DATU APSTRĀDE AR IBM SPSS MODELER RĪKA PALĪDZĪBU
DATA PROCESSING USING THE IBM SPSS MODELER TOOL

Autors: **Jānis ČIŽIKS**
E-pasts: janisciziks@inbox.lv
Zinātniskais darba vadītājs: Dr.sc.ing., profesors **Pēteris GRABUSTS**
E-pasts: peteris.grabusts@rta.lv
Rēzeknes Tehnoloģiju akadēmija, Inženieru fakultāte,
Atbrīvošanas aleja 115, 3. korpuss. Rēzekne, LV-4601, Latvija

Abstract. *IBM SPPS Modeler platform is the commercial rival of RapidMiner platform, characterized by a low entrance threshold for beginners. Nonsense for beginners is expressed by the "autopilot" modes. Auto models (Auto Numeric, Auto Classifier) distinguishes several possible patterns with different parameters, which identify them better. Not an experienced analyst, using such a solution, is able to develop an adequate model. The SPSS user interface is constantly improving, making the system intuitive to understand. For simple tasks, such as fuling, there is no need for preparation in principle. This makes IBM SPSS Modeler a good solution for data analysis for beginners.*
Keywords: *Analysis tools, IBM SPSS Modeler, chronic diseases, neural networks*

Ievads

Šobrīd visā pasaulē arvien lielāku nozīmi dažādos procesos ieņem lielie dati, attīstot gan mākslīgo intelektu, gan ļaujot pieņemt datos balstītus lēmumus. Tehnoloģiju laikmetā gandrīz jebkuram uzņēmumam uzkrāto datu apjoms ir milzīgs, tāpēc bieži rodas jautājums, ko ar šiem datiem darīt.

SPSS Modeler ir vadošais vizuālo datu zinātnes un mašīnmācības risinājums. Tas palīdz uzņēmumiem paātrināt laiku, lai novērtētu un sasniegtu vēlamos rezultātus, paātrinot datu zinātnieku darbības uzdevumus. Vadošās organizācijas visā pasaulē paļaujas uz IBM datu sagatavošanai un atklāšanai, prognozējošai analīzei, modeļu pārvaldībai un izvietošanai, kā arī mašīnu apguvei.

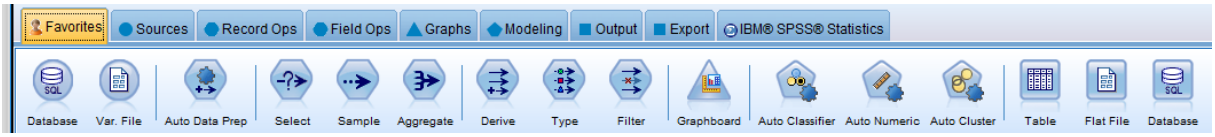
Visas *IMB SPSS Modeler* priekšrocības var aizēnot ar vienu trūkumu, kas samazina apjomīgu lietotāju skaitu. Galvenokārt runa ir par to, ka šī sistēma nav labākais līdzeklis lielo datu analīzei. Atribūti, kas padara SPSS viegli lietojamu ir pārāk ierobežoti liela mēroga pieejai, strādājot ar *Big Data* tehnoloģijām. Sliktākā gadījumā, t.i., pārslodzes gadījumā SPSS vienkārši pārstāj darboties. Neskatoties uz to, *IBM SPSS Modeler* joprojām ir populārs risinājums, jo tas ir viegli lietojams ar vienkāršu saskarni.

Tā kā tajā ir integrēts neironu tīklu risinājums, tas ir piemērotākais rīks lielo datu analīzei.

Pētījuma objekti un metodes
IBM SPSS Modeler iespējas

Programma ir integrēts neironu tīklu risinājums, tas ir piemērotākais rīks lielo datu analīzei.

Rīka izpildes vide tiek definēta kā straumējums (*stream).* Darba vides izvēlne parādīta 1. attēlā.

