Дәріс 14

Statistica, SPSS

Rapid Miner жүйесі барлық керекті форматтарды қолдамайды. Деректерді жүктеу үшін болады:

– Файлды тышқан көмегімен Repositories панеліне орналастыру
– Мәзірден File/Import Data таңдау.
– Операторлар панелінен қажетті операторды таңдау.

Соңғы нұсқасы ең әмбебап болып табылады. Operators панелінде операторлар ағашы бейнеленеді, әрбір түйін санмен (ұрпақтардағы оператор саны) белгіленеді.



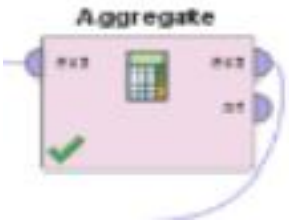
Operators панелі

Import/Data/Read ARFF таңдаймыз. Бұл үшін оған екі рет басуға немесе Process панеліне алып шығуға болады. Process панелі жұмыс кезінде негізгі болып табылады. Операторды таңдау кезінде ол оның үстінде пайда болады және процесс графына автоматты түрде қосылады.

Әрбір оператор сол жағынан кіріс түйіндері (бар болса), ал оң жағынан шығыс түйіндері (бар болса) орналасқан пиктограммамен бейнеленеді. Бір оператордың шығыстарын басқаның кірістерімен байланыстыруға болады, сонда оның жұмысының нәтижесі басқа оператор кірісіне келіп отырады. Main Process ортақ өрісінің (негізгі процесс) сондай-ақ кіріс және шығыс түйіндері болады. Атап айтқанда, Read ARFF операторының шығысы негізгі процесстің ортақ шығысымен автоматты түрде байланысқан. Операторлардың жұмыс нәтижесінің визуализациясы үшін, сәйкес операторлар негізгі процесс шығыстарымен байланысқан болуы керек. Әрбір оператор пиктограммасының астынға сол жақ бұрышында (Process панелінде) индикатор күйі болады:

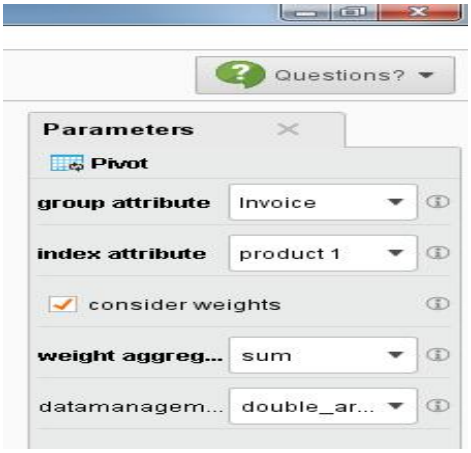
– Жасыл түс – оператор жұмыс істеп болды;
– Сары түс – оператор жұмысқа дайын;

– Қызыл түс – оператор жұмысқа дайын емес (қателер бар).



Process панеліндегі Read ARFF операторы

Біздің жағдайда оператор жұмысқа дайын емес, себебі деректері бар файл көрсетілмеген. Егер оператор пиктограммасына шертетін болсақ, онда Parameters панелінде оның параметрлері көрінеді. Мысалы, Read ARFF операторында толтыруды қажет ететін data file параметрі болады.



Parameters панелі

Қателерді, сондай-ақ Problems панелін пайдалана отырып түзеуге болады, онда қателердің тізімі көрсетілген. Тізімде үш баған бар:

1) проблемаларды сипаттау (Message) ;
2) жою тәсілі (Fixes);
3) бастама (Location).

Fixes өрісін сол жақ тышқан батырмасын екі рет басу арқылы қателерді түзеуге көшеміз (тышқанның оң жақ батырмасын басуға да болады және контекстік мәзірден керекті бөлімді таңдаймыз).

Негізгі процесс Run (F11) батырмасы арқылы іске қосылады. Жұмыс аяқталғаннан кейін бағдарлама нәтижелеріді көру (панель үстіндегі Results) режиміне ауысуға рұқсат сұрайды. Result Overview панелінде процесс нәтижелерінің тізімі келтіріледі. Сәйкес тармаққа басқанда есеп беруді бүктейді немесе өрістетеді.

Rapid Miner жүйесінде деректерді визуализациялаудың өте көп әртүрлі

тәсілдері жүзеге асырылған. Визуализация тәсілі Example Set панелінде Plot View режимінде Plotter белгісі бар батырмамен таңдалады. Жүйеде іске асырылған визуализацияның кейбір түрлерін келтірейік:

– Scatter – қарапайым екі және үш өлшемді проекция нүктелері;

– Bubble – әрбір объект, радиусы белгілер ішіндегі біреуінің мәніне сәйкес келетін шеңбер түрінде бейнеленеді;

– Parallels – әрбір объект түзу түрінде бейнеленеді (белгілер мәніне сәйкес келетін нүктелер арқылы өтеді);

– Series – қатарды бейнелеу;

– SOM – Кохонен картасы;

– Block – объектілер тіктөртбұрышпен бейнеленеді;

– Density – тығыздықтың бейнеленуі;

– Pie, Ring – секторлы диаграммалар;

– Bars – гистограммалар;

– Surface – беті және т.б.

