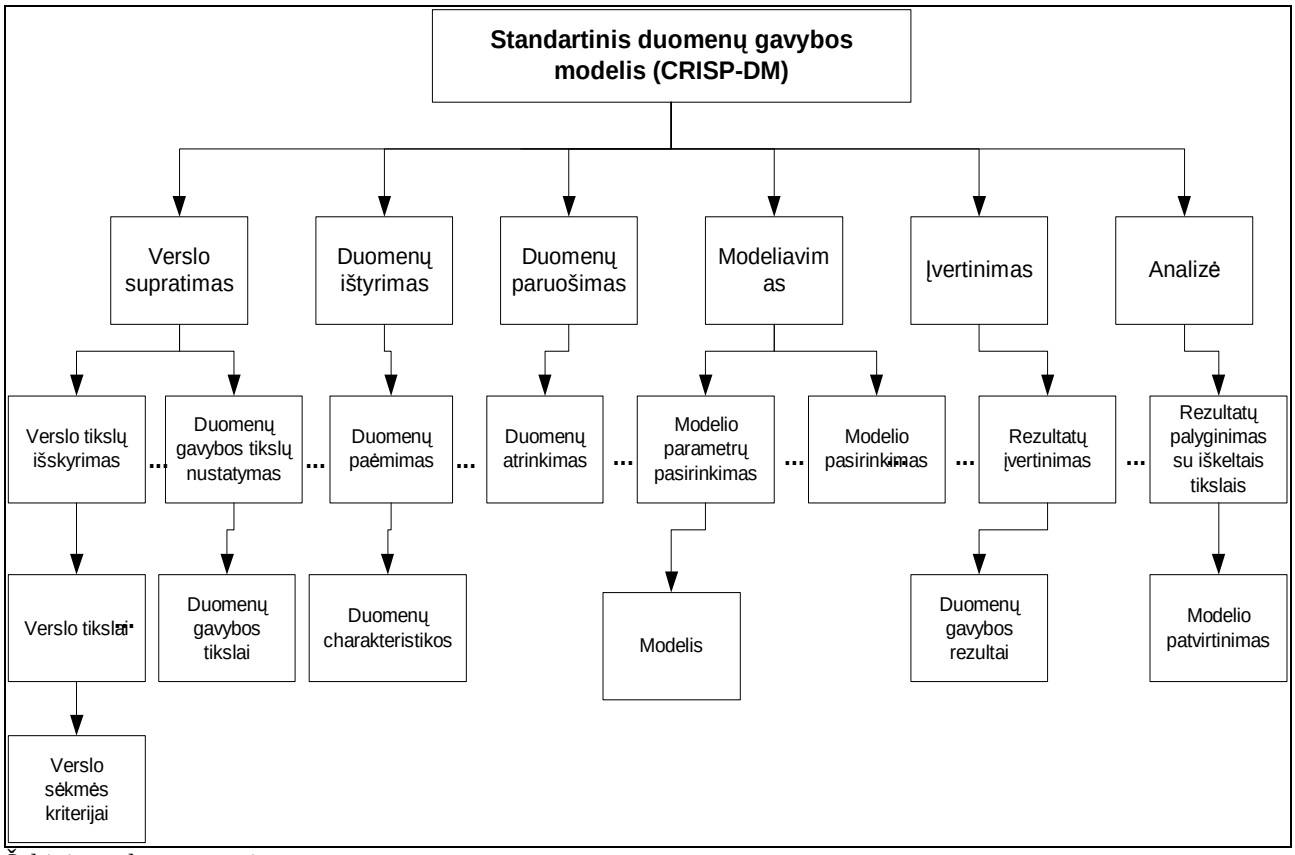
Šaltinis: sudaryta autoriaus pagal CRISP-DM (2002).

14 pav. Standartinis duomenų gavybos modelis

Duomenų gavybos procesas nėra linijinis procesas, tai uždaras ciklas, kai įvykdžius vieną žingsnį neišvengiamai reikia grįžti į prieš tai padarytus žingsnius. Tai svarbu siekiant užtikrinti proceso vientisumą, kokybišką duomenų gavybos proceso eigą bei norint išvengti galimų klaidų. Tai sunkus ir didelio pasiruošimo reikalaujantis procesas. Pagrindiniai duomenų gavybos kūrimo žingsniai yra šie:

1. Išanalizuoti verslo aplinką, apibrėžiant verslo problemą bei sukuriant veiklos modelį;
2. Ištirti organizacijos duomenis;
3. Pasiruošti duomenis modeliavimui;
4. Sukurti modelį;
5. Įvertinti modelį;
6. Įvertinti duomenų gavybos rezultatus;
7. Pateikti gautus rezultatus vartotojui.

Galima teigti, kad šis modelis yra hierarchinis modelis, nes kiekviena fazė turi jai priskirtus pagrindinius uždavinius (15 pav.). Kiekvienas konkretus uždavinys išplėtojamas į specializuotas užduotis, kurios galutiniame etape pateikia laukiamus rezultatus.



Šaltinis: sudaryta autoriaus.

15 pav. Modelio hierarchinis atvaizdavimas

2.1.1. Verslo aplinkos supratimas ir problemos išskyrimas

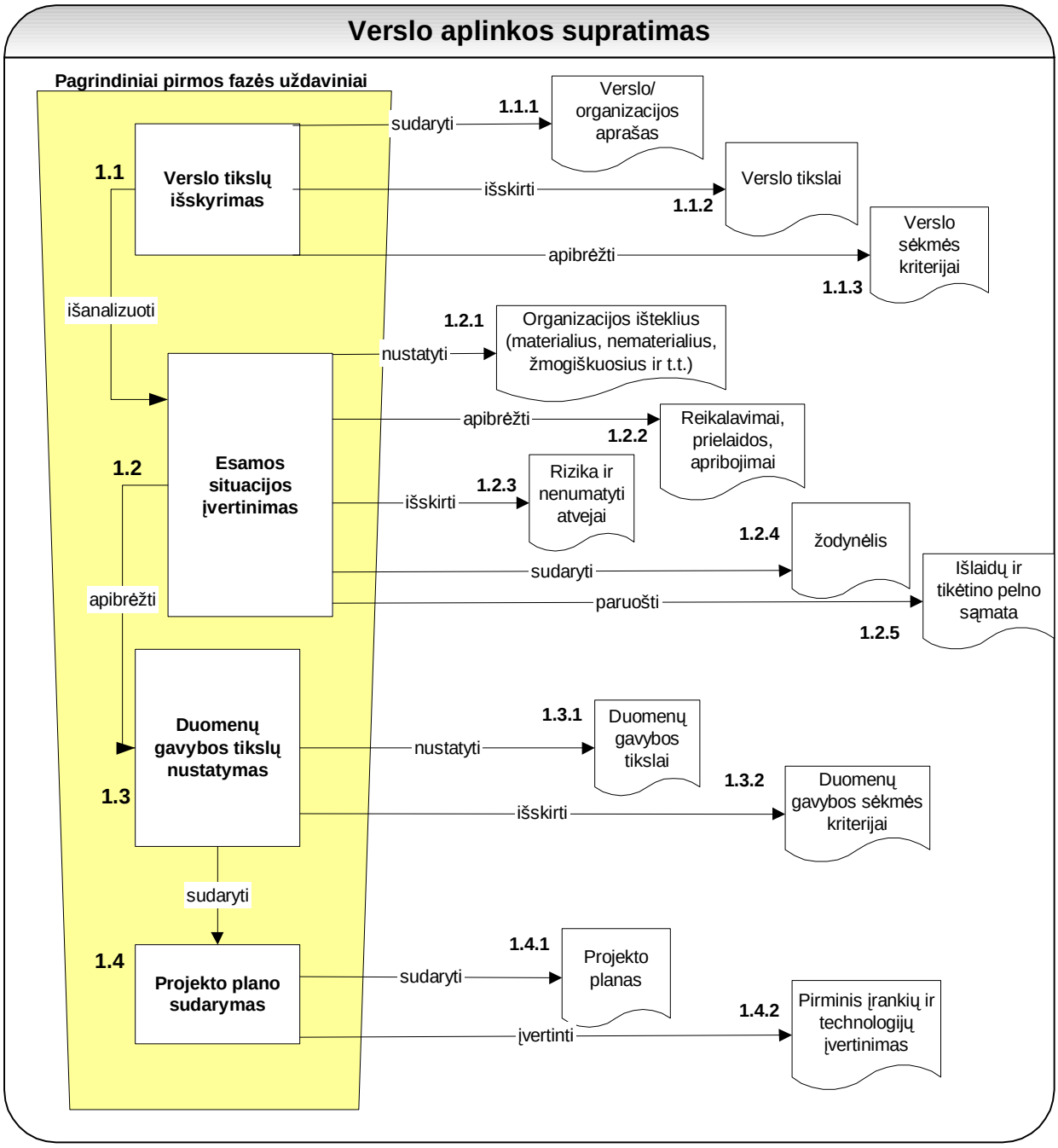
Duomenų gavybos procesui svarbu ištirti organizacijos, kurioje bus pritaikyta duomenų gavyba, aplinką. Pirmas analitiko tikslas yra nuodugniai ištirti verslo aplinką ir nustatyti ką iš tikrųjų klientas nori įvykdyti. Dažnai klientas turi daug įvairių tikslų ir apribojimų, kurie privalo būti tinkamai subalansuoti. Analitiko tikslas yra projekto pradžioje atskleisti svarbiausius faktorius, kurie gali paveikti galutinį rezultatą. Nepaisant šio žingsnio galutinis rezultatas turėtų būti sumaniai sukurti teisingi atsakymai į klaidingus klausimus.

Analitikas, šiame etape, turi ištirti organizacijos veiklą, informacijos judėjimo srautus, nustatyti ir apibrėžti pagrindines verslo problemas. Ši duomenų gavybos proceso fazė reikalauja kruopštaus ir tikslaus organizacijos ištyrimo. Tai viso proceso pagrindas. Suvokus pagrindinius organizacijos veiklos principus, galima identifikuoti problemas, kurios iškyla organizacijos veikloje. Identifikavus problemą svarbu nustatyti pagrindinius duomenų gavybos tikslus, apibrėžti laukiamus rezultatus, formuoti duomenų gavybos projektą.

Ši fazė apima dar keturis etapus, reikalingus sėkmingai verslo analizei:

- Verslo tikslų išskyrimas;
- Esamos situacijos įvertinimas;
- Duomenų gavybos tikslų nustatymas;
- Kuriamo projekto plano sudarymas.

16 paveiksle pavaizduoti pagrindiniai pirmojo duomenų gavybos proceso etapo uždaviniai. Kiekvienas iš nurodytų uždavinių turi savo išeigos rezultatus, kurių pagalba formuojamas duomenų gavybos projektas.



Šaltinis: sudaryta autoriaus pagal CRISP-DM (2002).

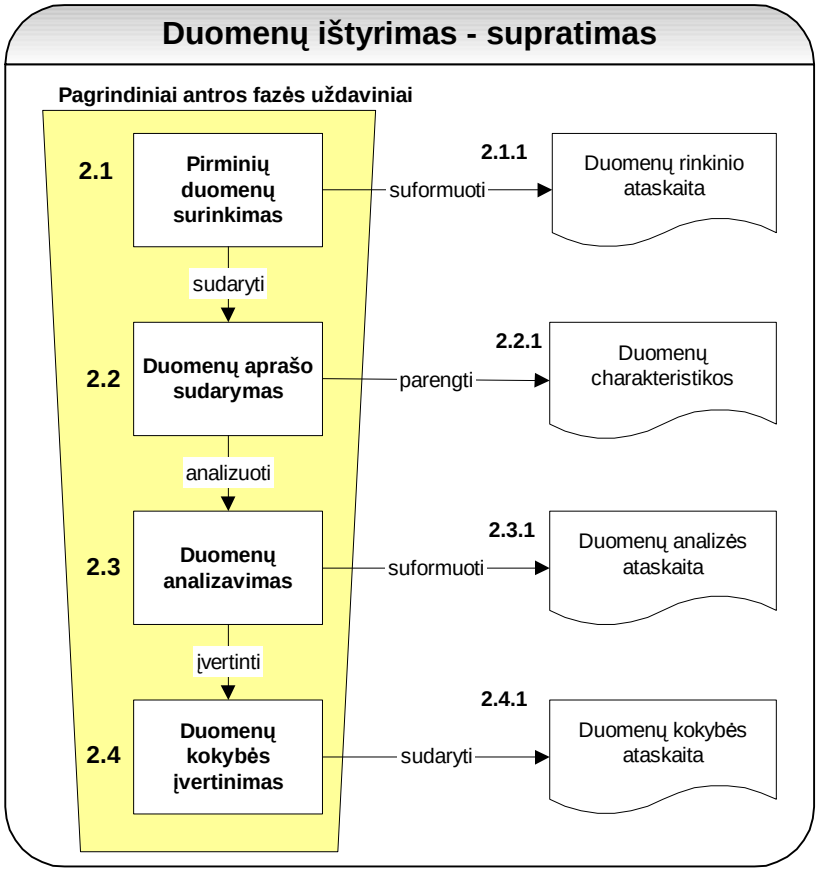
16 pav. Verslo aplinkos supratimo etapas

2.1.2. Duomenų ištyrimas - supratimas

Trečiasis svarbus etapas, tai organizacijos duomenų ištyrimas (17 pav.). Šio etapo pagrindinis tikslas identifikuoti svarbiausius duomenų gavybos procesui duomenis. Šie duomenys turi didžiausią įtaką tolimesniems procesams bei laukiamiems rezultatams. Analitikas turi išanalizuoti duomenų bazės struktūrą, lentelių ypatybes, kintamųjų charakteristikas. Išanalizuoti duomenis bei įvertinti jų kokybę bei tinkamumą tolimesnei projekto eigai.

Ši fazė apima dar keturis etapus, reikalingus tolimesnei projekto kūrimo eigai:

- Pirminių duomenų surinkimas;
- Duomenų aprašo sudarymas;
- Duomenų analizavimas;
- Duomenų kokybės įvertinimas.



Šaltinis: sudaryta autoriaus pagal CRISP-DM (2002).

17 pav. Duomenų ištyrimo – supratimo etapas

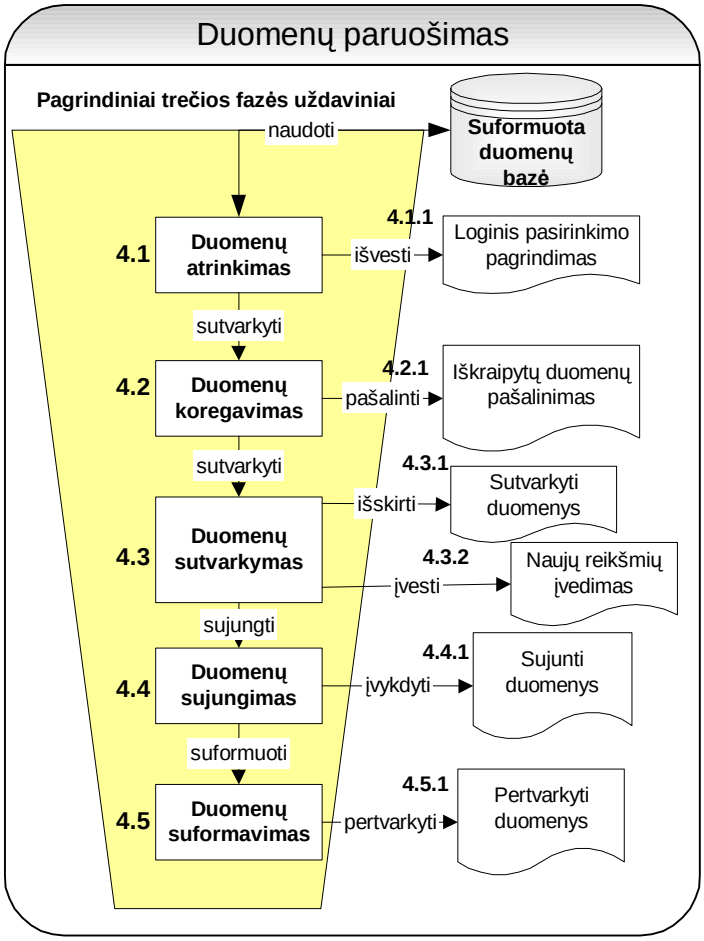
2.1.3. Duomenų paruošimas

Paruošti duomenis modeliavimui (18 pav.). Tai yra galutinis duomenų paruošimo žingsnis prieš sukuriant modelį. Duomenų paruošimas vienas iš svarbiausių duomenų gavybos etapų, nes tik tinkamai surinkus, išanalizavus ir sutvarkius duomenis, galima tikėtis sėkmingo duomenų gavybos

rezultato. Analitikas iš ankstesnių duomenų gavybos proceso etapų atsineša patyrimą, žinias, kurias panaudos duomenų paruošimo etape. Tai svarbu grupuojant duomenis reikalingam duomenų gavybos modeliui, paruošiant juos sėkmingai duomenų gavybai.

Yra penki pagrindiniai šio žingsnio etapai:

- Išrinkti reikšmingus duomenis;
- Sutvarkyti duomenis;
- Išdėstyti duomenis;
- Sujungti duomenis į duomenų gavybai reikalingas grupes;
- Pertvarkyti duomenis.



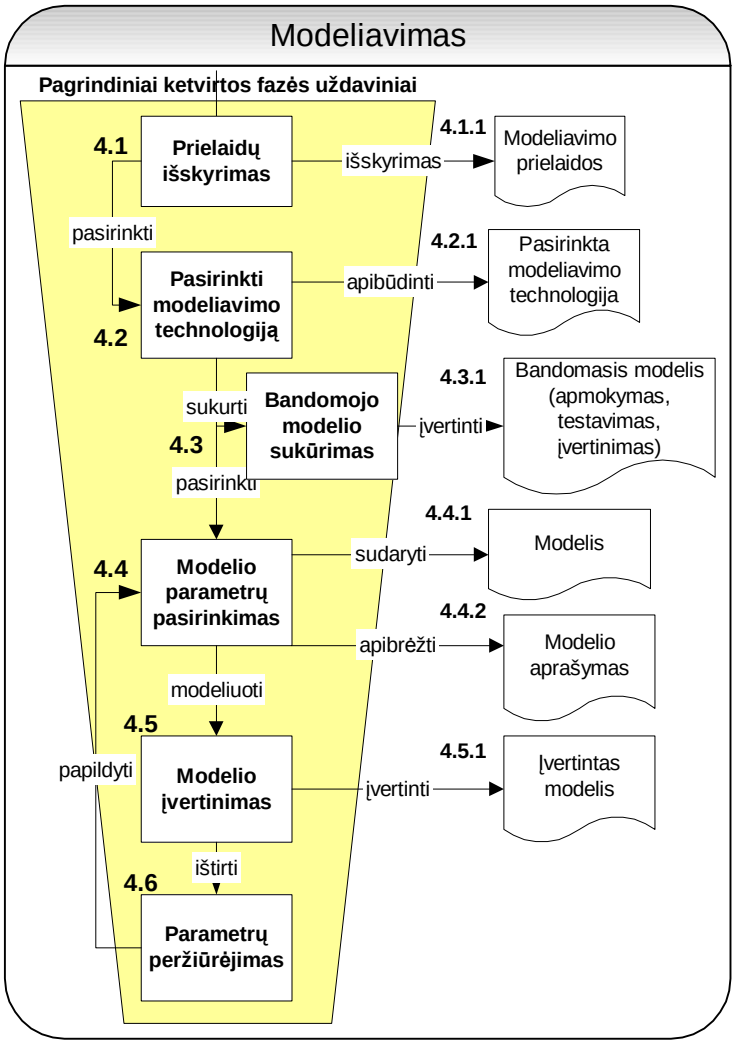
Šaltinis: sudaryta autoriaus pagal CRISP-DM (2002).

18 pav. Duomenų paruošimo etapas

2.1.4. Modelio sukūrimas - modeliavimas

Šio etapo tikslas sukurti modelį. Svarbiausias šio etapo uždavinys - įvertinus iškeltą dalykinės srities problemą ir panaudojus praeituose etapuose įgytą patirtį, sukurti modelis, kuris tinkamiausias verslo problemai išspręsti (19 pav.).

Pirmiausiai iškeliamos modeliavimo prielaidos, kurios turi atitikti iškeltus duomenų gavybos tikslus. Pagal modeliavimo prielaidas pasirenkamas duomenų gavybos algoritmas. Viso šio etapo metu būtina įvertinti prieš tai buvusių etapų rezultatus. Esant poreikiui, galima juos šiek tiek koreguoti. Pasirinktus modeliavimo metodą, svarbu pasirinkti ir modeliavimo parametrus. Pirmiausiai sudaromas bandomasis modelis – duomenų testavimui ir galutinio modelio rengimui.



Šaltinis: sudaryta autoriaus pagal CRISP-DM (2002).

19 pav. Modeliavimo etapas

Duomenų gavybos modelio parengimui ir testavimui duomenys padalinami bent į dvi dalis: vieni skirti modelio rengimui, o kiti modelio testavimui. Jei nebus naudojami skirtingi rengimo ir testavimo duomenys, tai modelis bus netikslus. Po to, kai modelis bus generuojamas, naudojant rengimo duomenų bazę, galima numatyti testavimo duomenų bazę. Atlikus bandomojo modelio testavimą ir įvertinus jo klaidas, dar kartą pasirenkami modelio parametrai ir formuojamas galutinis modelis.

Modelio įvertinimui naudojami visi duomenų bazėje esantys duomenys. Duomenys atsitiktinai dalinami į du lygius rinkinius. Pirmiausiai, modelis sudaromas pagal pirmą rinkinį, kad

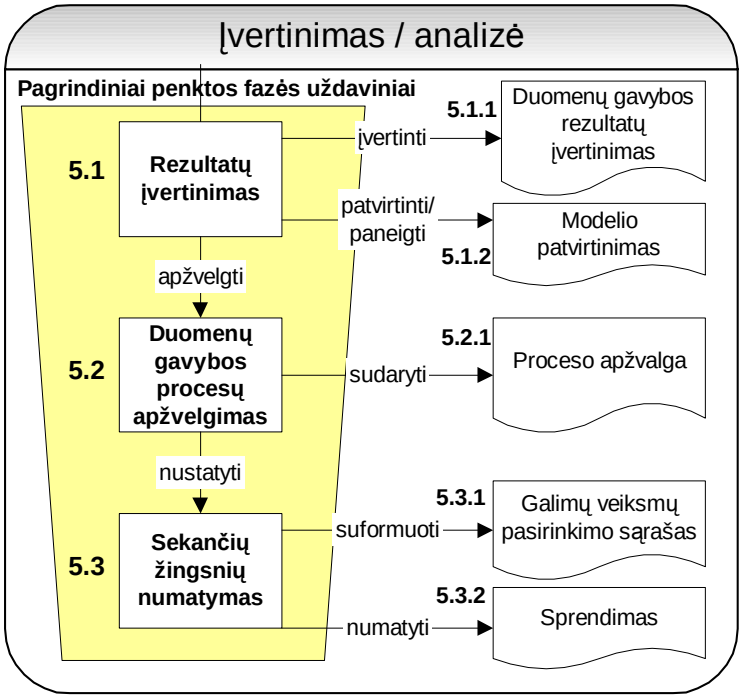
numatytų rezultatus antrame rinkinyje ir būtų galima suskaičiuoti klaidos dažnį. Tada modelis kuriamas pagal kitą rinkinį ir vėl apskaičiuojamas klaidos dažnis. Galiausiai, modelis kuriamas naudojant visus duomenis.

Sukūrus modelį, sudaromas jo aprašas. Kokie parametrai reikalingi modeliavimo etape, nurodomi pagrindiniai modeliavimo žingsniai, technologiniai sprendimai ir reikalavimai įrankiui. Taip pat aprašomas modelio veikimo principas bei laukiami rezultatai.

Sudarius modelį ir atlikus jo įvertinimą, atsižvelgiant į pateiktus pasiūlymus, atliekami numatyti pakeitimai. Tai gali būti modelio parametrų pakeitimas, duomenų rinkinio pakoregavimas ir t.t.

2.1.5. Modelio įvertinimas - patvirtinimas

Įvertinimas ir modelio patvirtinimas (20 pav.). Po modelio sukūrimo, vartotojas privalo įvertinti jo rezultatus ir interpretuoti jų reikšmę ir svarbą. Modelio įvertinimo žingsnyje apibūdinami sėkmės ir nesėkmės kriterijai, įgyvendinti lūkesčiai, netikėti rezultatai. Analitikas turi sutapatinti iškeltus duomenų gavybos tikslus ir gautus rezultatus, bei įvertinti ar pavyko modeliavimo etapas.



Šaltinis: sudaryta autoriaus pagal CRISP-DM (2002).

