

# *PASW Modeler 13.0*

> Stručný průvodce

SPSS®



# Obsah

1.	ÚVOD.....	3
2.	INSTALACE A SPUŠTĚNÍ PROGRAMU.....	5
3.	HLAVNÍ NABÍDKA PROGRAMU .....	6
4.	PANEL NÁSTROJŮ (TOOL BAR) .....	8
5.	ZÁKLADY PRÁCE V PASW MODELER.....	10
6.	VYTVÁŘENÍ PROUDŮ.....	12
7.	KOMENTÁŘE.....	16
8.	CACHE.....	17
9.	UZLY .....	18
10.	UZLY SOURCES.....	20
11.	UZLY RECORD OPERATIONS.....	22
12.	UZLY FIELD OPERATIONS.....	24
13.	UZLY GRAPHS .....	27
14.	UZLY MODELING .....	28
15.	UZLY OUTPUT .....	32
16.	UZLY EXPORT .....	33
17.	UZLY PASW STATISTICS .....	34
18.	CLEF .....	35
19.	SKRIPTY A PROGRAMOVÁNÍ .....	36
20.	TVORBA REPORTŮ A DOKUMENTACE .....	37
21.	SQL.....	38
22.	SYSTÉM NÁPOVĚD .....	40
23.	PŘÍKLAD.....	42

# 1. Úvod

---

PASW Modeler je nástroj pro data mining, který podporuje celý dataminingový proces a který svým uživatelům umožňuje rychlý přístup k datům, dále datové manipulace, konstrukci a ověřování modelů a jejich následné nasazení do reálného prostředí.

Program kombinuje pokročilé modelovací techniky se snadným způsobem ovládání, který umožňuje objevení a predikce užitečných informací v datech. Celý systém je navržen v souladu s metodikou CRISP-DM a podporuje klíčové aktivity, mezi které patří:

- tvorba zákaznických profilů a určení jejich hodnoty,
- detekce a predikce podvodů,
- detekce a predikce vazeb v datech z webu,
- predikce budoucích prodejních a růstových trendů,
- odhad účinnosti marketingových akcí,
- kreditní riziko,
- odhad rizik v monitorování procesů,
- predikce churnu,
- klasifikace, segmentace zákazníků,
- analýza velmi rozsáhlých dat,
- objevování skrytých vazeb a struktur.

Portfolio produktů PASW Modeler 13.0 zahrnuje:

**PASW Modeler** (dříve SPSS Clementine Client) – funkčně kompletní produkt, běžící na pracovní stanici uživatele (PC). Spouští se lokálně, nebo v distribučním modu společně s **PASW Modeler® Server**, čímž se podstatně zvýší výkon při zpracování velkých objemů dat. Systém je modulární podle typu pokrytých úloh:

- PASW Modeler Base (dříve Clementine Base)
- PASW Association (dříve Clementine Association)
- PASW Classification (dříve Clementine Classification)
- PASW Segmentation (dříve Clementine Segmentation)

**PASW® Modeler server 13** (dříve SPSS Clementine Server) – serverová verze PASW Modeler. Může být nastavena tak, aby běžela v distribučním modu ve spojení s jednou a více klientskými verzemi **PASW Modeler**.

**Clementine Batch** – dávkový (batch) mode pro spouštění opakujících se požadavků bez nutnosti zásahu uživatele a bez použití uživatelského rozhraní Clementine. Využití pouze spolu se serverovou verzí Clementine.

**PASW Modeler Solution Publisher** (dříve Clementine Solution Publisher) – přídatná komponenta pro export a spouštění proudů v externích aplikacích. Proudů jsou spouštěny pomocí PASW Modeler Solution Publisher Runtime. PASW Modeler Solution Publisher je instalován společně s PASW Modeler, jeho funkčnost je podmíněna samostatnou licencí.

**Cleo** – přídatná komponenta pro rychlé (on-line) uvádění modelů do praxe pomocí nástroje založeného na XML. Uživatelé mají simultánní přístup k modelu a skórují jednotlivé záznamy, skupiny záznamů nebo celou databázi přes uživatelské rozhraní internetového prohlížeče.

**PASW Text Analytics** (dříve Text Mining for Clementine) – přídatná komponenta pro analýzu nestrukturovaných textových informací a jejich konvertace do strukturovaného formátu.

**PASW Web Mining** (dříve Web Mining for Clementine) – přídatná komponenta pro analýzu dat z webu (web logů).

**CATS** – (Clementine Application Templates) – sada proudů, dat a postupů z konkrétních aplikací data miningu (CRM, Telco, Fraud, WebMining apod.), které slouží jako podklad pro vlastní aplikace.

**PES** – (Predictive Enterprise Services) – server pro správu životního cyklu modelu, pro uchování modelů, proudů a výstupů z PASW Modeler. Organizuje objekty podle různých kritérií (verze, témata, autoři, klíčová slova) a umožňuje jejich efektivní vyhledávání. PES zároveň podporuje uchovávání a spouštění objektů z dalších aplikací SPSS.


## 2. Instalace a spuštění programu

---

Instalace systému PASW Modeler probíhá v několika jednoduchých krocích:

- 1) CD PASW Modeler vložte do mechaniky a vyčkejte na automatické spuštění instalace.
- 2) Postupujte dle pokynů průvodce instalací.
- 3) Po instalaci se automaticky spustí License Authorization Wizard, nástroj pro vložení autorizačního nebo licenčního kódu. Pro více informací o autorizaci softwaru navštivte webové stránky <http://www.spss.cz/autorizace.htm>.

PASW Modeler spustíte několika způsoby:

- 1) Dvakrát klikněte na ikonu , kterou najdete na ploše (*desktop*),
- 2) PASW Modeler spustíte také z hlavního menu počítače pod tlačítkem *Start*.

### 3. Hlavní nabídka programu

---

Práce s PASW Modeler vyžaduje používání položek hlavního menu jen v minimálním rozsahu. Naprostou většinu akcí provedete přesouváním grafických prvků/ikon po pracovní ploše. Ikonám se v prostředí PASW Modeler říká uzly.

Hlavní menu je v následujícím uspořádání:

<b>File</b>	Vytvoření nového proudu, otevření existujícího proudu ze souboru nebo z úložiště, uzavření aktivního proudu, otevření/uzavření projektu, modelů, výstupů. Nastavení výchozích adresářů, otevření naposledy otevřených proudů, opuštění systému, export popisu proudu.
<b>Edit</b>	Kopírování/vkládání proudů, částí proudů nebo uzlů. Vyprázdnění obsahu výstupů a další manipulace s proudy nebo částmi proudů. Přiřazení dočasné paměti označenému uzlu ( <i>cache</i> ), viz kap. 7.
<b>Insert</b>	Vložení uzlu ze souboru nebo úložiště, vložení superuzlu ze souboru nebo úložiště, vložení části nebo celého proudu ze souboru nebo úložiště. Přidá na pracovní plochu libovolný uzel z nabídky na spodní liště pracovní plochy.
<b>View</b>	Nastavení viditelnosti jednotlivých částí pracovní plochy.
<b>Tools</b>	Velké množství nastavitelných prvků systému PASW Modeler: <ul style="list-style-type: none"><li>▪ <b>Server Login</b> – nastavení práce se serverem, připojení k serveru</li><li>▪ <b>Databases</b> – připojení k databázím</li><li>▪ <b>Repository</b> – nastavení práce s produktem <i>Predictive Enterprise Services (PES)</i></li><li>▪ <b>Encode Password</b> – zakódování hesla pro jeho utajení při volání Clementine Batch z příkazové řádky</li><li>▪ <b>Options</b> – nastavení systému<ul style="list-style-type: none"><li>○ <b>System Options</b> – přidělení dočasné paměti</li><li>○ <b>User Options</b> – uživatelské rozhraní PASW Modeler; výstrahy a upozornění, barevné rozlišení grafů, optimalizace proudů a SQL, PMML (Predictive Model Markup Language)</li><li>○ <b>Helper Applications</b> – nastavení přístupu k aplikacím, které mohou s PASW Modeler spolupracovat</li></ul></li><li>▪ <b>Stream Properties</b> – vlastnosti proudů<ul style="list-style-type: none"><li>○ <b>Options</b> – počet zobrazených desetinných míst ve výstupech, formát desetinného oddělovače, referenční datum, formát data a času a další nastavení chování proudů</li><li>○ <b>Layout</b> – velikost ikon, rozvržení pracovní plochy apod.</li><li>○ <b>Messages</b> – zobrazení informací o průběhu výpočtu</li><li>○ <b>Parameters</b> – nastavení uživatelských parametrů</li><li>○ <b>Deployment</b> – uložení proudu ve formě tzv. scénáře pro použití např. v Predictive Applications</li><li>○ <b>Script</b> – okno pro vytváření a spouštění skriptů, tj. vlastních programů, které automatizují práci s proudy</li><li>○ <b>Globals</b> – zobrazení globálních statistik. Výpočet provádí uzel <i>Set Globals</i> – viz kapitola 12</li><li>○ <b>Search</b> – hledání uzlu v proudu</li></ul></li></ul>

- **Comments** – seznam a možnosti editace komentářů k superuzlům a proudům
- **Annotations** – textový popis proudu
- **Enterprise View Connestions** – spojení do úložiště, které je organizováno programem *Predictive Enterprise Services*
- **Standalone Script** – dialogové okno pro psaní skriptů, které nejsou přiřazeny konkrétnímu proudu
- **Set Session Parameters** – nastavení parametrů platných pro všechny otevřené proudy
- **Manage Palettes** – úprava podoby nabídky na spodní liště pracovní plochy, výběr uzlů do záložek
- **Extensions** – uživatelská rozšíření pomocí CLEF
- **CEMI** – načtení CEMI uzlů do systému
- **Predictive Applications 4.x Wizard** – průvodce propojení PASW Modeler a Predictive Applications
- **Publish stream** – Export souboru ve speciálním formátu pro PASW Modeler Solution Publisher
- **Execute** – spuštění celého proudu
- **Execute Selection** – spuštění vybrané části proudu
- **Stop Execute** – přerušení vykonávání proudu

**SuperNode** Tvorba, editace a terminace superuzlů

**Window** Manipulace s okny

**Help** Menu obsahuje podrobný popis uzlů a demonstračních příkladů jak pracovat s PASW Modeler

## 4. Panel nástrojů (Tool Bar)

---

V PASW Modeler, stejně jako v mnoha jiných programech, přistoupíte k nejčastěji používaným funkcím rychle pomocí panelu nástrojů (Tool Bar).



Jednotlivé ikony vykonají tyto funkce (zleva):



**File ... New Stream** – založení nového proudu



**File ... Open Stream** – otevření existujícího proudu



**File ... Save Stream** – uložení aktivního proudu



**File ... Print** – tisk aktivního proudu



**Edit ... Cut** – vyjmutí vybrané oblasti do schránky



**Edit ... Copy** – kopírování vybrané oblasti do schránky



**Edit ... Paste** – vložení oblasti ze schránky na plochu



**Edit ... Undo** – krok zpět (poslední vykonaná akce)



**Edit ... Redo** – opětovné provedení zrušené akce; aktivní pouze v případě použití funkce *Undo*



**Tools ... Stream Properties ... Search** – hledání uzlu v proudu



**Tools ... Stream Properties** – editace vlastností aktivního proudu



**Preview SQL** – náhled na SQL kód; ikona je aktivní pouze v případě práce s databázovými zdroji (vytváření náhledu SQL musí být povoleno)



**Tools ... Execute** – spuštění aktivního proudu



**Tools ... Execute Selection** – spuštění vybrané části aktivního proudu



**Tools ... Stop Execution** – přerušení běhu aktivního proudu; ikona je aktivní pouze při spuštění proudu



**Tools ... Publish Stream** – publikování proudu pomocí PASW Modeler Solution Publisher, aktivní pouze po jeho zalicencování





**SuperNode ... Create Supernode** – přidání superuzlu do proudu; ikona je aktivní v případě označení vhodné skupiny uzlů



**SuperNode ... Zoom In** – pohled do superuzlu; ikona je aktivní v případě, že v proudu označíme superuzel



**SuperNode ... Zoom Out** – vystoupení ze superuzlu; ikona je aktivní v případě, že vstoupíme do existujícího superuzlu



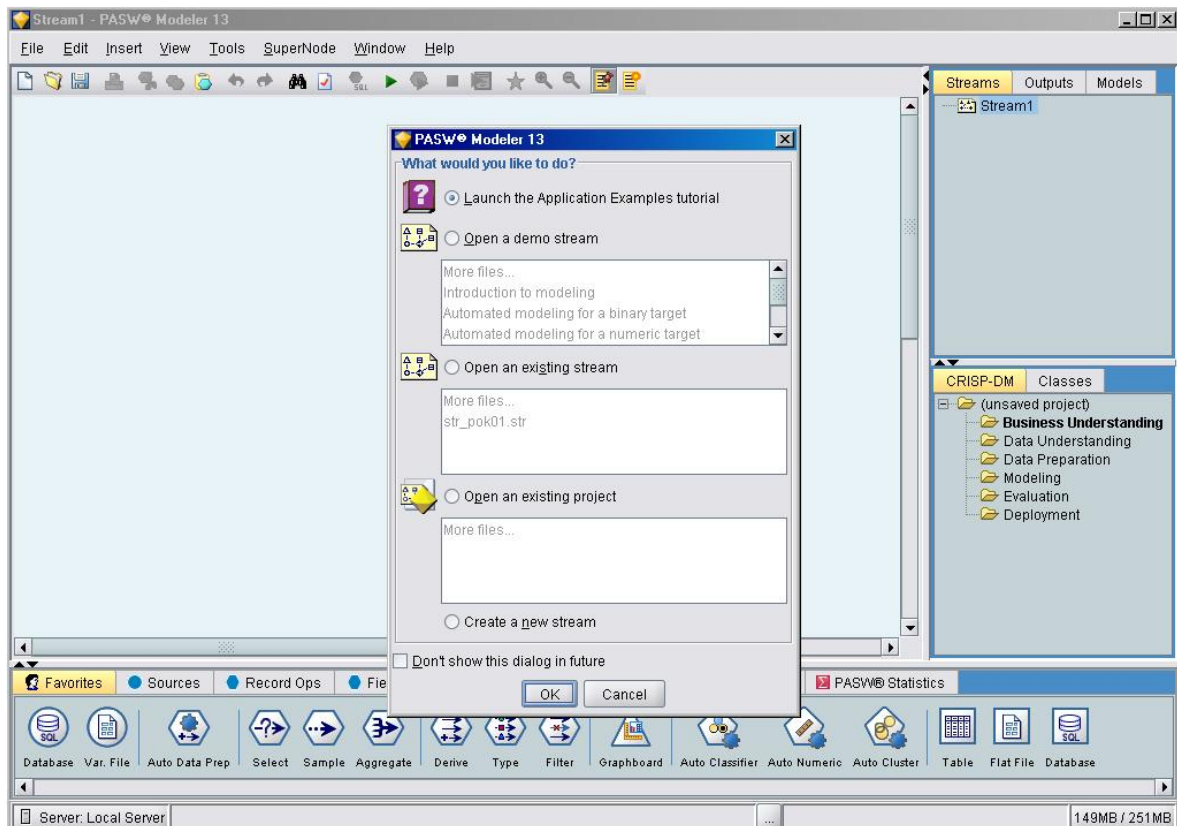
**Show/hide comments** – zobrazí/skryje komentáře v proudu



**Insert ... New Comment** – vloží nový komentář k označené skupině uzlů

## 5. Základy práce v PASW Modeler

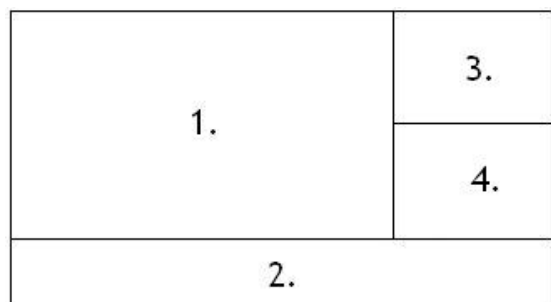
Po spuštění programu se objeví hlavní okno PASW Modeler spolu s úvodním dialogovým oknem (obr. 1), které nabízí otevření nového proudu nebo nápovědy, přístup k již existujícím proudům a projektům.