Бейнелерді үлкейтуге, орнын ауыстыруға, осьтер масштабын өзгертуге болады (қажетті батырмалар панельдің астыңғы сол жағында орналасқан). Export Image батырмасына басу бейнелерді дискіге сақтауға мүмкіндік береді (қолдайтын форматтар: eps, swf, emf, gif, pdf, ps, svg, raw, ppm, bmp, jpg, png, gif және т.б.).

RapidMiner бағдарламасы (алғашқы атауы "Yale") машиналық оқыту және деректерді талдау ортасы болып табылады, яғни оны қолданатын пайдаланушының деректерді талдауда жұмысы жеңілдейді. Оның орнына, оған деректерді талдау процесін граф түрінде «бояу» ұсынылады және оны орындауға жібереді. RapidMiner - де тізбек операторлары интерактивті граф түрінде ұсынылады және XML тілінде (eXtensible Markup Language, негізгі тіл жүйесі) түрінде көрсетіледі. Осы жүйесі Java тілінде жазылған және AGPL version 3 лицензиясы арқылы қолданылады. Барлық негізгі функцияларына қатынау үшін Java API арқылы мүмкіндік алады.

Қазір осы жүйеде 400-ден аса операторлар жүзеге асырылды. Оның ішінде:

1) Жіктеу алгоритмдері, регрессия, кластерлеу және қауымдастықтар іздеу, сондай-ақ мета-алгоритмдері жүзеге асқан үлгі болған операторлар (machine learning algorithms) іске асырылған;

2) WEKA операторлар жүйесі;

3) Өңдеу операторлары (дискритизация, сүзу, рұқсатнамаларды толтыру, өлшемін азайту және т. б.);

4) Белгілермен жұмыс жасау операторлары (селекция және генерация белгілері);

5) Мета-операторлар (мысалы, бірнеше параметрлер бойынша оңтайландыру операторы);

6) Сапаны бағалау операторлары (сырғымалы бақылау және т. б.);

7) Визуалдау операторлары (визуалдау тәсілдері жеткілікті түрде көп);

8) Мәліметтерді тиеу және сақтау операторлары (арнайы форматтардығы жұмысты қоса алғанда: arff, C4.5, csv, bibtex, деректер базасын және т. б.).

Іске қосу кезінде RapidMiner келесі мүмкіндіктерді ұсынады:

1) Процестің жаңа анықтамасын бастау (New).

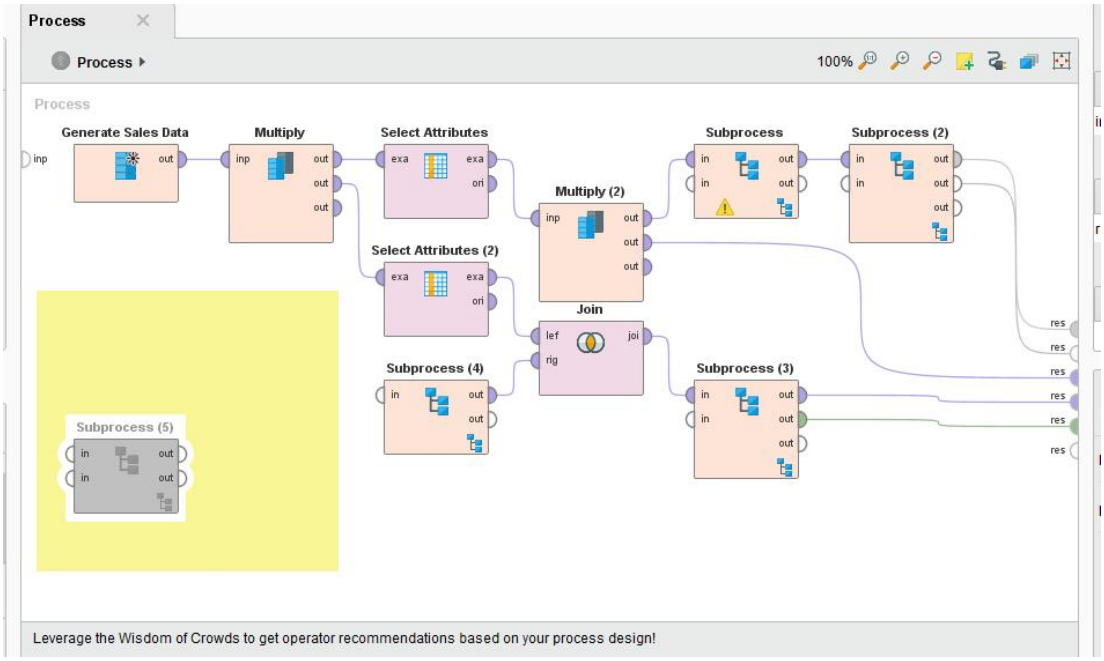
2. Қолда барды ашу (Open recent).1

3. Ашу (Ореп).