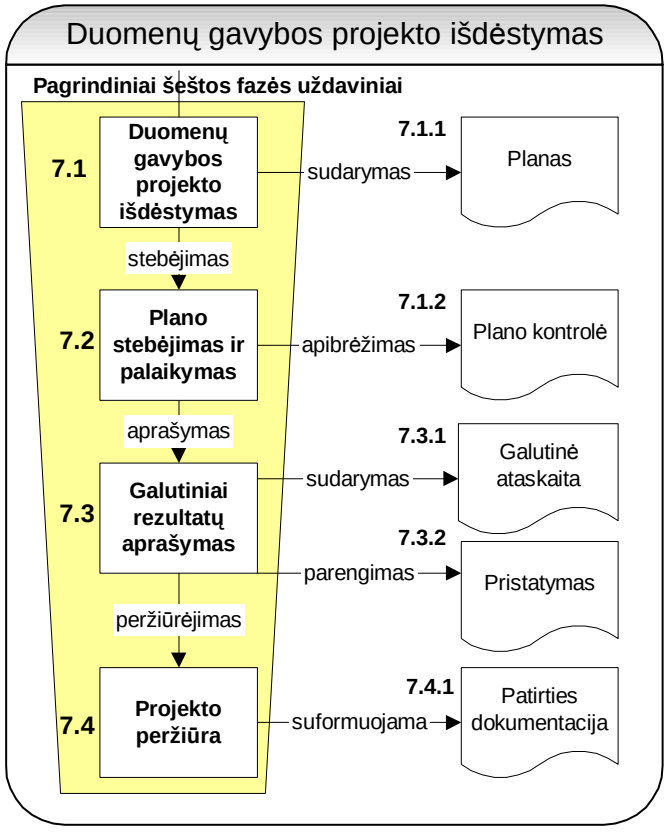
20 pav. Įvertinimo – analizės etapas

Šiame etape taip pat priimamas sprendimas ar pavyko duomenų gavybos procesas. Įvertinami rezultatai, numatomi tolimesni duomenų gavybos projekto žingsniai. Svarbu numatyti tolimesnius veiksmus duomenų gavybos projektui plėtoti.

2.1.6. Duomenų gavybos projekto išdėstymas - užbaigimas

Paskutiniame duomenų gavybos projekto etape įvertinamas duomenų gavybos projektas, jo rezultatai, apibendrinama projekto metu įgyta duomenų gavybos patirtis. Projekto dalyviai turi parenti šiuo dokumentus (21 pav.):

- parengti duomenų gavybos projekto planą, kurį bus galima panaudoti tolimesniems duomenų gavybos projektams;
- numatyti kontrolės veiksmus, kurie bus panaudoti, jeigu duomenų gavyba neteiks laukiamų rezultatų;
- nuspręsti duomenų gavybos projekto sudarymo dokumentus, kokiame pavidale jie bus pateikiami (ataskaitos, protokolai, pristatymai ir t.t.);
- išskirti tikslines organizacijos darbuotojų grupes, kuriems aktualus duomenų gavybos projektas bei numatyti kaip jiems bus pristatomas projektas;
- parengti projekto pristatymą tikslinėms grupėms;
- apklausti projekto dalyvius bei užfiksuoti jų įgytą patirtį, pastebėtus trūkumus ir privalumus;
- pateikti pasiūlymus ateičiai.



Šaltinis: sudaryta autoriaus pagal CRISP-DM (2002).

21 pav. Duomenų gavybos rezultatų įvertinimas

2.2. Duomenų gavybos proceso etapų specifikacijos

Kiekvienas duomenų gavybos proceso etapas (fazė) hierarchiškai paveldi specializuotus uždavinius, kuriuos įvykdžius gaunamas galutinis rezultatas. Kiekvienas etapas yra svarbus, nes jų rezultatų pagrindu formuojamas duomenų gavybos projektas ir įgyvendinamas duomenų gavybos uždavinys.

Toliau pateikiamos lentelės, kuriose pateikiamos kiekvieno duomenų gavybos etapo specifikacijos.

5 lentelėje pateikiamos pirmojo etapo – verslo aplinkos specifikacijos.

5 lentelė

Verslo aplinkos supratimo etapo specifikacijos

Pagrindiniai uždaviniai	Užduotys	Užduoties etapai
1. Verslo aplinkos supratimas ir problemos išskyrimas		
1.1. Verslo tikslų išskyrimas	1.1.1. Verslo organizacijos aprašo sudarymas	1.1.1.1. Formuoti organizacines diagramas, kurios parodo ryšius tarp skyrių, departamentų ir padalinių. Diagrama taip pat turi atvaizduoti darbuotojų pareigybes ir atsakomybes. 1.1.1.2. Nustatyti atsakingus asmenis ir jų roles. 1.1.1.3. Nustatyti vidinius ir išorinius vartotojus, vadovus. 1.1.1.4. Nustatyti dalykines sritis, kurios svarbios duomenų gavybos projektui (pvz.: Rinkodara, Pardavimai, Finansai). 1.1.1.5. Identifikuoti probleminę sritį ir ją aprašyti (pvz.: Rinkodara, pardavimai, elektroninė prekyba, gaminio kokybė ir t.t.). 1.1.1.6. Išsiaiškinti projekto kūrimo prielaidas (pvz.: Kokia motyvacija kurti projektą? Ar organizacijoje, jau naudojama duomenų gavyba? 1.1.1.7. Jeigu reikia, paruošti pristatymą ir pristatyti duomenų gavybos svarbą organizacijai. 1.1.1.8. Identifikuoti tikslines vartotojų grupes, kurioms svarbūs projekto rezultatai. 1.1.1.9. Identifikuoti vartotojo poreikius ir lūkesčius.
	1.1.2. Verslo tikslų išskyrimas	1.1.2.1. Neformaliai aprašyti problemą, kuri bus sprendžiama duomenų gavybos pagalba. 1.1.2.2. Kuo galima tiksliau apibrėžti visus iškilusius dalykinės srities klausimus, problemas. 1.1.2.3. Nustatyti kitus verslo – organizacijos poreikius (pvz.: organizacija nenori daugiau prarasti klientų ir t.t.) 1.1.2.4. Apibrėžti tikėtiną naudą.
	1.1.3. Verslo sėkmės kriterijų apibrėžimas	1.1.3.1.Išskirti verslo sėkmės kriterijus (pvz.: 10% pagreitinti atsakymų trukmę į reklaminę kampaniją bei 20% sumažinti tempą). 1.1.3.2. Identifikuoti rezultatus, tikėtinus sėkmės atveju. 1.1.3.3. Sėkmės kriterijai turi atitikti išskirtus verslo tikslus.

5 lentelės tęsinys

Pagrindiniai uždaviniai	Užduotys	Užduoties etapai
1.2. Esamos situacijos įvertinimas	1.2.1. Organizacijos išteklių įvertinimas	1.2.1.1. Įvertinti organizacijos techninę įrangą bei įvertinti jos tinkamumą DM. 1.2.1.2. Patikrinti techninės įrangos palaikymo specifikacijas ir nustatyti ar nekonfliktuos su duomenų gavybos įrankiais. 1.2.1.3. Nustatyti duomenų šaltinius ir duomenų šaltinių tipus. 1.2.1.4. Nustatyti žinių šaltinius ir jų tipus (dokumentai, ekspertai ir t.t.). 1.2.1.5. Aprašyti svarbiausių žinių kilmę (formaliai ir neformaliai). 1.2.1.6. Nustatyti duomenų gavybos projekto rėmėją. 1.2.1.7. Numatyti sistemos ir duomenų bazės administratorių, techninį personalą. 1.2.1.8. Numatyti rinkos analitiką, duomenų gavybos ekspertą ir statistikus. 1.2.1.9. Įvertinti visų numatytų specialistų tinkamumą tolimesnėms projekto kūrimo fazėms.
	1.2.2. Reikalavimų, prielaidų ir apribojimų apibrėžimas	1.2.2.1. Nustatyti tikslines grupes. 1.2.2.2. Surašyti tiksliai, suprantama visus reikalavimus duomenų gavybos projektui ir galutiniam rezultatui. 1.2.2.3. Surašyti reikalavimus saugumo užtikrinimui, juridiniams apribojimams, slaptumui, ataskaitų ir projekto planavimui. 1.2.2.4. Detaliai išsiaiškinti visas prielaidas (pvz.: pirkėjai, kurių amžius virš 50 yra svarbūs). 1.2.2.5. Sudaryti prielaidų sąrašą duomenų kokybei, išoriniams faktoriams.
	1.2.3. Rizikos ir nenumatytų atvejų išskyrimas	1.2.3.1. Nustatyti visas galimas verslo rizikas (pvz. konkurentai pirmieji pritaiko DM ir ištiria rinką). 1.2.3.2. Nustatyti visas galimas organizacines rizikas. 1.2.3.3. Nustatyti visas galimas finansines rizikas (pvz.: projektui neskiriamas finansavimas). 1.2.3.4. Nustatyti visas galimas technines rizikas. 1.2.3.5. Nustatyti visas galimas duomenų kokybės ir kitas rizikas (pvz.: prasta duomenų kokybė ir t.t.).
	1.2.4. Terminų - žodynėlio sudarymas	1.2.4.1. Formuoti specifinių terminių žodyną. 1.2.4.2. Aptarti terminių prasmes su dalykinės srities ekspertais. 1.2.4.3. Susipažinti su verslo terminologija.
	1.2.5. Išlaidų ir planuojamo pelno sąmata	1.2.5.1. Apytiksliai apskaičiuoti duomenų atrinkimo ir sukaupimo išlaidas. 1.2.5.2. Apytiksliai apskaičiuoti sprendimų įgyvendinimo ir plėtojimo išlaidas. 1.2.5.3. Apskaičiuoti galimą įgyvendinto projekto naudą (pvz.: padidėjo pirkėjų pasitenkinimas, pajamų ir apyvartos didėjimas ir t.t.). 1.2.5.4. Apskaičiuoti eksploatacijos išlaidas. 1.2.5.5 Įvertinti galimas netikėtas išlaidas (pvz.: papildomos išlaidos apmokymams, užtrukusiems darbams apmokėti, pakartotiniems darbams atlikti).
1.3. Duomenų gavybos tikslų nustatymas	1.3.1. DM tikslų nustatymas	1.3.1.1. Perkonstruoti dalykinės srities problemas į duomenų gavybos tikslus; 1.3.1.2. Nustatyti duomenų gavybos tipą (pvz.: klasifikavimas, atvaizdavimas, prognozė ar grupavimas).
	1.3.2. DM sėkmės kriterijų išskyrimas	1.3.2.1. Tiksliai nustatyti kriterijus modelio įvertinimui (pvz.: modelio tikslumas, sudėtingumas ir įgyvendinimo tikslumas). 1.3.2.2. Nustatyti sėkmės įvertinimo kriterijus.
1.4. Projekto plano sudarymas	1.4.1. Projekto plano sudarymas	1.4.1.1. Nustatyti pradinius plano procesus ir apsvarstyti jo įvykdymą. 1.4.1.2. Išdėstyti visus nustatytus tikslus ir pasirinktus metodus kartu, kurie sprendžia verslo klausimus ir atitinka sėkmės kriterijus. 1.4.1.3. Įvertinti sprendimo įgyvendinimui reikalingus resursus. 1.4.1.4. Nustatyti kritinius žingsnius. 1.4.1.5. Išskirti sprendimų vietas. 1.4.1.6. Išskirti peržiūros vietas. 1.4.1.7. Nustatyti svarbiausius pasikartojimus.
	1.4.2. Įrankių, technologinių priemonių pasirinkimas	1.4.2.1. Sudaryta įrankių ir metodų atrankos kriterijų sąrašą (ar bus naudojama esama įranga). 1.4.2.2. Išskirti potencialius įrankius ir techniką. 1.4.2.3. Įvertinti technikos tinkamumą. 1.4.2.4. Peržiūrėti ir išskirti technikos prioritetus.

Šaltinis: sudaryta autoriaus.

6 lentelėje pateikiamos antrojo duomenų gavybos etapo - duomenų ištyrimo specifikacijos.

6 lentelė

Duomenų ištyrimo – supratimo specifikacija

Pagrindiniai uždaviniai	Užduotys	Užduoties etapai
2. Duomenų ištyrimas ir supratimas		
2.1. Pirminių duomenų surinkimas	2.1.1. Duomenų rinkinio ataskaita	2.1.1.1. Numatyti, kuri informacija reikalinga (pvz.: tik svarbiausi atributai, reikalinga papildoma informacija). 2.1.1.2. Patikrinti, kuri iš reikalingos informacijos faktiškai prieinama. 2.1.1.3. Patikslinti atrankos kriterijus (pvz.: kurie atributai yra svarbiausi duomenų gavybai). 2.1.1.4. Atrinkti reikalingas lenteles/ failus. 2.1.1.5. Atrinkti duomenis iš lentelių/ failų.
2.2. Duomenų aprašo sudarymas	2.2.1. Duomenų charakteristikos	2.2.1.1. Patikrinti atributų naudingumą ir pasiekiamumą. 2.2.1.2. Patikrinti atributų tipus (skaitinis, simbolinis, taksonominis ir t.t). 2.2.1.3. Patikrinti atributo požymio vertes. 2.2.1.4. Įvertinti atributo koreliaciją. 2.2.1.5. Suprasti kiekvieno atributo prasmę ir reikšmę verslo prasmėje. 2.2.1.6. Kiekvienam atributui, jeigu reikia, apskaičiuoti svarbiausius statistinius duomenis (vidurkį, minimalias ir maksimalias reikšmes, standartinį nuokrypį ir t.t.). 2.2.1.7 Nustatyti atributų reikšmę duomenų gavybai. 2.2.1.8. Identifikuoti surinktus duomenis. 2.2.1.9. Informaciją gauti iš duomenų šaltinių. 2.2.1.10. Nustatyti lenteles ir ryšius tarp jų. 2.2.1.11. Patikrinti duomenų apimtis, pasikartojančius skaičius ir duomenų sudėtingumą.
2.3. Duomenų analizavimas	2.3.1. Duomenų analizės ataskaita	2.3.1.1. Detaliai išanalizuoti svarbių atributų savybes. 2.3.1.2. Išanalizuoti duomenų charakteristikas. 2.3.1.3. Suformuoti hipotezę ir nustatyti tolimesnius veiksmus. 2.3.1.4. Performuluoti iškeltą hipotezę į duomenų gavybos tikslą.
2.4. Duomenų kokybės įvertinimas	2.4.1. Duomenų kokybės ataskaita	2.4.1.1. Patikrinti atributus su skirtingomis reikšmėmis, tačiau turinčius tą pačią prasmę (pvz.: dietinis, mažesnis riebalų kiekis). 2.4.1.2. Patikrinti kintamųjų rašybą (pvz.: kartais reikšmės prasideda mažąją, o kartais didžiąja raide). 2.4.1.3. Patikrinti kintamųjų patikimumą (pvz.: visi reikšmių laukai turi tas pačias ar beveik tas pačias reikšmes).

Šaltinis: sudaryta autoriaus.

7 lentelėje pateikiamos trečiojo etapo - duomenų paruošimo specifikacijos.

7 lentelė

Duomenų paruošimo specifikacija

Pagrindiniai uždaviniai	Užduotys	Užduoties etapai
3. Duomenų paruošimas		
3.1. Duomenų atrinkimas	3.1.1. Loginis pasirinkimo pagrindimas	3.1.1.1. Surinkti duomenis iš visų šaltinių; 3.1.1.2. Įvertinti laukų koreliacijas ir tarpusavio ryšius, nusprendžiant ar laukas reikalingas; 3.1.1.3. Peržiūrėti duomenų atrankos kriterijus; 3.1.1.4. Atrinkti skirtingas duomenų grupes; 3.1.1.5. Sugalvoti atrankos būdus ir įrankius.
3.2. Duomenų koregavimas	3.2.1. Iškraipytų duomenų pašalinimas	3.2.1.1. Apgalvoti kaip elgtis su duomenimis, kurie turi neaiškias reikšmes; 3.2.1.2. Pataisyti, pašalinti ar ignoruoti iškraipytus duomenis; 3.2.1.3. Pergalvoti duomenų atrankos kriterijus.

7 lentelės tęsinys

Pagrindiniai uždaviniai	Užduotys	Užduoties etapai
3.3. Duomenų sutvarkymas	3.3.1. Sutvarkyti duomenys	3.3.1.1. Parinkti priemones grafiniam duomenų išdėstymui; 3.3.1.2. Nuspręsti, kuris įrankis yra tikslesnis, efektyvesnis.
	3.3.2. Naujų duomenų įvedimas	3.3.2.1. Parinkti priemones naujų duomenų įvedimui; 3.3.2.2. Atrinkti reikšmingus naujus duomenis ir jais papildyti duomenis.
3.4. Duomenų sujungimas	3.4.1. Sujungti duomenys	3.4.1.1. Apgalvoti duomenų sujungimo galimybes (pvz.: ar gali būti sujungti duomenys su išoriniais duomenimis); 3.4.1.2. Integruoti šaltinius ir išsaugoti rezultatą; 3.4.1.3. Peržiūrėti dar kartą duomenų paruošimo kriterijus.
3.5. Duomenų suformavimas	3.5.1. Pertvarkyti duomenys	3.5.1.1. Atlikti sintaksinius pakeitimus pagal duomenų gavybos modelio reikalavimus.

Šaltinis: sudaryta autoriaus.

8 lentelėje pateikiamos ketvirtojo etapo - modeliavimo specifikacijos.

8 lentelė

Modeliavimo specifikacija

Pagrindiniai uždaviniai	Užduotys	Užduoties etapai
4. Modeliavimas		
4.1. Prielaidų išskyrimas	4.1.1. Modeliavimo prielaidos	4.1.1.1. Apibrėžti modeliavimo prielaidas, įvertinant duomenis, tipus.
4.2. Pasirinkti modeliavimo technologiją	4.2.1. Pasirinkta modeliavimo technologija	4.2.1.1. Įvertinti modeliavimo prielaidas; 4.2.1.2. Palyginti modeliavimo prielaidas su duomenų aprašu (2.2.1.); 4.2.1.3. Įsitikinti ar prielaidos atitinka duomenų paruošimo etapo žingsnius.
4.3. Bandomojo modelio sukūrimas	4.3.1. Bandomasis modelis (apmokymas, testavimas, įvertinimas)	4.3.1.1. Patikrinti sudarytą projektą su kiekvienu duomenų gavybos tikslu; 4.3.1.2. Atrinkti svarbiausius žingsniu; 4.3.1.3. Paruošti duomenis testavimui.
4.4. Modelio parametrų pasirinkimas	4.4.1. Modelis	4.4.1.1. Nustatyti pradinius parametrus; 4.4.1.2. Aprašyti pasirinktų reikšmių priežastis.
	4.4.2. Modelio aprašymas	4.5.1.1. Aprašyti kiekvieno modelio charakteristikas, kurios bus panaudotos tolimesniuose etapuose; 4.5.1.2. Surašyti modeliavimo parametrus; 4.5.1.3. Sudaryti detalų modelio aprašą, pažymint svarbius jo žingsnius; 4.5.1.4. Parengti taisyklių, techninės informacijos sąrašus (priklausomai nuo naudojamo modelio); 4.5.1.5. Parengti modeliavimo aprašą ir modelio elgsenos supratimą; 4.5.1.6. Įvertinti galimus modeliavimo rezultatus.
4.5. Modelio įvertinimas	4.5.1. Įvertintas modelis	4.5.1.1. Įvertinti modelį pagal iškeltus įvertinimo kriterijus.
4.6. Parametrų peržiūrėjimas	4.6.1. Peržiūrėti parametrai	4.6.1.1. Patikrinti parametrus ir jeigu reikia (dėl geresnio modeliavimo rezultatų), juos pakeisti.

Šaltinis: sudaryta autoriaus.

9 lentelėje pateikiamos penktojo etapo modelio įvertinimo – patvirtinimo specifikacijos.

9 lentelė

Modelio įvertinimo – patvirtinimo specifikacija

Pagrindiniai uždaviniai	Užduotys	Užduoties etapai
5. Įvertinimas - analizė		
5.1. Rezultatų įvertinimas	5.1.1. Duomenų gavybos rezultatų įvertinimas	5.1.1.1. Suprasti ir įvertinti duomenų gavybos rezultatus; 5.1.1.2. Įvertinti duomenų gavybos rezultatus ir jų atitikimą iškeltiems duomenų gavybos tikslams; 5.1.1.3. Įvertinti ar duomenų gavybos rezultatai originali ir naudinga.
	5.1.2. Modelio patvirtinimas	5.1.2.1. Nustatyti ar rezultatai patvirtina iškeltus verslo sėkmės kriterijus; 5.1.2.2. Suformuoti išvadas ateities duomenų gavybos projektams.
5.2. Duomenų gavybos procesų apžvelgimas	5.2.1. Proceso apžvalga	5.2.1.1. Aprašyti duomenų gavybos proceso sėkmę; 5.2.1.2. Išanalizuoti duomenų gavybos procesą, kiekvieną jo etapą; 5.2.1.3. Identifikuoti nesėkmes; 5.2.1.4. Nustatyti klaidingus ir nereikalingus žingsnius; 5.2.1.5. Nustatyti galimus alternatyvius veiksmus, nenumatytus žingsnius procese; 5.2.1.6. Peržiūrėti duomenų gavybos rezultatus įvertinant juos pagal iškeltus verslo sėkmės kriterijus.
5.3. Sekančių žingsnių numatymas	5.3.1. Galimų veiksmų pasirinkimo sąrašas	5.3.1.1. Išanalizuoti rezultatų įvertinimą; 5.3.1.2. Apsvarstyti kiekvieno proceso galimus patobulinimus; 5.3.1.3. Patikrinti išliekamus išteklius; 5.3.1.4. Rekomenduoti alternatyvius procesus; 5.3.1.5. Patobulinti projekto planą.
	5.3.2. Sprendimas	5.3.2.1. Suklasifikuoti galimus tolimesnius veiksmus; 5.3.2.2. Atrinkti vieną iš galimų veiksmų; 5.3.2.3. Aprašyti pasirinkimo priežastis.

Šaltinis: sudaryta autoriaus.

10 lentelėje pateikiamos šeštojo etapo – duomenų gavybos rezultatų įvertinimo specifikacijos.

10 lentelė

Duomenų gavybos projekto išdėstymo – užbaigimo specifikacija

Pagrindiniai uždaviniai	Užduotys	Užduoties etapai
6. Plano išskleidimas		
6.1. Duomenų gavybos projekto išdėstymas	6.1.1. Planas	6.1.1.1. Apibendrinti rezultatus; 6.1.1.2. Nustatyti plano vertę. 6.1.1.3. Nuspręsti kiekvienos aiškio informacijos rezultatą; 6.1.1.4. Nuspręsti kokiu būdu pateikti rezultatus vartotojams; 6.1.1.5. Nustatyti kiekvieno modelio ir programinės įrangos rezultatus; 6.1.1.6. Nustatyti galimas duomenų gavybos problemas.
6.2. Plano stebėjimas ir palaikymas	6.2.1. Plano kontrolė	6.2.1.1. Patikrinti dinaminius aspektus (pvz.: kas gali pakeisti sąlygas?); 6.2.1.2. Pagalvoti kaip kontroliuoti planą; 6.2.1.3. Nuspręsti, kuriuo atveju duomenų gavybos rezultatai ar modeliai nebus naudojami; 6.2.1.4. Nustatyti kriterijus (tikslumas, nauji duomenys, srities pasikeitimas, modelio atnaujinimas, naujų duomenų įterpimas į duomenų gavybos projektą); 6.2.1.5. Tiksliai nustatyti problemas, kurios bus sprendžiamas modelio pagalba; 6.2.1.6. Sudaryti kontrolės ir palaikymo planą.

10 lentelės tęsinys

Pagrindiniai uždaviniai	Užduotys	Užduoties etapai
6.3. Galutiniai rezultatų aprašymas	6.3.1. Galutinė ataskaita	6.3.1.1. Nuspręsti, kokios ataskaitos reikalingos (skaidrių demonstracija, valdymo išdėstymas, detalizuotos išvados, modelių paaiškinimai ir t.t.); 6.3.1.2. Išanalizuoti kaip pirminiai duomenų gavybos tikslai gali būti įgyvendinti; 6.3.1.3. Nustatyti tikslines grupes, kuriems reikalingos ataskaitos; 6.3.1.4. Nuspręsti ataskaitų struktūrą ir turinį; 6.3.1.5. Atrinkti išvadas – duomenis, kurios sudarys ataskaitas; 6.3.1.6. Parašyti ataskaitą.
	6.3.2. Pristatymas	6.3.2.1. Parengti baigiamąjį pristatymą (o gal ir baigiamąją ataskaitą); 6.3.2.2. Pasirinkti iš baigiamosios ataskaitos punktus, kurie sudarys pristatymą.
6.4. Projekto peržiūra	6.4.1. Patirties dokumentacija	6.4.1.1. Apklausti projekto dalyvius apie jų įgytą patirtį projekto metu; 6.4.1.2. Apklausti galutinio rezultato vartotojus, ar projektas pateisino lūkesčius; 6.4.1.3. Apibendrinti grįžtamąjį žingsnį ir aprašyti įgytą patirtį; 6.4.1.4. Išanalizuoti procesą (ar viskas vyko sėkmingai, klaidos, išmoktos pamokos); 6.4.1.5. Aprašyti duomenų gavybos procesą ir įgytą patirtį bei jos naudą ateityje.