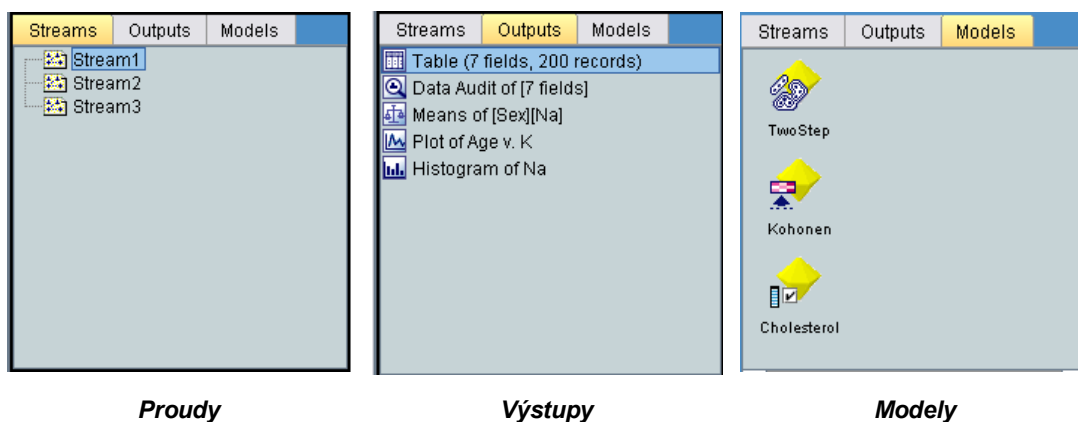


Obr.1: Hlavní okno programu PASW Modeler 13.0

Hlavní okno je možné rozdělit na čtyři základní části:

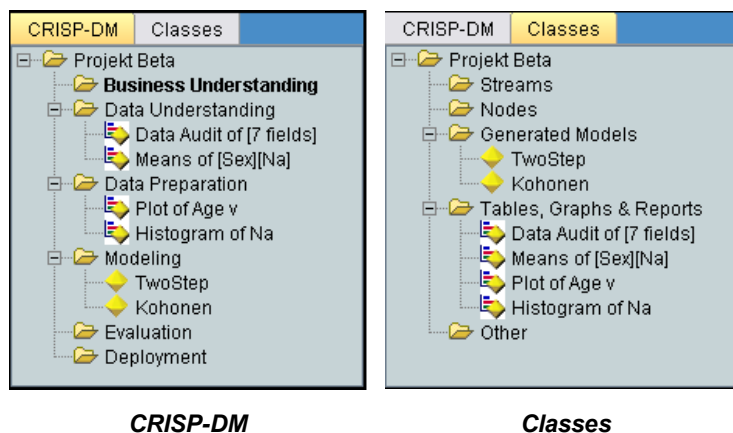
1. **Pracovní plocha** (*stream canvas*) – prostor, ve kterém vytvoříte proudy a provedete požadované akce
2. **Lišta s paletami uzlů** (*palettes*) – seznam všech uzlů seskupených podle tematických oblastí
3. **Správce výstupů** – prostor pro řízení a prohlížení proudů, výsledků a modelů. Obsahuje tři záložky:
  - a. **Streams (proudy)** – seznam všech aktuálně otevřených proudů
  - b. **Outputs (výstupy)** – seznam všech typů výstupů z PASW Modeler
  - c. **Models (modely)** – vytvořené modelyPříklad *správce výstupů* je na obr. 2.





Obr. 2: Správce výstupů

4. **Správce projektu** – prostor, ve kterém se systematickým způsobem ukládají výsledky práce v členění podle standardní metodologie CRISP-DM (obr. 3). Obsahuje seznam všech aktuálně otevřených proudů, také řídí ukládání/otevírání nových proudů.



Obr. 3: Správce projektu

## 6. Vytváření proudů

PASW Modeler je program, který se ovládá pomocí vizuálních nástrojů. Práce s PASW Modeler je hlavně práce s daty. V nejjednodušší formě lze postup charakterizovat třemi základními kroky: *načtení dat*, *datové manipulace* a *výstup*. Každý krok je složen z jedné nebo více procedur, které jsou reprezentovány ikonami nazývanými *uzly (nodes)*.

Uzel představuje základní jednotku procesu při práci s daty, např. načtení dat z textového souboru, filtrování chybějících hodnot, spojení s jiným souborem, modelování logistickou regresí či uložení výsledků do databáze.

Jednotlivé uzly se vzájemně spojují do proudů, které pak představují posloupnost operací.

### 6.1 Přidání uzlu do proudu

Proud vytvoříte či doplníte jedním z postupů:

- kliknutím označte koncový uzel v proudu, ke kterému budete nový uzel napojovat, a dvojím kliknutím na požadovaný uzel na paletě jej připojte,
- uzel přetáhněte myší na pracovní plochu,
- jednou klikněte na požadovaný uzel (označení uzlu) a následně klikněte na pracovní plochu,
- v hlavním menu zvolte *Insert* a vyberte uzel.

### 6.2 Odstranění uzlu z proudu

Uzly z plochy odstraníte:

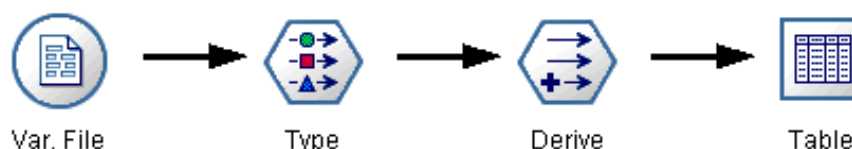
- klikněte na uzel a stiskněte klávesu *Delete*,
- klikněte na uzel a vyvolejte místní menu, ve kterém vyberete *Delete*,
- klikněte na uzel a v hlavním menu zvolte *Edit* → *Delete* → *Selected Nodes*.

### 6.3 Spojení uzlů do proudu

Uzly zapojíte do proudů tak, aby bylo možné vykonávat sekvence operací, které na sebe logicky navazují. Orientované spojnice uzlů reprezentují předávání dat a metadat mezi uzly. Základní způsoby spojení dvou uzlů do proudu:

- kliknutím označte poslední uzel v proudu a potom dvakrát klikněte na požadovaný uzel na paletě,
- oba spojované uzly přesuňte na pracovní plochu, kurzor umístěte na výchozí uzel, stiskněte prostřední tlačítko myši (kolečko) a táhněte směrem k uzlu cílovému. V momentě, kdy je kurzor nad cílovým uzlem, tlačítko myši uvolněte,
- klikněte na výchozí uzel a vyvolejte místní menu, ve kterém vyberete *Connect*, pak klikněte na cílový uzel.

Ukázka spojených uzlů je na obr. 4.



Obr. 4: Spojení uzlů do proudu

## 6.4 Odpojení uzlů

Způsobů, jakými uzly z proudu odpojíte, je opět více:

- označte uzel v proudu a *pravým* tlačítkem myši vyvolejte místní menu a vyberte položku *Disconnect*,
- dvakrát klikněte prostředním tlačítkem (kolečkem) myši na uzel, který chcete odpojit,
- označte uzly, které chcete odpojit, *pravým* tlačítkem myši vyvolejte místní menu a zvolte položku *Disconnect Nodes* (více uzlů najednou označíte buď tažením myši nad požadovanou oblastí nebo použitím klávesy *Ctrl*).

## 6.5 Editace uzlů

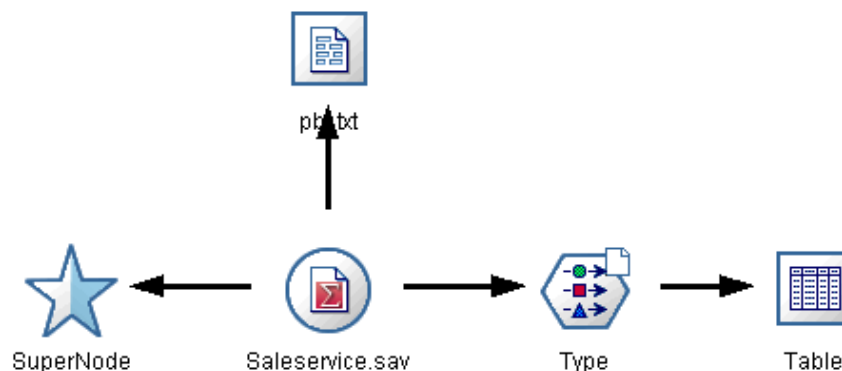
Po napojení uzlu do proudu je většinou nutná jejich editace. Jejím cílem je nastavení konkrétních parametrů pro konkrétní úlohu. Např. ve skupině uzlů *Sources* nastavujeme cestu k datovému zdroji, v uzlech *Field Ops.* výběr proměnných, kterých se operace týkají, v uzlech *Modeling* parametry jednotlivých modelů apod.

Vlastní editování názvu uzlu a jeho popis provedete v záložce *Annotations*. Jde o nepovinný, ale doporučený krok, který zpřehledňuje práci s proudy. Přístupy k editaci uzlů:

- označte myší požadovaný uzel, *pravým* tlačítkem myši vyvolejte doplňkové menu, zvolte položku *Edit*, ze záložek zvolte požadovanou část editace,
- dvakrát klepněte na požadovaný uzel.

## 6.6 Superuzly uzlů

Pro větší přehlednost uspořádání velkého počtu uzlů na pracovní ploše je zavedena možnost tvorby tzv. superuzlů (SuperNode). Jde o zapouzdření vybrané skupiny uzlů do jednoho většího superuzlu.



Obr. 5: Proud se superuzlem

Vytvoření superuzlu:

- myší označte uzly, které chcete do superuzlu zahrnout (stiskněte levé tlačítko myši a pohybem po ploše označte najednou požadované uzly),
- *pravým* tlačítkem myši vyvolejte místní menu,
- vyberte položku *Create SuperNode* nebo v panelu nástrojů klikněte na odpovídající ikonu (viz oddíl 5).

Náhled do superuzlu:

- označte myší superuzel, *pravým* tlačítkem myši vyvolejte místní menu, zvolte položku *Zoom In*,
- označte myší superuzel a z panelu nástrojů zvolte tlačítko *Zoom In*,
- editujte vlastnosti superuzlu a v dialogu zvolte tlačítko *Zoom In*,
- v seznamu proudů vyberte superuzel.

Zrušení superuzlu:

- označte myší superuzel,
- pravým tlačítkem myši vyvolejte místní menu,
- zvolte položku *Expand*.

Zamknutí superuzlu heslem:

- editujte superuzel a tlačítkem *Lock Node* vlevo nahoře vyvolejte dialog zadávání hesla,
- zadejte (*Password*) a potvrďte (*Confirm Password*) heslo.



**Obr. 6: Zamknutí superuzlu heslem**

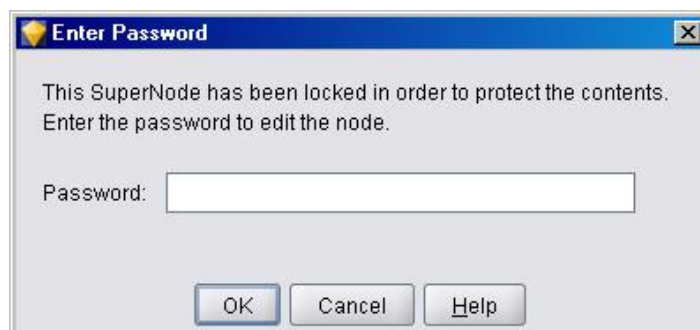
- v ikoně zamknutého superuzlu se objeví malý *zámeček*



SuperNode

**Obr. 7: Ikona zamknutého superuzlu**

- po zavření a znovuotevření proudu nebude možné do zamknutého superuzlu bez hesla nahlédnout, při pokusu o nahlédnutí se objeví dialog požadující zadání hesla




**Obr. 8: Dialog zadání hesla pro vstup do zamknutého superuzlu**

- jakmile zadáte heslo, bude superuzel volně přístupný

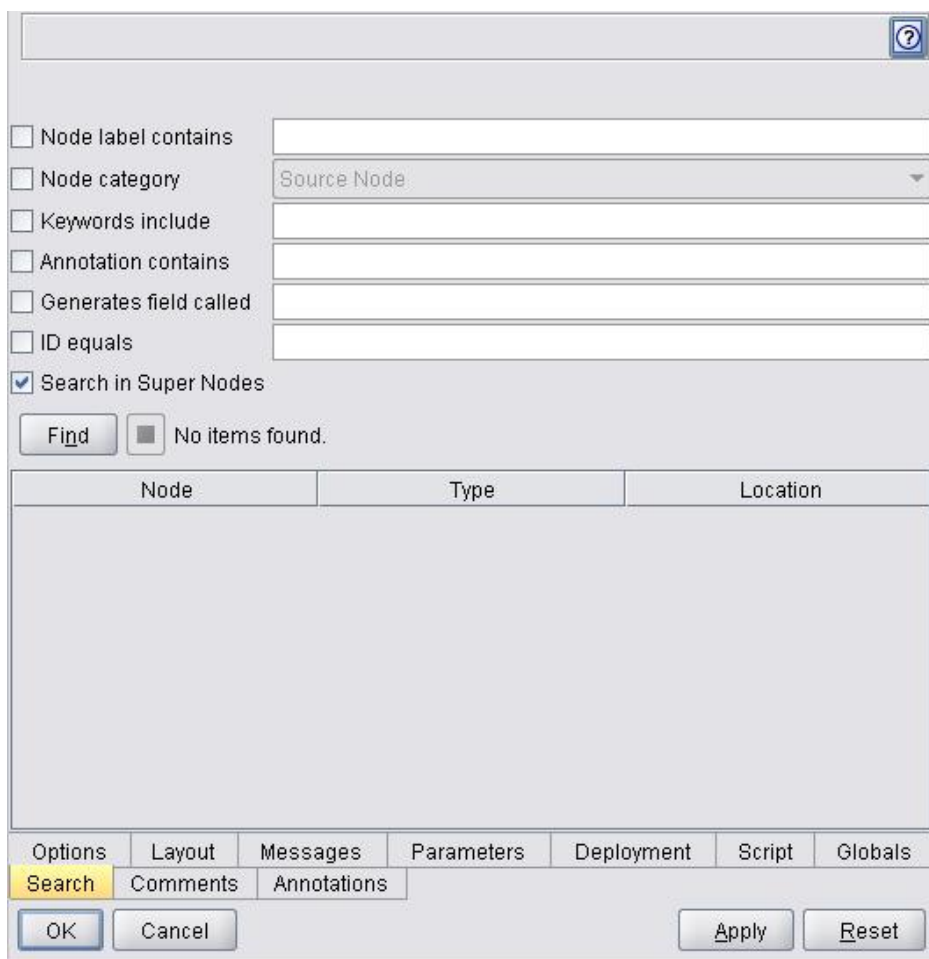
## 6.7 Vyhledávání uzlů v proudu

Vyhledávání uzlů je užitečné v rozsáhlých proudech. Vyhledávací formulář (obr. 9) lze vyvolat několika způsoby:

- pomocí ikony *Search for nodes in the current stream* ,
- pomocí hlavního menu *Tools → Stream Properties → Search*.

Vyhledávat lze podle různých kritérií, která lze kombinovat:

- popis uzlu,
- kategorie uzlu,
- klíčová slova,
- anotace k uzlu,
- název odvozované proměnné,
- ID uzlu (ID vytvoří PASW Modeler automaticky při vzniku uzlu),
- vyhledávání/nevyhledávání v superuzlech.




Node	Type	Location
------	------	----------

**Obr. 9: Formulář vyhledávání uzlu v proudu**


## 7. Komentáře

Komentáře slouží ke zpřehlednění pracovní plochy a proudů. V PASW Modeler lze přiřadit komentáře k jednotlivým uzlům či ke skupinkám uzlů. Dále je možné umístit komentář na pracovní plochu bez vazby ke konkrétnímu uzlu. Ukázka komentářů je na obr. 10.

Umístění komentáře na pracovní plochu:

- komentář na pracovní plochu lze umístit několika způsoby:
  - a. kdekoli na volném místě pracovní plochy klikněte pravým tlačítkem a v místním menu vyberte volbu *New Comment*,
  - b. v hlavním menu zvolte *Insert → New Comment*,
  - c. stiskněte tlačítko *Insert a new comment*  na nástrojové liště,
- na pracovní ploše se objeví nový komentář, do kterého lze rovnou zapisovat, neboť se nachází v editačním modu,
- pro ukončení zadávání klikněte myší kamkoli na volné místo pracovní plochy.

Umístění komentáře k uzlu/skupince uzlů:

- označte myší uzel/skupinku uzlů,
- komentář k uzlu/skupince uzlů lze umístit několika způsoby:
  - a. pravým tlačítkem myši na některém označeném uzlu vyvolejte místní menu a vyberte volbu *New Comment*,
  - b. v hlavním menu zvolte *Insert → New Comment*,
  - c. stiskněte tlačítko *Insert a new comment*  na nástrojové liště,
- na pracovní ploše se objeví nový komentář s čárkovanou vazbou ke všem označeným uzlům, do kterého lze rovnou zapisovat, neboť se nachází v editačním modu,
- pro ukončení zadávání klikněte myší kamkoli na volné místo pracovní plochy.

Správa komentářů:

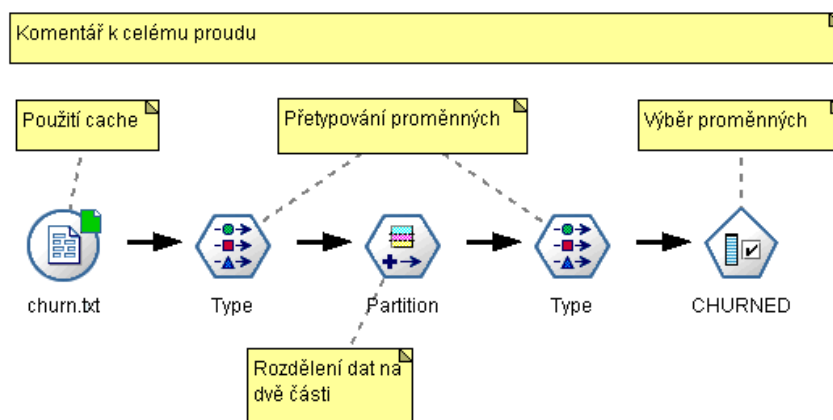
- komentáře lze společně spravovat pomocí vlastností proudu v menu *Tools → Stream Properties → Comments*

Editace komentářů:

- komentáře lze editovat v editačním modu, který získáte dvojklikem levým tlačítkem myši na příslušném komentáři

Skrytí/odkrytí komentářů:

- komentáře lze na pracovní ploše skrýt/odkryt pomocí ikony  na nástrojové liště



Obr. 10: Ukázka komentářů

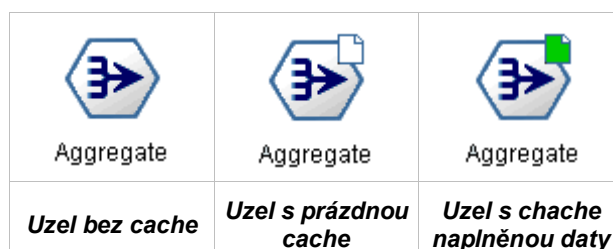


## 8. Cache

K optimalizaci běhu proudů se používá systém dočasných pamětí (*cache*). Cache lze použít na všech uzlech kromě terminálních (koncových) uzlů. Po nastavení cache a spuštění proudů je dočasná paměť naplněna daty. Při opakovaném spouštění jsou pak data čtena přímo z cache, nikoli z původních zdrojů. Smyslem používání cache je hlavně nezatěžování sítě, vyšší rychlost při práci s daty, ale také uchování mezivýsledků.