4. Ашу үлгісі (Open Template).

5. Demo - ны іске қосу (Online Tutorial).

"New" опцияларын таңдау кезінде Repository Browser терезесі пайда болады, онда деректерді өңдеу процесіне қажетті файлды таңдау қажет. Осыдан кейін, негізгі терезесі пайда болады



Rapid Miner бағдарламасының негізгі терезесі

Панельдің үстіңгі оң жағында үш батырма орналасқан, олар үш негізгі режимдері арасында ауысып қосылып отыруға мүмкіндік береді: жаңадан құру (F8 пернесі), нәтижені көру (F9 пернесі) және сәлемдесуге көшу (5 опцияны таңдау). Сондай-ақ ауысу View/Perspectives мәзірі арқылы бола алады



Жұмыс режимдері арасында ауысу батырмалары

Мәзірдегі View/Show View опциясы арқылы көрінетін панелдерді таңдауға болады, оларды негізгі терезеде өз талғамыңыз бойынша қоюға болады. Назар аударыңыз! Жұмыс істеу үшін әрқашан әдепкі қалпы бойынша панельдер орналасып тұрмайды. Атап айтқанда, пайдаланушыға қажет панельдер: Process, Operators. Бағдарламаның панельдерін атап өтейік:

1) Process. Осы панельде пайдаланушы негізгі процесс бойынша "сурет салады", яғни оған операторларды апарып қояды және оларды бір-бірімен байланыстырады. Көптеген пайдалы функцияларды мәтіндік мәзір (панельде оң жақ тышқан батырмасын басу арқылы) арқылы шақырылады.

2) Overview. "Салынған процесс" көрінеді (панель Process панелінде шарлау үшін керек, онда барлық графтар орналаса алмайды, сондықтан Overview панелінде тек жарты бөлігі ғана бейнеленеді).

3) Operators. Мұнда каталогтар жүйесінде операторлары сақталады, оларды Process панеліне сүйреп апаруға болады. Help панелінде операторды таңдау кезінде негізгі ақпарат бейнеленеді.

4) Tree. Салынған процесті ағаш түрінде ұсыну.

5) XML. Салынған процесті XML-код түрінде ұсыну.

6) Parameters. Process панелінде операторды бөлу кезінде мұнда оның параметрлері көрсетіледі. Параметрлерге өзгерістер енгізу қалыпты жұмыс үшін қажет болады (атап айтқанда, деректерді жүктеу операторына деректері бар файл атын көрсету керек).

7) Problems. Мұнда процесстің іске қосылуына мүмкіндік бермейтін қателер көрсетіледі.

8) Repositories. Осы панельдерде процестер мен деректердегі файлдарға навигация жүргізіледі.

9) Context. Кіру және шығу процесін анықтау.

10) Help. Қажетті ақпаратты көрсету.

11) Comment. Панель «белгілер үшін».

12) Log. Оқиғалар журналы.

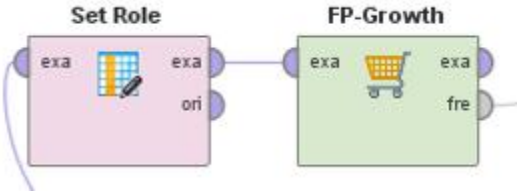
13) Remote Processes. «Қашықтағы процестермен» жұмыс істеу(жаңа бастаған пайдаланушы үшін бұл панель қажет емес).

14) System Monitor. Жүйелік ресурстар көрсетіледі (пайдаланылатын жады).

15) Result Overview. Жұмыс процесінің нәтижелерін көрсету. Results режимінде графиктерді көрсету үшін қосымша панельдер көрсетіледі.

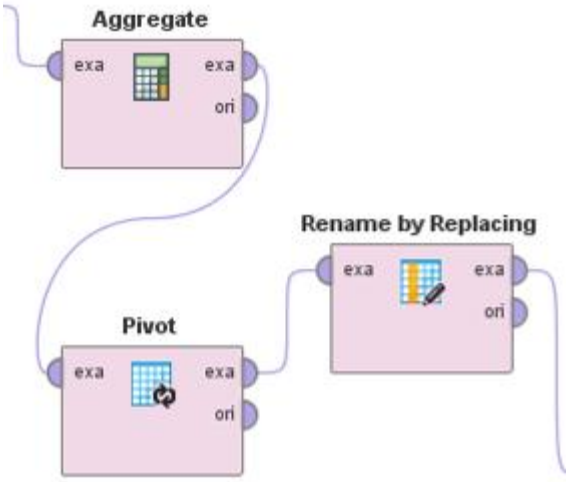
Классификатор жұмысын суреттеу үшін Operators панелінен Process панеліне қандай да бір классификаторды алып шығайық. Мысалы, байесов классификаторы (орналасуы Modeling/Classification and Regression/Bayesian Modeling). Деректерді жүктеу операторының шығысын классификациялау операторының кірісімен, ал классификациялау операторының шығысын негізгі