Šaltinis: sudaryta autoriaus.

2.3. Pasiūlymai duomenų gavybos procesui patobulinti

Šiame skyriuje pateikiami pasiūlymai kaip patobulinti duomenų gavybos procesą, pastebėti proceso trūkumai sėkmingam duomenų gavybos projekto vykdymui.

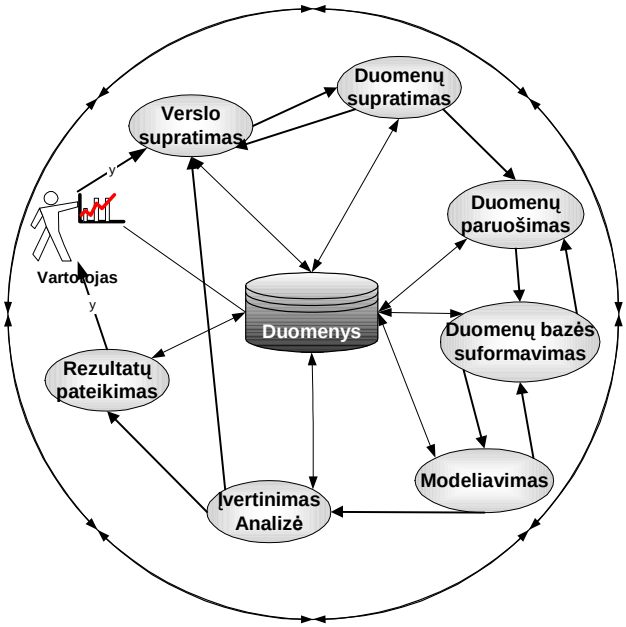
2.3.1. Duomenų gavybos gyvavimo ciklo schemos patobulinimas

Projektuojant duomenų gavybos proceso modelį remtasi CRISP-DM standartine kūrimo metodika [21]. Duomenų gavybos proceso modelis - tai vientisas duomenų gavybos gyvavimo ciklas, kurį sudaro 6 svarbios fazės – etapai:

1. Verslo aplinkos ištyrimas, probleminės verslo srities išskyrimas;
2. Organizacijos duomenų ištyrimas;
3. Duomenų paruošimas modeliavimui;
4. Modeliavimas;
5. Modelio įvertinimas;
6. Duomenų gavybos rezultatų įvertinimas;
7. Gautų duomenų gavybos rezultatų pateikimas vartotojui.

Kiekvienas etapas glaudžiai susijęs su kitu, prieš tai buvusiu etapu. Taip pat labai svarbu, kad vieno etapo sėkmė tiesiogiai proporcinga duomenų gavybos rezultato sėkmei. Išanalizavus

modelį, siūloma papildyti jį duomenų bazės suformavimo etapu, vartotojo vaidmeniu bei įterpti papildomus duomenų tikrinimo laukus tarp etapų (22 pav.).



Šaltinis: sudaryta autoriaus.

22 pav. Standartinis duomenų gavybos modelis

2.3.2. Duomenų bazės suformavimas

Kita svarbi duomenų gavybos procese fazė yra tinkamai sutvarkyta duomenų bazė, kurios duomenys bus naudojami tolimesnei duomenų gavybos eigai. Ši fazė svarbi tolimesnei eigai bei duomenų tyrimo ir paruošimo fazėms. Šios trys fazės atima daugiausiai laiko ir pastangų. Pagal CRISP-DM modelio sudarytojus, duomenų paruošimo ir ištyrimo fazės sudaro 80 % viso projekto kūrimo laiko. Duomenų bazės suformavimas svarbus tuo, kad duomenų gavybai reikalingi kokybiški ir tvarkingi duomenų rinkiniai.

Dažniausiai organizacijose duomenys yra saugomi įvairiuose duomenų šaltiniuose, todėl šiame etape svarbu juos integruoti į vieną duomenų bazę. Esant dideliems duomenų kiekiams tampa sunku juos valdyti ir analizuoti. Geriau yra kurti atskirus, struktūrizuotus ir susistemintus duomenų centrus ar duomenų grupes. Duomenų gavybos procese, kiekvienas etapas reikalauja pakartotino analitinio duomenų įvertinimo.

Duomenų bazės kūrimo procese galėtų būti šie veiksmai:

a. duomenų atrinkimas iš įvairių duomenų šaltinių;
b. duomenų apipavidalinimas;
c. duomenų kokybės įvertinimas ir duomenų gryninimas;
d. sujungimas ir integracija;
e. Metaduomenų konstravimas.

2.4. Duomenų gavybos proceso metodikos išvados

1. Duomenų gavybos procesą, remiantis CRISP-DM standartu, sudaro šeši pagrindiniai etapai: verslo aplinkos supratimas, duomenų ištyrimas, duomenų paruošimas, modeliavimas, modelio įvertinimas ir duomenų gavybos projekto įvertinimas.
2. Išanalizavus duomenų gavybos procesą, sudarytos duomenų gavybos proceso specifikacijos. Išskirti žingsniai, kuriuos reikia atlikti duomenų gavybos proceso etapų metu.
3. Įvertinus duomenų gavybos proceso gyvavimo ciklą, siūloma papildyti modelį šiais aspektais: pridėti duomenų bazės suformavimo etapą, į gyvavimo ciklą įterpti vartotojo vaidmenį bei kiekvieną etapą labiau susieti su duomenims.

3. EKSPERIMENTINIS SKYRIUS

Eksperimentinėje dalyje pateikiama duomenų gavybos metodika, realizuota panaudojant MS SQL server 2005 Analysis Services programinę įrangą [23]. Duomenų gavybos kūrimo proceso atvaizdavimui sukurti 6 laboratoriniai darbai, atvaizduojantys pagrindinius duomenų gavybos principus:

1. Verslo aplinkos analizė ir probleminės srities išskyrimas;
2. Programinės įrangos analizė ir pasiruošimas projektui;
3. Duomenų paruošimas projektui;
4. Dimensijų, reikalingų duomenų gavybai, kūrimas;
5. Duomenų gavybos modelio pritaikymas ir panaudojimas;
6. Duomenų gavybos modelių plėtojimas;
7. Duomenų gavybos modelių palyginimas.

Kuriant ir realizuojant duomenų gavybos projektą, remsimės CRISP-DM duomenų proceso modelio standartu. Tai nuoseklus ir išsamus būdas, skirtas duomenų gavybos projekto kūrimui. Šį standartą sudaro šeši pagrindiniai etapai:

- verslo aplinkos supratimas, probleminės srities išskyrimas;
- duomenų ištyrimas – supratimas;
- duomenų paruošimas;
- modeliavimas;
- modeliavimo rezultatų įvertinimas;
- duomenų gavybos projekto įvertinimas.

Laboratorinius darbus sudaro šios pagrindinės struktūrinės dalys:

- teorinė dalis;
- laboratorinio darbo tikslas ir siekiamas rezultatas;
- darbo eiga;
- papildomos savarankiškos užduotys.

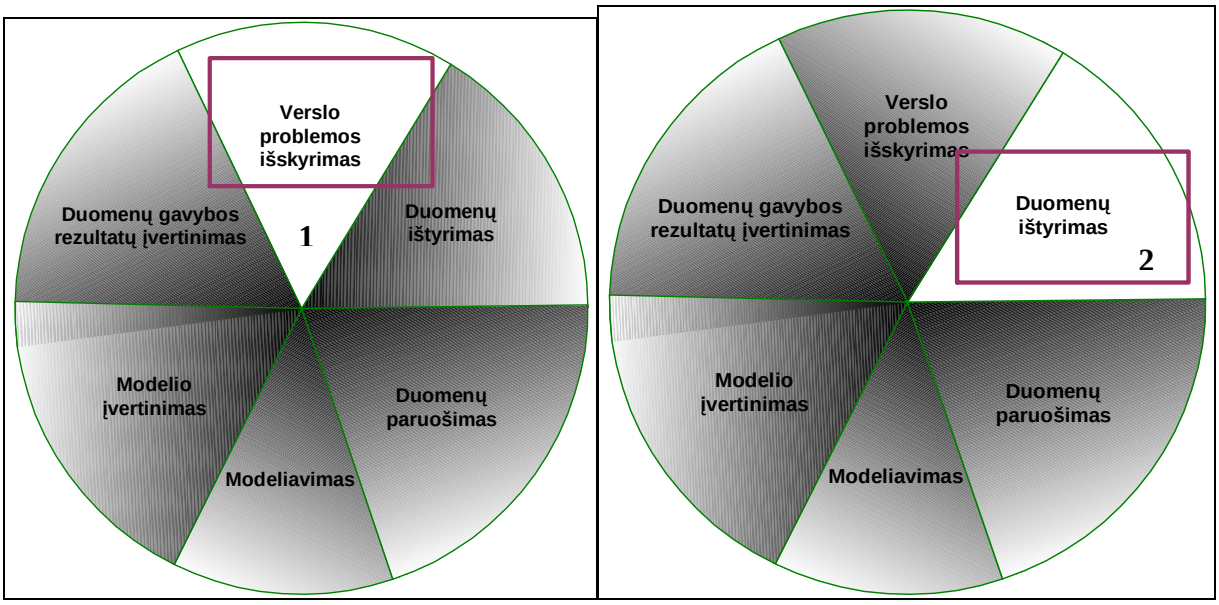
Šių laboratorinių darbų tikslas išdėstyti duomenų gavybos procesą panaudojant tam tikras taikomąsias programas. Sukurtus darbus panaudoti moksliniams tikslams, duomenų gavybos sistemų kurso praktiniams darbams.

3.1. Pirmas laboratorinis darbas – organizacijos verslo aplinkos pažinimas

Teorinė dalis:

Pradedant duomenų gavybos procesą svarbu išanalizuoti verslo poreikius, išskirti pagrindines verslo problemas, numatyti duomenų gavybos poreikius ir išskirti duomenų gavybos tikslus (23 pav.). Dažniausiai organizacijose iškyla šios problemos:

- ko organizacija nori ir kokias problemas mato savo veikloje?
- kokios informacijos trūksta sprendimų priėmimui?
- kokie yra organizacijos darbuotojų lūkesčiai iš duomenų gavybos?



Šaltinis: sudaryta autoriaus.

23 pav. Pirmasis ir antrasis duomenų gavybos proceso etapai

Darbo tikslas:

Pirmojo laboratorinio darbo tikslas – remiantis CRISP – DM standartu, išanalizuoti organizacijos verslo procesus ir išskirti duomenų gavybos tikslus.

Užduotis:

Pirmojo laboratorinio darbo uždaviniai:

- sudaryti tarptautinės organizacijos, užsiimančios prekyba aprašą bei išskirti rinkodaros ir pardavimų funkcijas;
- suformuluoti pagrindines verslo problemas;
- sudaryti ER diagramą, esybių lygmenyje.
- išskirti duomenų gavybos tikslus.

Darbo eiga:

Duomenų gavybos projektui naudosime pavyzdinę MS AdventureWorks DW duomenų bazę. Įmonę galima apibūdinti kaip organizaciją, užsiimančia tarptautine prekyba laisvalaikio prekėmis. Įmonė produktus reklamuoja internetinės svetainės pagalba. Sugalvokite, kokios pagrindinės įmonės valdymo funkcijos, produkcijos grupes, verslo problemas. Sudarykite įmonės aprašymą (24 pav).

Organizacijos pavadinimas: UAB "XXX".
Veiklos apibūdinimas: prekyba laisvalaikio prekėmis, tokiomis kaip dviračiai, klanų slidinėjimo įranga ir t.t.
Pagrindinės valdymo funkcijos:

1) **Bendrasis valdymas.**
   a) Duoda nurodymus finansų, marketingo, transportavimo, personalo valdymo skyriams.
   b) Gauna išsamias ataskaitas ir kitą reikalingą informaciją iš minėtų skyrių.
   c) Priima reikalingus sprendimus bei analizuoja gaunamą informaciją bei darbo veiklos procesus.
   d) Gauna marketingo tyrimo rezultatus ir priima sprendimus bei teikia pasiūlymus.
2) **Marketingas**.
   a) Užsakymų priėmimas ir valdymas.
   b) Reklamos valdymas.
   c) Atliekami rinkos tyrimai, norint išsiaiškinti įmonės padėtį rinkoje, vartotojų poreikius, paslaugų poreikį rinkoje (daugiau vietiniai ar tarptautiniai pervežimai)
   d) Klientų paieška.
3) **Finansų valdymas**.
   a) Kontroliuoja piniginius srautus.
   b) Tvarko buhalterinę apskaitą.
   c) Analizuoja finansinę įmonės būklę.

Pagrindinis veiklos produktas: laisvalaikio prekės: slidinėjimo įranga, dviračiai, lauko teniso įranga.
Numatytos probleminės sritys: naujų potencialių klientų paieška, rinkos analizė, klientų pirkimo elgsenos analizė, produktų poreikio analizė.

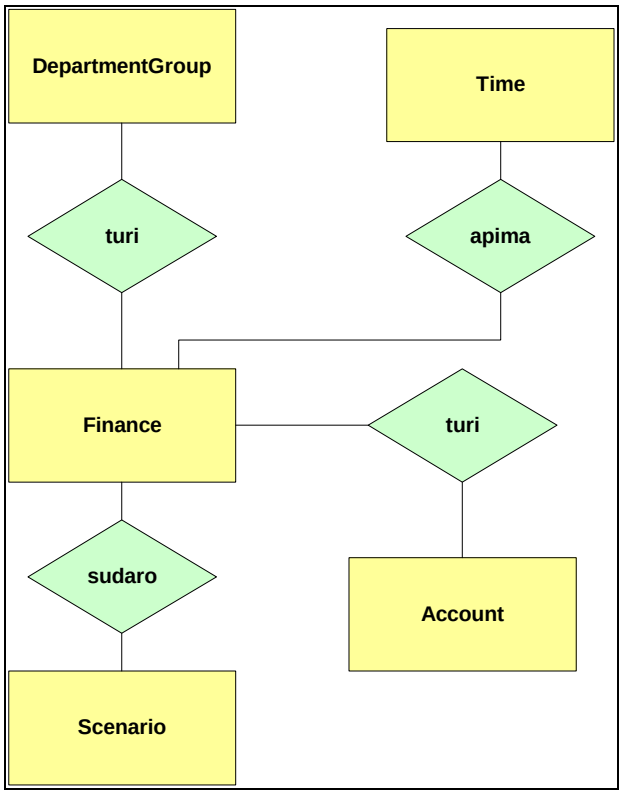
Šaltinis: sudaryta autoriaus.

24 pav. Veiklos aprašas

Įmonės aprašymą gali sudaryti šios pagrindinės dalys:

- organizacijos pavadinimas;
- veiklos apibūdinimas;
- pagrindinės valdymo funkcijos;
- pagrindinis veiklos produktas;
- numatytos probleminės veiklos sritys.

Esybių lygmenyje sudaryti diagramą, kuri atvaizduotų principinę duomenų bazės schemą. Tam naudokite pavyzdinę duomenų bazę MS AdventureWorks DW. 25 paveiksle pateikiamas ER diagramos fragmentas [24].



Šaltinis: sudaryta autoriaus.

25 pav. ER diagramos fragmentas

3.2. Antras laboratorinis darbas – programinės įrangos analizė ir pasiruošimas projektui

Teorinė dalis:

Programinės įrangos analizei naudosime sistemos PATTERN modelį - tikslų medį (The Relevance Tree). Modelis naudojamas norint aiškiai apibrėžti tiriamos problemos (šiuo atveju – programinės įrangos pasirinkimo) tikslų struktūrą ir įvertinti visų potikslių (programinės įrangos kriterijų) santykinę svarbą . Šis tikslų medis sudaromas remiantis konstravimo metodu "iš viršaus žemyn" (top-down).

Santykinės svarbos koeficientai r(i) nustatomi visoms tikslų medžio viršūnėms bei kiekvienam tikslų medžio lygiui atskirai. Santykinės svarbos koeficientai išreiškiami kiekybine išraiška. Tai pagrindinis sistemos PATTERN privalumas. Santykinės svarbos koeficientas įvertina vienos ar kitos programinės įrangos išskirtinumą, kitų to paties lygio priemonių atžvilgiu. Koeficientai parenkami pagal kriterijaus svarbą programinės įrangos pasirinkime. Kriterijai pasirenkami keliais etapais – turais, o kiekvienas turas susideda iš kelių lygmenų. Pirmajame etape nustatomi vertinimo kriterijai A, B, X , …V ir kiekvieno iš kriterijų svoriai Q(x). Antrajame etape

užpildoma 11 lentelė, nurodomas potikslių santykinis svoris S. Viename lygyje esančių viršūnių a, b, …j, ..., n santykinių svorių suma turi būti lygi vienetui (pagal kiekvieną iš kriterijų) [25].

11 lentelė

Kriterijų išskyrimas ir įvertinimas

Kriterijus	**Kriterijaus svoris**	**i-ojo lygio potiksliai**					
		a	**b**	**…**	**j**		**n**
A	***Q (A)***	S(A,a)	S(A,b)		S(A,j)		S(A,n)
B	***Q(B)***	S(B,a)	S(B.b)		S(B,j)		S(B.n)
X	***Q(X)***	S(X,a)			S(X,j)		S(X,n)
V	***Q(V)***	S(V,a)			S(V,j)		S(V,n)
		r(a, i)	**r(b, i)**		**r(j, i)**		**r(n, i)**

Šaltinis: sudaryta prof. S. Gudo (2008).

Baigus apklausos turus, apskaičiuojami potikslių santykinės svarbos koeficientai r(j, i). Tikslų medžio i-ojo lygio potikslio j santykinės svarbos koeficientas r(j,i) yra kriterijaus svorių ir atitinkamo potikslio santykinio svorio S(x,j) sandaugų suma [25].

Darbo tikslas:

Remiantis tikslų analizės sistema PATTERN sudaryti programinės įrangos pasirinkimo analizę. Nors duomenų gavybos proceso realizacijai naudosime MS SQL server 2005 Analysis Services, svarbu mokėti įvertinti programinės įrangos tinkamumą duomenų gavybai.

Užduotis:

- pasirinkti keturis duomenų gavybos programinės įrangos produktus ir atlikti jų analizę, panaudojant PATTERN šabloną.
- įdiegti MS SQL server 2005 Analysis Services ir importuoti pavyzdinę MS AdventureWorks DW duomenų bazę.

Darbo eiga:

Tikslų medžio sudarymui pasirenkame keturis duomenų gavybos produktus, kuriuos vertinsime pagal pasirinktus kriterijus, suteikiant atitinkamus svorius. 12 lentelėje pateikiami trys programiniai produktai, kurie bus vertinami pagal tikslų medį. Kriterijais galima pasirinkti įvairias, programinės įrangos funkcionalumą apibūdinančias savybes:

- vartotojo sąsaja;
- duomenų šaltinio integravimas;

- duomenų formato išsaugojimo pasirinkimas;
- algoritmų pasirinkimas;
- duomenų gavybos rezultatų atvaizdavimas;
- ir kiti svarbūs kriterijai.

12 lentelė

Tikslų medžio struktūra

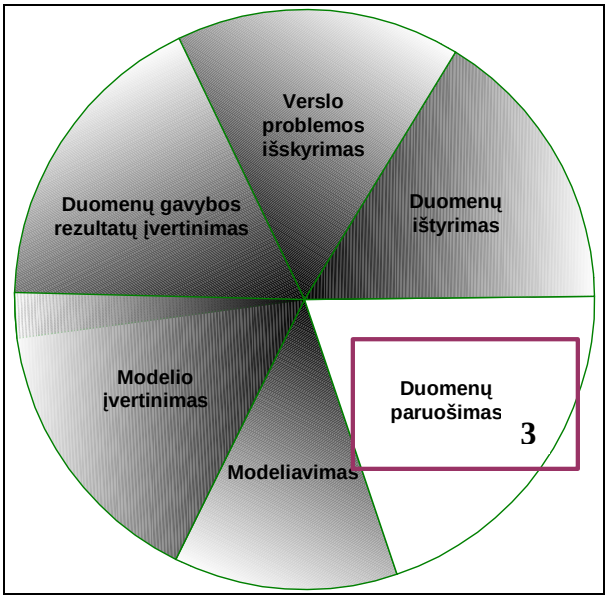
Kriterijus	**Kriterijaus svoris**	**2-ojo lygio potiksliai**		
		Vartotojo sąsaja	**Duomenų šaltinio integravimas**	**Algoritmų pasirinkimas**
Oracle	0,4	0,6	0,3	0,1
Clementine	0,2	0,6	0,1	0,3
MS SQL server 2005	0,4	0,4	0,5	0,5
Santykinės svarbos koeficientai		**0,58**	**0,22**	**0,20**

Šaltinis: sudaryta autoriaus.

3.3. Trečias laboratorinis darbas – duomenų paruošimas

Teorinė dalis:

Duomenų paruošimo etapas gali būti siejamas su duomenų šaltinio nustatymu, duomenų gavybos projekte (26 pav.). Tai vienas svarbiausių duomenų gavybos etapų, nes sėkmingas jo įgyvendinimas, sąlygoja gerus duomenų gavybos rezultatus.



Šaltinis: sudaryta autoriaus.

26 pav. Duomenų paruošimo etapas

Darbo tikslas:

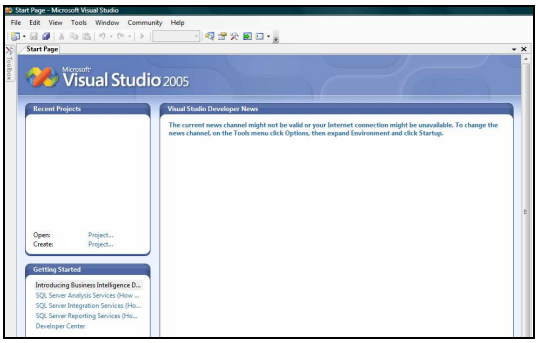
Darbo tikslas yra MS SQL server 2005 Analysis Services programinio produkto pagalba sukurti duomenų gavybos projektą, nurodant duomenų šaltinį, kurio pagrindu bus kuriamas projektas.

Užduotis:

- Duomenų gavybos projekto sukūrimas.

Darbo eiga:

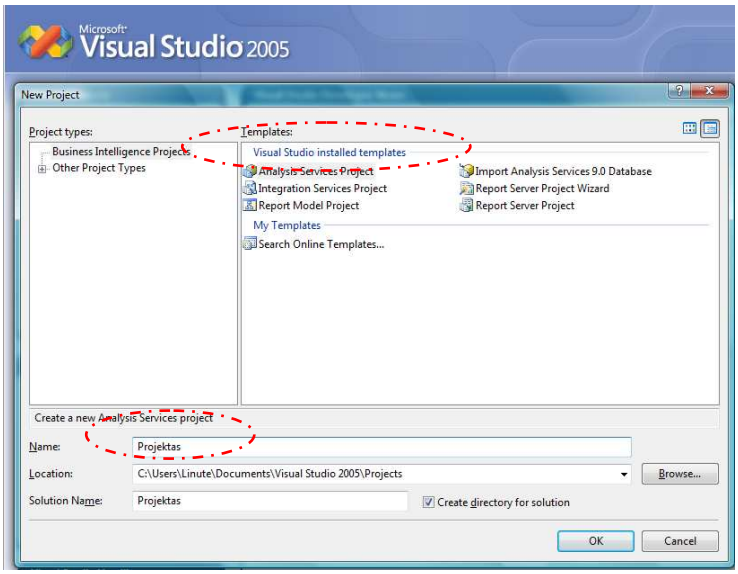
Programos paleidimas: **Start →Programs → Microsoft SQL server 2005 → SQL Server Business Intelligence Development Sturio** (27 pav.).



Šaltinis: sudaryta autoriaus.

27 pav. SQL Server Business Intelligence Development Studio aplinkos langas

Naujam projektui sukurti pasirenkame: **File → New → Project**. Atsivėrusiame lange pakeičiame projekto pavadinimą ir pasirenkame **Analysis Service Project** ruošinio šabloną (28 pav.). Taip pat pasirenkama projekto išsaugojimo vieta.



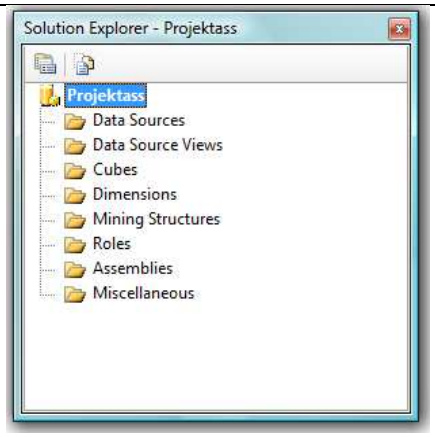
Šaltinis: sudaryta autoriaus.