Ve spojení s uzlem *Sample* si můžete velmi pohodlně uložit náhodný výběr z velkého rozsahu dat a opětovně je použít k další práci. Čtete-li data z databáze, která následně agregujete, je efektivnější opatřit uzel **Aggregate** dočasnou pamětí a nezatěžovat tak databázi neustálým načítáním stejných dat.

Uzel s dočasnou pamětí poznáte podle malé ikony *dokumentu* v pravém horním rohu. Je-li cache naplněna daty, je ikona *dokumentu* zbarvena do zelena (viz obr. 11).



Obr. 11: Cache

Nastavení cache:

- označte myší požadovaný uzel,
- pravým tlačítkem myši vyvolejte místní menu,
- zvolte položku *Cache*,
- z doplňkového menu vyberte položku *Enable*.

Zrušení nastavené cache:

- označte myší požadovaný uzel,
- pravým tlačítkem myši vyvolejte místní menu,
- zvolte položku *Cache*,
- z doplňkového menu vyberte položku *Disable*.

Vyprázdnění plné cache (uplatňováno také v případě, kdy chceme změnit obsah cache novými daty):

- označte myší požadovaný uzel,
- pravým tlačítkem myši vyvolejte místní menu,
- zvolte položku *Cache*,
- z doplňkového menu vyberte položku *Flush*.

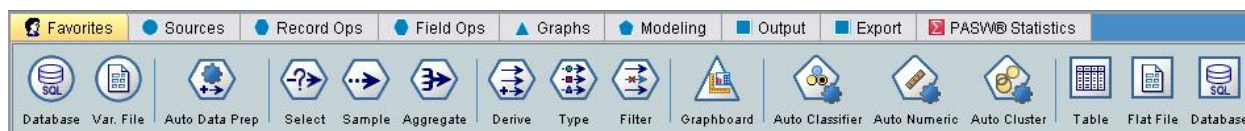
Uložení cache:

Naplněnou cache lze uložit, a to ve formátu PASW Statistics data file (.sav). Při znovuootevření proudů se tak nemusí znovu plnit cache ze zdrojových uzlů, ale data se načítají přímo do cache:

- označte myší požadovaný uzel,
- pravým tlačítkem myši vyvolejte místní menu,
- zvolte položku *Cache*,
- z doplňkového menu vyberte položku *Save Cache*,
- zvolte cestu a název souboru.

## 9. Uzly

Práce v PASW Modeler se provádí pomocí uzlů, jež jsou pojmenovány podle typu operací, které vykonávají. Všechny uzly jsou uspořádány do několika skupin podle podobnosti vykonávané činnosti. Jejich ikony jsou uloženy v paletách v dolní části obrazovky (obr. 12).



Obr. 12: Uspořádání uzlů v PASW Modeler

**Favorites** Skupina nejčastěji používaných uzlů. Obsah této skupiny uživatel volí sám. Uživatelské nastavení provedeme v hlavním menu *Tools* → *Manage Palettes*.

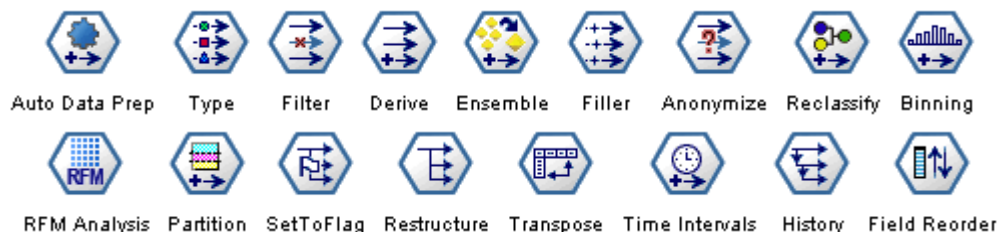
**Sources** Skupina uzlů zajišťujících přístup k datům.



**Record Ops** Skupina uzlů pro datové manipulace směřující k případům a podsouborům. Jejich nejčastější použití je ve fázi přípravy dat. Zahrnují uzly pro provádění výběrů, agregací, spojování a další.



**Field Ops** Skupina uzlů pro datové manipulace směřující k proměnným. Jejich nejčastější použití je ve fázi přípravy dat. Zahrnují uzly pro odvozování nových proměnných, rozdělení souboru na testovací a trénovací množinu a další.

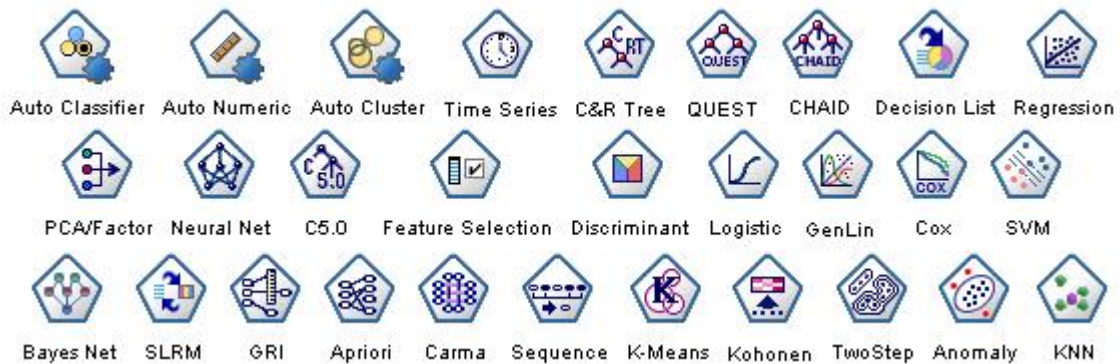


**Graphs** Skupina uzlů pro grafické zobrazení vztahů a pro evaluaci modelů.



## Modeling

Skupina uzlů pro modelování. Bohatá nabídka modelovacích technik od jednoduché lineární regrese až po sofistikované dataminingové techniky.



## Output

Skupina výstupních uzlů používaných pro tabulární a textovou prezentaci výsledků jednotlivých proudů.



## Export

Skupina výstupních uzlů používaných pro export výstupů do externích souborů, databází či aplikací.










## PASW Statistics



Skupina uzlů, které umožňují užší spolupráci s aplikací PASW Statistics.




## 10. Uzly Sources

Při tvorbě proudů je první fází načtení dat z určitého zdroje. Uzel skupiny Sources je nutný pro přebírání dat, proto se vždy vyskytuje v kterémkoliv proudě. Načítáme různé zdroje a formy.

 Enterprise View	<b>Enterprise View</b> <ul style="list-style-type: none"> <li>- napojení na Predictive Enterprise Repository (PES)</li> <li>- načtení dat z PES</li> <li>- předání modelu do PES ve formě scénáře</li> </ul>
 Database	<b>Databáze</b> <ul style="list-style-type: none"> <li>- načtení dat z aplikací typu MS Access, dBASE, SAS, ORACLE, Sybase apod.</li> <li>- použití je popsáno v několika základních krocích:             <ol style="list-style-type: none"> <li>1) instalujte ODBC driver a nakonfigurujte datový zdroj (data source) k odpovídající databázi. Instalační sada SPSS Data Access Pack obsahuje sadu driverů použitelných pro různé typy aplikací</li> <li>2) v uzlu <i>Database</i> proveďte spojení s požadovanou databází buď načtením konkrétní tabulky (<i>Table Mode</i>) nebo SQL příkazem (<i>SQL query</i>)</li> </ol> </li> </ul>
 Var. File	<b>Var. File</b> <ul style="list-style-type: none"> <li>- načítání dat z textových souborů s libovolným oddělovačem, ale pevným počtem polí</li> <li>- nastavení uzlu vyžaduje určit cestu ke zdrojovému souboru a typ použitého oddělovače dat</li> </ul>
 Fixed File	<b>Fixed File</b> <ul style="list-style-type: none"> <li>- načítání textových souborů s pevnou délkou polí</li> <li>- jde o pole, která nejsou zcela neomezená, ale začínají na stejné pozici a obsahují fixní počet znaků</li> </ul>
 Statistics File	<b>Statistics File</b> <ul style="list-style-type: none"> <li>- import dat uložených ve formátu .sav, tj. data uložená v PASW Statistics. Při importu je třeba určit zejména:             <ol style="list-style-type: none"> <li>1) umístění souboru</li> <li>2) nastavení způsobu čtení názvů proměnných (<i>Read names and labels</i> – načte názvy proměnných i popisy; <i>Read labels as names</i> – načte popisy proměnných z PASW Statistics jako jejich názvy)</li> <li>3) nastavení způsobu čtení hodnot (<i>Read data and labels</i> – popisy hodnot pak zobrazujeme v grafech a ostatních výstupech; <i>Read data as labels</i> – místo numerických hodnot načítá jejich popisy a často tak dojde k jejich typové konverzi)</li> </ol> </li> </ul>
 Data Collection	<b>Data Collection</b> <ul style="list-style-type: none"> <li>- import dat z produktů PASW Data Collection</li> <li>- čtení dat a metadat</li> </ul>
 SAS File	<b>SAS File</b> <ul style="list-style-type: none"> <li>- import dat uložených ve formátech:             <ul style="list-style-type: none"> <li>▪ SAS for Windows/OS2 (.sd2)</li> <li>▪ SAS for UNIX (.ssd)</li> <li>▪ SAS Transport File (.tpt)</li> <li>▪ SAS version 7/8/9 (.sas7bdat)</li> </ul> </li> </ul>








 Excel	<b>Excel</b> <ul style="list-style-type: none"> <li>- načítání dat z programu MS Excel</li> <li>- specifikace listu a rozpětí načítaných buněk</li> <li>- načítání názvů proměnných z prvního řádku</li> </ul>
 User Input	<b>User Input</b> <ul style="list-style-type: none"> <li>- načítání fixní množiny vlastních dat</li> <li>- umožňuje návrh dat uživatelem a jejich přímé použití v proudech. Spolu s hodnotami definujeme název proměnné a datový typ</li> <li>- automatické generování z jiných uzlů</li> </ul>



Poznámka:

Po rozkliknutí obsahují uzly palety *Sources* vlevo nahoře tlačítko  Preview, kterým lze vyvolat rychlý tabulkový náhled na data, aniž by bylo potřeba k datům ručně přidávat uzel **Table**.


## 11. Uzly Record Operations

Skupina *Record Operations* zahrnuje uzly, které jsou navrženy pro různé typy manipulací s daty (se záznamy v řádcích datové matice).

 Select	<b>Select</b> <ul style="list-style-type: none"> <li>- výběr nebo vyřazení definované podmnožiny záznamů z datové matice</li> <li>- výběr je prováděn na základě podmínky (podmínek) určené v poli <i>Conditions</i></li> </ul>
 Sample	<b>Sample</b> <ul style="list-style-type: none"> <li>- výběr množiny záznamů, které buď vyřadíme, nebo předáme dále do proudu</li> <li>- výběr je prováděn z mnoha důvodů, mezi nečastější patří redukce velkého objemu dat nebo vytváření náhodných výběrů</li> <li>- komplexní výběry (skupinky a oblasti)</li> </ul>
 Balance	<b>Balance</b> <ul style="list-style-type: none"> <li>- vážení souboru</li> <li>- v uzlu nastavíme tzv. <i>faktor a podmínku</i> pro přiřazení. Vážení je v proudu založeno na duplikování hodnot v datové matici a to v závislosti na nastavených podmínkách</li> <li>- uzel <i>Balance</i> může být také generován automaticky z grafů distribuce (<i>Distribution</i>) nebo histogramu (<i>Histogram</i>), a to metodami nafukování (<i>boosting</i>) nebo redukce (<i>reduce</i>)</li> </ul>
 Aggregate	<b>Aggregate</b> <ul style="list-style-type: none"> <li>- nahrazuje posloupnost určitých hodnot jejich souhrnnou charakteristikou – průměr, součet, maximum, minimum apod.</li> <li>- v poli <i>Key fields</i> určete proměnnou, podle které budete agregovat, v poli <i>Aggregate field</i> určete proměnné, které budete agregovat</li> </ul>
 RFM Aggregate	<b>RFM Aggregate</b> <ul style="list-style-type: none"> <li>- příprava transakčních dat pro RFM analýzu</li> <li>- výstupem jsou data pro jednotlivé zákazníky, pro které je spočtena doba od posledního nákupu (<i>recency</i>), četnost nákupů (<i>frequency</i>) a celková utracená částka (<i>monetary</i>)</li> </ul>
 Sort	<b>Sort</b> <ul style="list-style-type: none"> <li>- vzestupné nebo sestupné řazení případů podle vybrané jedné nebo více proměnných</li> <li>- používáme často v případech, kdy hledáme maximum v jednotlivých pozorováních</li> </ul>
 Merge	<b>Merge</b> <ul style="list-style-type: none"> <li>- spojování dat z více zdrojů. Při spojování dat existují v PASW Modeler dva způsoby:           <ol style="list-style-type: none"> <li>1. <b>spojování podle pořadí (order)</b> – slučování odpovídajících záznamů ze všech zdrojů podle pořadí. Před použitím této metody je třeba data vhodně seřadit.</li> <li>2. <b>spojování na základě klíče (keys)</b> – specifikace způsobu spojení dat:               <ul style="list-style-type: none"> <li>- <b>vnitřní spojení (inner join)</b> – podmínkou spojení je výskyt záznamu ve všech spojovaných tabulkách</li> <li>- <b>úplně vnější spojení (full outer join)</b> – spojení zahrne všechny záznamy vyskytující se alespoň v jedné ze spojovaných tabulek</li> </ul> </li> </ol> </li> </ul>

	<ul style="list-style-type: none"> <li>- <b>částečné vnější spojení (partial outer join)</b> – podmínkou spojení je výskyt záznamu v jedné nebo více vybraných tabulkách</li> <li>- <b>inverzní spojení (anti-join)</b> – spojení zahrnuje záznamy, které nejsou ve vybrané tabulce, ale vyskytují se alespoň v jedné z napojovaných tabulek</li> </ul>
 Append	<b>Append</b> <ul style="list-style-type: none"> <li>- přidávání záznamů se stejnou strukturou</li> <li>- předpokladem funkčnosti uzlu je použití stejných datových typů</li> </ul>
 Distinct	<b>Distinct</b> <ul style="list-style-type: none"> <li>- odstranění duplicitních záznamů dvojím způsobem: <ol style="list-style-type: none"> <li>1) ponechání prvního záznamu a všechny následující duplicity odstraníme (<i>Include</i>)</li> <li>2) ponechání jen duplicitních záznamů v proudu – vhodné pokud chceme analyzovat duplicity (<i>Discard</i>)</li> </ol> </li> </ul>




Poznámka:

Po rozkliknutí obsahují uzly palety *Record Operations* vlevo nahoře tlačítko  Preview, kterým lze vyvolat rychlý tabulkový náhled na data, aniž by bylo potřeba k datům ručně přidávat uzel **Table**.
















## 12. Uzly Field Operations

Skupina *Field Operations* zahrnuje uzly, které jsou navrženy pro různé typy manipulací se záznamy ve sloupcích datové matice.