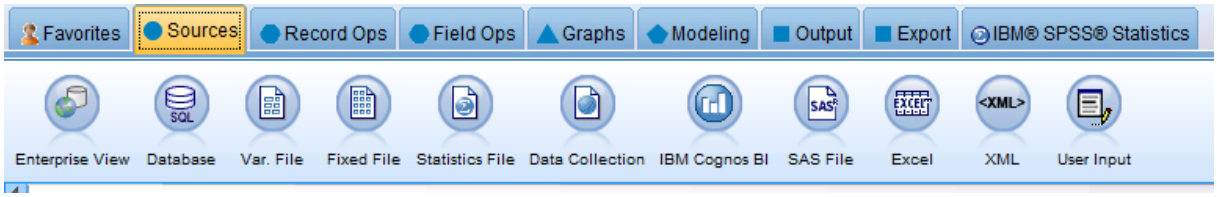


1. attēls. **Darba vides galvenie elementi**

Sadalē *Favorites* tiek definēti biežāk izmantojamie rīka elementi, kas visbiežāk tiek izmantoti dažādu konstrukciju realizēšanā.

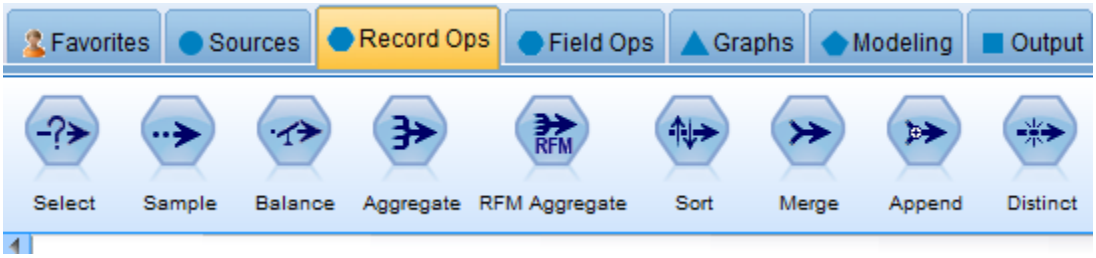
http://dx.doi.org/10.17770/het2019.23.4388

Datu apstrādes sākumā ir nepieciešams izvēlētie datu avotus (*Sources)*. Tie var būt gan Excel dati, vai SPSS avoti, teksta dati utt. (sk. 2. attēlu).



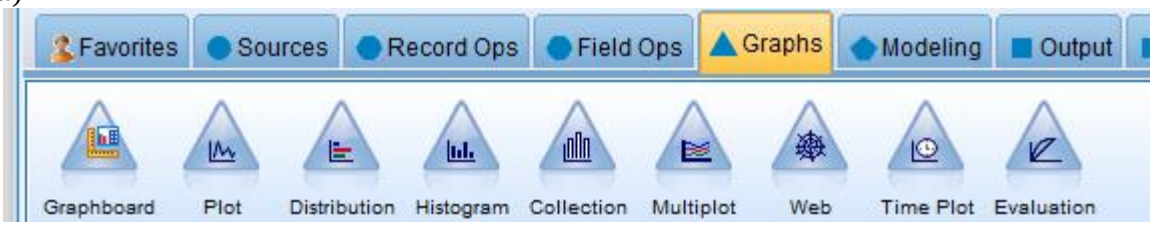
2. attēls. **Datu avotu izvēle**

Pēc datu izvēles ir nepieciešamība tos atdalīt pēc noteikta kritērija datu anlīzes nolūkā. Tas tiek darīts sadalē *Record Ops* (sk. 3. att.).



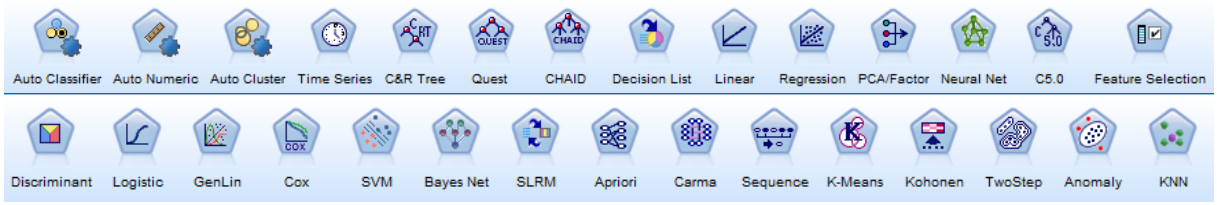
3. attēls. **Datu avotu sadalīšanas iespējas**

Modelēšanas rezultātā ir iespēja rezultātu vizualizēšanā izvēlēties grafisko analīzi (sk. 4. att.)