процесстің шығысымен байланыстырамыз



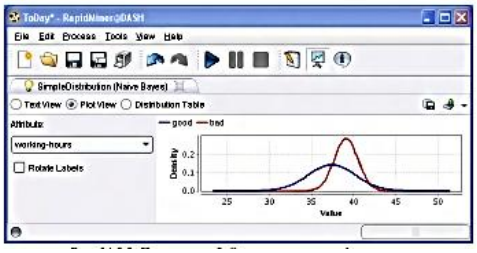
Байесов классификаторы – деректерді жүктеу тізбегі

Naive Bayes операторында екі шығысы болатынын байқайық: model (классификатор сипаттамасы, оны байланыстыру керек) және example set (таңдау, керек жағдайда оны басқа классификатор кірісіне беруге болады). Классификациялау операторы бұрынғыдай жұмысқа дайын емес (қызыл индикатормен жанып тұр), себебі мақсатты белгісі көрсетілмеген. Бұл қатенің жолында Problems панелінде Fixes өрісіне басу арқылы қажетті белігі таңдаймыз. Нәтижесінде Process панелінде жаңа Set Role операторы пайда болады, оның параметрінде мақсатты белгі жазылып тұрады



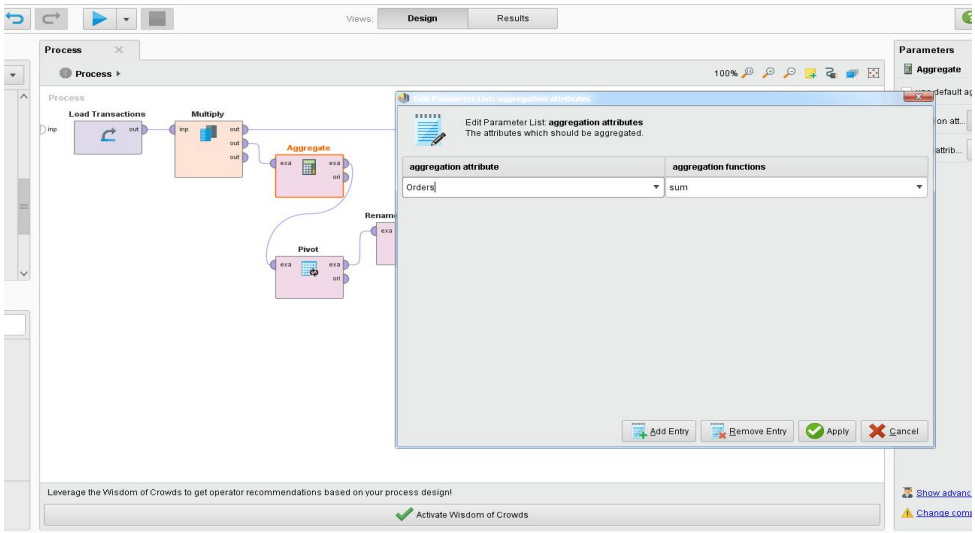
Деректерді жүктеу тізбегі

Біз классификаторды құрдық, бірақ оны іске қоспадық. Бұл процедураны блоктарға бөлу қажет. Ол үшін деректерді жүктеу процессіне Evaluation/Validation/X-Validation операторын қосамыз. Негізінен, процесстің негізгі панеліне басқа операторларды қосу қажет емес (қалған операторлар «Validation операторының ішіне» қосылады), сондықтан процесстің негізгі панелінде тек қана деректерді жүктеу операторы мен Validation операторын қалдырамыз. Read ARFF операторының шығысы Validation операторының кірісіне жалғанады. Осыдан кейін байесов классификаторы құрылады



Процессті іске қосу және байесов классификаторының құрылуы

Қосылған Validation операторының пиктограммасы астынғы оң жақ бұрышта «екі көк терезе» белгісін ие болады, бұл осы операторды «ашуға» және «толтыруға» болады дегенді білдіреді. Пиктограммаға екі рет басамыз, бұл кезде екі өріс ашылады: оқыту (training) және тестілеу (testing). Бірінші терезеге классификациялау операторын орналастырайық, мысалы, нейрондық жүйе (Modeling/Classification and Regression/Neural Net Training/Neural Net), ал екіншісіне – классификациялау сапасын бағалайтын оператор, мысалы, Performance Measurement/Performance. Қалған қажет операторларды қатені түзей отырып қосуға болады



Validation операторын толтыру

Parameters панелінде әрбір оператор параметрін өзгертуге болады (алдымен оны Process панелінде бөліп көрсету керек). Кейде классификатор параметрін де өзгертуге болады. Validation операторы параметрінде блок санын өзгертуге болады.