28 pav. Pradedamo kurti projekto langas

Sukūrus projektą, atsiveria pagrindinis programos langas, kuriame bus vykdomi tolimesni duomenų apdorojimo, paruošimo ir gavybos veiksmai. Duomenų gavybos projekto kūrimui, galima pasinaudoti 13 lentelėje pateiktomis funkcijomis.

13 lentelė

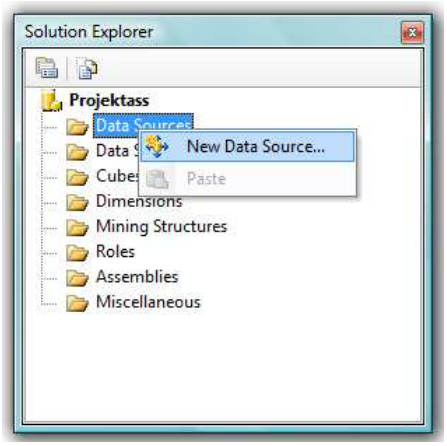
Projekto kūrimo funkcijos

Grafinis vaizdas	Funkcijos apibūdinimas
	• **Projekto pavadinimas;** • **Data Source – duomenų šaltinis;** • **Data Source Views – duomenų šaltinio peržiūra;** • **Cubes – duomenų kubas;** • **Dimensions – dimensijų kūrimas;** • **Mining Structures – duomenų gavybos struktūros;** • **Roles – vartotojams suteikiamos teises;**

Šaltinis: sudaryta autoriaus.

Duomenų šaltinio įkėlimas į projekto aplinką:

Susikūrus projektą, tolimesniems veiksmams reikalingas duomenų šaltinis. Duomenų šaltinį galima priskirti iškvietus komandą **New Data Source** arba pasirinkus **Project → New Data Source** (29 pav.).

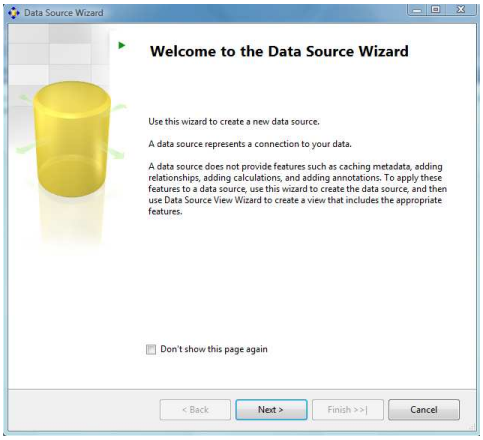


Šaltinis: sudaryta autoriaus.

29 pav. Duomenų šaltinio įkėlimas

Analysis Service suteikia vartotojui galimybę pasinaudoti pagalba vartotojui. Duomenų vedlys – tai pagalbinė priemonė, skirta padėti vartotojui teisingai atlikti duomenų gavybos projekto veiksmus. Kiekvienas projekto kūrimo žingsnis paaiškintas vartotojui. Šio vedlio pagalba sukursime duomenų gavybos projekto šabloną, nurodysime jungtis į duomenų šaltinį bei įgyvendinsime trečiąjį duomenų gavybos proceso etapą – duomenų paruošimą.

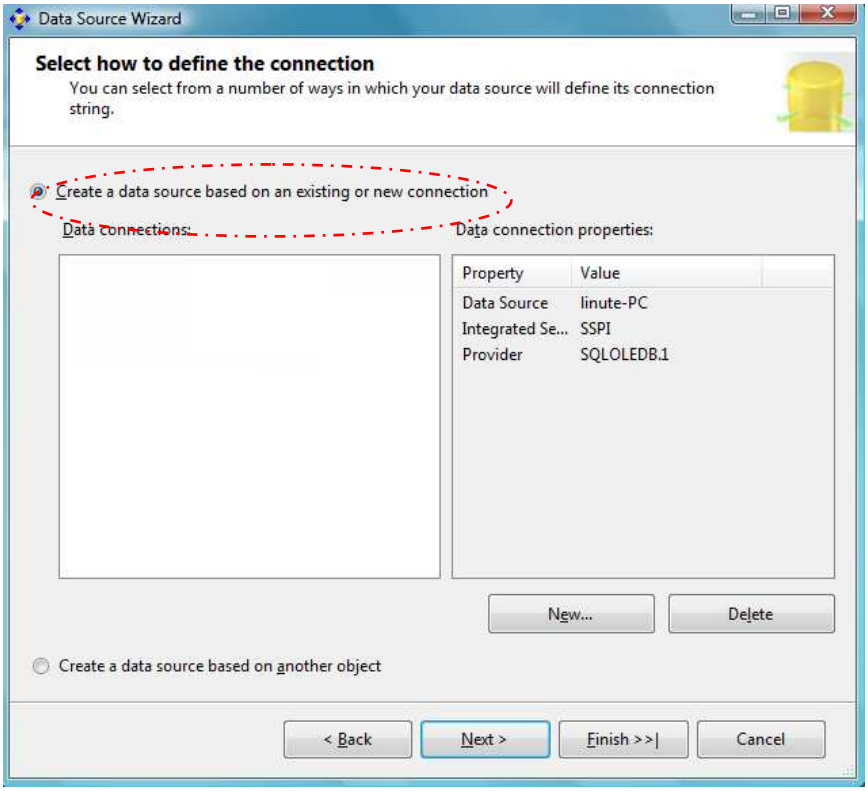
Pasirinkus **New Data Source** atsiveria programos vedlys (30 pav.), kuriame reikia pasirinkti konkretų duomenų šaltinį. Spaudžiame **Next**.



Šaltinis: sudaryta autoriaus.

30 pav. Programos vedlio langas

Pasirenkame **Create a data source based on an existing or new connection** (31 pav.). Yra galimybė pasinaudoti ir kitame projekte naudojamu duomenų šaltiniu (pasirinkus **Create a data source based on another object**). Spaudžiame **New**.



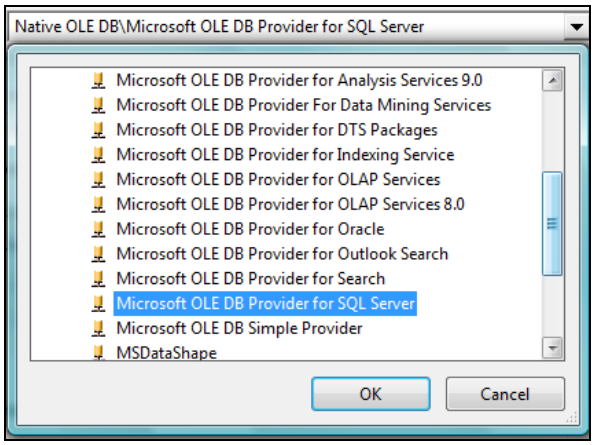
Šaltinis: sudaryta autoriaus.

31 pav. Duomenų šaltinio įkėlimas į programinę aplinką

MS SQL leidžia prisijungti prie įvairių duomenų šaltinių. Galime nurodyti duomenų bazę esančią MS SQL server aplinkoje arba duomenų šaltinį, turintį ODBC jungtį. Šiuo atveju

naudojame pavyzdinę Microsoft duomenų bazę **AdventureWorks DW**, kuri yra MS SQL serverio aplinkoje. Prie duomenų bazės jungsimės per **Native OLE DB** platformą (32 pav.).

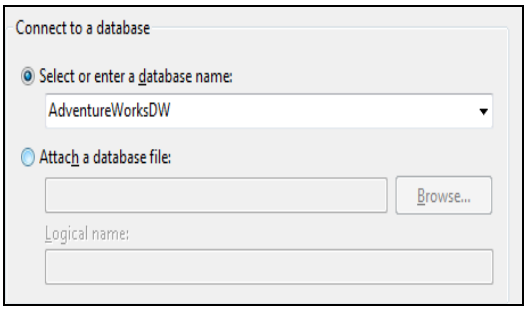
Pirmiausiai pasirenkame duomenų bazės prijungimui reikalingą platformą:



Šaltinis: sudaryta autoriaus.

32 pav. Native OLE DB platformos pasirinkimas

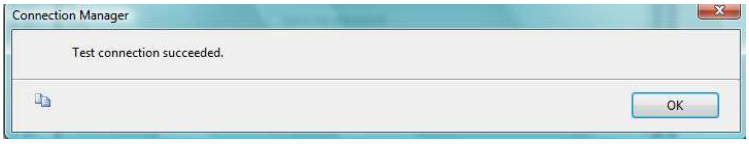
Pasirinkamus tinkamą platformą, toliau nurodome duomenų bazę prie kurios reikia prisijungti. Pasirenkame duomenų šaltinį, kadangi naudosime MS AdventureWorks DW duomenų bazę, nurodome jos vardą (33 pav.).



Šaltinis: sudaryta autoriaus.

33 pav. Nurodoma duomenų bazė

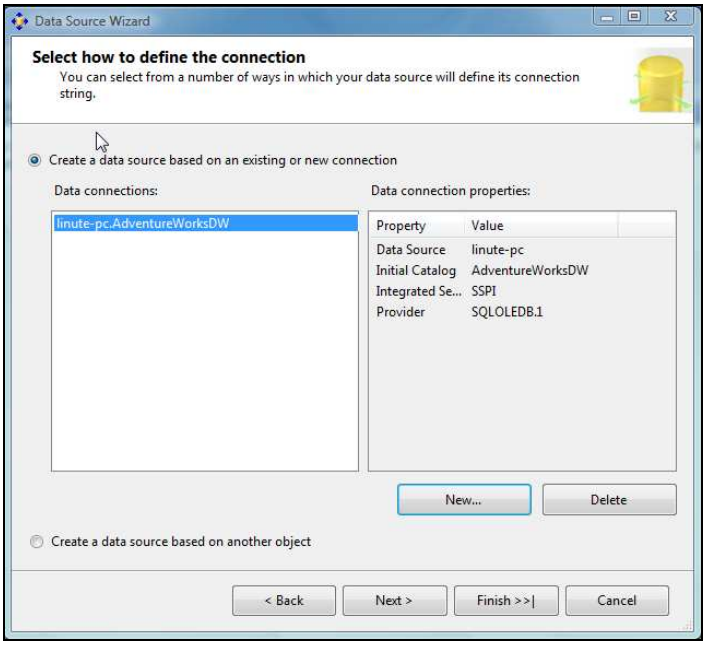
Nurodžius visus reikalingus parametrus galima atlikti jungimosi testą, jam atlikti pasirenkame **Test Connection**. Tinkamai atliktus jungties į duomenų šaltinį veiksmus, išmetamas pranešimas **Test connection succeeded** (34 pav.). Neteisingai atlikus duomenų šaltinio jungties nustatymo veiksmus, išmetamas pranešimas **Test connection failed**.



Šaltinis: sudaryta autoriaus.

34 pav. Jungties patikrinimas

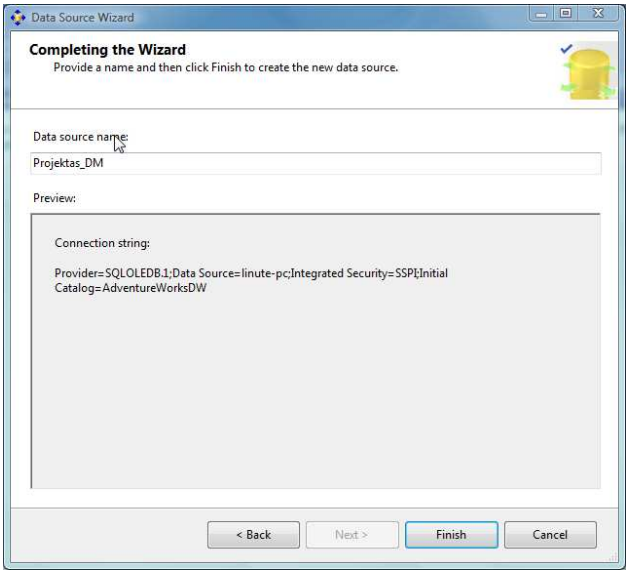
Sukuriamas kreipimasis į duomenų šaltinį (35 pav.).



Šaltinis: sudaryta autoriaus.

35 pav. Naujos jungties į duomenų šaltinį sukūrimas

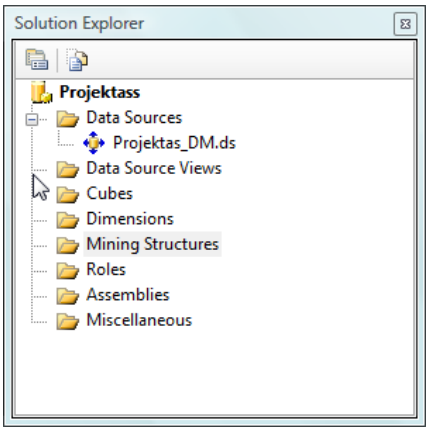
Susiformavus jungčiai į duomenų šaltinį spaudžiame **Next**, toliau pasirenkame prisijungimo tipą - **Use the Service Account**, spaudžiame vėl **Next**. Suteikiame duomenų šaltiniui **Projektas_DM** (paaišk. *Projektas data mining*) pavadinimą ir užbaigiame procesą **Finish** mygtuku (36 pav.). Taip sukuriamas duomenų gavybos projekto ruošinys, turintis kreiptį į konkretų duomenų šaltinį.



Šaltinis: sudaryta autoriaus.

36 pav. Pavadinimo sukūrimas

Tinkamai atlikus visus anksčiau nurodytus veiksmus ir sėkmingai įvykdžius duomenų šaltinio įkėlimą, programinėje aplinkoje matome 37 paveiksle pateiktą vaizdą.

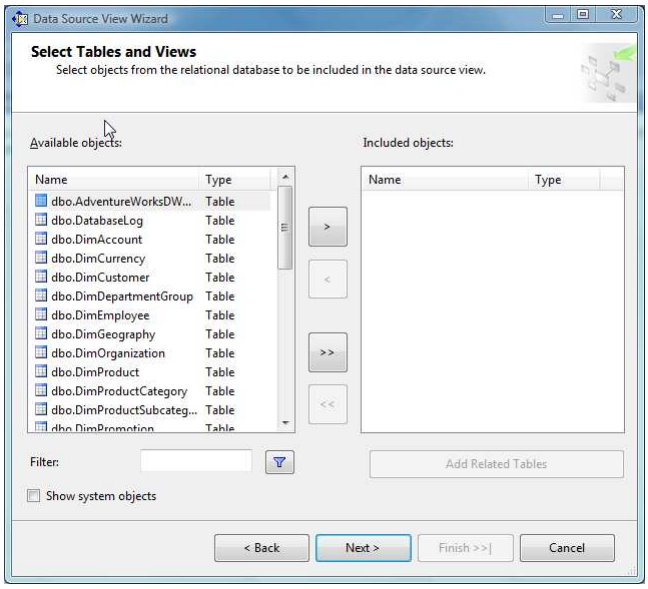


Šaltinis: sudaryta autoriaus.

37 pav. Sėkmingas duomenų šaltinio įkėlimas

Duomenų šaltinio įkėlimas į projekto aplinką:

Tolimesnis projekto kūrimo etapas – tai duomenų šaltinio peržiūros failo sukūrimas (Data Source Views). Dešinio pelės klavišo pagalba pasirenkama komanda **New Data Source View.** Atsiveria programos vedlio langas, kurio pagalba pasirenkame objektą (šiuo atveju sukurtą duomenų gavybos projektą - Projektas_DM), spaudžiame **Next** ir atsiveria langas, kuriame reikia pasirinkti duomenų bazės lenteles reikalingas peržiūrai (38 pav.).

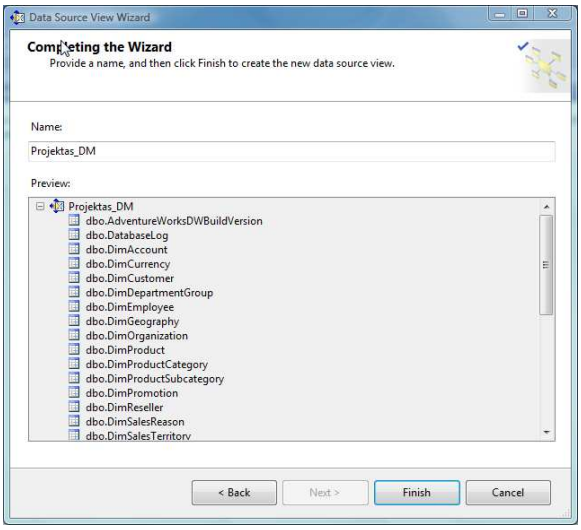


Šaltinis: sudaryta autoriaus.

38 pav. Duomenų bazių lentelių įkėlimas į peržiūros aplinką

Pasirenkame visas reikalingas MS AdventureWorks DW duomenų bazės lenteles, spaudžiame **Next** ir atsiveria langas, kuriame matome visas pasirinktas DB lenteles. Suteikiame

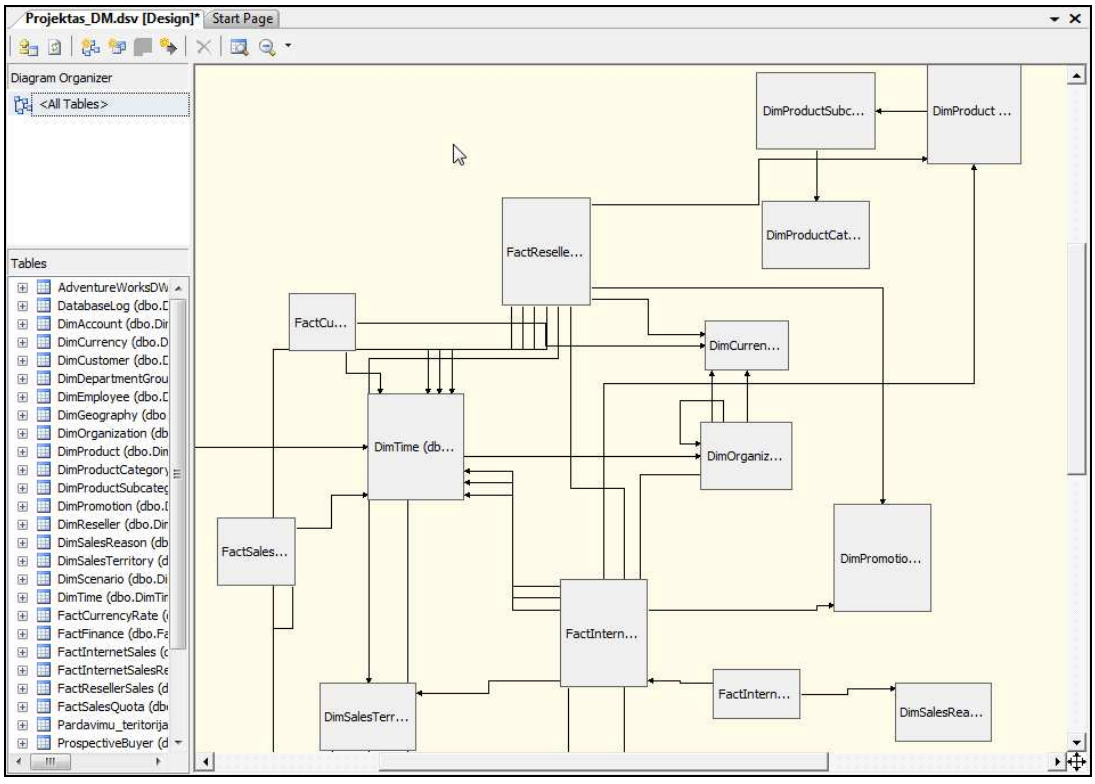
peržiūros failui pavadinimą (paliekame tą patį, kuris buvo suteiktas duomenų gavybos projektui - **Projektas_DM**) (39 pav.).



Šaltinis: sudaryta autoriaus.

39 pav. Paskutinis duomenų šaltinio peržiūros sukūrimo etapas

Įrašius duomenų šaltinio peržiūros pavadinimą spaudžiame **Finish**. Jeigu sėkmingai sukūrėte peržiūros failą, matome jį programinėje aplinkoje (40 pav.).



Šaltinis: sudaryta autoriaus.

40 pav. Duomenų šaltinio peržiūra

3.4. Ketvirtas laboratorinis darbas – dimensijų reikalingų duomenų gavybai kūrimas

Teorinė dalis:

MS SQL Analysis Services suteikia galimybę sukurti šiuos dimensijų tipus: viešąsias ir lokalias dimensijas. Pagrindinis šių dimensijų skirtumas, kad lokalios dimensijos yra sukuriamas kiekvienam duomenų kubui atskirai, tuo tarpu viešosios dimensijos gali būti naudojamos keliems kubams, taip sutaupoma laiko ir vietos. Pakanka nurodyti kreipinį į reikalingą duomenų kubui dimensiją.

Pirmiausia sukursime standartines ir pagrindines dimensijas, kurios bus reikalingos tolimesnei projekto vykdymo eigai.

Dimensijos – tai duomenų ląstelės, išdėstytos pagal koordinačių ašis. Dimensijų pagalba formuojamas duomenų kubas, kuris naudojamas duomenų analizei ar duomenų gavybai. Dimensijos gali būti kelių rūšių:

- laiko dimensija;
- standartinė dimensija;
- serverio laiko dimensija.

Taip pat dimensijos gali būti skirstomos pagal dimensijų tipus, kurios charakterizuoja dimensijos paskirtį ir atvaizduojančius elementus:

- organizacijos dimensija;
- produkto dimensija;
- reklamos dimensija;
- skaičiuojamoji dimensija;
- laiko dimensija;
- standartinė dimensija.

Darbo tikslas:

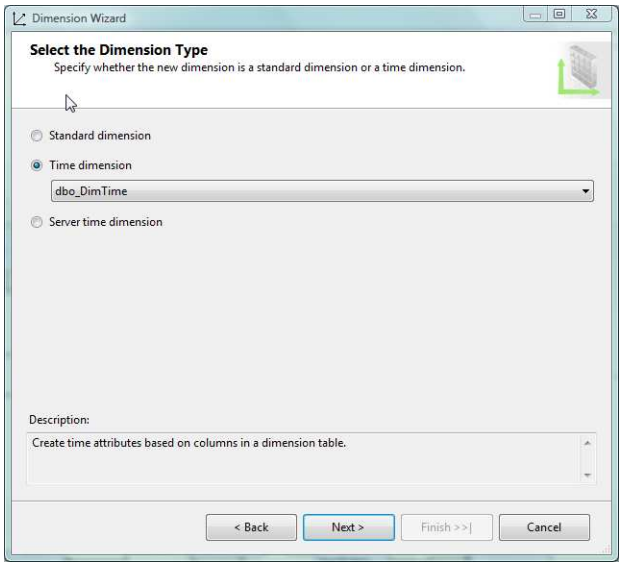
Sukurti dimensijas, kurios bus reikalingos duomenų gavybai ir suprasti duomenų išsidėstymą koordinačių ašyje.

Užduotis:

Ketvirtojo laboratorinio darbo užduotis yra sudaryti laiko dimensiją bei savarankiškai sukurti pardavimų internetinėje aplinkoje, prekybos tinklo dimensijas.

Laiko dimensijos kūrimas:

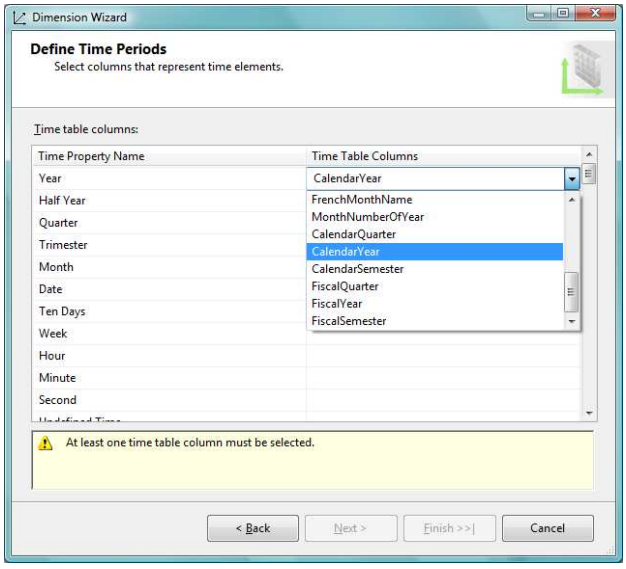
Dimensijų kūrimui pasirenkame **Project → New dimension** arba dešinio pelės klavišo pagalba pasirenkama **New dimension**. Atsiveria programos vedlio langas, kuriame prašoma pasirinkti dimensijos tipą (41 pav.). Kadangi kuriame laiko dimensiją, todėl pasirenkame **Time dimension** tipą bei reikiamą duomenų bazės lentelę.



Šaltinis: sudaryta autoriaus.