4. attēls. **Rezultātu vizualizēšanas iespējas**

Pats galvenais ir Modeler piedāvātais bagātīgais modelēšanas rīku klāsts (sk. 5. att.), kas iekļauj sevī gan neironu tīklu iespējas, gan klasterizāciju, lēmumu kokus gan citas iespējas.



5. attēls. **Modelēšanas rīku klāsts**

2. Datu izlases apraksts

Datu analīzei tika izvēlēta hronisko slimību indikatoru datu bāze (65356 ieraksti), kas ietver **124** dažādus indikatorus slimību aprakstos. Respondentiem tika uzdoti dažādi jautājumi par viņu slimības izpausmēm, kas arī tiek turpmāk analizēts.

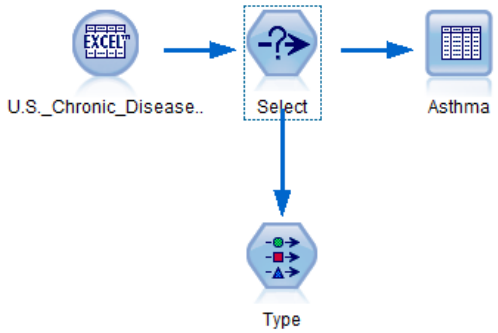
3. Datu izlases pamatotība

Autors saviem pētījuma nolūkiem izvēlējās konkrēto hronisko slimību – astma, jo dažādo pētījumu rezultāti liecina, ka tā ir būtiska problēma Latvijas apstākļos.

Astma ir elpceļu saslimšana. Tās simptomus izraisa iekaisums, kura rezultātā elpceļi ir pietūkuši, sašaurināti un pārmērīgi jutīgi pret kairinātājiem. Tas izraisa atkārtotas sēkšanas lēkmes, elpas trūkumu, smaguma sajūtu krūšu kurvī un klepu. Vieglākas lēkmes var pārvarēt bez ārstēšanas, bet ārstēšana parasti palīdz tikt ar tām galā daudz ātrāk. Piemērota ārstēšana var arī samazināt lēkmju atkārtošanās risku. Ja ir nopietna lēkme, ir jāmeklē neatliekamā palīdzība.

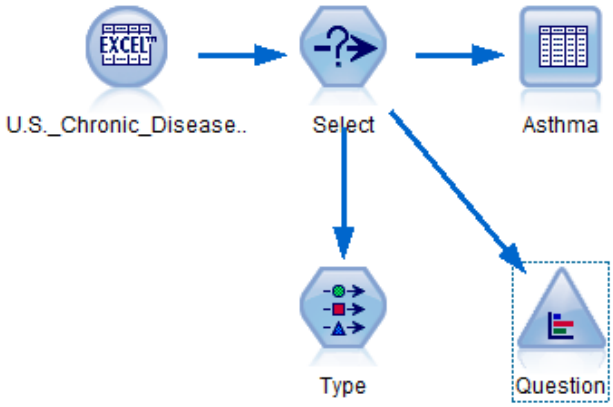
4. Modeļa izstrāde

Modeļa izstrāde sākas ar datu avotu izvēli, astmas slimnieku atlasi no lielās datu bāzes. Lauku tipus un atlasīto datu struktūru Modeler veic automātiski. Iesāktais modelis attēlots att. 6. Kā datu atlases kritērijs tika izvēlēts vaicājums *Topic = "Asthma"*.



6. attēls. **Sākotnējo datu atlases**

Projekta priekšizpētes stadijā tika manuāli izskaitļots astmas slimnieku skaits datubāzē un to statistiskie rādītāji. Modeler piedāvā to automatizēt, pievienojot statistikas atskaites grafiskā veidā atbilstoši lauks *Question* vērtībai. Šajā laukā Excel tabulā apkopota informācija par pacientiem uzdotajiem jautājumiem par viņu veselības stāvokli (sk. att. 7).