 Auto Data Prep	<p><b>Auto Data Preparation (ADP)</b></p> <ul style="list-style-type: none"> <li>- tento uzel vám pomůže projít základní kroky postupu přípravy dat, což zrychlí budování modelu a zvýší sílu predikce. Na záložce <i>Objectives</i> můžete zvolit jednu ze čtyř základních možností: <ol style="list-style-type: none"> <li>1. <b>Balance speed and accuracy (vyvážení rychlosti a přesnosti)</b> – při této volbě procedura transformuje data s přihlédnutím k tomu, aby vytvořené modely byly vyvážené z hlediska rychlosti zpracování a přesnosti predikcí</li> <li>2. <b>Optimize for speed (optimalizace pro rychlost)</b> – tato volba je vhodná při práci s rozsáhlými datovými soubory</li> <li>3. <b>Optimize for accuracy (optimalizace pro přesnost)</b> – tato volba se zaměřuje na co možná nejlepší predikce z dat, které jsou k dispozici</li> <li>4. <b>Custom analysis (vlastní nastavení)</b> – manuální doplnění nastavení na záložce <i>Settings</i>; tato volba se vybere automaticky jakmile uživatel provede jakékoli změny v nastavení jedné ze tří předchozích voleb</li> </ol> </li> </ul>
 Type	<p><b>Type</b></p> <ul style="list-style-type: none"> <li>- vlastnosti jednotlivých proměnných v datové matici nastavte ve zdrojovém uzlu (záložka <i>Types</i>) nebo odděleně v uzlu <i>Type</i>. Jde o tyto vlastnosti: <ol style="list-style-type: none"> <li>1. <b>Type (typ proměnné)</b> – rozlišují se tyto typy: <i>Range</i> (číselná proměnná), <i>Discrete</i> (obecná kategorizovaná proměnná), <i>Flag</i> (dichotomická proměnná), <i>Set</i> (nominální proměnná), <i>Ordered Set</i> (ordinální proměnná), <i>Typeless</i> (proměnná bez udaného typu). Zvolíte-li typ <i>Default</i> nebo <i>Discrete</i>, PASW Modeler sám po průchodu daty uzlem navrhne vhodný typ proměnné.</li> <li>2. <b>Values (hodnoty)</b> – přípustné hodnoty a nastavení způsobu čtení dat s volbami: <ol style="list-style-type: none"> <li>a. Read: přípustné hodnoty budou načteny v okamžiku spuštění proudu</li> <li>b. Read +: přípustné hodnoty budou načteny a připojeny ke stávajícím (pokud existují)</li> <li>c. Pass: přípustné hodnoty nejsou čteny z dat</li> <li>d. Current: ponechání stávajících hodnot</li> <li>e. Specify: uživatelské nastavení přípustných hodnot</li> </ol> </li> <li>3. <b>Missing Values (chybějící hodnoty)</b> – pro specifikaci hodnot, které budou zpracovány jako <i>Blanks</i> (prázdné pole)</li> <li>4. <b>Check (kontrola)</b> – kontrola hodnot pole podle přípustných hodnot ze sloupce <i>Values</i></li> <li>5. <b>Directions (směr)</b> – rozlišení proměnných při modelování na vstupní (<i>Input</i>) a výstupní (<i>Output</i>). K dispozici je rovněž volba oba směry (<i>Both</i>) a bez směru (<i>None</i>). Směr rozdělení (<i>Partition</i>) určuje proměnnou, podle které rozdělujeme soubor na trénovací a testovací podmnožinu. Volbu <i>Split</i> lze použít pouze pro proměnné typu <i>Set</i>, <i>Ordered Set</i> a <i>Flag</i> a umožňuje vytvořit pomocí jednoho proudu více modelů - pro každou možnou hodnotu kategorizované proměnné, která má směr <i>Split</i>, jeden model.</li> </ol> </li> </ul>
 Filter	<p><b>Filter</b></p> <ul style="list-style-type: none"> <li>- filtrování nebo vyřazení specifikovaných proměnných</li> <li>- přepisování názvů polí (proměnných)</li> </ul>










 <p>Derive</p>	<p><b>Derive</b> Umožňuje odvozovat nové proměnné:</p> <ul style="list-style-type: none"> <li>- <b>Formula</b> – odvozena z libovolného CLEM výrazu, tj. na základě vzorce zapsaného syntaxí jazyka CLEM. Zápis vzorců usnadňuje kalkulátor (<i>Expression Builder</i>), který je součástí dialogového okna uzlu <b>Derive</b></li> <li>- <b>Flag</b> – odvození proměnné typu <i>Flag</i>, odpovídající specifikované podmínce</li> <li>- <b>Set</b> – odvození kategorizované proměnné typu <i>Set</i> (nominální proměnná)</li> <li>- <b>State</b> – sekvenční stavová proměnná, jejíž hodnoty „On“ a „Off“ závisí na nastavení podmínky</li> <li>- <b>Count</b> – odvozená proměnná sekvenčně načítá počet případů, vyhovujících zadané podmínce</li> <li>- <b>Conditional</b> – hodnoty nové proměnné závisí na splnění jedné z podmínek. Odvození metodou podmíněného výpočtu <i>If ... Then</i></li> </ul>
 <p>Ensemble</p>	<p><b>Ensemble</b></p> <ul style="list-style-type: none"> <li>- kombinuje výsledky dvou nebo více modelů pro získání přesnější predikce</li> <li>- používá se zejména v kombinaci s uzly vytvářejícími modely automaticky</li> </ul>
 <p>Filler</p>	<p><b>Filler</b></p> <ul style="list-style-type: none"> <li>- změna hodnot. Změny hodnot se provádí pomocí jazyka CLEM, např. <code>@BLANK (@FIELD)</code>.</li> <li>- změna formátu uložení dat (<i>storage</i>): např. z typu <i>integer</i> na typ <i>string</i></li> </ul> <p>V poli <b>Replace</b> volíme způsob nahrazení:</p> <ul style="list-style-type: none"> <li>- <b>Based on condition</b> – stanovíme podmínku pro nahrazení hodnot. Zápis podmínky v okně <i>Condition</i></li> <li>- <b>Always</b> – nahrazení všech hodnot. Výhodné při změně formátu uložení dat</li> <li>- <b>Blank values</b> – nahrazení předem definovaných chybějících hodnot (<i>user missing</i>)</li> <li>- <b>Null values</b> – nahrazení prázdných buněk (<i>system missing</i>)</li> <li>- <b>Blank and null values</b> – nahrazení všech chybějících hodnot a prázdných buněk</li> </ul>
 <p>Anonymize</p>	<p><b>Anonymize</b></p> <ul style="list-style-type: none"> <li>- nabízí utajení popisů proměnných a jejich hodnot pro potřeby ochrany dat při jejich poskytnutí třetí straně (např. technické podpoře SPSS)</li> <li>- podle zařazení uzlu do proudu je nutné změnit nastavení dalších uzlů</li> </ul>
 <p>Reclassify</p>	<p><b>Reclassify</b></p> <ul style="list-style-type: none"> <li>- reklasifikace hodnot kategorizované proměnné</li> <li>- slučování kategorií kategorizované proměnné</li> <li>- reklasifikace do nové proměnné (<i>New field</i>) nebo do stávající proměnné (<i>Existing field</i>)</li> </ul>
 <p>Binning</p>	<p><b>Binning</b></p> <ul style="list-style-type: none"> <li>- kategorizace spojité proměnné do předem určeného počtu kategorií (<i>No. Of Bins</i>)</li> <li>- kategorizace spojité proměnné do kategorií určené délky (<i>Bin width</i>)</li> <li>- další metody kategorizace včetně optimální</li> </ul>

 RFM Analysis	<b>RFM Analysis</b> <ul style="list-style-type: none"> <li>- analýza doby od posledního nákupu (recency), četnosti nákupů (frequency) a celkové utracené částky (monetary)</li> <li>- jednotlivé RFM parametry jsou rozděleny do kategorií a pro každého zákazníka se spočítá skóre – pravděpodobnost zařazení do kategorie</li> </ul>
 Partition	<b>Partition</b> <ul style="list-style-type: none"> <li>- rozdělení datového souboru na <i>trénovací</i>, <i>testovací</i> a <i>validační</i> množinu</li> <li>- odvození proměnné typu <i>Partition</i>, ve které je rozdělení zaznamenáno</li> <li>- pro jednoduché ověření kvality získaných modelů</li> <li>- nastavení pevné hodnoty generátoru náhodných čísel</li> </ul>
 SetToFlag	<b>Set To Flag</b> <ul style="list-style-type: none"> <li>- restrukturalizace kategorizované proměnné na příznaky kategorií (<i>Flag</i>)</li> <li>- např. pro analýzu nákupních košíků, spotřebitelského chování, pavučinový graf (<i>Web</i>)</li> </ul>
 Restructure	<b>Restructure</b> <ul style="list-style-type: none"> <li>- odvození nových proměnných na základě hodnot stávajících proměnných nebo příznaků</li> <li>- flexibilnější varianta uzlu <b>SetToFlag</b></li> </ul>
 Transpose	<b>Transpose</b> <ul style="list-style-type: none"> <li>- záměna řádků a sloupců datové matice</li> </ul>
 Time Intervals	<b>Time Intervals</b> <ul style="list-style-type: none"> <li>- specifikace časových intervalů a automatické generování jejich popisů u časových dat (<i>TimeLabel</i>)</li> <li>- automatické generování indexu <i>TimeIndex</i> identifikujícího každé pozorování časové řady</li> <li>- datové úpravy časových řad – <i>agregace</i>, <i>doplňování</i></li> </ul>
 History	<b>History</b> <ul style="list-style-type: none"> <li>- restrukturalizace sekvenčních (časových) dat</li> <li>- odvození nových proměnných metodou časového zpoždění</li> <li>- klíčová nastavení: <ul style="list-style-type: none"> <li>- <i>Offset</i>: zpoždění, po kterém začínáme sledování</li> <li>- <i>Span</i>: časové okno</li> </ul> </li> </ul>
 Field Reorder	<b>Field Reorder</b> <ul style="list-style-type: none"> <li>- ruční seřazení proměnných v matici</li> <li>- automatické řazení proměnných podle typu, názvu nebo uložení</li> </ul>

Poznámka:









Po rozkliknutí obsahují uzly této palety vlevo nahoře tlačítko  Preview, kterým lze vyvolat rychlý tabulkový náhled na data, aniž by bylo potřeba k datům ručně přidávat uzel **Table**.









## 13. Uzly Graphs







 Graphboard	<b>Graphboard</b> <ul style="list-style-type: none"> <li>- prostředí pro tvorbu různých druhů grafů (např. sloupcový, koláčový, bodový)</li> <li>- podle typu vybraných proměnných nabízí vhodné grafy</li> </ul>
 Plot	<b>Plot</b> <ul style="list-style-type: none"> <li>- dvourozměrný a třírozměrný bodový graf</li> <li>- grafické zobrazení vztahu proměnných</li> <li>- možnost rozlišení hodnot podle dalších (kategorizovaných) proměnných</li> <li>- proložení grafu křivkou</li> </ul>
 Distribution	<b>Distribution</b> <ul style="list-style-type: none"> <li>- sloupcový graf rozložení kategorizované proměnné</li> <li>- rozlišení hodnot podle další (kategorizované) proměnné</li> <li>- proporční škála</li> </ul>
 Histogram	<b>Histogram</b> <ul style="list-style-type: none"> <li>- rozložení hodnot číselné proměnné</li> <li>- obarvení podle hodnot kategorizované proměnné (<i>Color</i>)</li> <li>- separátní grafy (<i>Panel</i>), animovaná grafika (<i>Animation</i>)</li> <li>- automatické i manuální nastavení os</li> </ul>
 Collection	<b>Collection</b> <ul style="list-style-type: none"> <li>- rozložení hodnot číselné proměnné (<i>Collect</i>) ve vztahu k další proměnné (<i>Over</i>)</li> <li>- rozlišení podle hodnot jiné proměnné (<i>Overlay</i>)</li> <li>- separátní grafy, obarvené grafy, animované grafy</li> <li>- automatické i manuální nastavení os</li> </ul>
 Multiplot	<b>Multiplot</b> <ul style="list-style-type: none"> <li>- vztah jedné proměnné na ose X (<i>X field</i>) a více proměnných na ose Y (<i>Y fields</i>)</li> <li>- rozlišení podle hodnot jiné proměnné (<i>Overlay</i>)</li> <li>- proložení grafu vlastní funkcí</li> </ul>
 Web	<b>Web</b> <ul style="list-style-type: none"> <li>- zobrazení vztahů mezi více kategorizovanými proměnnými</li> <li>- síla linky odpovídá síle vztahu</li> <li>- uplatnění při analýze nákupních košíků, hledání podvodů, analýze textů atd.</li> </ul>
 Time Plot	<b>Time Plot</b> <ul style="list-style-type: none"> <li>- zobrazení jedné nebo více časových řad</li> <li>- použití s uzlem Time Intervals pro vytvoření proměnné <i>TimeLabel</i></li> </ul>
 Evaluation	<b>Evaluation</b> <ul style="list-style-type: none"> <li>- evaluační grafy pro hodnocení kvality a vzájemnou komparaci predikčních modelů</li> <li>- <i>gains chart</i>, <i>lift chart</i>, <i>response chart</i>, <i>profit chart</i>, <i>ROI chart</i></li> <li>- zahrnutí křivky optimálního modelu</li> <li>- uživatelsky nastavitelné parametry pro evaluaci</li> </ul>

## 14. Uzly Modeling










 Auto Classifier	<b>Auto Classifier</b> <ul style="list-style-type: none"> <li>- součást modulu Classification</li> <li>- rychlý výběr optimálního modelu pro binární cílovou proměnnou zvolenými modely</li> <li>- automatické generování nejlepších zvolených modelů na paletu</li> <li>- urychluje proces hledání optimálního modelu</li> </ul>
 Auto Numeric	<b>Auto Numeric</b> <ul style="list-style-type: none"> <li>- součást modulu Classification</li> <li>- rychlý výběr optimálního modelu pro číselnou cílovou proměnnou zvolenými modely</li> <li>- automatické generování nejlepších zvolených modelů na paletu</li> <li>- urychluje proces hledání optimálního modelu</li> </ul>
 Auto Cluster	<b>Auto Cluster</b> <ul style="list-style-type: none"> <li>- součást modulu Classification</li> <li>- rychlý výběr optimálního modelu identifikace skupin případů s podobnými charakteristikami</li> <li>- automatické generování nejlepších zvolených modelů na paletu</li> </ul>
 Time Series	<b>Time Series</b> <ul style="list-style-type: none"> <li>- analýza časových dat</li> <li>- Expert Modeler – automatický výběr nejvhodnějšího modelu</li> <li>- ARIMA modely</li> <li>- exponenciální vyrovnávání</li> </ul>
 C&R Tree	<b>C&amp;R Tree (Classification &amp; Regression Tree)</b> <ul style="list-style-type: none"> <li>- binární rozhodovací strom</li> <li>- metoda náhradníků při práci s chybějícími hodnotami (<i>surrogates</i>)</li> <li>- kritéria pro zastavení růstu</li> <li>- pruning – zjednodušování struktury stromu, která podstatně nezhoršuje přesnost predikce (prořezávání stromu)</li> <li>- zahrnutí nákladů chybné klasifikace</li> <li>- nastavení maximální úrovně hloubky</li> </ul>
 QUEST	<b>QUEST (Quick, Unbiased, Efficient Statistical Tree)</b> <ul style="list-style-type: none"> <li>- rychlý binární rozhodovací strom</li> <li>- metoda náhradníků při práci s chybějícími hodnotami (<i>surrogates</i>)</li> <li>- pruning – zjednodušování struktury stromu, která podstatně nezhoršuje přesnost predikce (prořezávání stromu)</li> <li>- nastavení maximální úrovně hloubky</li> <li>- zahrnutí nákladů chybné klasifikace</li> </ul>
 CHAID	<b>CHAID (Chi-squared Automatic Interaction Detection)</b> <ul style="list-style-type: none"> <li>- metody CHAID a Exhaustive CHAID</li> <li>- interaktivní růst stromu</li> <li>- expertní nastavení algoritmu (<i>Alpha for splitting, Alpha for merging, konvergenční kritéria</i>)</li> <li>- zahrnutí nákladů chybné klasifikace</li> </ul>

 Decision List	<b>Decision List</b> <ul style="list-style-type: none"> <li>- identifikace podskupin nebo segmentů vykazujících vyšší pravděpodobnost sledované cílové kategorie než celá populace</li> <li>- sada rozhodovacích pravidel</li> <li>- interaktivní prostředí pro vývoj modelů</li> </ul>
 Regression	<b>Regression</b> <ul style="list-style-type: none"> <li>- lineární regrese</li> <li>- metody zahrnutí proměnných – <i>Enter, Forward, Backward, Stepwise</i></li> <li>- Durbin-Watsonova statistika</li> <li>- index determinace</li> <li>- konfidenční intervaly</li> </ul>
 PCA/Factor	<b>PCA/Factor (Principal Components Analysis/Factor)</b> <ul style="list-style-type: none"> <li>- analýza hlavních komponent, faktorová analýza</li> <li>- 6 metod extrakce faktorů</li> <li>- faktorová analýza kovarianční/korelační matice</li> <li>- rotace faktorů <i>Varimax, Quartimax, Equamax, Promax, Direct Oblimin</i></li> </ul>
 Neural Net	<b>Neural Net</b> <ul style="list-style-type: none"> <li>- součást modulu Classification</li> <li>- supervizované neuronové sítě</li> <li>- metody stanovení topologie (<i>Quick, Dynamic, Multiple, Prune, RFBN, Exhaustive prune</i>)</li> <li>- kritéria pro zastavení: <i>Default</i> (ve chvíli suboptimálního trénovacího stavu), <i>Accuracy</i> (přesnost), <i>Cycles</i> (počet cyklů), <i>Time</i> (uplynutí doby)</li> <li>- analýza sensitivity</li> </ul>
 C5.0	<b>C5.0</b> <ul style="list-style-type: none"> <li>- součást modulu Classification</li> <li>- tvorba rozhodovacího stromu nebo rozhodovacích pravidel</li> <li>- metoda boosting pro vytváření rozhodovacích lesů</li> <li>- cross-validace</li> <li>- zahrnutí nákladů chybné klasifikace</li> <li>- pruning – zjednodušování struktury stromu, která podstatně nezhoršuje přesnost predikce</li> </ul>
 Feature Selection	<b>Feature Selection</b> <ul style="list-style-type: none"> <li>- součást modulu Classification</li> <li>- algoritmus pro orientační výběr nejdůležitějších prediktorů pro modelování</li> <li>- důležitost (<i>importance</i>) měřená podle různých přístupů (<i>Chi-kvadrát, Cramérovo V, Lambda, t-testy</i>)</li> <li>- každý prediktor je označen jedním z klasifikačních stupňů, které lze nastavit</li> </ul>
 Discriminant	<b>Discriminant</b> <ul style="list-style-type: none"> <li>- diskriminační analýza</li> <li>- metody výběru proměnných <i>Enter</i> a <i>Stepwise</i></li> <li>- určení apriorních pravděpodobností</li> </ul>
 Logistic	<b>Logistic</b> <ul style="list-style-type: none"> <li>- součást modulu Classification</li> <li>- logistická regrese – <i>binární, multinomická</i></li> <li>- metody zahrnutí proměnných – <i>Enter, Forward, Backward, Stepwise</i></li> <li>- predikce pravděpodobností všech kategorií</li> <li>- klasifikační tabulka, <i>goodness of fit</i> statistiky, vlastní model s interakcemi</li> </ul>