41 pav. Laiko dimensijos pasirinkimas

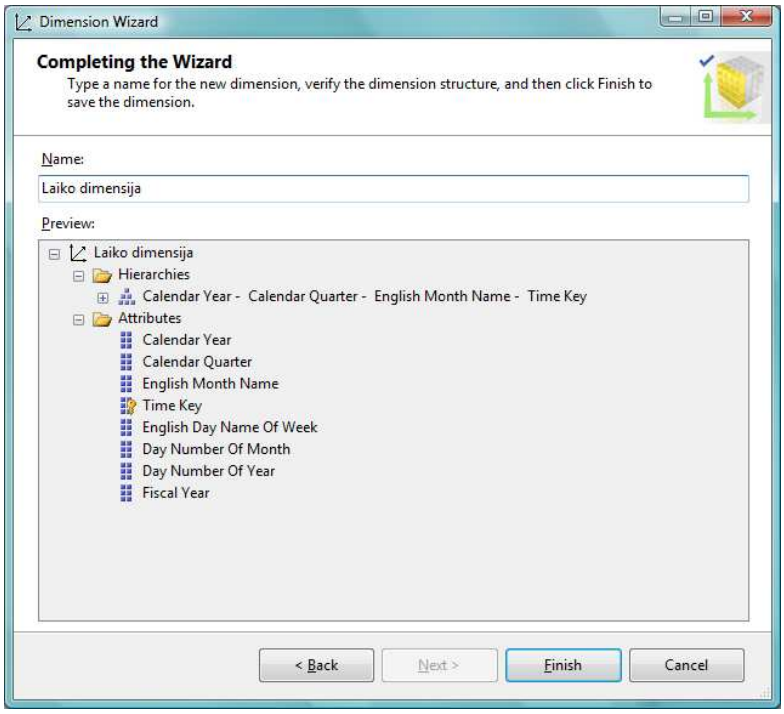
Pasirinkę reikalingus atributus, spaudžiame **Next**. Atsiveria programos langas, kuriame reikia lentelės atributams priskirti atitinkamas laiko reikšmes, pagal kurias bus vykdomi duomenų peržiūros pjūviai. Pagal pasirinktas laiko reikšmes bus automatiškai sugeneruotos duomenų hierarchijos (42 pav.).



Šaltinis: sudaryta autoriaus.

42 pav. Laiko hierarchijų pasirinkimas

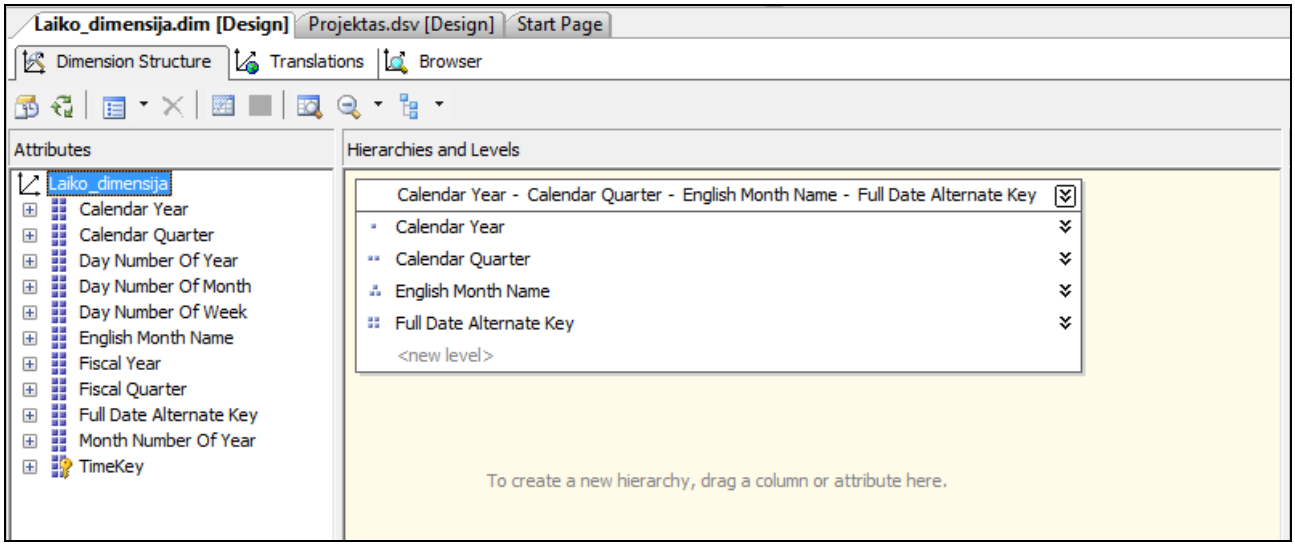
Priskyrus laiko periodus konkrečioms duomenų reikšmėms, gauname hierarchinę laiko dimensijos struktūrą ir ją pavadiname Laiko dimensija arba Dim Time (43 pav.).



Šaltinis: sudaryta autoriaus.

43 pav. Dimensijos sukūrimas

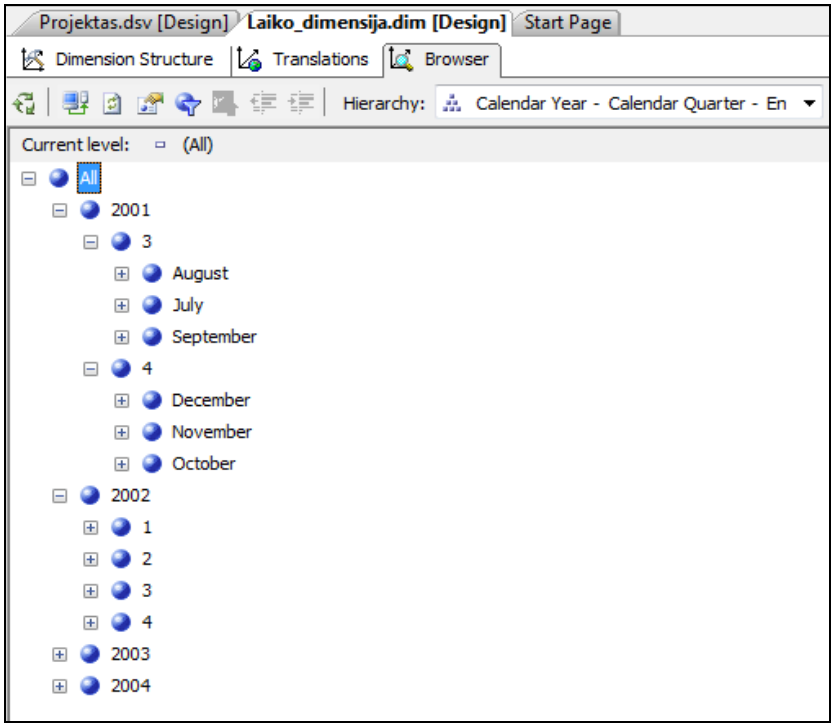
Suformuotą laiko dimensiją matome projekto lange (44 pav.). Norint vykdyti duomenų peržiūrėjimą, reikia apdoroti sukurtą dimensiją. Duomenų apdorojimo metu vyksta reikalingų duomenų atranka lentelėse ir sudaroma hierarchinė išdėstymo struktūra. Apdorojimo procesui įvykdyti paspaudžiame **Process**.



Šaltinis: sudaryta autoriaus.

44 pav. Laiko dimensijos hierarchijos lygmenys

Įvykdžius apdorojimo procesą ir pasirinkus peržiūros langą **Browser**, matome suformuotą laiko hierarchiją. Šiame lange galime matyti laikotarpių išsidėstymą, atlikti duomenų filtravimą. Gaunamas laiko medis, kurio viršūnėse metai, medžio šakose – ketvirčiai, kurie smulkiau išdėstomi į mėnesius (45 pav.). Galima pasirinkti ir kitokį laiko hierarchinį išdėstymą.



Šaltinis: sudaryta autoriaus.

45 pav. Laiko dimensijos struktūros atvaizdavimas

Savarankiška užduotis:

- sukurti pardavimų internete, prekybos tinklo dimensijas.

3.5. Penktas laboratorinis darbas – duomenų gavybos modelio pritaikymas ir panaudojimas

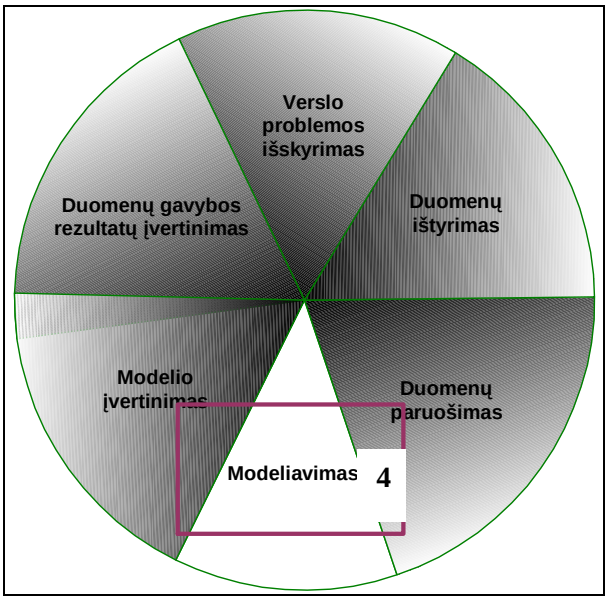
Teorinė dalis:

MS SQL server 2005 Analysis Services leidžia vartotojui duomenų gavybai naudoti šiuos algoritmus:

- Microsoft Decision Trees Algorithm – sprendimų medžio algoritmas;
- Microsoft Clustering Algorithm – grupavimo algoritmas;
- Microsoft Naive Bayes Algorithm – Naive Bayes algoritmas;
- Microsoft Association Algorithm – asociacijų taisyklių algoritmas;
- Microsoft Sequence Clustering Algorithm – sekų grupavimo algoritmas;
- Microsoft Time Series Algorithm – laiko eilučių algoritmas;
- Microsoft Neural Network Algorithm (SSAS) – neuroninių tinklų algoritmas;

- Microsoft Logistic Regression Algorithm – logistinės regresijos algoritmas;
- Microsoft Linear Regression Algorithm – tiesinės regresijos algoritmas.

Pagal CRISP-DM standartą – tai ketvirtasis duomenų gavybos etapas. Šis etapas svarbus, nes nuo jo priklauso duomenų gavybos rezultatų ir viso projekto sėkmė. Pateiktame 46 paveiksle parodyta šio etapo vieta visame procese.



Šaltinis: sudaryta autoriaus.

46 pav. Modeliavimas

Nustačius duomenų gavybos tikslus ir juos įvertinus, reikia pasirinkti duomenų gavybos modelį. Kiekvienas modelis sprendžia tam tikro tipo uždavinius, kurie pateikti 14 lentelėje.

14 lentelė

Duomenų gavybos algoritmai verslo problemų sprendimui

Užduotis	**Duomenų gavybos algoritmas**
Pavienių elementų požymių numatymui. Šių algoritmų pagalba gali būti sprendžiami tokie uždaviniai, kaip išsiaiškinti ar elektroninės reklamos gavėjai pirks produktus, perskaitę reklamą ir susipažinę su įmonės produkcija.	Microsoft Decision Trees Algorithm Microsoft Naive Bayes Algorithm Microsoft Clustering Algorithm Microsoft Neural Network Algorithm (SSAS)
Nenutrūkstamų procesų prognozavimui. Šiais algoritmais gali būti bandoma nuspėti ateinančių metų pardavimus.	Microsoft Decision Trees Algorithm Microsoft Time Series Algorithm
Numatyti požymių sekas. Galima analizuoti įmonės internetinės svetainės lankomumą.	Microsoft Sequence Clustering Algorithm
Nustatyti tikslines grupes, kurios daugiausia naudojasi konkrečia paslauga ar perka tam tikrą produktų grupę. Duomenų gavyboje siūloma analizuoti rinką, nusprendžiant, kokius produktus pasiūlyti vartotojams.	Microsoft Association Algorithm Microsoft Decision Trees Algorithm
Nustatyti panašius dėsningumus, priklausomybes tarp grupių. Tai gali būti analizuojama geografinė priklausomybė vartotojų elgsenai.	Microsoft Clustering Algorithm Microsoft Sequence Clustering Algorithm

Pagrindinės sąvokos, reikalingos duomenų gavybos procesui formuoti:

- **Case** - pagrindinė duomenų gavybos modelio esybė (lentelė), kurios pagrindu formuojamas modelis. Ją charakterizuoja atributai, detaliau apibūdinantys informacijos sudėtį;
- **Attributes** - atributai gali skirtis nuo modelio struktūros tipo, tai gali būti raktas, įvedimo laukas, prognozuojama reikšmė bei įvedimo ir prognozavimo reikšmė kartu;
- **Nested case** - papildoma esybė, reikalinga duomenų gavybos modeliui. Tai gali būti papildoma informacijos lentelė, papildanti pagrindinės informacijos turinį;
- **Prediction** - prognozuojama reikšmė.

Darbo tikslas:

Penktojo laboratorinio darbo tikslas yra sudaryti dviejų tipų duomenų gavybos modelius, panaudojant šiuos algoritmus: sprendimų medžio ir grupavimo.

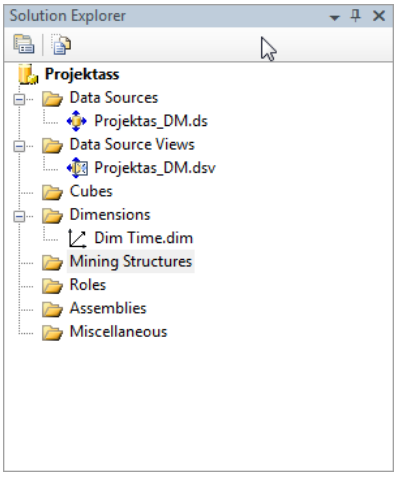
Užduotis:

Darbo tikslui įgyvendinti iškelti ir suformuoti šie uždaviniai:

- Sprendimų medžio modelio formavimas. Organizacijos rinkodaros skyrius nori nustatyti pirkėjų savybes, nuo kurių priklausys produktų pirkimas ateityje. Duomenų bazėje saugoma demografinė informacija, apibūdinanti kiekvieną klientą. Panaudokite MS Desicion Tree algoritmą pirkėjų elgsenos savybėms nustatyti.
- Grupavimo modelio formavimas. Sugrupuoti klientus, kurie gyvena vienoje geografinėje zonoje bei perka vienos grupės produktus internetinėje svetainėje.

Darbo eiga:

Duomenų gavybos modelio struktūros formavimas vykdomas pasirinkus **Mining Structures → New Mining Structures** (47 pav.). MS SQL server 2005 Analysis Services suteikia galimybę atrasti naujus ryšius tarp duomenų algoritmų pagalba.



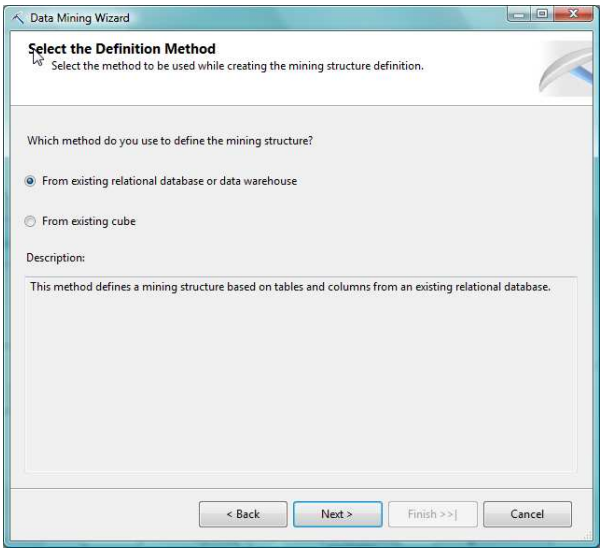
Šaltinis: sudaryta autoriaus.

47 pav. Duomenų gavybos struktūros pasirinkimas

Sprendimų medžio algoritmo pritaikymas duomenų gavybai:

Atsiradusiame programos vedlio lange pasirenkame duomenų šaltinio vietą. **Analysis Services** leidžia duomenų gavybos struktūras kurti pasinaudojant duomenų sandėlio ar reliacinės duomenų bazės duomenimis bei sukurto OLAP kubo duomenimis (48 pav.).

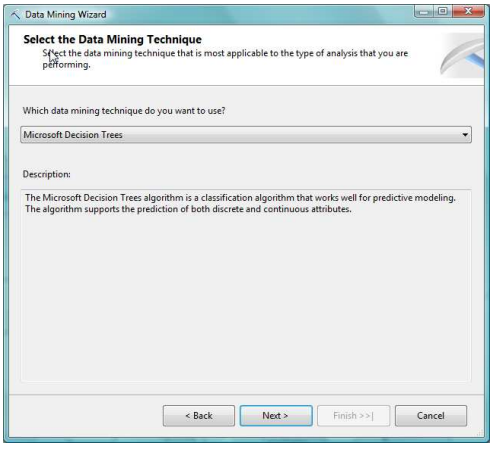
Šiuo atveju naudosime duomenų bazės šaltinį. Toliau spaudžiame **Next**.



Šaltinis: sudaryta autoriaus.

48 pav. Duomenų šaltinio pasirinkimas

Sekančiame žingsnyje pasirenkame duomenų gavybos algoritmą, kurio pagalba bus apdorojami duomenys. Pasirenkame Microsoft Decision Trees - sprendimų medžio algoritmas, kuris atlieka klasifikavimo funkcijas, ir spaudžiame **Next** (49 pav.). Atsivėrusiame **Select Data Source Views** lange matome duomenų šaltinį ir jo duomenų lenteles. Spaudžiame **Next**.

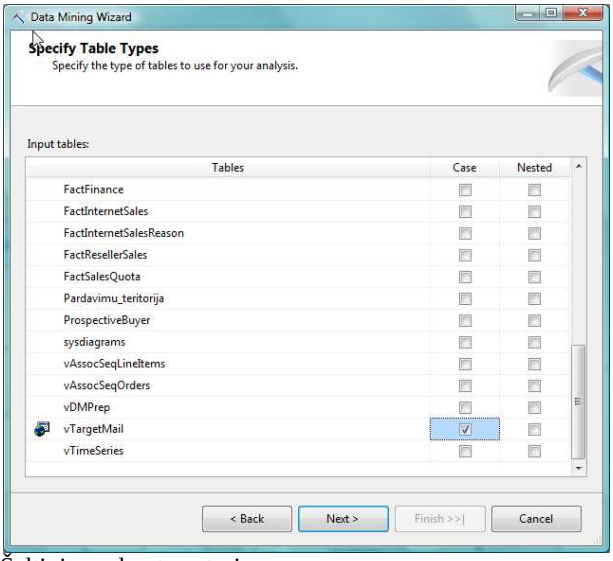


Šaltinis: sudaryta autoriaus.

49 pav. Duomenų gavybos algoritmo pasirinkimas

Iš **Specify Table Types** pasirenkame duomenų lentelę, kurioje bus ieškoma pasislėpusių sąryšių tarp duomenų (50 pav.). Šiuo atveju žiūrėsime tikslines elektroninių laiškų siuntimo grupes.

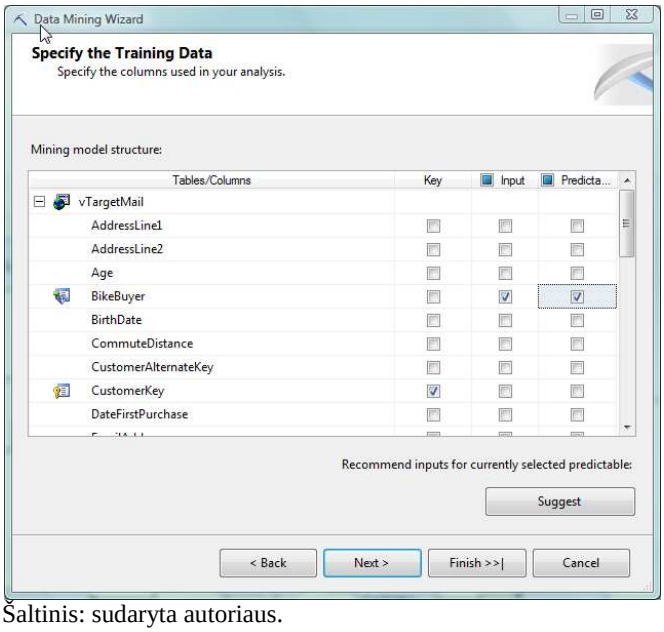
Tikslas yra išsiaiškinti, kokie potencialūs klientai gali būti surasti elektroninių laiškų pagalba, siunčiant reklaminius skelbimus. Iš lentelių sąrašo pasirenkame **vTargetMail** duomenų lentelę. Spaudžiame **Next**.



Šaltinis: sudaryta autoriaus.

50 pav. Lentelės reikalingos duomenų gavybai pasirinkimas

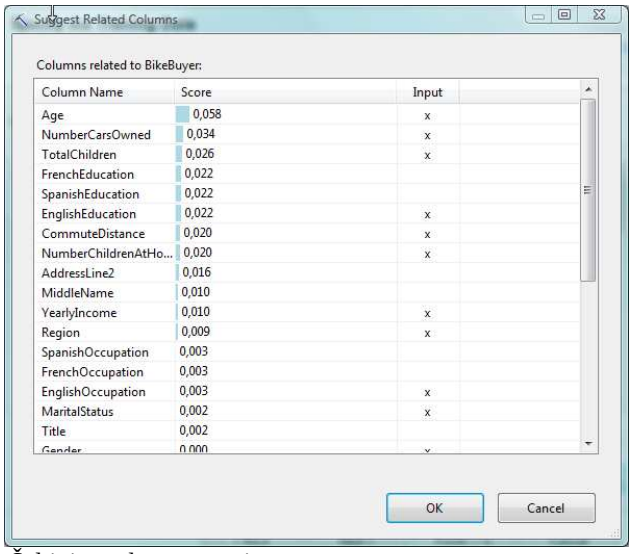
Sekantis žingsnis, nurodyti pirminį raktą, kuris dažniausiai automatiškai generuojamas, pagal pasirinktos lentelės pirminį raktą. Šiuo atveju pasirenkame **CustomerKey** (51 pav.). Toliau svarbu pasirinkti įvedamąsias reikšmes. Pasirinktu modeliu tirsime, koks ryšys tarp potencialių dviračių pirkėjų ir reklaminės kampanijos dalyvių. Todėl **Input** ir **Predictable** pasirenkame **BikeBuyer**.



Šaltinis: sudaryta autoriaus.

51 pav. Raktinių ir reikšminių laukų pasirinkimas

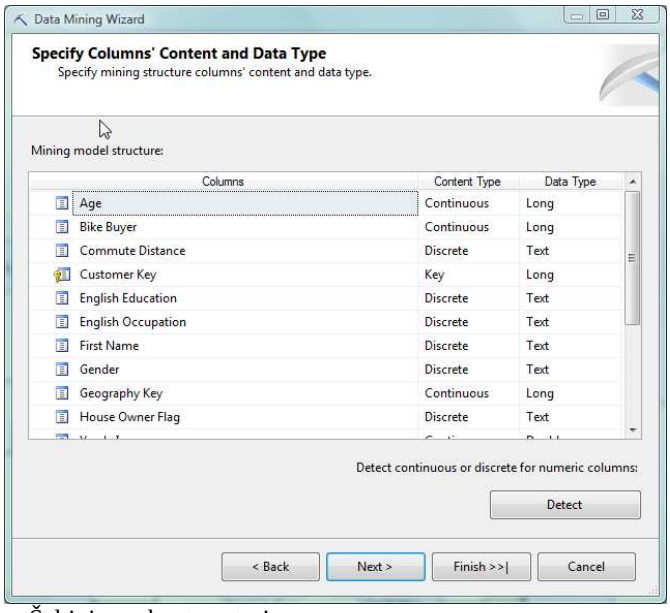
Sekančiame žingsnyje patikriname, ar nėra raktų pasirinkime klaidų. Kadangi sistema patikrina galimus nesutapimus, todėl spaudžiame **Suggest**. **Suggest Related Columns** lange pasirenkama duomenų analizei svarbius parametrus ir pažymime **Inputs** reikšmėmis. Spaudžiame **OK** (52 pav.).



Šaltinis: sudaryta autoriaus.

52 pav. Reikalingų reikšmių modeliui pasirinkimas

Specify Column‘s Content and Data Type lentelėje spaudžiame **Detect** (53 pav.). Šia komanda algoritmas patikrina ar stulpelyje duomenys turi nepertraukiamas, ar pavienes reikšmes.