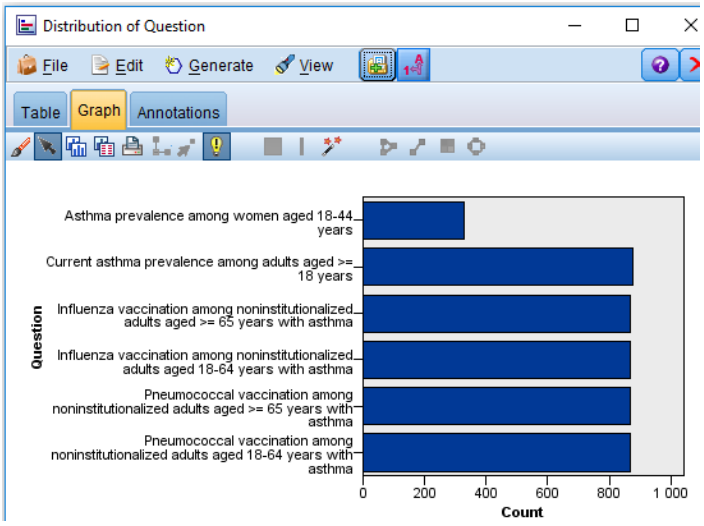


7. attēls. **Grafiskās informācijas uzdošana modelī**

Startējot modeli pašreizējā stāvoklī – tika iegūtas šādas statistikas atskaites, kas atbilda manuāli skaitļotajam (sk. 8.-9. att.).

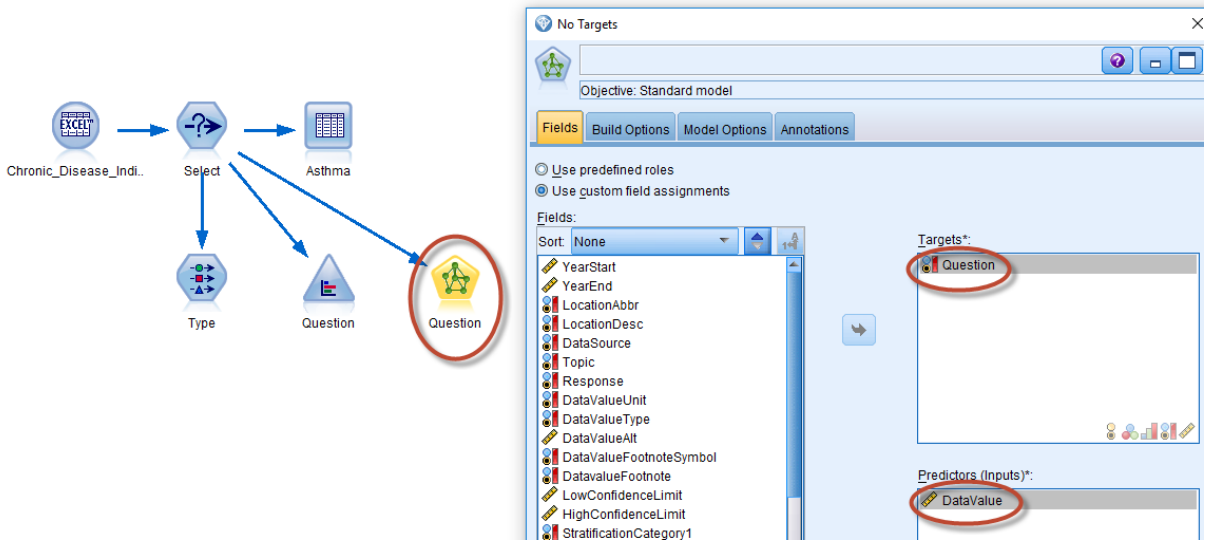
Value	Proport...	%	Count
Asthma prevalence among women aged 18-44 years		6.98	327
Current asthma prevalence among adults aged >= 18 years		18.69	875
Influenza vaccination among noninstitutionalized adults aged 18-64 yea...		18.58	870
Influenza vaccination among noninstitutionalized adults aged >= 65 yea...		18.58	870
Pneumococcal vaccination among noninstitutionalized adults aged 18-...		18.58	870
Pneumococcal vaccination among noninstitutionalized adults aged >= ...		18.58	870

8. attēls. **Astmas slimnieku procentuālais sadalījums pēc uzdotā jautājuma**



9. attēls. **Astmas slimnieku skaits sadalījums pēc uzdotā jautājuma**

Modelī pievieno neironu tīklu iespējas un izvēlas manuālu lauku tipu izvēli un kā mērķi (*Target*) paņem lauku *Question*. Neironu tīkla ieejā tiek padotas lauka *DataValue* vērtības. Modeļa realizācija parādīta 10. att.