 GenLin	<b>GenLin</b> <ul style="list-style-type: none"> <li>- součást modulu Classification</li> <li>- zobecněné lineární modely</li> <li>- volba pravděpodobnostního rozdělení cílové proměnné a linkové funkce</li> <li>- volba metody pro odhad parametrů modelu</li> </ul>
 Cox	<b>Cox</b> <ul style="list-style-type: none"> <li>- součást modulu Classification</li> <li>- model pro predikci doby, která uplyne do určité události</li> <li>- bere do úvahy cenzorovaná pozorování (u nichž v historii sledovaná událost nenastala)</li> </ul>
 SVM	<b>SVM (Support Vector Machine)</b> <ul style="list-style-type: none"> <li>- součást modulu Classification</li> <li>- klasifikace dat pomocí metody <i>support vector machine</i></li> <li>- vhodná metoda pro velký počet prediktorů</li> </ul>
 Bayes Net	<b>Bayes Net</b> <ul style="list-style-type: none"> <li>- součást modulu Classification</li> <li>- grafický model znázorňující proměnné a vztahy mezi nimi</li> <li>- robustní metoda</li> </ul>
 SLRM	<b>SLRM (Self-Learning Response Model)</b> <ul style="list-style-type: none"> <li>- součást modulu Classification</li> <li>- vytváří model, který může být automaticky aktualizován s přibývajícími daty bez nutnosti přepočítávat model vždy pro veškerá data</li> </ul>
 GRI	<b>GRI (Generalized Rule Induction)</b> <ul style="list-style-type: none"> <li>- detekce asociačních pravidel na bázi podmínek <i>If ... Then</i></li> <li>- práce s číselnými i kategorizovanými proměnnými ve směru <i>In</i></li> <li>- výstupní proměnné (<i>Out</i>) kategorizované</li> <li>- výběr (maximálně) stanoveného počtu pravidel</li> <li>- vybírá implikace s nejvyšší mírou informace (<i>míra J</i>)</li> <li>- negeneruje predikční model</li> </ul>
 Apriori	<b>Apriori</b> <ul style="list-style-type: none"> <li>- součást modulu Association</li> <li>- detekce asociačních pravidel na bázi podmínek <i>If ... Then</i></li> <li>- pro symbolické (kategorizované) proměnné</li> <li>- předpokladem funkčnosti je jedna nebo více proměnných směru <i>In</i> a jedna nebo více proměnných směru <i>Out</i></li> <li>- vybírá na základě pokrytí a spolehlivosti</li> <li>- pokročilé filtrování pravidel na základě apriorní a aposteriorní pravděpodobnosti</li> <li>- vstupy ve formě tabulkových nebo transakčních dat</li> </ul>
 Carma	<b>Carma</b> <ul style="list-style-type: none"> <li>- součást modulu Association</li> <li>- detekce asociačních pravidel na bázi podmínek <i>If ... Then</i></li> <li>- nerespektuje specifikaci směru proměnných (<i>In, Out</i>)</li> <li>- nastavení příčin a důsledků</li> <li>- vstupy ve formě tabulkových nebo transakčních dat</li> </ul>

 Sequence	<b>Sequence</b> <ul style="list-style-type: none"> <li>- součást modulu Association</li> <li>- detekce vztahů mezi sekvenčními (časovými) daty</li> <li>- detekce asociačních pravidel na bázi podmínek <i>If ... Then</i></li> <li>- sekvence je časově uspořádaný seznam položek, které je možné považovat za posloupnost predikovatelných událostí</li> <li>- předpoklady: identifikátor (<i>ID field</i>), volitelná časová proměnná (<i>Time field</i>) a jedno či více obsahových polí (<i>content fields</i>)</li> </ul>
 K-Means	<b>K-Means</b> <ul style="list-style-type: none"> <li>- algoritmus pro seskupování s předem určeným počtem klastrů</li> <li>- kritéria pro zastavení algoritmu</li> <li>- vyžaduje jen proměnné typu <i>In</i>, proměnné typu <i>Out</i>, <i>Both</i> a <i>None</i> jsou ignorovány</li> </ul>
 Kohonen	<b>Kohonen</b> <ul style="list-style-type: none"> <li>- součást modulu Segmentation</li> <li>- speciální typ neuronových sítí pro seskupování případů</li> <li>- kritéria zastavení algoritmu</li> <li>- specifikace velikosti dvourozměrné mapy klastrů</li> <li>- nastavení parametrů učení <i>Neighborhood</i>, <i>Initial Eta</i>, <i>Cycles</i></li> </ul>
 TwoStep	<b>TwoStep</b> <ul style="list-style-type: none"> <li>- součást modulu Segmentation</li> <li>- dvoustupňové seskupování</li> <li>- vyžaduje jen proměnné typu <i>In</i>, proměnné typu <i>Out</i>, <i>Both</i> a <i>None</i> jsou ignorovány</li> <li>- automatický návrh optimálního počtu klastrů</li> </ul>
 Anomaly	<b>Anomaly</b> <ul style="list-style-type: none"> <li>- součást modulu Segmentation</li> <li>- explorační metoda pro detekci neobvyklých či odlehklých pozorování</li> <li>- identifikace na základě indexu neobvyklosti (<i>anomaly index</i>)</li> <li>- uživatelsky nastavitelné procento neobvyklých případů nebo jejich počtu</li> </ul>
 KNN	<b>KNN (K Nearest Neighbor)</b> <ul style="list-style-type: none"> <li>- součást modulu Classification</li> <li>- analýza nejbližšího souseda – metoda klasifikace případů založená na jejich vzájemné podobnosti</li> <li>- mírou podobnosti je vzdálenost mezi případy</li> <li>- specifikace počtu (<i>K</i>) nejbližších sousedů</li> </ul>







## 15. Uzly Output

 Table	<b>Table</b> <ul style="list-style-type: none"> <li>- tabulkové zobrazení datové matice</li> </ul>
 Matrix	<b>Matrix</b> <ul style="list-style-type: none"> <li>- zobrazení vztahů mezi dvěma proměnnými v kontingenční tabulce</li> <li>- možnost sumarizace třetí spojité proměnné</li> <li>- řádková, sloupcová procenta</li> </ul>
 Analysis	<b>Analysis</b> <ul style="list-style-type: none"> <li>- kvalitativní analýza predikčních modelů</li> <li>- vzájemné porovnání modelů</li> </ul>
 Data Audit	<b>Data Audit</b> <ul style="list-style-type: none"> <li>- rychlý pohled na kvalitu dat</li> <li>- popisné statistiky, histogramy, grafy distribucí</li> <li>- analýza chybějících hodnot</li> <li>- analýza odlehklých pozorování</li> <li>- generování uzlů pro čištění dat</li> </ul>
 Transform	<b>Transform</b> <ul style="list-style-type: none"> <li>- transformační funkce rozdělení</li> <li>- náhled na rozložení hodnot transformovaných proměnných</li> <li>- generování plnohodnotných grafů a nových proměnných uzlem <i>Derive</i> a <i>Filler</i></li> </ul>
 Statistics	<b>Statistics</b> <ul style="list-style-type: none"> <li>- základní popisné statistiky</li> <li>- průměr, součet, minimum, maximum</li> <li>- rozpětí, směrodatná odchylka</li> <li>- míry korelace</li> </ul>
 Means	<b>Means</b> <ul style="list-style-type: none"> <li>- t-test pro dva nezávislé výběry</li> <li>- t-test pro dva závislé výběry</li> <li>- analýza rozptylu – test shody středních hodnot ve více nezávislých výběrech</li> </ul>
 Report	<b>Report</b> <ul style="list-style-type: none"> <li>- tisk výstupů ve formě reportů obsahujících strukturovaný text a data</li> <li>- specifikuje se formát reportu použitím šablony (<i>Template</i>)</li> <li>- zápis podmínek v jazyce CLEM</li> </ul>
 Set Globals	<b>Set Globals</b> <ul style="list-style-type: none"> <li>- uzel prochází data a počítá souhrnné statistiky, které mohou být použity ve výrazech jazyka CLEM</li> <li>- zobrazení hodnot v menu <i>Tools</i> → <i>Stream Properties</i> → <i>Globals</i></li> </ul>








## 16. Uzly Export

---

 Database	<b>Database</b> <ul style="list-style-type: none"><li>- ukládání výstupů do databází</li><li>- vytvoří novou tabulku (<i>Create table</i>) nebo provede zápis do existující (<i>Insert into table</i>)</li><li>- použití podmíněno existencí ODBC zdroje</li></ul>
 Flat File	<b>Flat File</b> <ul style="list-style-type: none"><li>- export datové matice do textového formátu s oddělovači</li></ul>
 Statistics Export	<b>Statistics Export</b> <ul style="list-style-type: none"><li>- export datové matice do .sav formátu</li></ul>
 Data Collection Export	<b>Data Collection Export</b> <ul style="list-style-type: none"><li>- export datové matice do softwaru PASW Data Collection pro výzkum trhu</li></ul>
 SAS Export	<b>SAS Export</b> <ul style="list-style-type: none"><li>- export datové matice ve formátu SAS, čitelný programy SAS a kompatibilními produkty</li><li>- k dispozici jsou tři formáty:<ul style="list-style-type: none"><li>- SAS for Windows/OS2</li><li>- SAS for UNIX</li><li>- SAS Version 7/8</li></ul></li></ul>
 Excel	<b>Excel</b> <ul style="list-style-type: none"><li>- export datové matice ve formátu .xls</li><li>- automatické otevření Excelu (jsou podporovány verze MS Excel 97 až 2007)</li></ul>

## 17. Uzly PASW Statistics

 Statistics File	<b>Statistics File</b> <ul style="list-style-type: none"> <li>- import dat uložených ve formátu .sav, tj. data uložená v PASW Statistics. Při importu je třeba určit zejména: <ol style="list-style-type: none"> <li>4) umístění soboru</li> <li>5) nastavení způsobu čtení názvů proměnných (<i>Read names and labels</i> – načte názvy proměnných i popisy; <i>Read labels as names</i> – načte popisy proměnných z PASW Statistics jako jejich názvy)</li> <li>6) nastavení způsobu čtení hodnot (<i>Read data and labels</i> – popisy hodnot pak zobrazujeme v grafech a ostatních výstupech; <i>Read data as labels</i> – místo numerických hodnot načítá jejich popisy a často tak dojde k jejich typové konverzi)</li> </ol> </li> </ul>
 Statistics Transform	<b>Statistics Transform</b> <ul style="list-style-type: none"> <li>- datové transformace programu PASW Statistics v rámci PASW Modeler</li> <li>- definice syntaxovým jazykem PASW Statistics včetně ověření správnosti syntaxového zápisu</li> <li>- využití podmíněno platnou instalací PASW Statistics</li> </ul>
 Statistics Model	<b>Statistics Model</b> <ul style="list-style-type: none"> <li>- umožňuje používat modelovací procedury programu PASW Statistics</li> <li>- vytvořené modely lze používat v PASW Modeler obvyklým způsobem</li> <li>- využití podmíněno platnou instalací PASW Statistics</li> </ul>
 Statistics Output	<b>Statistics Output</b> <ul style="list-style-type: none"> <li>- volání procedur PASW Statistics, které generují výstupy</li> <li>- podmínkou je instalace PASW Statistics</li> <li>- výsledky jsou zobrazeny buď v okně prohlížeče, nebo uloženy do výstupního formátu PASW Statistics</li> </ul>
 Statistics Export	<b>Statistics Export</b> <ul style="list-style-type: none"> <li>- export datové matice do .sav formátu</li> </ul>

## 18. CLEF

---

**CLEF** (*Component-Level Extension Framework*) je mechanismus, který umožňuje rozšířit funkcionalitu PASW Modeleru o uživatelem definované rozšíření. Jeho součástí bývá sdílená knihovna – např. algoritmus vlastního modelu – která se přidá do PASW Modeler a je dostupná buď z menu, nebo z palety jako nový uzel.

K tomu jsou potřeba určité informace o uživatelském programu, např. název, parametry, které se budou předávat, způsob, jakým má PASW Modeler zobrazovat výsledky atd. Tyto informace se zapisou do specifikačního souboru ve formátu XML. PASW Modeler tyto informace přeloží do podoby nového menu nebo nového uzlu.

CLEF nahrazuje dosavadní CEMI, které v dalších verzích nebude podporováno.

Výhody použití CLEF:

- flexibilní a robustní prostředí pro integraci vlastních procedur do PASW Modeler
- přidané moduly vypadají a chovají se jako nativní moduly PASW Modeler
- přidané uzly se vykonávají skoro tak rychle a efektivně jako nativní uzly PASW Modeler

CLEF zahrnuje tyto specifikace:

- specifikační soubor
- uzly
- datový model
- vstupní a výstupní soubory
- uživatelské rozhraní (API = Application Programming Interfaces)

### Specifikační soubor

Každé rozšíření musí zahrnovat XML soubor `extension.xml` nazývaný specifikační soubor. Tento soubor obsahuje základní informace o rozšíření, určuje externí zdroje, podle potřeby definuje objekty jako uzel, diamant modelu atd.

### Uzly

Je-li součástí rozšíření definice nového uzlu, je nutné určit jeho typ (např. zda bude uzel generovat model nebo pouze transformovat data). Po vytvoření specifikačního souboru a potřebných Java tříd a sdílených knihoven se specifikační soubor překopíruje do umístění, ze kterého jej PASW Modeler může číst. Po spuštění PASW Modeler se nový uzel objeví na příslušné paletě.

### Datový model

Datový model reprezentuje strukturu dat, která protékají proudem. Pro každý nový uzel se musí určit, jak bude pracovat s daty, které do něj vtékají, a jak je bude předávat dál.

### Vstupní a výstupní soubory

Před použitím nového uzlu se vytvoří jeden nebo více dočasných souborů. Jedná se o tzv. vstupní soubory. Další dočasné soubory se vytvoří v průběhu výpočtu uzlu a ty se nazývají výstupní. CLEF umožňuje specifikaci práce s těmito soubory.

### Uživatelské rozhraní (API)

V závislosti na účelu rozšíření je potřeba připravit uživatelské rozhraní (API). Jedná-li se o jednoduché datové transformace, lze veškeré nezbytné procesy definovat pouze ve specifikačním souboru. Jedná-li se o pokročilejší požadavky, je nutné použít jedno či více z následujících API-rozhraní:

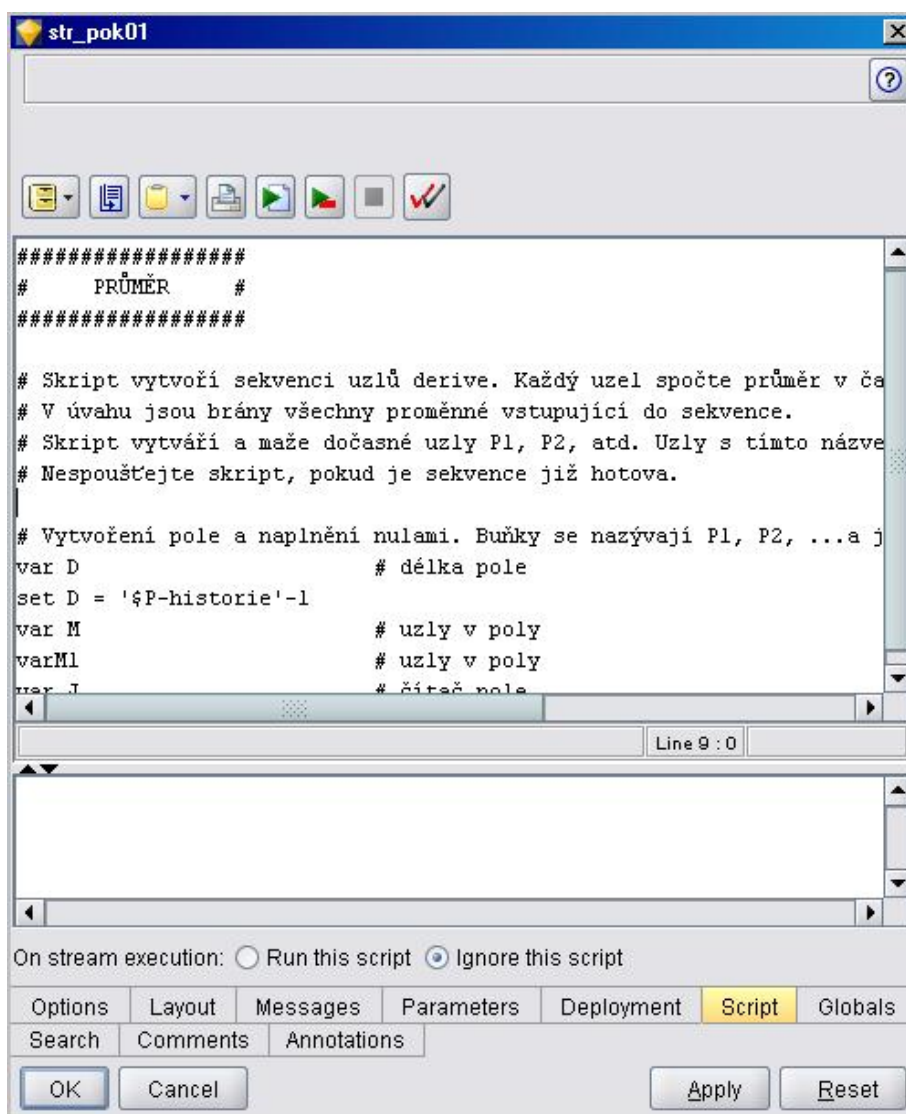
- CLEF client-side API
- CLEF server-side API
- Predictive Server API (PSAPI)

## 19. Skripty a programování

PASW Modeler podporuje uživatelské programy a skripty. Většinou jde o kratší programy, které usnadňují práci se systémem nebo automatizují často se opakující akce. Pomocí skriptů lze řídit postupné vykonávání proudů, nastavovat vlastnosti uzlů, specifikovat automaticky prováděné sekvence akcí atd.