Rapid Miner жүйесінде операторларды ауыстыру үдерісі өте ыңғайлы орналасқан. Ол үшін блоктар бойынша бақылаудың орнына сапаны бақылайтын басқа бір үдерісін ұйымдастыру қажет. X-Validation оператор пиктограммасына

тышқанның оң жақ батырмасына басамыз, Replace Operator опциясын және блок бойынша бақылауды ауыстырғымыз келетін оператор атын таңдаймыз, мысалы, Split Validation (таңдауды оқыту мен бақылау деп бөлу).

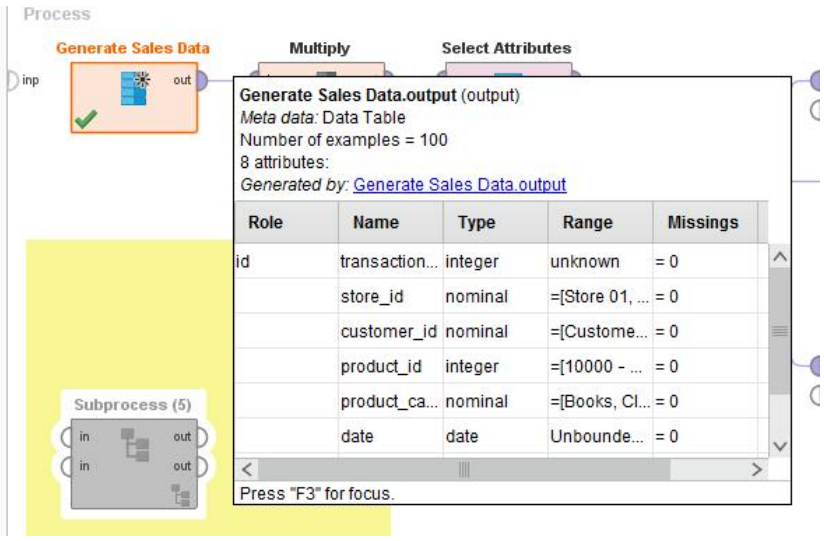
Кластеризациялау тапсырмасын шешуді бұл жобада көрсетілген. CSV типті деректері бар тексттік файл болсын. Осындай түрдегі файлды жүктеу үшін Import/Data/Read CSV операторы керек. Бұл файлды Repositories панеліне тышқанмен алып шығуға болмайды, себебі бұл операция мәтіндік файлдар үшін қолдауға ие болмайды. Бұл кедергіні келесі жолмен өтуге болады:

– Файлды Excel бағдарламасында ашу,
– «Книга Microsoft Excel» сияқты сақтау,
– Құрылған файлды Repositories панеліне қажетті каталогқа орналастыру.

Үшінші қимылды орындаған кезде файлды жіберетін, кейбір қасиеттерге белгі (идентификатор белгісі және т.б) қоятын көмекші ашылады.

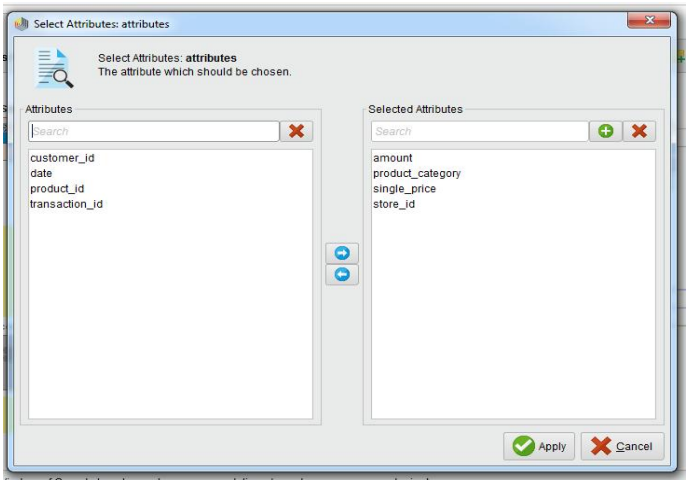
Бағдарламаның орындалуы

Берілген бағдарлама блоктар түрінде ұсынылған, әрбір блок кіріс деректерін түрлендіреді немесе оларды өңдейді. Біздің жағдайда тапсырманың бастапқы деректері Rapid Miner ішінде бағдарламалы түрде генерацияланады. Ол үшін Generate Sales Data операторын қолданамыз, оның параметріне мысал ретінде 100 санын береміз. Кіріс деректері, нәтижені ұсынудың қарапайымдылығы үшін бағдарламаның ішіндегі генератор көмегімен кездейсоқ алынған. Бағдарламада екі генератор ұсынылған. Бірінші генератор сатылымдардың кездейсоқ кестесін құру үшін арналған, оның ішінде бағандағы мәндер генерацияланады: транзакция id, дүкен id, сатып алушы id, өнім коды, өнім категориясы, сатылымдар саны және тауардың бағасы