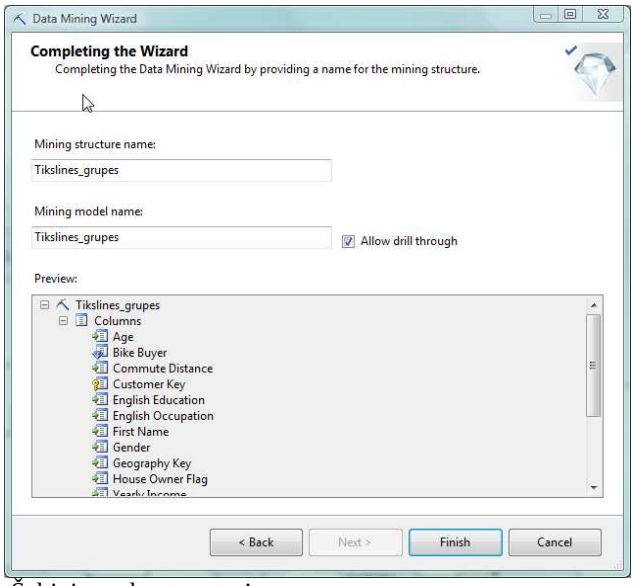


Šaltinis: sudaryta autoriaus.

53 pav. Duomenų tipų ir turinio tipo patikrinimas

Paskutiniame lange pakeičiame duomenų gavybos modelio ir struktūros pavadinimus. Pavadinimus geriausiai suteikti, atsižvelgiant į uždavinio sąlygą, nes tarp gausybės modelių

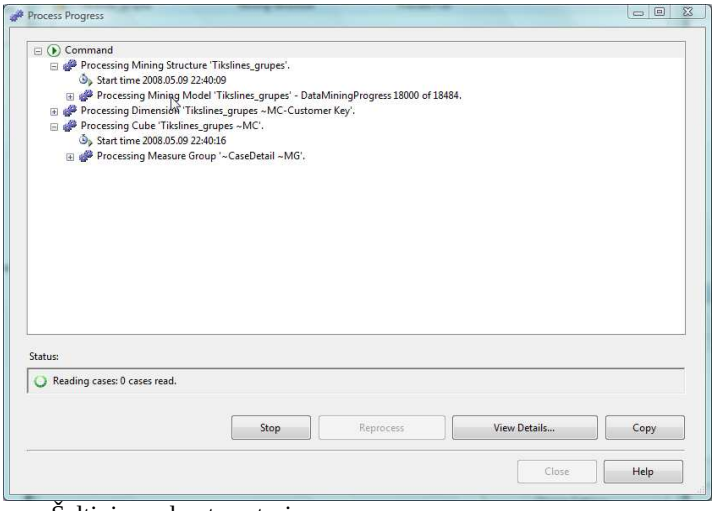
lengviau atrasti reikalingą modelį. Šiuo atveju pasirenkame **Tikslines_grupes** pavadinimą (54 pav.). Taip pat pažymime galimybę apmokyti duomenis, tam pasirenkame **Allow drill through**.



Šaltinis: sudaryta autoriaus.

54 pav. Pavadinimo suteikimas gavybos modeliui

Duomenų gavybos modelio kūrimo užbaigimui spaudžiame **Finish**. Sėkmingai sukūrus modelį, programinėje aplinkoje atsiveria langas su modelio struktūra. Norint peržiūrėti gautus rezultatus, pirmiausiai reikia patikrinti sukurtą duomenų gavybos modelį. Klaidų tikrinimui pasirenkame **Database → Process** (55 pav.)**.**

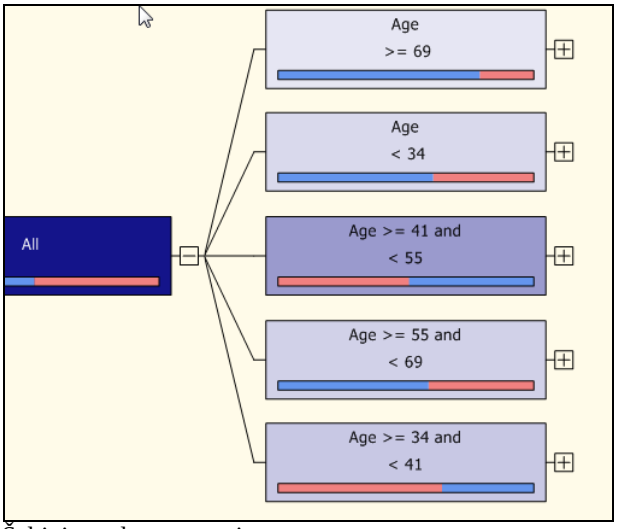


Šaltinis: sudaryta autoriaus.

55 pav. Klaidų patikrinimas

Patikrinus sukurto modelio klaidas galime toliau vykdyti gautų rezultatų peržiūrą. Sprendimų medis pateikia pirkėjų grupių pasiskirstymą pagal amžių. Suklasifikuoti duomenys pateikiami medžio struktūra. Pirkėjai pagal amžiaus grupes suskirstyti į penkis segmentus (medžio šakas) (56 pav.). Tolimesnei analizei galima papildomai išskleisti medžio struktūrą ir pamatyti

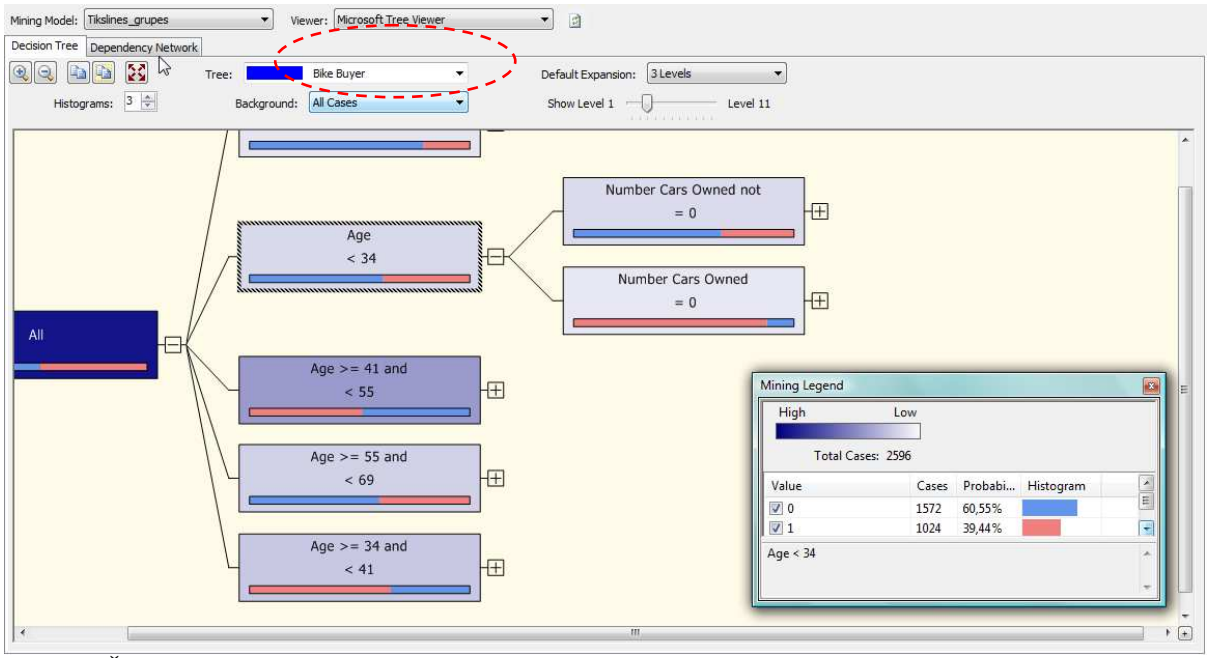
kitus, svarbius kriterijus, lemiančius reklaminės kompanijos sėkmę ir produkcijos apyvartos didėjimą.



Šaltinis: sudaryta autoriaus.

56 pav. Duomenų gavybos rezultatas

Tikslesnei duomenų analizei išskleidžiamas sprendimų medis ir pagal gautus rezultatus galima teigti, kad didžiausią dėmesį reklaminiams skelbimams teikia jaunesnio amžiaus asmenys, kurie turi mažiau nei 3 mašinas, augina šiuo metu vaikučius ir atstumas tarp darbo ir namų ir mažesnis. Šie asmenys yra potencialūs dviračių pirkėjai (57 pav.).

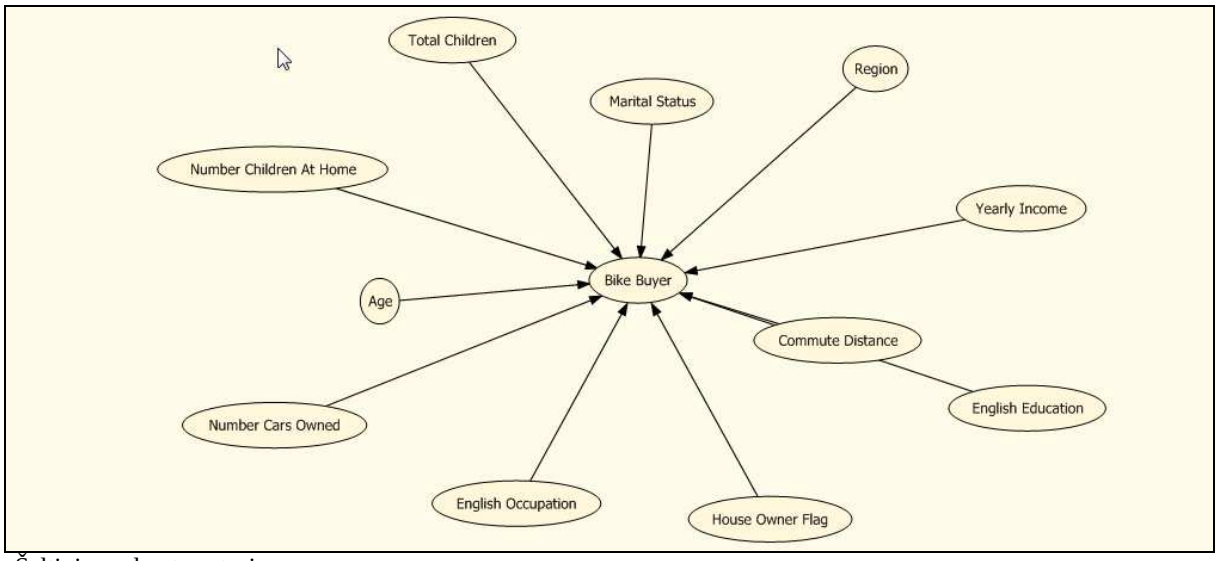


Šaltinis: sudaryta autoriaus.

57 pav. Duomenų gavybos modelio siūlomų funkcijų peržiūra

Priklausomybių tinkle matome, kokie kriterijai turi didžiausią įtaką pirkėjui. 58 paveiksle pateiktame grafike, didžiausią įtaką turi atstumas tarp namų ir darbo, amžius, šeimyninė padėtis.

Likusieji kriterijai turi mažesnę įtaką pirkėjo elgsenai. Sprendimų medžio algoritmo pagalba, sugrupuojami pirkėjai pagal šiuo kriterijus ir priklausomybių tinkle, pateikiamas šių kriterijų svarbumas, vertinant potencialius dviračių pirkėjus.

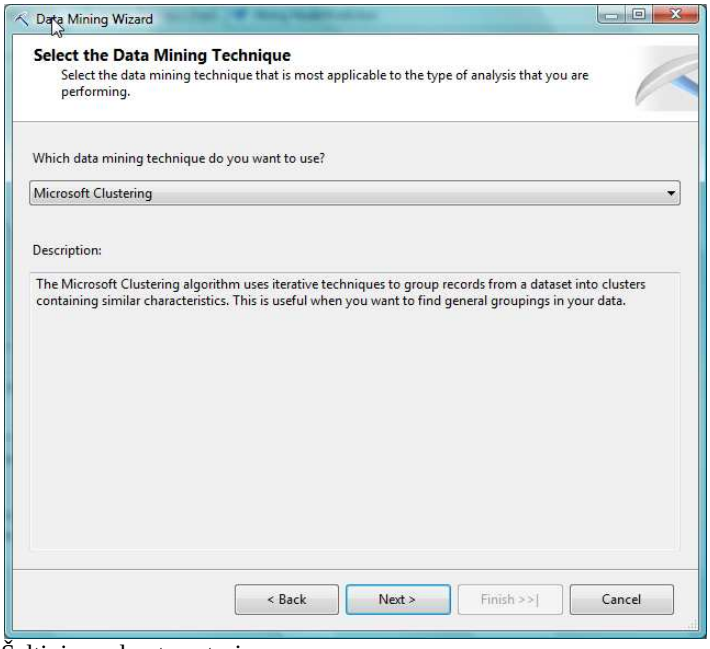


Šaltinis: sudaryta autoriaus.

58 pav. Priklausomybių tinklas

Grupavimo algoritmo pritaikymas duomenų gavybai:

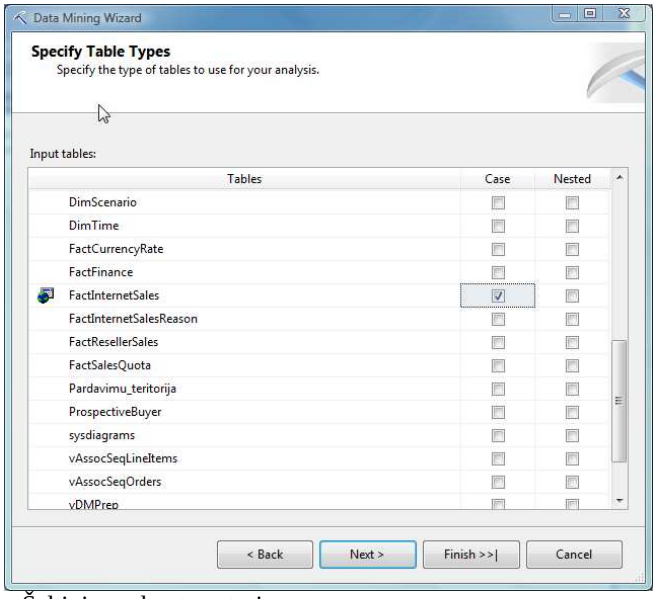
Tolimesnei duomenų gavybos analizei, panaudosime grupavimo algoritmą. Atsiradusiame programos vedlio lange pasirenkame duomenų šaltinio vietą bei duomenų gavybos algoritmą. Šiuo atveju pasirenkame Microsoft Clustering metodą (59 pav.).



Šaltinis: sudaryta autoriaus.

59 pav. Grupavimo algoritmo pasirinkimas

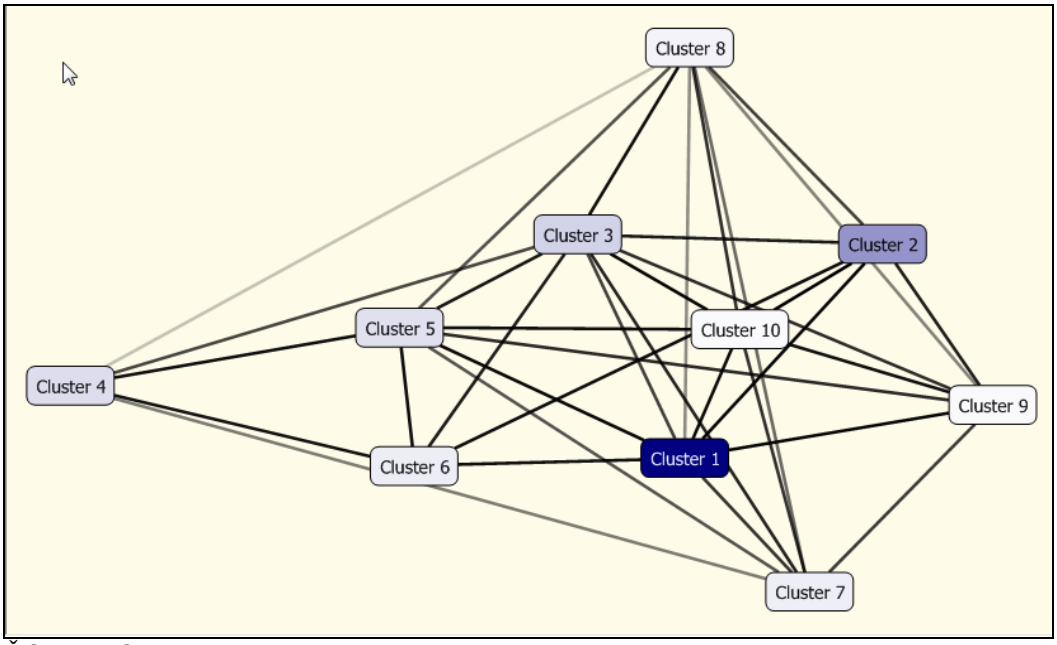
Iš **Specify Table Types** pasirenkame **InternetSales** duomenų lentelę (60 pav.). Ši lentelė pasirenkama, norint išanalizuoti internetinėje svetainėje perkamų produktų pirkėjų grupes pagal geografines zonas. Spaudžiame **Next**.



Šaltinis: sudaryta autoriaus.

60 pav. Lentelės pasirinkimas

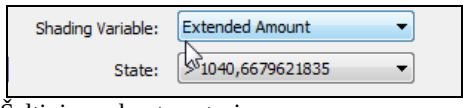
Tolimesniuose etapuose atliekame tuos pačius žingsnius, kaip ir prieš tai esančio modelio kūrime. Turite gauti panašų modelį, kuris pavaizduotas 61 paveiksle. Jame matomos sugrupuotos klientų grupės pagal geografines zonas. Pasirenkant klasterius (geografines zonas), parodomas jo išsidėstymas tinkle.



Šaltinis: sudaryta autoriaus.

61 pav. Grupavimo algoritmo modelis

Norint valdyti ir sukioti grupavimo algoritmą, 62 paveiksle pavaizduotos pasirinkimo galimybės. Vartotojas gali stebėti sudarytą grupių modelį pasirinkdamas vis kitus parametrus: gyvenamosios vietos regioną, išleistą produktų užsakymams sumą ir t.t. Šių funkcijų pagalba galima analizuoti, kurio regiono gyventojai daugiau naudojasi elektronine prekyba, kurie išleidžia didesnę pinigų sumą užsakymams.



Šaltinis: sudaryta autoriaus.

62 pav. Pasirinkimo langas

3.6. Šeštas laboratorinis darbas – duomenų gavybos modelių plėtojimas

Darbo tikslas:

Darbo tikslas išanalizuoti Naive Bayes ir Asociacijų taisyklių algoritmus bei jų sudaromus duomenų gavybos modelius. Savarankiškai suformuoti laiko eilučių algoritmu paremtą duomenų gavybos struktūrą ir išsiaiškinti gavybos rezultatus.

Užduotis:

- Naive Bayes modelio formavimas. Rinkodaros skyrius pageidauja išanalizuoti elektroninės kampanijos sėkmę ir įvertinti, kokie klientai yra potencialūs pirkėjai. Taip pat išsiaiškinti, kokie pagrindiniai klientų elgesio motyvai yra svarbiausi, reaguojant į elektroninę reklamą. Šį uždavinį įgyvendinti analizuojant dviračių pirkėjų duomenis.
- Asociacijų taisyklių modelio formavimas. Organizacija pageidauja išsiaiškinti internetinės paslaugų užsakymo svetainės funkcionalumą bei nori suformuoti produktų grupes, kurios yra paklausios tarp pirkėjų.
- Savarankiškai suformuoti Laiko eilučių modelį bei įvertinti gavybos rezultatus.

Darbo eiga:

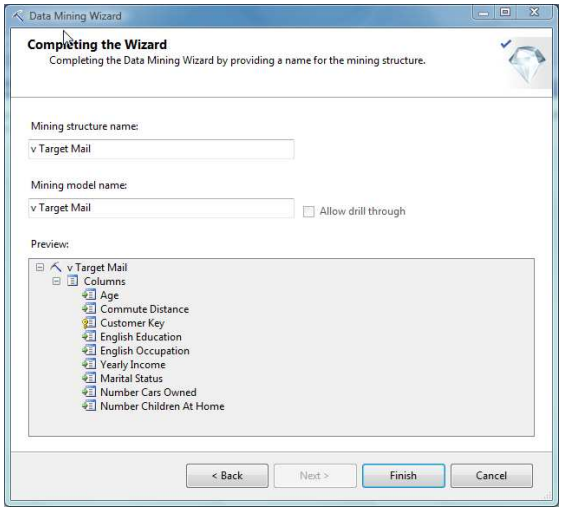
Naive Bayes algoritmo pritaikymas duomenų gavybai

Modeliui formuoti pasirenkame Naive Bayes duomenų gavybos algoritmą. Duomenų gavybos modelis kuriamas **Target Mail** duomenų lentelės pagalba, kurią pažymime **Specify Table Types** lentelėje. Pasirenkama laukus, kurie svarbūs duomenų gavybos rezultatui, tai:

- Age - amžius;
- Commute distance - atstumas;
- English education - išsilavinimas;

- English occupation - profesija;
- Marital status – šeimyninė padėtis;
- Number Cars Owned – mašinų skaičius šeimoje;
- Number Children At home – vaikų skaičius namuose.

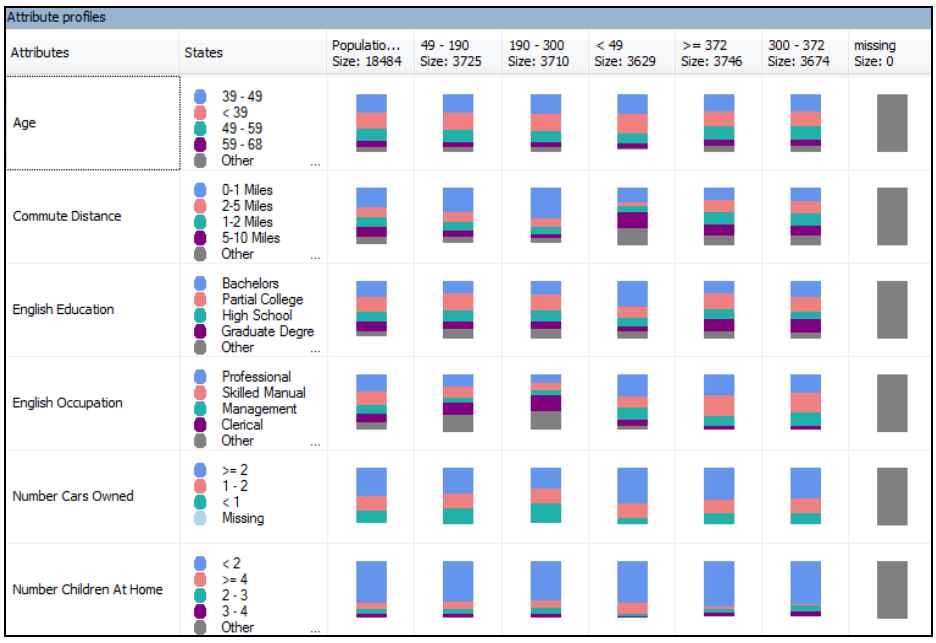
Pasirinkus visus modelio kūrimui reikalingus kriterijus, modelį išsaugome pavadinimu **Target Mail** (63 pav.).



Šaltinis: sudaryta autoriaus.

63 pav. Modelio išsaugojimas

Suformuojama Naive Bayes modelio struktūra, kurios pagalba galima matyti klientų populiacijos išsidėstymą pagal pasirinktus kriterijus. Matome, kokio amžiaus klientai domisi organizacijos produkcija, pagrindinius klientų pirkimo įpročius lemiančius kriterijus: vaikų skaičius, mašinų kiekis, atstumas nuo namų iki darbo, išsilavinimas ir kiti kriterijai (64 pav.).



Šaltinis: sudaryta autoriaus.

64 pav. Naive Bayes modelio rezultatai

Šio modelio pagalba suformuojamas priklausomybių sąrašas nuo tam tikrų pastebėtų elgsenos motyvų. Pagal pateiktą 65 paveikslą matome, kad dviračių pirkimą įtakoja vaikų, esančių namuose, skaičius. Todėl galima teigti, kad tėvai auginantys vaikučius namuose, dažniau perka dviračius nei tėvai, kurių vaikai užaugę. Visi kiti pirkėjų elgsenos kriterijai išdėstomi mažėjimo tvarka. Įvertinkite visus galimus kriterijus, sąlygojančius dviračių pirkimo elgseną.

Characteristics for < 49

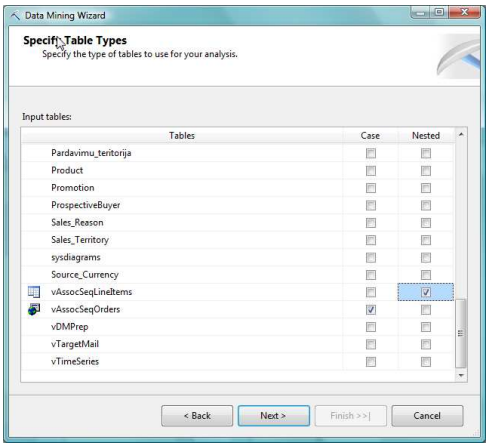
Attributes	Values	Probability
Number Children At Home	< 2	
Number Cars Owned	>= 2	
English Education	Bachelors	
English Occupation	Professional	
Age	39 - 49	
Age	< 39	
Commute Distance	5-10 Miles	
Number Cars Owned	1 - 2	
Commute Distance	0-1 Miles	
Commute Distance	10+ Miles	
English Occupation	Management	
Number Children At Home	>= 4	
English Education	Partial College	
English Occupation	Skilled Manual	
Age	49 - 59	
English Education	High School	
English Occupation	Clerical	
Commute Distance	1-2 Miles	
Number Cars Owned	< 1	
Age	59 - 68	
English Education	Graduate Degree	
Commute Distance	2-5 Miles	
English Education	Partial High School	
English Occupation	Manual	

Šaltinis: sudaryta autoriaus.