PASW Modeler rozlišuje:

1. **StandAlone script** – skripty, které nejsou asociovány se žádným proudem a mohou tak manipulovat s více proudy. Jsou uloženy v externím textovém souboru.
2. **Stream script** – jsou součástí konkrétního proudu a s tímto proudem také uloženy. Ke skriptu se dostaneme v menu *Tools* → *Stream Properties* → *Script*.
3. **SuperNode script** – speciální skripty, které řídí obsah superuzlů. Záložka *Script* je aktivní pouze u *koncových* superuzlů.



Obr. 13: Ukázka skriptu

## 20. Tvorba reportů a dokumentace

PASW Modeler snadno vytváří reporty a dokumentaci z probíhajících projektů. Reporty se generují okamžitě a jsou ihned zobrazovány v dialogovém okně *Project Properties*, ze kterého mohou být vytisknuty nebo uloženy ve formátu HTML.

Před vlastním generováním reportu specifikuje uživatel objekty, které chce do reportu zahrnout.

### 20.1. Generování reportu


Report obsahuje informace z jednotlivých částí projektu, které jsou uloženy ve složce projektů. Ten je vytvářen ze dvou přístupů:

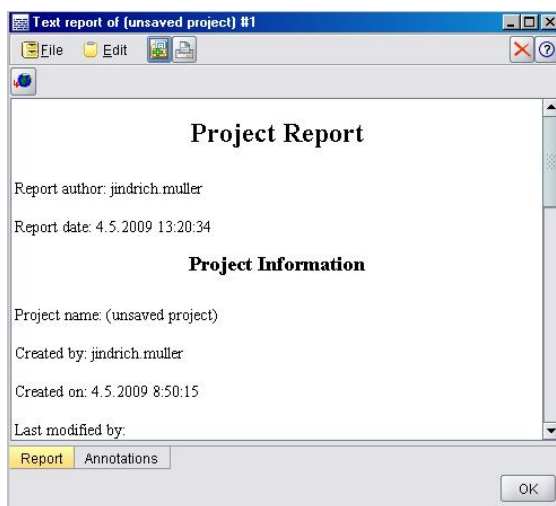
- přístup *CRISP-DM*;
- objektový přístup (*Classes*), tj. podle složek (proudy, uzly, modely, tabulky, grafy, ostatní).

Report zahrnuje popisy jednotlivých objektů umístěných do projektu, proto je vhodné těmto popisům věnovat patřičnou pozornost.

#### Postup při generování reportu

- v okně manager výstupů (složka projektů) vyvolejte pravým tlačítkem doplňkové menu a vyberte položku *Project Properties*,
- zvolte záložku *Report*,
- zvolte strukturu reportu – *CRISP-DM* nebo *Classes*,
- vyberte metodu zahrnutí objektů do reportu:
  - *Only object marked for inclusion in the report* – pouze označené,
  - *Only items marked for exclusion from the report* – pouze označené k vymazání,
  - *All folders and objects* – celý obsah,
- vyberte další možnosti řazení – podle typu objektu, názvu objektu, data vložení,
- stiskněte tlačítko *Generate Report*,
- vše potvrďte tlačítkem OK.

Generovaný projekt je zobrazen v dialogovém okně ve formátu HTML. Do internetového prohlížeče převedete report ikonou  (*Launch in external browser*) nebo report můžete vytisknout.



Obr. 14: Okno s generovaným reportem

## 21. SQL

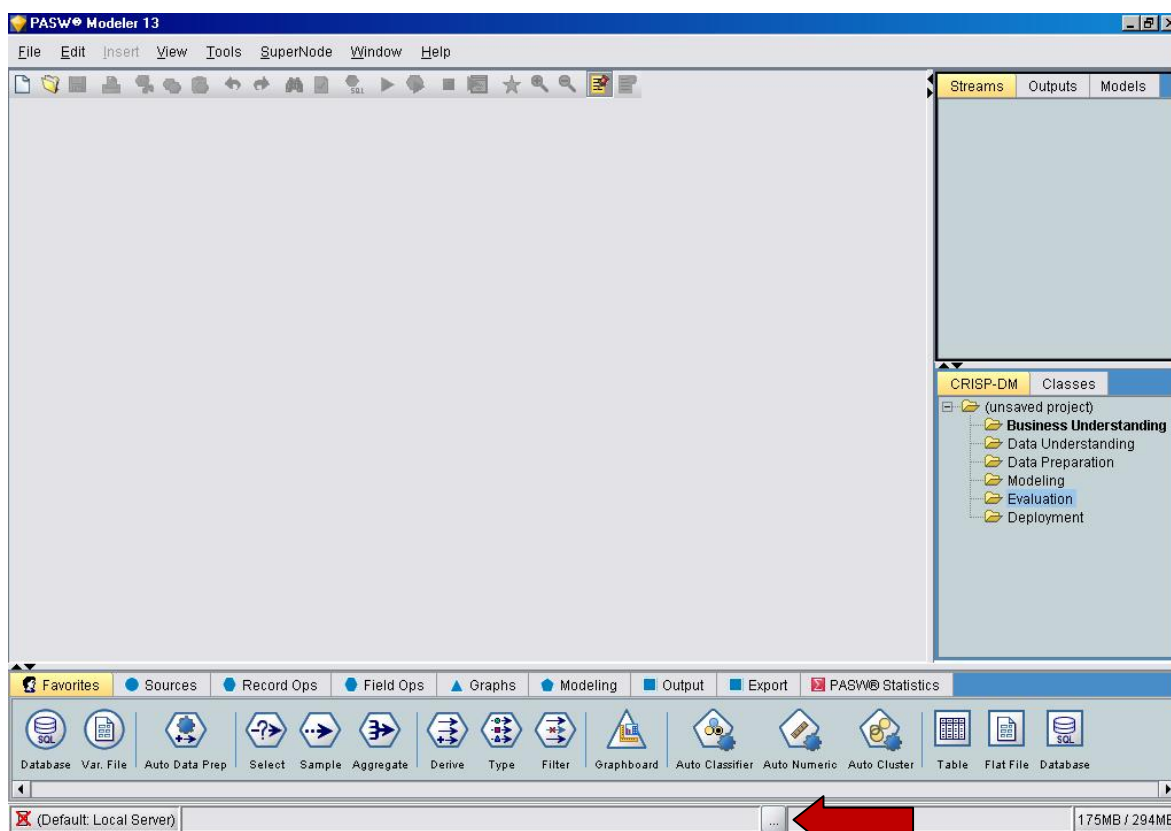
Jazyk **SQL** (*Structured Query Language*) je používán pro zápis požadavků na databázový server. Jednou z nejsilnějších stránek PASW Modeler je právě schopnost delegování úloh na databáze. Jedná se především o kroky při přípravě dat, skórování v databázi připraveným postupem a v některých případech lze dokonce do SQL převést i modelování. Generováním SQL kódu a delegováním úloh na stranu databáze lze mj. pořídit náhodný výběr, seřadit proměnné, odvodit nové proměnné a provést mnoho dalších operací. V případě rozsáhlých datových souborů tento tzv. *pushback* výrazně posiluje výkon celého nástroje.

Optimalizace pomocí jazyka SQL spolupracuje se všemi standardními databázemi (např. Oracle, SQL Server, DB2) s využitím driverů *Pack* distribuovaných společně s PASW Modeler.

Pro zobrazení kódu SQL je třeba provést nastavení systému. V menu *Tools* → *Options* → *User Options* přejděte na záložku *Optimization*. V dolní části dialogového okna je třeba zaškrtnout nabídku – alespoň jednu z voleb:

- *Display SQL in the messages log during stream execution*
- *Display SQL generation details in the messages log during stream preparation*

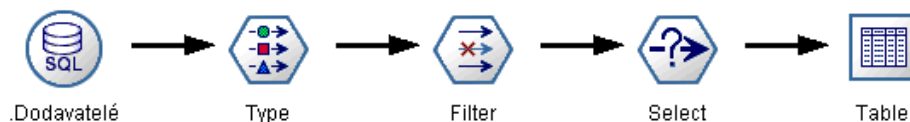
Generátor SQL kódu je aktivován stisknutím ikony *Preview SQL Generation for the Selection*. Vlastní kód si prohlédnete v okně *Messages*, které je vyvoláno ikonou *Show stream messages* umístěnou v dolní části hlavní obrazovky (viz šipka na obr. 15).



Obr. 15: Ikona stream messages

**Příklad:**

Z databáze dodavatelů je potřeba vybrat pouze italské firmy. Počet proměnných bude omezen jen na *Země* a *Firma* (uzel **Filter**) a podmínku výběru hodnot zapíšeme v uzlu **Select**. Celý proud vypadá např. takto:



Po jeho spuštění se fialově obarví uzly, jejichž akce jsou delegovány na databázi.



Zápis SQL příkazu přečteme v okně *Messages*:

```
Executing SQL: SELECT T0.`Firma` AS Firma, T0.`Země` AS Země FROM  
Dodavatelé AS T0 WHERE (T0.`Země` = 'Itálie');
```

Pro provedení modelování přímo v databázi vybereme používanou databázi v menu *Tools* → *Options* → *Helper Applications*. K dispozici jsou databáze společností Oracle, Microsoft, IBM a další. Na záložce PASW Statistics se nastavuje také spojení PASW Modeler s PASW Statistics, desktopovou nebo serverovou verzí. Po určení spojení s databází se paleta uzlů rozšíří o záložku *Database Modeling* obsahující modelovací algoritmy dostupné ve vybrané databázi.



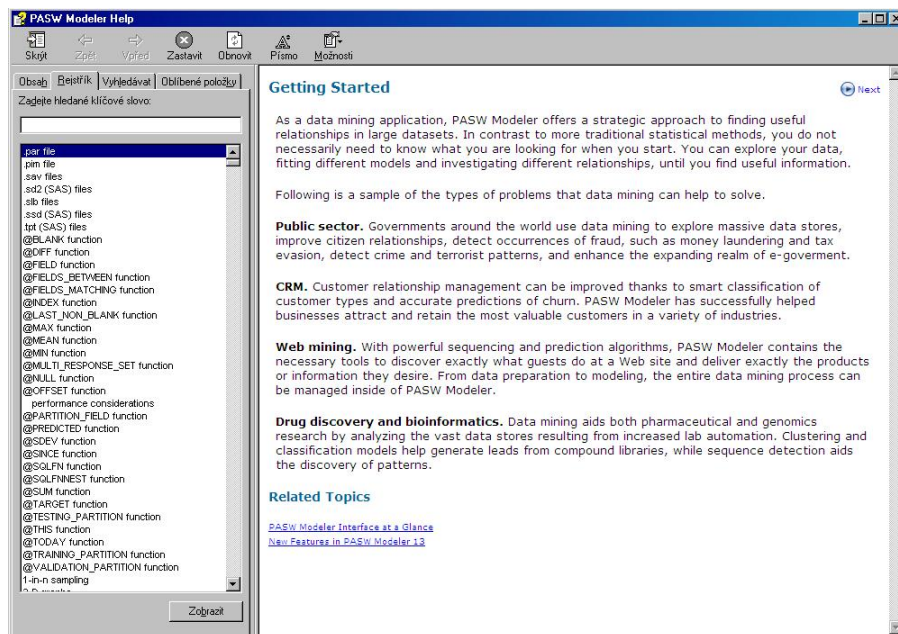
## 22. Systém nápověd

Nápověda k programu je v PASW Modeler dostupná v hlavním menu

**Help**

**Help Topics**

PASW Modeler otevře nové okno obsahující seznam nápověd (obr. 16).

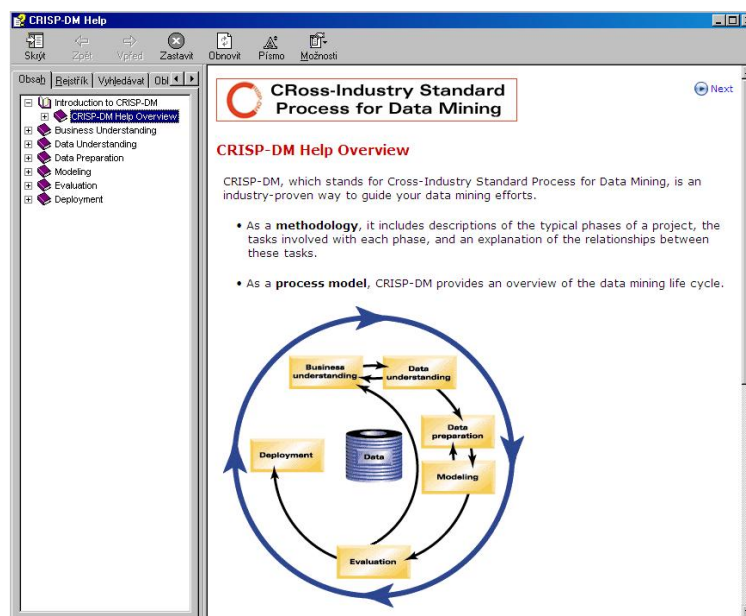


Obr. 16: Nápověda k programu PASW Modeler

Kliknutím na položku v levé části okna dojde k načtení odpovídající položky menu. Nápověda k metodologii CRISP-DM je dostupná v menu

**Help**

**CRISP-DM Help**



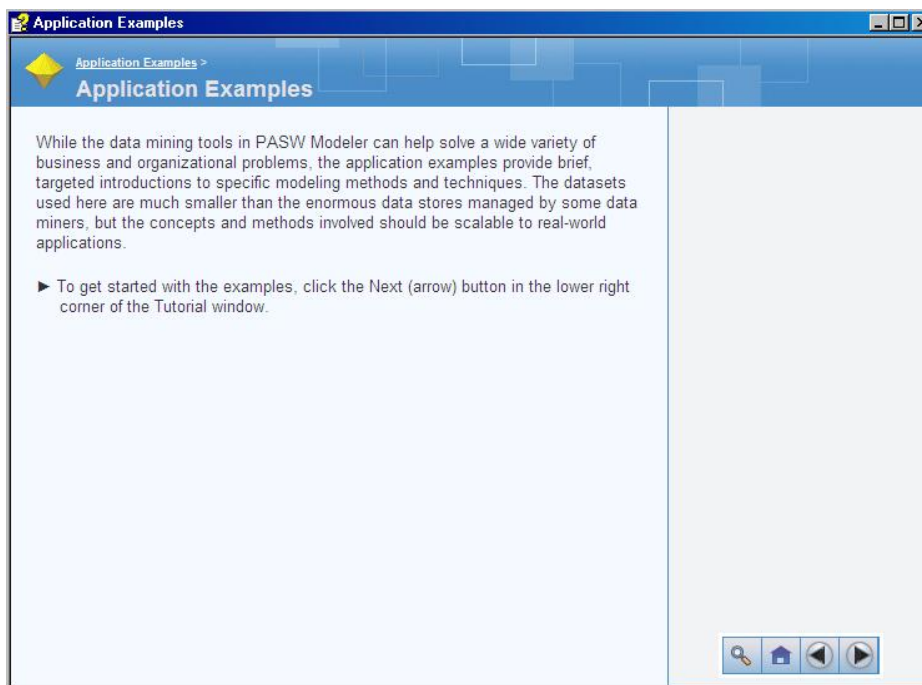
Obr. 17: Nápověda k metodice CRISP-DM



Kromě výše uvedených nápověd je k dispozici také podrobný průvodce aplikačními příklady v programu PASW Modeler (obr. 18). Obsahuje vzorové příklady a provádí všemi kroky metodiky CRISP-DM. Přístup k průvodci je v hlavním menu volbou

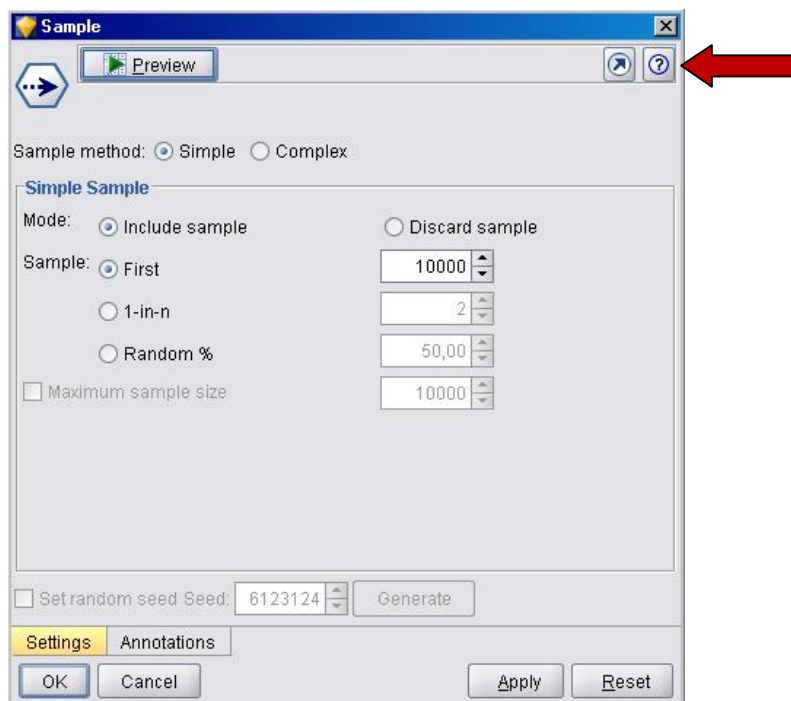
## Help

### Application Examples



Obr. 18: Průvodce programem PASW Modeler

Systém nápověd můžete rovněž vyvolat při vlastní práci s uzly. V dialogovém okně každého uzlu je k dispozici místní nápověda, která specifikuje kroky nastavení konkrétního uzlu.