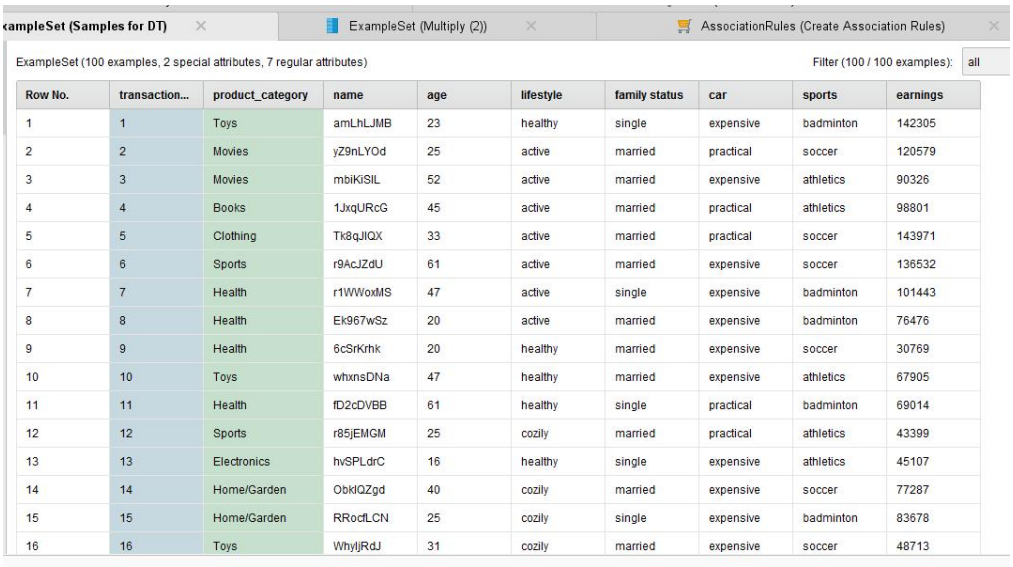
Кіріс деректерінің генерациялануы

Одан әрі кіріс деректерінің атрибуттары таңдалады, ол үшін Select Attributes операторы қолданылады



Select Attributes оператор қасиетінде қажетті атрибутты таңдау

Екінші генератор көмегімен деректер қорының кестесі құрылады. Оның ішінде өнім категориясы, сатып алушының аты-жөні, жасы, қандай өмір салтымен айналысатыны, еңбек ақысы көрсетіледі