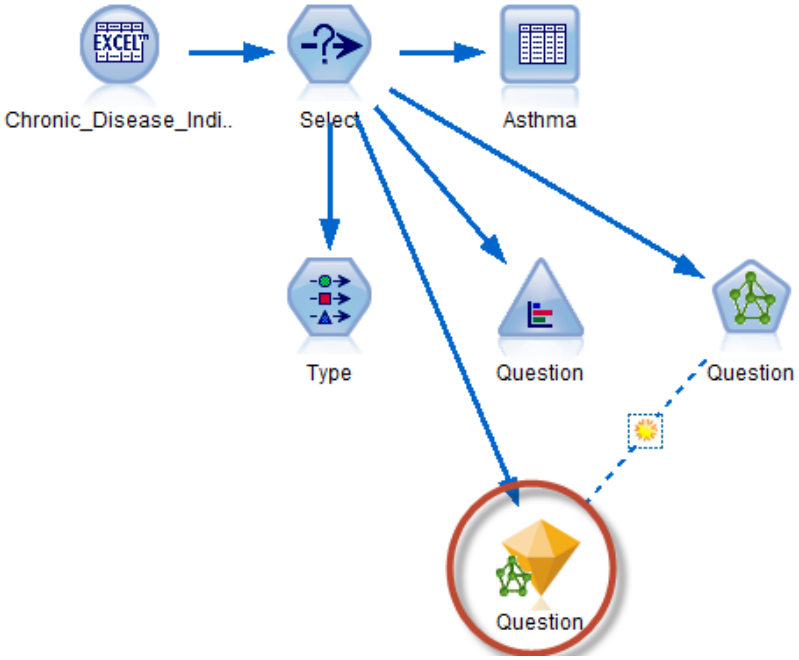


10. attēls. **Neironu tīkla realizācija**

Neironu tīkla modelis ar *Modeler* rīka palīdzību ir veiksmīgi nokonfigurēts un gatavs izpildei.

Rezultāti un to izvērtējums

Startējot neirona tīkla modeli tika iegūts šāds modeļa konfigurācijas gala rezultāts (sk. 11. att.).



11. attēls. **Neironu tīkla modeļa realizācijas gala rezultāts**

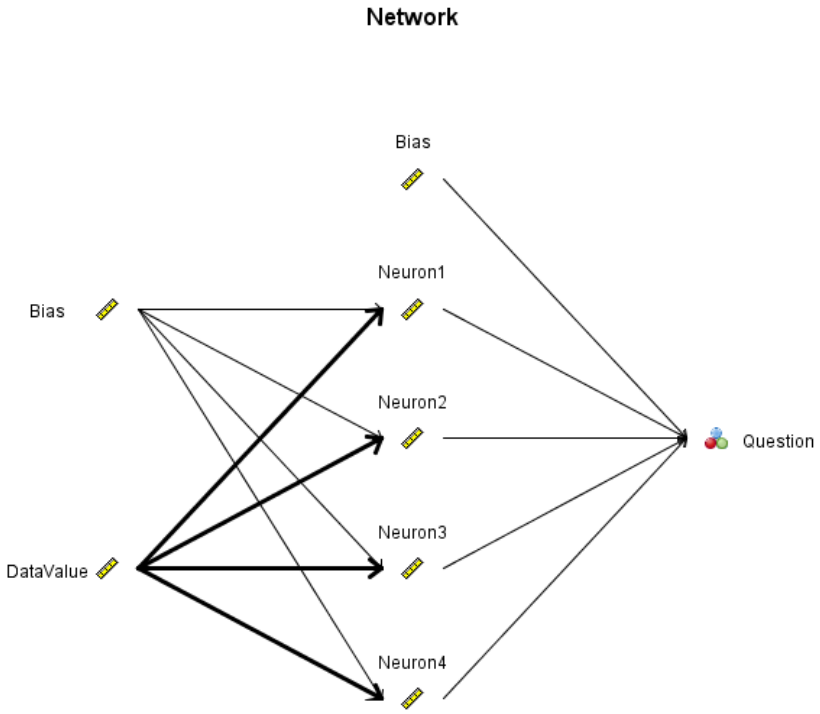
Izpildes rezultātā tika dots kopējais novērtējums par modeļa darbību (sk. 12. att.), pēc kura var secināt, ka pēc noklusējuma tika izvēlēts neironu tīkla algoritms daudzslāņu perceprons (*Multilayer perceptron)*ar 4 slēptā slāņa neironiem, kā mērķa funkciju izmantojot lauka *Question* vērtības.