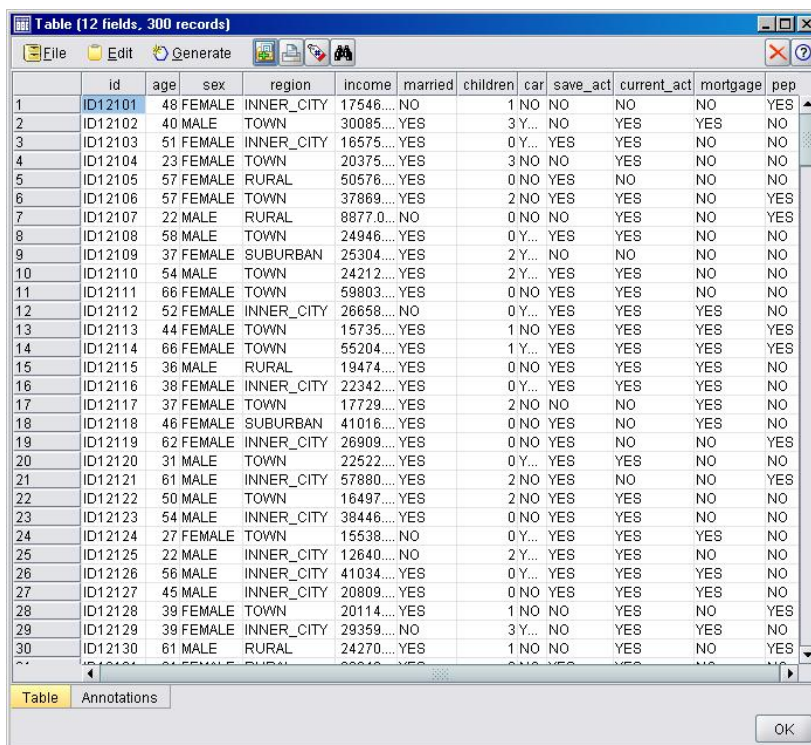
Obr. 19: Vyzvolání nápovědy při práci s uzlem

## 23. Příklad

Finanční instituce nabízí službu *Vytvoření osobního plánu správy cenných papírů PEP* (*Personal Equity Plan*). Někteří zákazníci o tuto službu již projevili zájem, jiní nikoli. Finanční instituce chce vytvořit model, který by jí pomohl v databázi vytypovat ty zákazníky, kterým se vyplatí poslat nabídkový leták, tj. ty, kteří jsou z hlediska této služby pro firmu perspektivní a o nabízenou službu budou mít pravděpodobně zájem.

Data jsou uložena v adresáři */Demos/SnapshottrainN.db*, který je součástí instalace PASW Modeler. Datová matice je na obr. 20 a obsahuje tyto proměnné:

<b>Id</b>	identifikátor zákazníka
<b>Age</b>	věk v letech
<b>Sex</b>	pohlaví
<b>Region</b>	region, ve kterém zákazník žije: INNER CITY, RURAL, SUBURBAN, TOWN
<b>Income</b>	celkový příjem
<b>Married</b>	ženatý/vdaná s kategoriemi YES a NO
<b>Children</b>	počet dětí
<b>Car</b>	vlastnictví auta s kategoriemi YES a NO
<b>Save_act</b>	příznak spořicího účtu
<b>Current_act</b>	příznak běžného účtu
<b>Mortgage</b>	příznak hypotéky
<b>PEP</b>	zájem o službu PEP s kategoriemi YES a NO

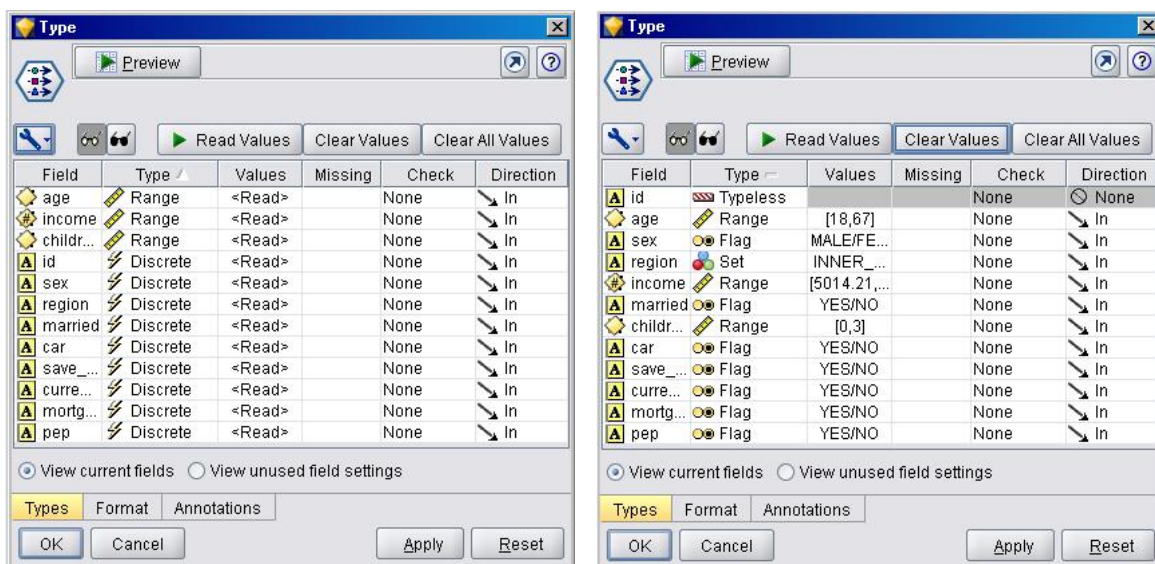


	id	age	sex	region	income	married	children	car	save_act	current_act	mortgage	pep
1	ID12101	48	FEMALE	INNER_CITY	17546....	NO	1	NO	NO	NO	NO	YES
2	ID12102	40	MALE	TOWN	30085....	YES	3	Y...	NO	YES	YES	NO
3	ID12103	51	FEMALE	INNER_CITY	18575....	YES	0	Y...	YES	YES	NO	NO
4	ID12104	23	FEMALE	TOWN	20375....	YES	3	NO	NO	YES	NO	NO
5	ID12105	57	FEMALE	RURAL	50576....	YES	0	NO	YES	NO	NO	NO
6	ID12106	57	FEMALE	TOWN	37869....	YES	2	NO	YES	YES	NO	YES
7	ID12107	22	MALE	RURAL	8877.0...	NO	0	NO	NO	YES	NO	YES
8	ID12108	58	MALE	TOWN	24946....	YES	0	Y...	YES	YES	NO	NO
9	ID12109	37	FEMALE	SUBURBAN	25304....	YES	2	Y...	NO	NO	NO	NO
10	ID12110	54	MALE	TOWN	24212....	YES	2	Y...	YES	YES	NO	NO
11	ID12111	66	FEMALE	TOWN	59803....	YES	0	NO	YES	YES	NO	NO
12	ID12112	52	FEMALE	INNER_CITY	26658....	NO	0	Y...	YES	YES	YES	NO
13	ID12113	44	FEMALE	TOWN	15735....	YES	1	NO	YES	YES	YES	YES
14	ID12114	66	FEMALE	TOWN	55204....	YES	1	Y...	YES	YES	YES	YES
15	ID12115	36	MALE	RURAL	19474....	YES	0	NO	YES	YES	YES	NO
16	ID12116	38	FEMALE	INNER_CITY	22342....	YES	0	Y...	YES	YES	YES	NO
17	ID12117	37	FEMALE	TOWN	17729....	YES	2	NO	NO	NO	YES	NO
18	ID12118	46	FEMALE	SUBURBAN	41016....	YES	0	NO	YES	NO	YES	NO
19	ID12119	62	FEMALE	INNER_CITY	26909....	YES	0	NO	YES	NO	NO	YES
20	ID12120	31	MALE	TOWN	22522....	YES	0	Y...	YES	YES	NO	NO
21	ID12121	61	MALE	INNER_CITY	57880....	YES	2	NO	YES	NO	NO	YES
22	ID12122	50	MALE	TOWN	16497....	YES	2	NO	YES	YES	NO	NO
23	ID12123	54	MALE	INNER_CITY	38446....	YES	0	NO	YES	YES	NO	NO
24	ID12124	27	FEMALE	TOWN	15538....	NO	0	Y...	YES	YES	YES	NO
25	ID12125	22	MALE	INNER_CITY	12640....	NO	2	Y...	YES	YES	NO	NO
26	ID12126	56	MALE	INNER_CITY	41034....	YES	0	Y...	YES	YES	YES	NO
27	ID12127	45	MALE	INNER_CITY	20809....	YES	0	NO	YES	YES	YES	NO
28	ID12128	39	FEMALE	TOWN	20114....	YES	1	NO	NO	YES	NO	YES
29	ID12129	39	FEMALE	INNER_CITY	29359....	NO	3	Y...	NO	YES	YES	NO
30	ID12130	61	MALE	RURAL	24270....	YES	1	NO	NO	YES	NO	YES

Obr. 20: Datová matice

Data jsou uložena v textovém formátu a pro jejich načtení je použit uzel **Var. File**. Dvojným kliknutím uzel otevřete a definujete cestu k datům. K uzlu je navíc připojena *cache*, abyste nemuseli data neustále načítat z textového souboru. Ostatní nastavení ponechte beze změn.

V dalším kroku je vhodné předat systému informaci o datových typech jednotlivých proměnných. Proto do proudu napojte uzel **Type**, který otevřete a tlačítkem *Read Values* hodnoty načtete (viz obr. 21). U proměnné **children** změňte typ z *Range* na *Ordered Set*.

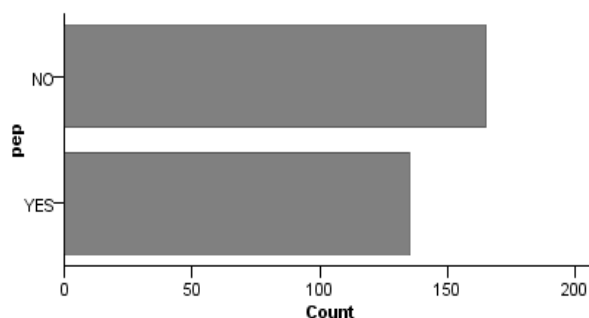
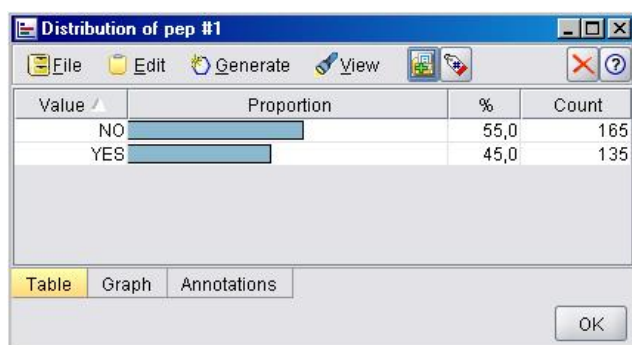


Obr. 21: Před načtením (vlevo) a po načtení (vpravo) hodnot v uzlu Type

V dalším kroku se již podíváme na rozložení hodnot cílové proměnné **PEP** pomocí uzlu **Distribution**, ve kterém zvolte do pole **Field** odpovídající proměnnou. Celý proud má v tomto okamžiku tuto podobu:



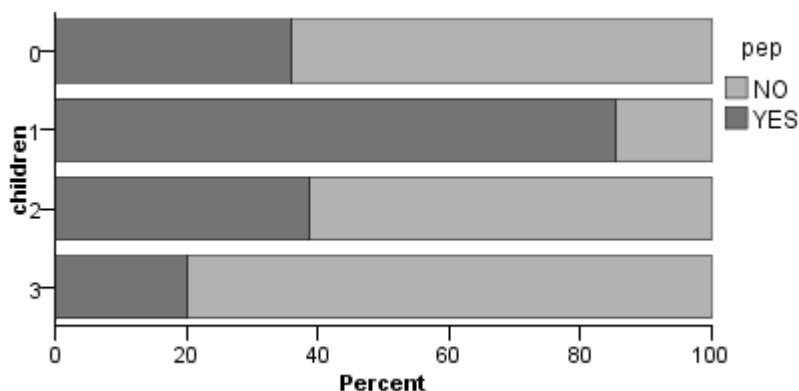
Proud spusťte a podívejte se na výsledek. Z grafu na obr. 22 je vidět, že 55 % zákazníků o službu zájem nemá, zbytek zájem má.



Obr. 22: Distribuce hodnot cílové proměnné (vpravo ukázka prezentační grafiky)

Před vlastním modelováním je vhodné provádět průzkumovou analýzu dat a hledat různé závislosti či vztahy. Podíváme se tedy na další grafy. Zajímavá může být např. informace, jak vypadá rozložení zájmu o **PEP** v jednotlivých kategoriích proměnné **Children** – počet dětí.

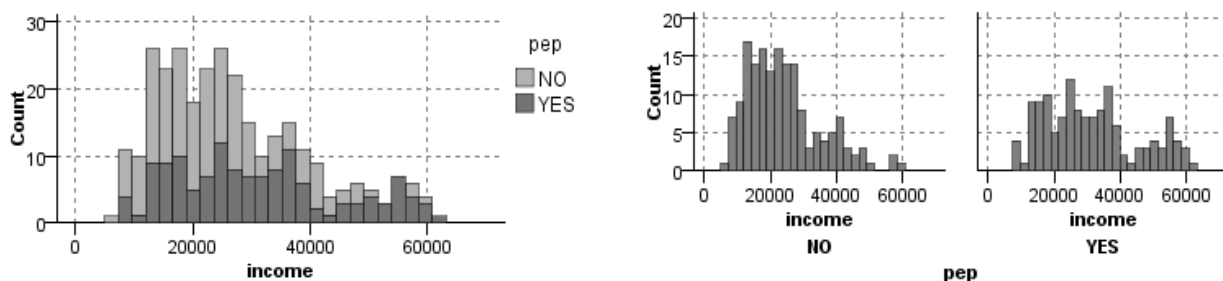
Vyberte uzel/graf **Distribution** a do pole **Field** vložte proměnnou **Children**. Do pole **Color** vložte cílovou proměnnou **PEP** a zaškrtněte pole **Normalize by color**. Tím je dosaženo normalizace, tj. stejné délky všech sloupců, díky čemuž lépe vyniknou relativní podíly kategorií sledované proměnné.



Obr. 23: Distribuce hodnot cílové proměnné v kategoriích proměnné počet dětí

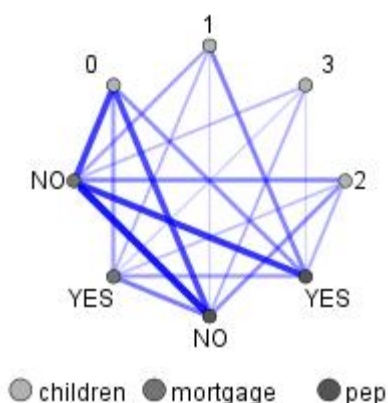
Na obr. 23 je zřejmý vysoký podíl zájemců o *PEP* v kategorii *1 dítě*. Obdobným způsobem zjistíme závislost zájmu o službu *PEP* a velikosti příjmu (*Income*).

Graf pro spojitou proměnnou vytvoříte pomocí uzlu **Histogram**, barevné rozlišení zvolte podle proměnné *PEP*. Zde je zřejmá tendence vzrůstajícího zájmu o *PEP* s růstem příjmu.