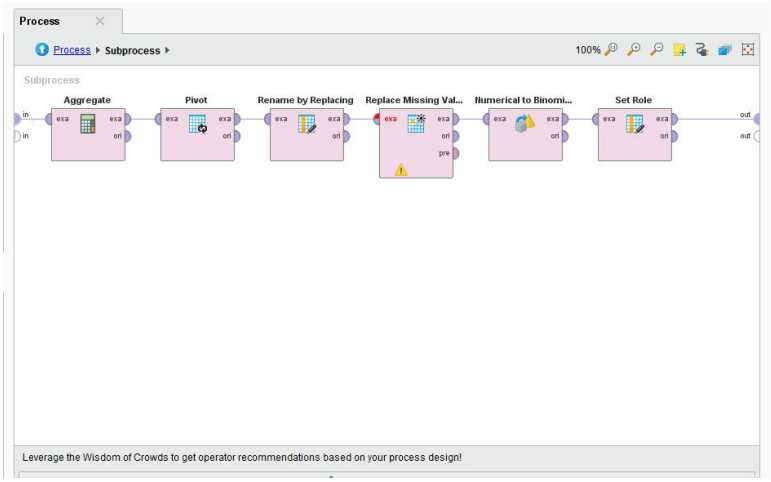


Row No.	transaction...	product_category	name	age	lifestyle	family status	car	sports	earnings
1	1	Toys	amLhLJMB	23	healthy	single	expensive	badminton	142305
2	2	Movies	yZ9nLYOd	25	active	married	practical	soccer	120579
3	3	Movies	mbiKiSIL	52	active	married	expensive	athletics	90326
4	4	Books	1JxqURcG	45	active	married	practical	athletics	98801
5	5	Clothing	Tk8qJIQX	33	active	married	practical	soccer	143971
6	6	Sports	r9AcJZdU	61	active	married	expensive	soccer	136532
7	7	Health	r1WWoxMS	47	active	single	expensive	badminton	101443
8	8	Health	Ek967wSz	20	active	married	expensive	badminton	76476
9	9	Health	6cSrKrhk	20	healthy	married	expensive	soccer	30769
10	10	Toys	whxnsDNa	47	healthy	married	expensive	athletics	67905
11	11	Health	fD2cDVBB	61	healthy	single	practical	badminton	69014
12	12	Sports	r85jEMGM	25	cozily	married	practical	athletics	43399
13	13	Electronics	hvSPLdrC	16	healthy	single	expensive	athletics	45107
14	14	Home/Garden	ObkIQZgd	40	cozily	married	expensive	soccer	77287
15	15	Home/Garden	RRocfLCN	25	cozily	single	expensive	badminton	83678
16	16	Toys	WhytjRdJ	31	cozily	married	expensive	soccer	48713

Бағдарламада құрылған деректер қоры

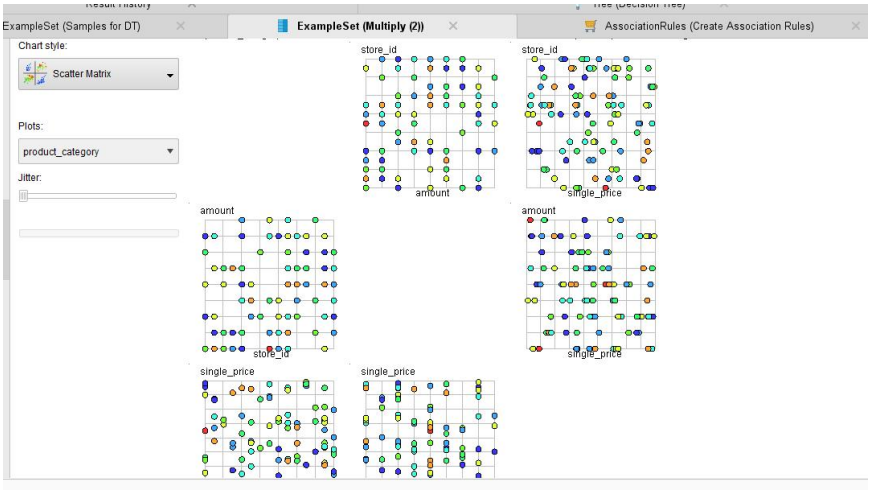
Бағдарламада кластерлеу тапсырмасын орындау үшін операторлар тізбегін құрамыз. Ол үшін келесі операторлар керек: Aggregate – тауарды қалдыру керек па жоқ па шешімін қабылдайды, Pivot – сатылымдардың барлық қосындысын санайды және олармен «sum(orders)» бағанын құрады, Rename by

Replacing – тауар бойынша тапсырыстар санын сұрыптайды, Replace Missing Values – сатылымдар саны бар бағанды қайта жазады, Numerical to Binominal – кестедегі деректерді нөлге ауыстырады, Set Role – деректерді оңдық жүйеден бинарлыға ауыстырады



Кластерлеу құру тізбегі

Нәтижесінде кластерлеу операторлар тізбегі құрылады. Оны Run батырмасы арқылы орындалуға жіберген кезде кластерлеу нәтижесін аламыз. Rapid Miner-да модельдердің визуализациясы үшін барлық құралдар келтірілген: дендограммалар, кластерлер каталогтары және т.б.

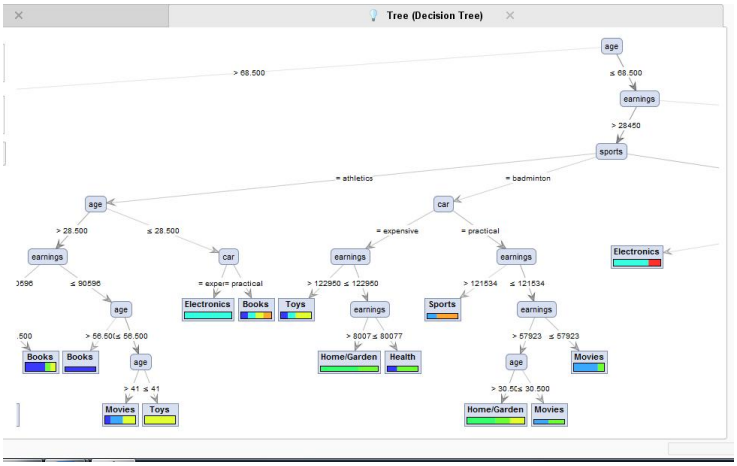


Кластерлеудің визуализациясы

Кластерлі талдаудың нәтижесі болып алғашқы жиын элементтерін қамтитын кластерлер жиыны болып табылады. Ол келесі сұрақтардың

шешімдерін қарастырады: алынған кластерге бөліну кездейсоқ емес па; бөліну сенімді және деректер таңдауында тұрақты бола ма; кластерлеу нәтижесі мен айнымалы арасында өзара қарым-қатынас бар ма.