65 pav. Priklausomybių sąrašas

Asociacijų taisyklių algoritmo pritaikymas duomenų gavybai:

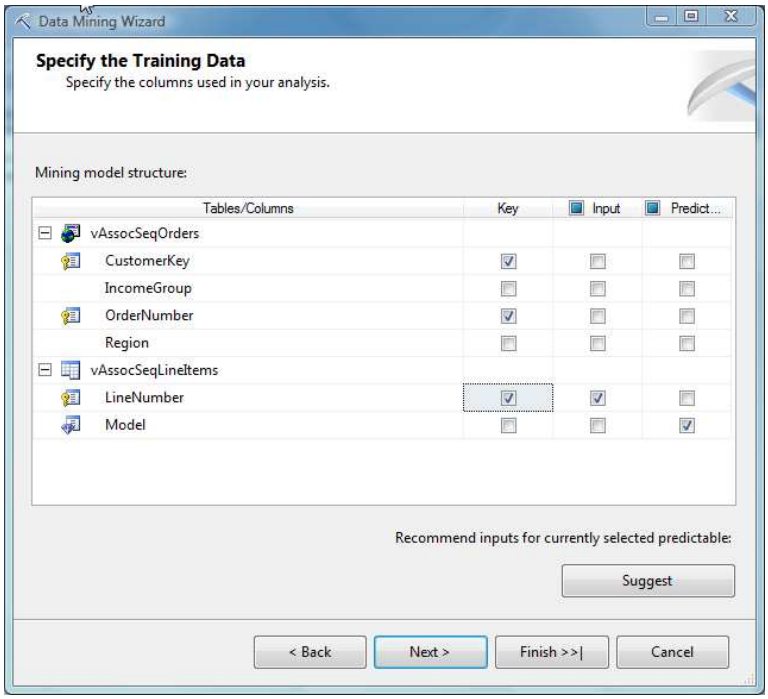
Modeliui formuoti pasirenkame Asociacijų taisyklių duomenų gavybos algoritmą. Duomenų gavybos modelis bus formuojamas **Specify Table** pasirinkus lenteles, kurios bus naudojamos analizuojant užsakytų prekių grupes (66 pav.).



Šaltinis: sudaryta autoriaus.

66 pav. Lentelių pasirinkimas

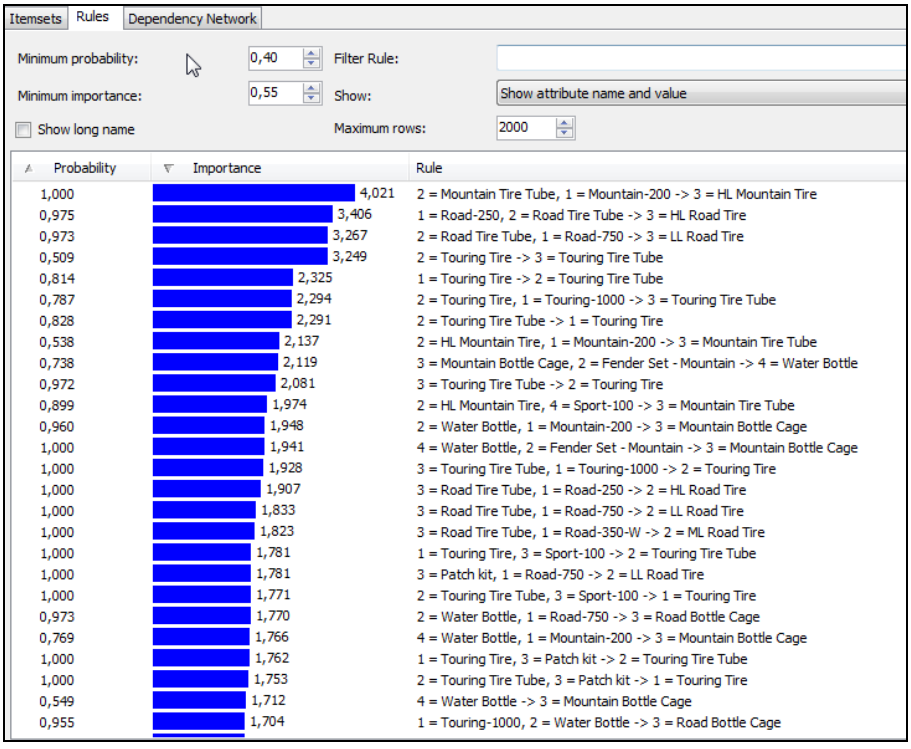
Pasirenkamas **Model** elementas, pagal jį bus kuriamos duomenų gavybos taisyklės (67 pav.).



Šaltinis: sudaryta autoriaus.

67 pav. Modelio struktūra

Pasirinkus visus parametrus sugeneruojamos taisyklės, pagal kurias galima spręsti produktų grupių vartojimo svarbą ir kuriems produktams pirkėjai teikia pirmenybę (68 pav.).

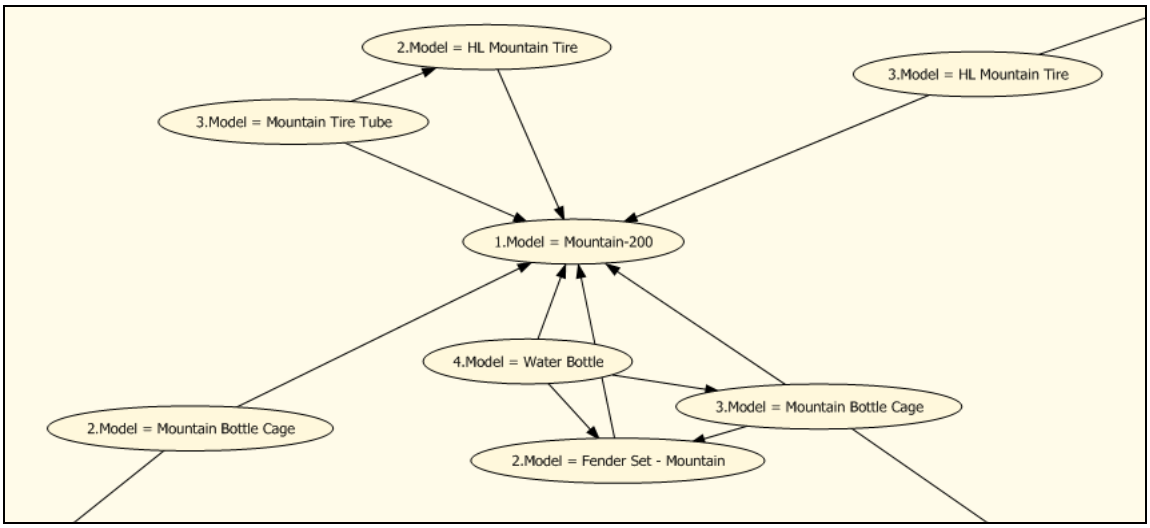


Probability	Importance	Rule
1,000	4,021	2 = Mountain Tire Tube, 1 = Mountain-200 -> 3 = HL Mountain Tire
0,975	3,406	1 = Road-250, 2 = Road Tire Tube -> 3 = HL Road Tire
0,973	3,267	2 = Road Tire Tube, 1 = Road-750 -> 3 = LL Road Tire
0,509	3,249	2 = Touring Tire -> 3 = Touring Tire Tube
0,814	2,325	1 = Touring Tire -> 2 = Touring Tire Tube
0,787	2,294	2 = Touring Tire, 1 = Touring-1000 -> 3 = Touring Tire Tube
0,828	2,291	2 = Touring Tire Tube -> 1 = Touring Tire
0,538	2,137	2 = HL Mountain Tire, 1 = Mountain-200 -> 3 = Mountain Tire Tube
0,738	2,119	3 = Mountain Bottle Cage, 2 = Fender Set - Mountain -> 4 = Water Bottle
0,972	2,081	3 = Touring Tire Tube -> 2 = Touring Tire
0,899	1,974	2 = HL Mountain Tire, 4 = Sport-100 -> 3 = Mountain Tire Tube
0,960	1,948	2 = Water Bottle, 1 = Mountain-200 -> 3 = Mountain Bottle Cage
1,000	1,941	4 = Water Bottle, 2 = Fender Set - Mountain -> 3 = Mountain Bottle Cage
1,000	1,928	3 = Touring Tire Tube, 1 = Touring-1000 -> 2 = Touring Tire
1,000	1,907	3 = Road Tire Tube, 1 = Road-250 -> 2 = HL Road Tire
1,000	1,833	3 = Road Tire Tube, 1 = Road-750 -> 2 = LL Road Tire
1,000	1,823	3 = Road Tire Tube, 1 = Road-350-W -> 2 = ML Road Tire
1,000	1,781	1 = Touring Tire, 3 = Sport-100 -> 2 = Touring Tire Tube
1,000	1,781	3 = Patch kit, 1 = Road-750 -> 2 = LL Road Tire
1,000	1,771	2 = Touring Tire Tube, 3 = Sport-100 -> 1 = Touring Tire
0,973	1,770	2 = Water Bottle, 1 = Road-750 -> 3 = Road Bottle Cage
0,769	1,766	4 = Water Bottle, 1 = Mountain-200 -> 3 = Mountain Bottle Cage
1,000	1,762	1 = Touring Tire, 3 = Patch kit -> 2 = Touring Tire Tube
1,000	1,753	2 = Touring Tire Tube, 3 = Patch kit -> 1 = Touring Tire
0,549	1,712	4 = Water Bottle -> 3 = Mountain Bottle Cage
0,955	1,704	1 = Touring-1000, 2 = Water Bottle -> 3 = Road Bottle Cage

Šaltinis: sudaryta autoriaus.

68 pav. Asociacijų taisyklės

Taip pat galima pamatyti priklausomybių tinklą, kuris atvaizduoja produkto modelio priklausomybę nuo kito modelio (69 pav.). Pagal šiuos duomenis galime nuspręsti, kokios produktų grupės turi įtakos pirkėjų elgesiui.

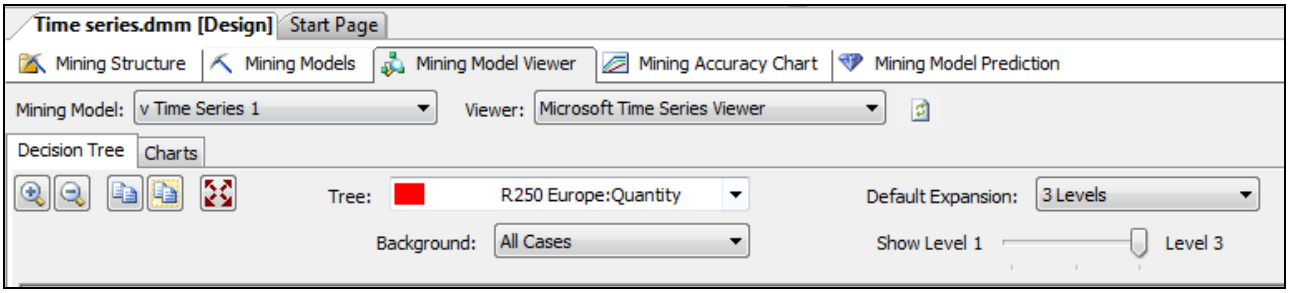


Šaltinis: sudaryta autoriaus.

69 pav. Asociacijų priklausomybių tinklas

Savarankiška užduotis:

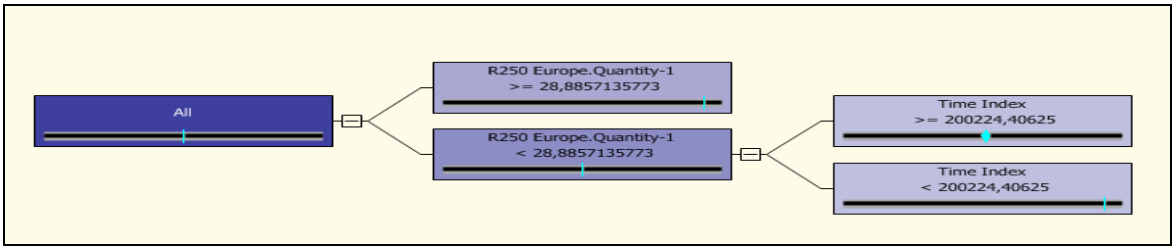
Suformuoti Laiko eilučių modelį, panaudojant Time Series duomenų lentelę. Turite gauti Laiko eilučių duomenų gavybos struktūrą ir šio modelio aplinką (70 pav.).



Šaltinis: sudaryta autoriaus.

70 pav. Laiko eilučių modelio aplinka

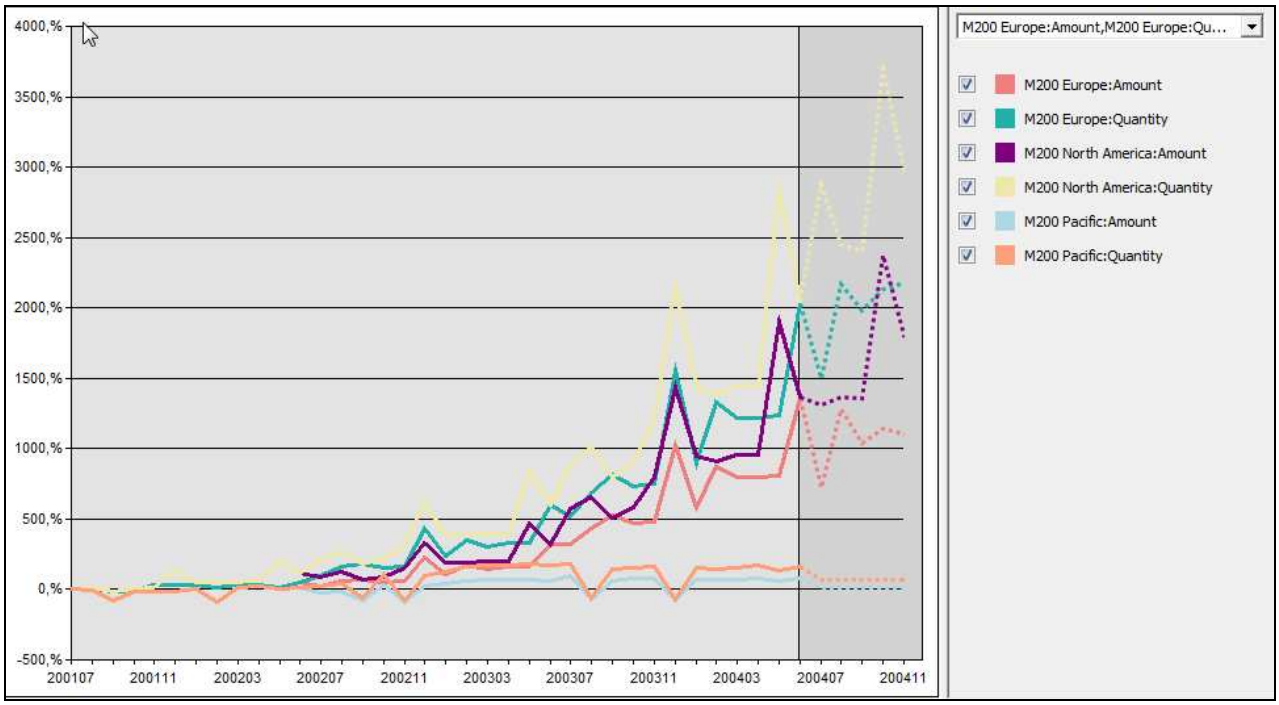
Suformuojamas panašus medis kaip Sprendimų medžio modelyje (71 pav.).



Šaltinis: sudaryta autoriaus.

71 pav. Laiko eilučių modelio rezultatas

Papildomai prie šio gavybos modelio pateikiami grafikai, parodantys produkcijos kokybės ir sumų išsidėstymą pagal regionus, tam tikru laiko momentu (72 pav.).



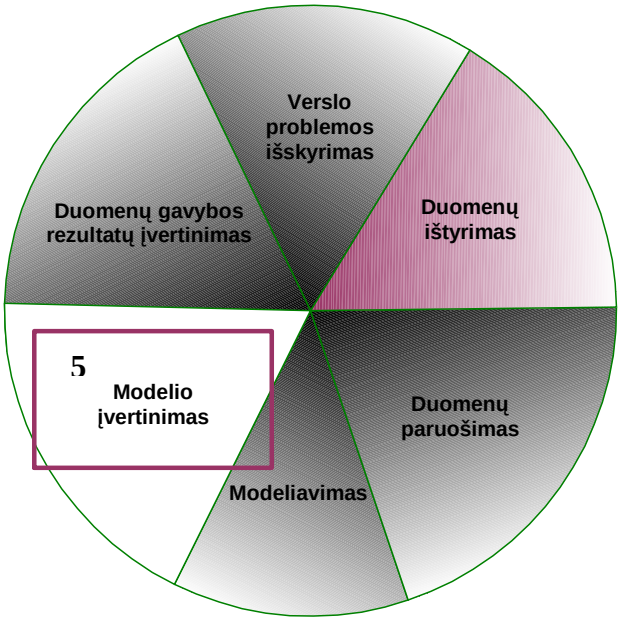
Šaltinis: sudaryta autoriaus.

72 pav. Laiko eilučių modelio atvaizdavimas

3.7. Septintas laboratorinis darbas – duomenų gavybos modelių palyginimas

Darbo tikslas:

Darbo tikslas įgyvendinti duomenų gavybos proceso penktąjį etapą – duomenų gavybos modelio įvertinimą (73 pav.).



Šaltinis: sudaryta autoriaus.

73 pav. Modelio įvertinimas

Darbo tikslas:

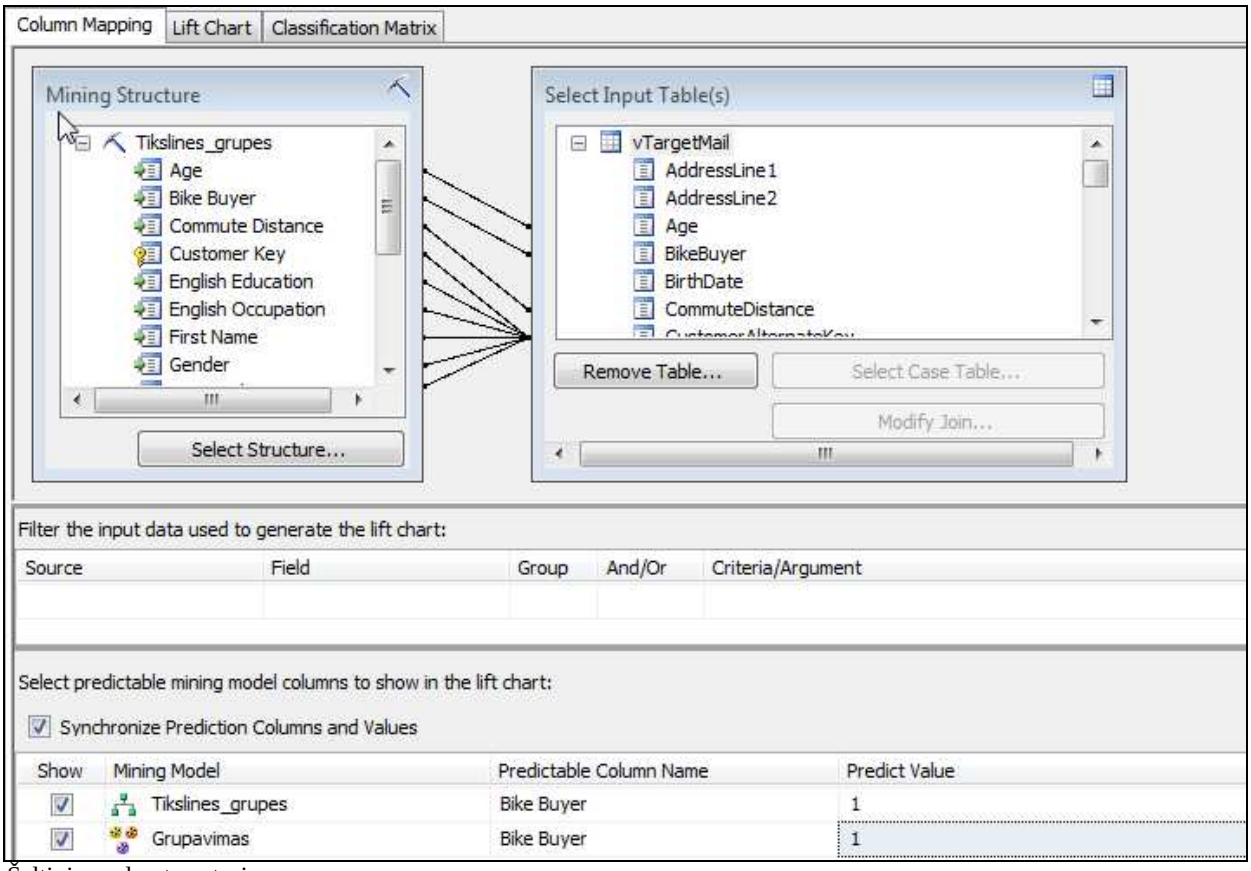
Darbo tikslas palyginti duomenų gavybos modelius tarpusavyje, nustatant, kuris modelis yra tikslesnis ir tinkamesnis duomenų gavyboje.

Užduotis:

- Sukurti grupavimo modelį, panaudojant **Target Mail** duomenų lentelę ir ją palyginti su jau suformuotu duomenų gavybos modeliu – **Tikslinės grupės**.
- Savarankiškai sugalvoti dar dviejų modelių palyginimą ir gautus rezultatus aprašyti.

Prognozuojamas rezultatas:

Sukūrus grupavimo modelį, atsidarome **Mining Accurancy Chart** lauką (74 pav.). **Column Mapping** lange, pasirenkame gavybos struktūrą **Tikslinės Grupės**. Pasirenkame įvedimo lentelę **Target Mail**. Norint palyginti duomenų gavybos modelius, būtina, kad tarp gavybos struktūros ir įvedimo lentelės atsirastų ryšiai. **Predictable mining model columns** lauke, turi matytis abu pasirinkti duomenų gavybos modeliai. Šiuo atveju Grupavimas ir Tikslines Grupes.

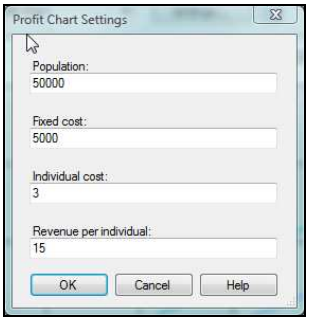


Šaltinis: sudaryta autoriaus.

74 pav. Dviejų modelių pasirinkimas

Pasirenkami parametrai (75 pav.):

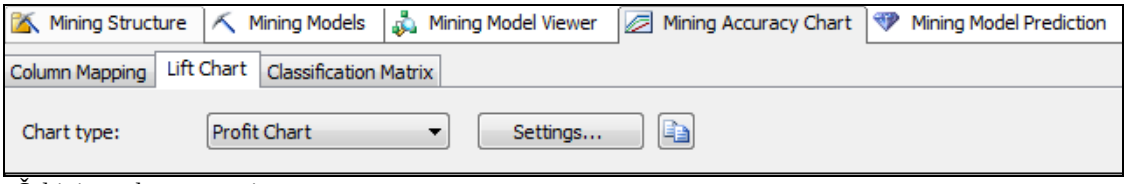
- Populiacija;
- Fiksuotos išlaidos;
- Individualios išlaidos.



Šaltinis: sudaryta autoriaus.

75 pav. Parametrų pasirinkimas

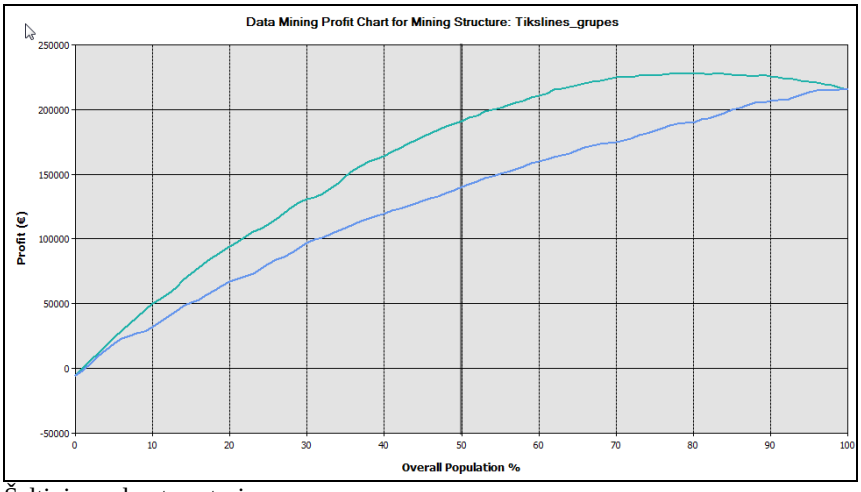
Duomenų gavybos modelių palyginimo aplinka pateikta 76 paveiksle. Vartotojui suteikiama galimybė pasirinkti grafiko tipą bei nustatyti duomenų gavybos rezultato parametrus.