Model Summary

Target	Question
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	4

12. attēls. **Kopsavilkums par modeļa darbību**

Pati neironu tīkla struktūra parādīta 13. att. (Saite *Bias* ir neironu tīkla komponente, kas faktiski ir kaut kāda konstante vai lineāra funkcija neironu tīkla algoritma darbības nolūkā).



13. attēls. **Modeļa neironu tīkla struktūra**

Paši svarīgākie neironu tīkla darbības rezultāti parādīti 14. att.

Classification for Question

Overall Percent Correct = 72,1%

Observed	Predicted					
	Asthma prevalence among women aged 18-44 years	Current asthma prevalence among adults aged >= 18 years	Influenza vaccination among noninstitutionalized adults aged 18-64 years with asthma	Influenza vaccination among noninstitutionalized adults aged >= 65 years with asthma	Pneumococcal vaccination among noninstitutionalized adults aged 18-64 years with asthma	Pneumococcal vaccination among noninstitutionalized adults aged >= 65 years with asthma
Asthma prevalence among women aged 18-44 years	0,0%	96,8%	3,2%	0,0%	0,0%	0,0%
Current asthma prevalence among adults aged >= 18 years	0,0%	98,9%	1,1%	0,0%	0,0%	0,0%
Influenza vaccination among noninstitutionalized adults aged 18-64 years with asthma	0,0%	1,2%	45,9%	3,6%	49,3%	0,0%
Influenza vaccination among noninstitutionalized adults aged >= 65 years with asthma	0,0%	0,0%	2,3%	88,3%	4,3%	5,1%
Pneumococcal vaccination among noninstitutionalized adults aged 18-64 years with asthma	0,0%	1,6%	40,3%	2,4%	55,6%	0,0%
Pneumococcal vaccination among noninstitutionalized adults aged >= 65 years with asthma	0,0%	0,0%	0,5%	6,9%	1,0%	91,5%

14. attēls. **Neironu tīkla darbības klasifikācijas rezultāti**

Secinājumi

Var secināt, ka neironu tīkls kopumā (72,1%) gadījumā pareizi klasificējis atbilžu rezultātus starp aptaujāmiem.

Būtiskākie secinājumi:

- astmas izplatība starp sievietēm 18-44 gadu vecumā ir 96,8%;
- pašreizējā astmas slimnieku izplatība pieaugušo vidē, kas vecāki par 18 gadiem, ir 98,9%;
- pretgripas vakcināciju ir veikuši 88,3% astmas slimnieki vecumā virs 65 gadiem;
- vakcināciju pret pneimokoku ir veikuši 91,5% astmas slimnieki vecumā virs 65 gadiem.

Literatūra

1. https://www.ibm.com/products/spss-modeler
2. https://gengo.ai/datasets/18-free-life-sciences-medical-datasets-for-machine-learning/?utm_campaign=c&utm_medium=quora&utm_source=rei
3. https://www.cdc.gov/mmwr/pdf/rr/rr6401.pdf

4. http://www.astmaalergija.lv/?id_p=1&id=2
5. https://www.g2crowd.com/products/ibm-spss-modeler/reviews
6. https://www.youtube.com/watch?v=_0YtxWUfACI
7. http://www.spss.com.hk/software/modeler/