Классификация тапсырмасын шешу үшін Decision Tree операторы қолданылады, мұнда сатып алушылардың жас ерекшелігі бойынша қандай тауарларға артықшылықтарын беретіні, тапқан табыстары мен өмір сүру деңгейі бойынша қандай тауарларды көбірек алатыны жөнінде шешім қабылданған. Мұндағы шешім ағаштарынның нәтижесіндегі әтүрлі түстер бұл – әртүрлі класстар, яғни, бір ғана түс болса, онда деректер тек бір класстан алынған. Екі түс болса, онда екі классқа жататынын білдіреді. Берілген суретте шешімдер ағашы граф түрінде бейнеленген. Оның сипаттамасын бөлек терезеден көруге болады (Қосымша А).



Шешімдер ағашы

AssociationRules нәтижелері бар кестесінде деректерді талдау негізіндегі тауарлар мен олардың сапасы ұсынылған. Сапа баған бойынша орналастырылған: тауардың сатылу жеңілдігі, тауардан түсетін пайда, тауарды сатуға кететін шығындар. Соңғы бағанда тауарды сатылымға қалдыру туралы шешім көрсетіледі, коэффициент саны үлкен болған сайын тауарды қалдыру ықтималдығы жоғары болады

Tree (Decision Tree)

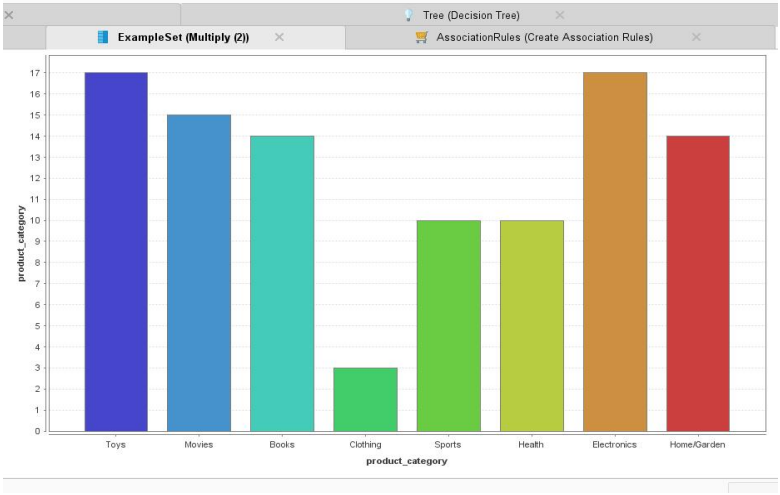
ExampleSet (Multiply (2))

AssociationRules (Create Association Rules)

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
486	Toys	Movies, Clothing	0.133	0.200	0.680	-1.200	0.044	1.500	1.083
487	Toys	Home/Garden, Clothing	0.133	0.200	0.680	-1.200	0.044	1.500	1.083
488	Toys	Electronics, Clothing	0.133	0.200	0.680	-1.200	0.044	1.500	1.083
489	Toys	Sports, Health	0.133	0.200	0.680	-1.200	0	1	1
490	Toys, Sports	Clothing	0.067	0.200	0.800	-0.600	0	1	1
491	Toys	Health, Clothing	0.133	0.200	0.680	-1.200	0.044	1.500	1.083
492	Books, Sports	Health	0.067	0.200	0.800	-0.600	-0.111	0.375	0.583
493	Movies, Home/Gard...	Health	0.067	0.200	0.800	-0.600	-0.111	0.375	0.583
494	Toys, Health	Movies, Home/Garden	0.067	0.200	0.800	-0.600	-0.089	0.429	0.667
495	Movies, Home/Gard...	Clothing	0.067	0.200	0.800	-0.600	0	1	1
496	Toys	Movies, Electronics, Clothi...	0.133	0.200	0.680	-1.200	0.044	1.500	1.083
497	Toys	Movies, Books, Health	0.133	0.200	0.680	-1.200	0.044	1.500	1.083
498	Toys	Movies, Sports, Health	0.133	0.200	0.680	-1.200	0	1	1
499	Toys, Sports	Movies, Clothing	0.067	0.200	0.800	-0.600	0.022	1.500	1.083
500	Movies, Toys, Sports	Clothing	0.067	0.200	0.800	-0.600	0	1	1

Тауарлардың сатылуы туралы болжам

Сонымен қатар шешімдер ағашындағы граф түрінде берілген нәтижені диаграмма түрінде ұсынуға болады. Мұнда өнім категориясымен өнім көлемі алынады. Яғни, тұтынушылардың қай өнімді көбірек қажет ететіні көрсетіледі. Біздің жағдайда ойыншықтар мен электроника бұйымдарына деген сұраныс өте жоғары болды



Сатылымдар көлемінің диаграммасы