Šaltinis: sudaryta autoriaus.

76 pav. Modelio aplinka

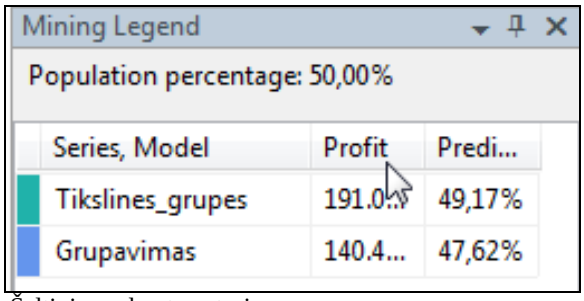
Gaunamas duomenų gavybos naudingumo grafikas (77 pav.), kuriame matomas modelių sugretinimas ir įvertinimas. Galima teigti, kad naudingesnis yra modelis – **Tikslinės grupės**, tai rodo nubraižyta kreivė. Šiame grafike matome, kad lyginami dviejų duomenų gavybos struktūrų rezultatai, kurie pateikiami procentine išraiška.



Šaltinis: sudaryta autoriaus.

77 pav. Modelių naudingumo palyginimas

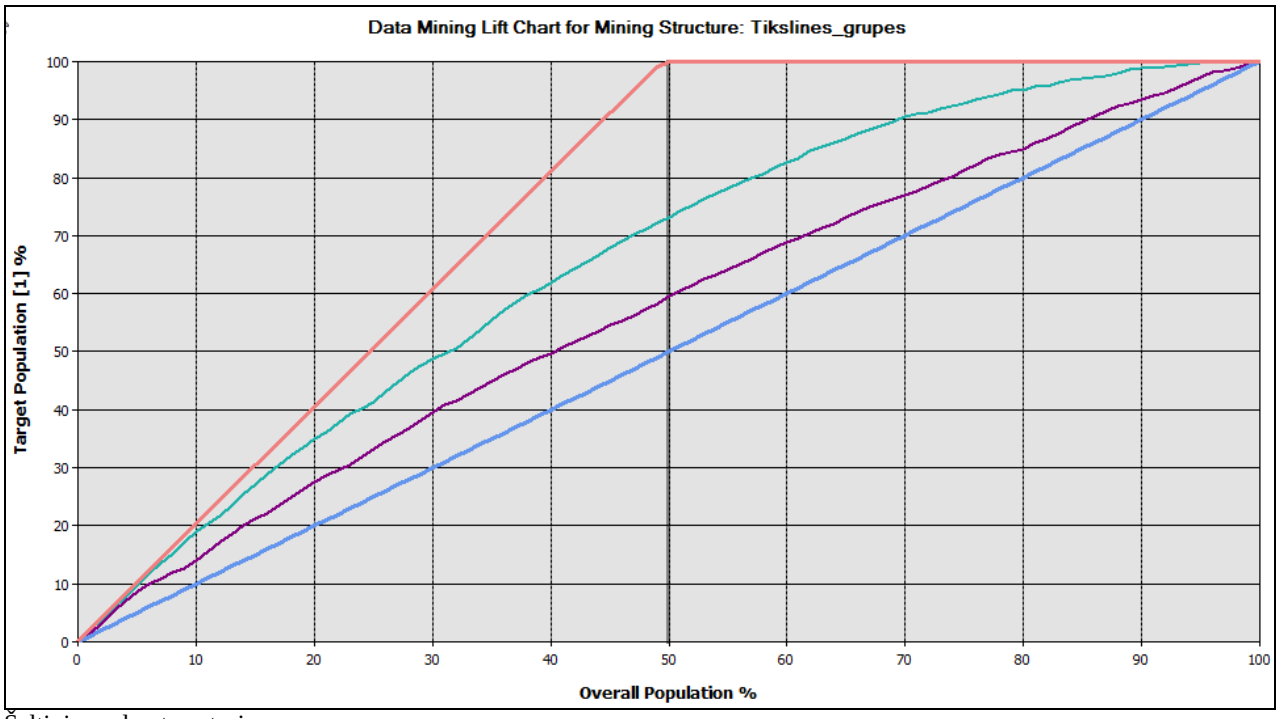
Gavybos legendoje atsispindi naudingumas ir prognozavimo tikslumas (78 pav.), pateikiami kiekybinė naudingumo išraiška bei procentinė palyginimo išraiška.



Šaltinis: sudaryta autoriaus.

78 pav. Gavybos legenda

Norint įvertinti sudarytų modelių tikslumą ir naudingumą duomenų gavybos procese, svarbu palyginti tuos modelius, įvertinti jų naudą duomenų gavybos rezultatui. Gavus šį rezultatą, verslo analitikas gali pasirinkti tikslesnį modelį probleminei sričiai analizuoti. 79 paveiksle pateikiamas keturių modelių palyginimas: grupavimo modelio struktūra, tikslinių grupių modelio struktūra, atsitiktinis sugeneruotas modelis bei idealaus duomenų gavybos modelio variantas.



Šaltinis: sudaryta autoriaus.

79 pav. Modelių struktūrų palyginimas

Gavybos legendoje, pavaizduotoje 80 paveiksle, pateiktas keturių modelių palyginimas. Pateiktoje rezultatų suvestinėje matome, kad lyginant dvi modelio struktūras (kurias mes pasirinkome), matome, kad aukštesniais balais įvertintas tikslinių grupių modelis. Šio modelio yra geresni ir kiti įvertinimo kriterijai. Tikslinių grupių modelio patikimumas sudaro 89 procentus,

grupavimo modelis – 76 procentus, o tuo tarpu atsitiktinai sugeneruotas modelis – 68 procentus. Idealus modelis sudaro 100 procentų.

Mining Legend

Population percentage: 68,00%

Series, Model	Score	Target population	Predict probability
Tikslines_grupes	0,87	88,84%	31,52%
Grupavimas	0,75	75,71%	40,85%
Random Guess Model		68,00%	
Ideal Model for: Tikslines_grupes, Grupavimas		100,00%	

Šaltinis: sudaryta autoriaus.

80 pav. Modelių įvertinimo rezultatų palyginimas

3.8. Eksperimentinio skyriaus išvados

1. MS SQL server 2005 Analysis Services programinio produkto pagalba realizuoti duomenų gavybos proceso laboratoriniai darbai;
2. Duomenų gavybos modeliavimui panaudota MS AdventureWorks DW pavyzdinė duomenų bazė;
3. Sudaryti šie laboratoriniai darbai:
    - Verslo aplinkos analizė ir probleminės srities išskyrimas;
    - Programinės įrangos analizė ir pasiruošimas projektui;
    - Duomenų paruošimas projektui;
    - Dimensijų, reikalingų duomenų gavybai, kūrimas;
    - Duomenų gavybos modelio pritaikymas ir panaudojimas;
    - Duomenų gavybos modelių plėtojimas;
    - Duomenų gavybos modelių palyginimas.

IŠVADOS IR PASIŪLYMAI

1. Darbe išanalizuoti pagrindiniai duomenų gavybos principai: duomenų gavybos proceso gyvavimo ciklas, duomenų gavybos algoritmai, programinės įrangos produktai.
2. Duomenų gavybos procesą sudaro šešios fazės: verslo aplinkos supratimas; duomenų ištyrimas – supratimas; duomenų paruošimas duomenų gavybos projektui; modeliavimas; modelio įvertinimas; duomenų gavybos rezultatų atskleidimas ir įvertinimas. Šiuos etapus siūloma papildyti dar vienu etapu – duomenų bazės suformavimu, kuris svarbus duomenų gavybos procese.
3. Šiame darbe buvo pasirinkta Microsoft programinė įranga, nes VUKHF yra įdiegusi MS Windows operacinę sistemą bei turi teises naudotis MS SQL server 2005 produktu. Studentai asmeniniam naudojimui, gali parsisiųsti ją nemokamai iš oficialaus Microsoft tinklapio.
4. Darbe panaudota pavyzdinė Microsoft duomenų bazė AdventureWorks DW, nes duomenų gavybos procesui būtini dideli kiekiai duomenų bei tinkamai sudaryta DB.
5. Duomenų gavybai atvaizduoti panaudotas MS SQL server 2005 Analysis Services programinis produktas, įvertintos šio įrankio suteikiamos duomenų gavybos galimybės. Panaudojant pavyzdinę duomenų bazę AdventureWorks DW sudaryti duomenų gavybos modeliai.
6. Eksperimentinėje dalyje pateikiami septyni laboratoriniai darbai, kurie parodo duomenų gavybos proceso eigą, duomenų gavybos modelių galimybes, analizuoti gautus modelius.
7. Sukurti laboratoriniai darbai puiki mokymosi priemonė, galinti papildyti „Verslo informacinių sistemų“ paskaitų modulį. Šių darbų dėka studentai galėtų susipažinti su duomenų gavybos procesu, sistemomis ir galimybėmis, atliekant paruoštas užduotis. Taip pat svarbu, kad studentai susipažins su vis populiarėjančia informacinių technologijų sritimi bei labiau pritaikyti savo žinias prie rinkos pokyčių.

LITERATŪRA

1. InfoPlanIT. (2007) Online Analytical Processing (OLAP) and Business Intelligence. [interaktyvus]. InfoPlanIT organization. [žiūrėta 2007 m. gruodžio 20d.]. Prieiga per internetą: www.infoplanit.com/Uploads/InfoPlanIT_OLAPandBI.pdf
2. RAS, Garnet. (2007) Magic Quandrant for customer data mining. [interaktyvus] Gartner coorporation. [žiūrėta 2008 m. balandžio 18 d.]. Prieiga per internetą: www.gartner.com/DisplayDocument?id=488171
3. SEIFERT, Jeffrey. (2004) Data mining overview. [interaktyvus] Federation of American Scientists. [žiūrėta 2008 m. balandžio 14d.]. Prieiga per internetą: www.fas.org/sgp/crs/intel/RL31798.pdf
4. GROSSMAN, Robert. (2000) Supporting the Data Mining Process with Next Generation Data Mining Systems. [interaktyvus] Cornell University. [žiūrėta 2008 m. balandžio 18 d.]. Prieiga per internetą: http://esj.com/article.aspx?ID=8139813919PM
5. ZAIANE, Osmar. (2004) A brief history of data mining. [interaktyvus] University of Alberta. [žiūrėta 2008 m. kovo 8 d.]. Prieiga per internetą: http://www.data-mining-software.com/data_mining_history.htm
6. Microsoft. (2004) Business Intelligence in Europe: Relevant and Valuable Information for better decisions inreal time enterprise context. [interaktyvus] Microsoft Gartner. [žiūrėta 2008 m. vasario 12 d.]. Prieiga per internetą: https://msdb.ru/Downloads/Docs/Events/Materials/210403/business_intelligence_in_europe.pdf
7. THEARLING, Kurt. (2008) An Introduction to Data Mining. [interaktyvus] Data Mining and Analytic Technologies. [žiūrėta 2008 m. vasario 12 d.]. Prieiga per internetą: http://www.thearling.com/text/dmwhite/dmwhite.htm
8. SIMUTIS, Rimvydas. (2008) Dirbtiniai neuroniniai tinklai, metodinė paskaitų medžiaga.
9. Wikipedia. (2008) Data mining [interaktyvus]. Wikipedia Foundation Inc. [žiūrėta 2008 m. balandžio 10 d.]. Prieiga per Internetą: http://en.wikipedia.org/wiki/datamining.
10. IT ekspertas. (2006) OLAP duomenų bazės [interaktyvus]. IT ekspertas. [žiūrėta 2008 m. kovo 20 d.]. Prieiga per Internetą: http://itekspertas.projektas.lt/index.php?option=com_content&task=view&id=94&Itemid53
11. HAMEL, Lutz. (2007) Cluster analysis. [interaktyvus] The Australian National University. [žiūrėta 2008 m. balandžio 18 d.]. Prieiga per internetą: datamining.anu.edu.au/student/math3346_2006/clusters-2x3.pdf

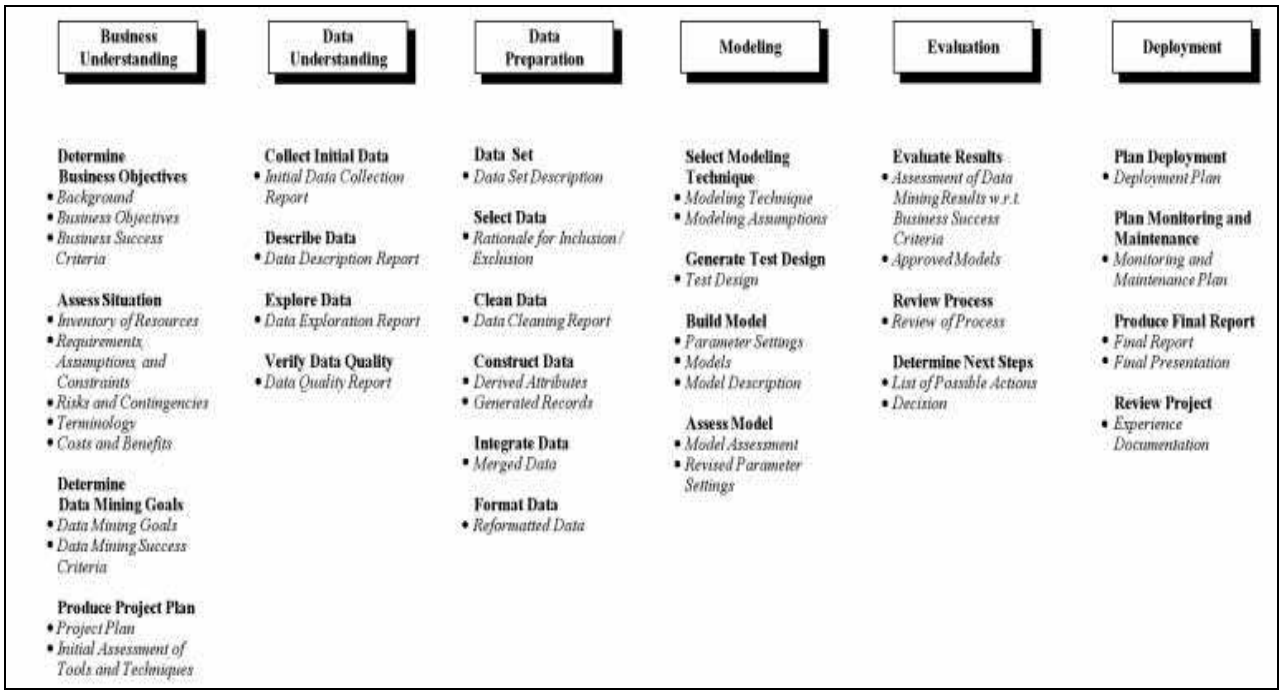
12. GROSSMAN, Robert. (2002) A Top-Ten List for Data Mining. [interaktyvus] SIAM News, Volume 34, Number 5, 1-2 psl. [žiūrėta 2008 m. vasario 18 d.]. Prieiga per internetą: http://www.siam.org/pdf/news/544.pdf
13. Microsoft. (2008) Feature differencies between OLAP ir data mining. [interaktyvus] Microsoft organization [žiūrėta 2008 m. kovo 22 d.]. Prieiga per internetą: http://office.microsoft.com/en-us/excel/HP101774371033.aspx#Feature%20differences%20between%20OLAP%20and%20non-OLAP%20source%20data
14. Informatique SA. (2001) Data mining and OLAP. [interaktyvus] Lausanne, Switzerland, [žiūrėta 2008 m. kovo 22 d.]. Prieiga per internetą: http://dml.cs.byu.edu/~cgc/docs/dm/Reading/DMvsOLAP-ELCA-WP.pdf
15. HAMEL, Lutz. (2001) A Brief Tutorial on Database Queries, Data Mining, and OLAP. [interaktyvus] The Australian National University. [žiūrėta 2008 m. vasario 12 d.]. Prieiga per internetą: http://homepage.cs.uri.edu/faculty/hamel/pubs/hamel-197-manuscript-final.pdf
16. BRANDIN, Chris. (2007) Contrasting Xpriori Insight with Traditional Statistical Analysis, Data Mining and Online Analytical Processing. [interaktyvus] Science of priori. [žiūrėta 2008 m. vasario 12 d.] Prieiga per internetą: http://www.xpriori.com/resources/insight/tp_contrasting-insight_v1.1.pdf
17. MOHAMMAD, Rob. (2007) Case Projects in data warehousing and data mining.[interaktyvus] University of Houston-Clear Lake. [žiūrėta 2008 m. vasario 20 d.] Prieiga per internetą: http://www.iacis.org/iis/2007_iis/PDFs/Rob_Ellis.pdf
18. ELDER, John. (1998) A Comparison of Leading Data Mining Tools. [interaktyvus] Data Mining and Pattern Discovery. [žiūrėta 2008 m. kovo 20d.]. Prieiga per internetą: http://www.datamininglab.com/pubs/kdd98_elder_abbott_nopics_bw.pdf
19. DW Review. (2006) Data mining Solutions. [interaktyvus] DwRewiew. [žiūrėta 2008 m. kovo 20 d.]. Prieiga per internetą - http://www.dwreview.com/datamining.html
20. KDnuggest. (2008) Data mining tools. [interaktyvus] KD stands for Knowledge Discovery. [žiūrėta 2008 m. balandžio 9 d.]. Prieiga per internetą: http://www.kdnuggets.com/polls/2006/data_mining_analytic_tools.htm
21. CHAPMAN, Pete, CLINTON, Julian, SHEARER, Colin, WIRTH, Rüdiger. (2002) CRISP-DM process model. [interaktyvus] CRoss-Industry Standard Process for Data Mining. [žiūrėta 2008 m. balandžio 12 d.]. Prieiga per internetą: http://www.crisp-dm.org/CRISPWP-0800.pdf

22. KUMAR, Sunil, BABU, Hari. (2007) Data mining technology for problem solving and knowledge creation. [interaktyvus] Tata Steel. [žiūrėta 2008 m. balandžio 12 d.]. Priega per internetą: http://www.tatasteel.com/webzine/tatatech43/page_35.htm
23. Microsoft. (2008) Analysis Services. [interaktyvus]Microsoft corporation. [žiūrėta 2008 m. kovo 20 d.]. Prieiga per internetą www.microsoft.com
24. SEKLIUKIS, Vitolis, GUDAS, Saulius, GARŠVA, Gintautas. (2003) Informacijos sistemos ir duomenų bazės. Kaunas: Technologija. 316 p. ISBN 9955-09-486-9
25. GUDAS, Saulius. (2008) Vadybos informacinės sistemos, metodinė paskaitų medžiaga.
26. EDELSTEIN, Herb. (2000) Data Mining: Exploiting the Hidden Trends in Your Data. [interaktyvus] University of Glasgou. [žiūrėta 2008 m. kovo 22 d.]. Prieiga per internetą: http://www.psy.gla.ac.uk/~steve/pr/edel.html
27. Oracle. (2008) Oracle software products. [interaktyvus] Oracle coorporation. [žiūrėta 2008 m. kovo 22 d.]. Prieiga per internetą: http://www.oracle.com
28. SSPS. (2008) SSPS software products. [interaktyvus] Statistical Package for the Social Sciences. [žiūrėta 2008 m. kovo 22 d.]. Prieiga per internetą: http://www.ssps.com
29. SAS. (2008) SAS software. [interaktyvus]. Statistical analysis software. [žiūrėta 2008 m. kovo 20 d.] Prieiga per internetą www.sas.com
30. THEARLING, Kurt. (2003) An Introduction to Data Mining. [interaktyvus] Data Mining and Analytic Technologies. [žiūrėta 2008 m. kovo 20 d.] Prieiga per internetą: http://www.thearling.com/dmintro/dmintro.pdf.
31. COLLIER, Ken. (1999) A Methodology for Evaluating and Selecting Data Mining Software. [interaktyvus] International Conference on System Sciences. [žiūrėta 2008 m. balandžio 25 d.] Prieiga per internetą: http://www.comp.dit.ie/.../DT228BSI/Methodology%20for%20Evaluating%20and%20Selecting%20Data%20Mining%20Software.pdf

PRIEDAI

1 PRIEDAS

CRISP duomenų proceso modelis



Tikslų medžio pavyzdys

Criteria	Weight	Knowledge Seeker		Data Mind		Model 1		Clementine		Darwin	
Performance	*0.15*	Rating	Score	Rating	Score	Rating	Score	Rating	Score	Rating	Score
Platform Variety	0.2	3	0.6	4	0.8	3	0.6	3	0.6	2	0.4
Software Architecture	0	3	0	5	0	3	0	3	0	2	0
Heterogeneous Data Access	0.15	3	0.45	3	0.45	3	0.45	3	0.45	2	0.3
Data Size	0.1	3	0.3	2	0.2	3	0.3	3	0.3	3	0.3
Efficiency	0.1	3	0.3	3	0.3	3	0.3	2	0.2	2	0.2
Interoperability	0.1	3	0.3	3	0.3	3	0.3	3	0.3	3	0.3
Robustness	0.35	3	1.05	3	1.05	3	1.05	1	0.35	2	0.7
Category Score			3		3.1		3		2.2		2.2
Functionality	*0.2*	Rating	Score	Rating	Score	Rating	Score	Rating	Score	Rating	Score
Algorithmic Variety	0.2	3	0.6	3	0.6	5	1	5	1	4	0.8
Prescribed Methodology	0.15	3	0.45	4	0.6	4	0.6	4	0.6	3	0.45
Model Validation	0.2	3	0.6	4	0.8	4	0.8	4	0.8	4	0.8
Data Type Flexibility	0.15	3	0.45	3	0.45	3	0.45	2	0.3	3	0.45
Algorithm Modifiability	0.05	3	0.15	3	0.15	4	0.2	4	0.2	4	0.2
Data Sampling	0.05	3	0.15	3	0.15	3	0.15	3	0.15	3	0.15
Reporting	0.2	3	0.6	4	0.8	5	1	4	0.8	4	0.8
Model Exporting	0	3	0	2	0	5	0	4	0	5	0
Category Score			3		3.55		4.2		3.85		3.65
Usability	*0.2*	Rating	Score	Rating	Score	Rating	Score	Rating	Score	Rating	Score
User Interface	0.2	3	0.6	3	0.6	3	0.6	4	0.8	2	0.4
Learning Curve	0.15	3	0.45	3	0.45	2	0.3	1	0.15	1	0.15
User Types	0.15	3	0.45	3	0.45	5	0.75	2	0.3	2	0.3
Data Visualization	0.2	3	0.6	2	0.4	2	0.4	2	0.4	1	0.2
Error Reporting	0.15	3	0.45	3	0.45	2	0.3	1	0.15	2	0.3
Action History	0.15	3	0.45	2	0.3	3	0.45	5	0.75	3	0.45
Domain Variety	0	3	0	2	0	4	0	4	0	4	0
Category Score			3		2.65		2.8		2.55		1.8
Ancillary Task Support	*0.25*	Rating	Score	Rating	Score	Rating	Score	Rating	Score	Rating	Score
Data Cleansing	0.2	3	0.6	3	0.6	5	1	4	0.8	4	0.8
Value Substitution	0.1	3	0.3	2	0.2	4	0.4	4	0.4	3	0.3
Data Filtering	0.15	3	0.45	3	0.45	4	0.6	5	0.75	5	0.75
Binning	0.05	3	0.15	4	0.2	3	0.15	2	0.1	2	0.1
Deriving Attributes	0.1	3	0.3	3	0.3	5	0.5	5	0.5	5	0.5
Randomization	0.05	3	0.15	3	0.15	3	0.15	3	0.15	3	0.15
Record Deletion	0.05	3	0.15	3	0.15	4	0.2	4	0.2	4	0.2
Handling Blanks	0.2	3	0.6	3	0.6	3	0.6	2	0.4	3	0.6
Metadata Manipulation	0	3	0	2	0	4	0	3	0	3	0
Result Feedback	0.1	3	0.3	3	0.3	3	0.3	5	0.5	4	0.4
Category Score			3		2.95		3.9		3.8		3.8
Other Criteria	*0.2*	Rating	Score	Rating	Score	Rating	Score	Rating	Score	Rating	Score
Cost	1	3	3	2	2	2	2	3	3	2	2
insert others	0	3	0	0	0		0		0		0
Category Score			3		2		2		3		2
Weighted Average			3		2.842		3.225		3.16		2.77