Obr. 24: Histogram proměnné *Income* podle *PEP* (vpravo alternativní podoba grafu)

Zajímavé je zobrazení vztahů pomocí pavučinového grafu, který se vytváří uzlem **Web**. Linky spojující jednotlivé kategorie proměnných ukazují sílu vztahu mezi nimi. Čím tlustější linka, tím častější jsou odpovídající kombinace hodnot v datech. Volitelné je nastavení absolutních či relativních hodnot. V uzlu jsou standardním způsobem voleny proměnné, které chceme zobrazit. Velmi silně je zastoupená kombinace hodnot *MORTGAGE*: No, *PEP*: No, *CHILDREN*: 0.



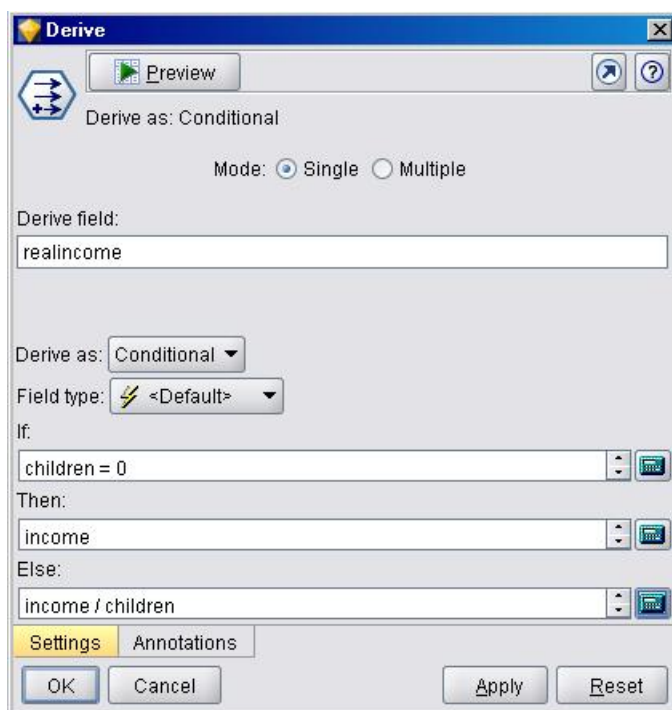
Obr. 25: Pavučinový graf

Nyní odvodíme proměnnou, ve které bude uvedena hodnota *příjmu na jedno dítě*. V případě bezdětné domácnosti pak budeme chtít, aby se v řádku objevila původní hodnota

proměnné *Income*. K odvození použijeme uzel **Derive**, ve kterém provedeme konkrétní nastavení (uzel **Derive**, umístěný na paletě *Field Ops*, zařadíte za uzel **Type**).

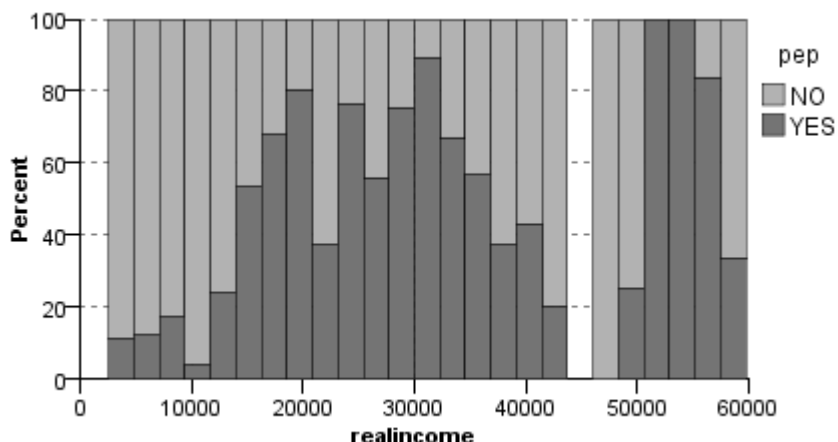
Do řádku *Derive field* zapište název nové proměnné, např. *realincome*. V poli *Derive as* je třeba nastavit způsob odvození, resp. typ odvozované proměnné. Výpočet nové proměnné je podmíněn počtem dětí, proto tedy zvolte *Derive as Conditional*. Pokud je domácnost bezdětná, bude hodnota nové proměnné *realincome* rovna hodnotě proměnné *income*, proto do podmínky *If* zapište *children = 0* a do větve *Then* pak *income*.

Pokud jsou v domácnosti děti, pak hodnota nové proměnné *realincome* bude rovna hodnotě proměnné *income* vztahované na jedno dítě, proto do druhé větve *Else* zapište podíl *income / children*.



Obr. 26: Nastavení v uzlu *Derive*

Nyní již lze znázornit rozložení příjmu na jedno dítě včetně obarvení podle cílové proměnné *PEP* pomocí uzlu **Histogram** (na záložce *Options* normalizujte podle barvy). Z obr. 27 je zřejmé, že s růstem příjmu na jedno dítě se mění podíl zákazníků, kteří mají o *PEP* zájem. V nejvyšších příjmových pásmech je tento podíl nejvyšší.

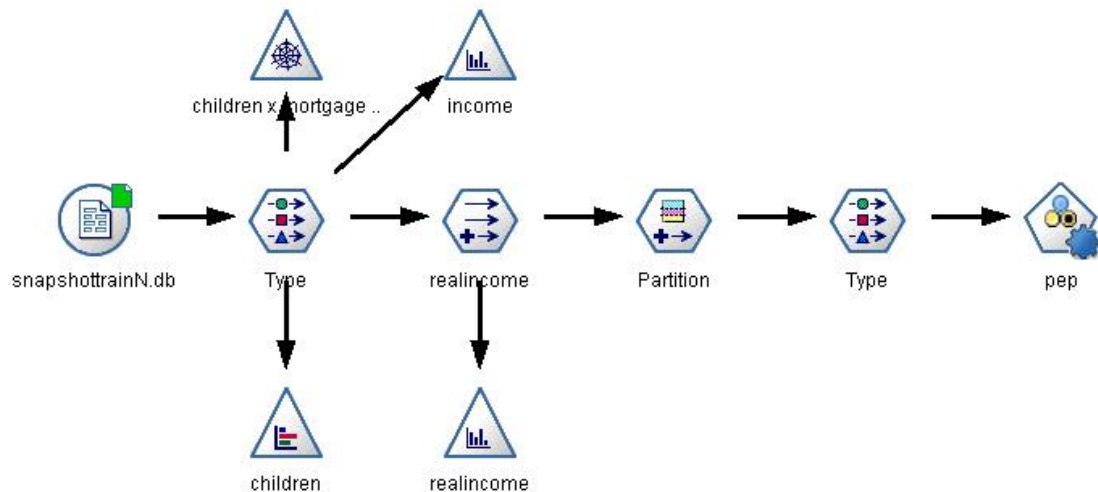


Obr. 27: Rozložení příjmu na 1 dítě obarveno podle cílové proměnné *PEP*

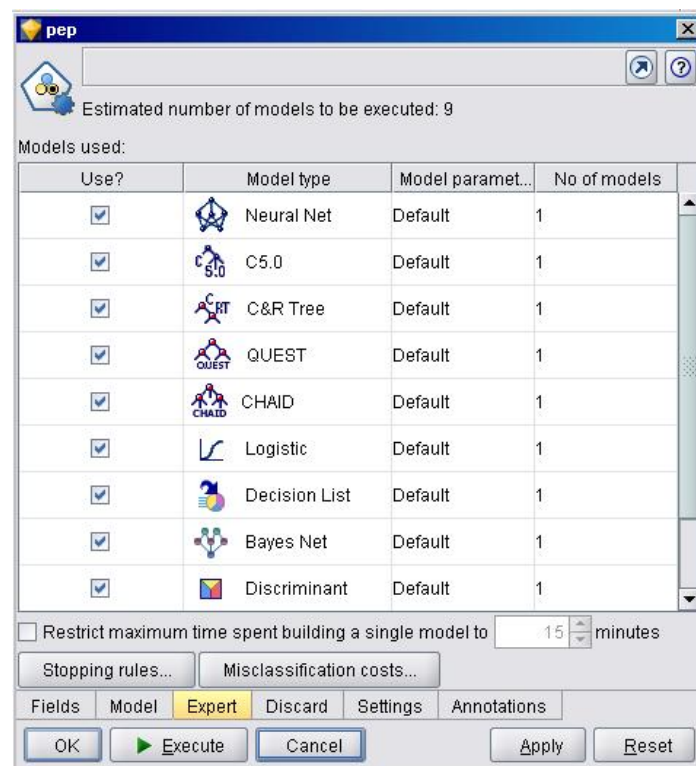


Při výběru optimálního modelu pro binární cílovou proměnnou lze využít uzel **Auto Classifier**, a to zejména v případech, kdy není znám model předem a kdy je třeba urychlit proces jeho hledání. Uzel je zapojen do proudu za nový uzel **Type**, ve kterém je změněn u proměnné **PEP** směr (*direction*) na **Out**.

Před vlastním modelováním je datový soubor rozdělen na dvě části. K rozdělení je použit uzel **Partition**, ve kterém se určí poměr velikosti trénovací a testovací množiny. Vyberte velikosti skupin např. v poměru 70:30. V poli *Training partition size* je nastavena hodnota 70 a v poli *Testing partition size* hodnota 30. Celý proud má nyní tuto podobu:



Při hledání optimálního modelu budeme testovat modely: neuronová síť, rozhodovací stromy *C5.0*, *C&RT*, *QUEST*, *CHAID*, logistická regrese, decision list, bayesovská síť a diskriminační analýza. Tato volba je dostupná v uzlu **Auto Classifier** na záložce **Expert**:



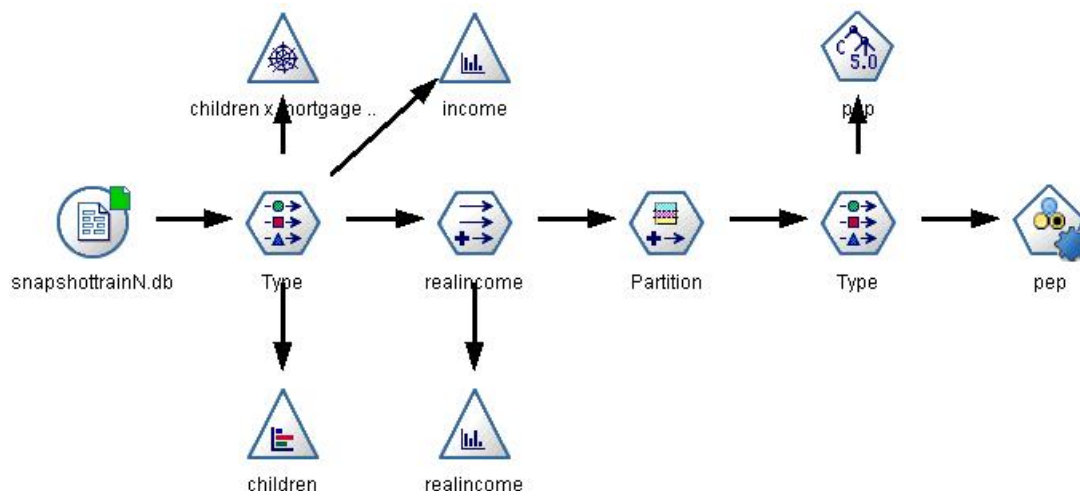
**Obr. 28: Nastavení v uzlu Auto Classifier**

Výsledkem procedury je seznam modelů seřazený podle zvoleného kritéria (např. *Profit*, *Lift*, *Overall accuracy*). V našem případě uzel **Auto Classifier** navrhuje tři modely, z nichž podle většiny kritérií vychází nejlépe rozhodovací strom C5.0.

Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift (Top 30%)	Overall Accuracy (%)	No. Fields Used	Area Under Curve
<input checked="" type="checkbox"/>		C5.1	< 1	190	49	1,92	91,49	7	0,93
<input checked="" type="checkbox"/>		C&R Tree 1	< 1	180	43	1,92	85,11	8	0,9
<input checked="" type="checkbox"/>		Logistic regression 1	< 1	155	36	2,04	82,98	11	0,91

Obr. 29: Seznam modelů

Pro vlastní modelování tedy budeme používat rozhodovací strom C5.0. Uzel naleznete na paletě *Modeling*. Uzel zapojte do proudu za uzel **Type** vedle uzlu **Auto Classifier**. Dbejte na to, aby v uzlu **Type** byl nastaven směr (*direction*) proměnné *PEP* na *Out*. Celý proud má tuto podobu:

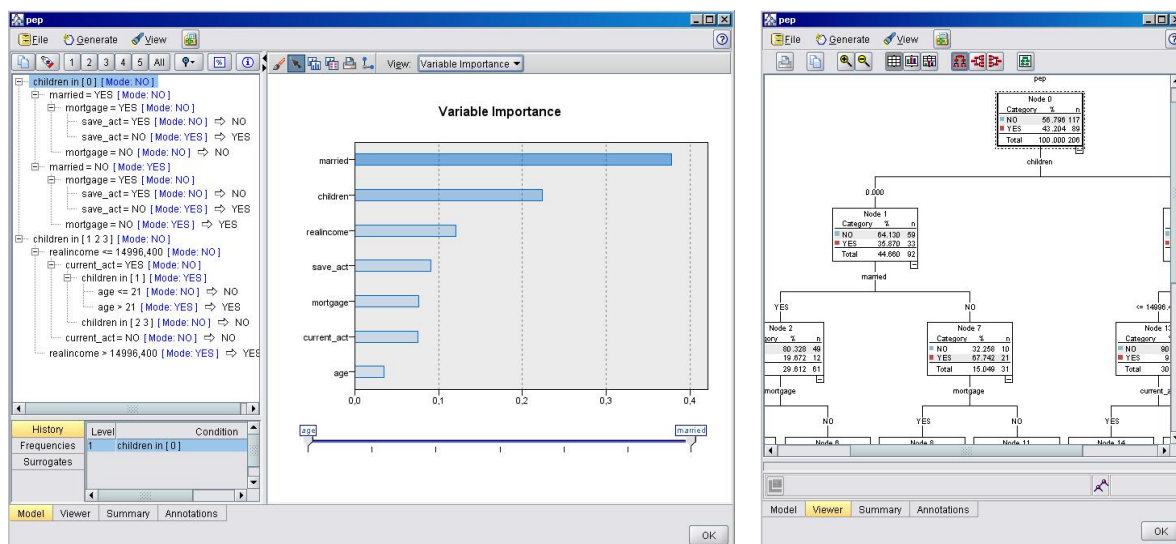


Hotový model C5.0 ukládá do datové matice dvě nové proměnné:

- **\$C-pep** – odhad kategorie cílové proměnné, tj. odhad toho, zda bude zákazník mít o službu *PEP* zájem nebo ne
- **\$CC-pep** – odhad pravděpodobnosti správného určení kategorie cílové proměnné

Výsledný model (žlutý drahokam) se po chvíli objeví ve správci výstupů na záložce *Models*. Označením modelu pravým tlačítkem myši (kliknutím) se vyvolá doplňkové menu, stiskněte *Browse* a dostanete se do modelu, ve kterém je možné prohlížet jeho vlastnosti:

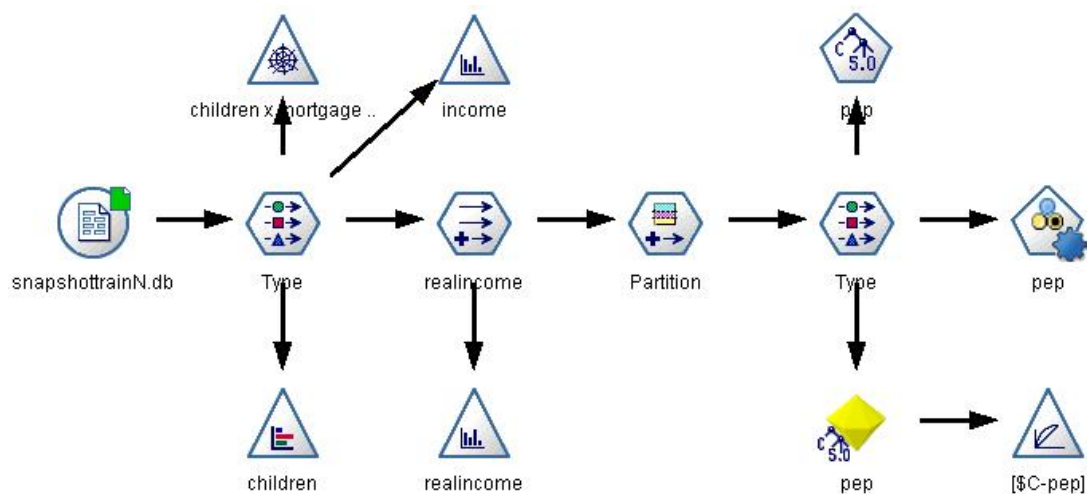
- **Model** – zobrazení stromové struktury od kořenového uzlu až po konečné listy a znázornění grafu důležitosti proměnných. Ve stromové struktuře je volitelné nastavení pro zobrazení četností koncových uzlů včetně pravděpodobnosti správné predikce.
- **Viewer** – grafické zobrazení rozhodovacího stromu
- **Summary** – sumář modelu
- **Annotation** – popisy a poznámky k modelu



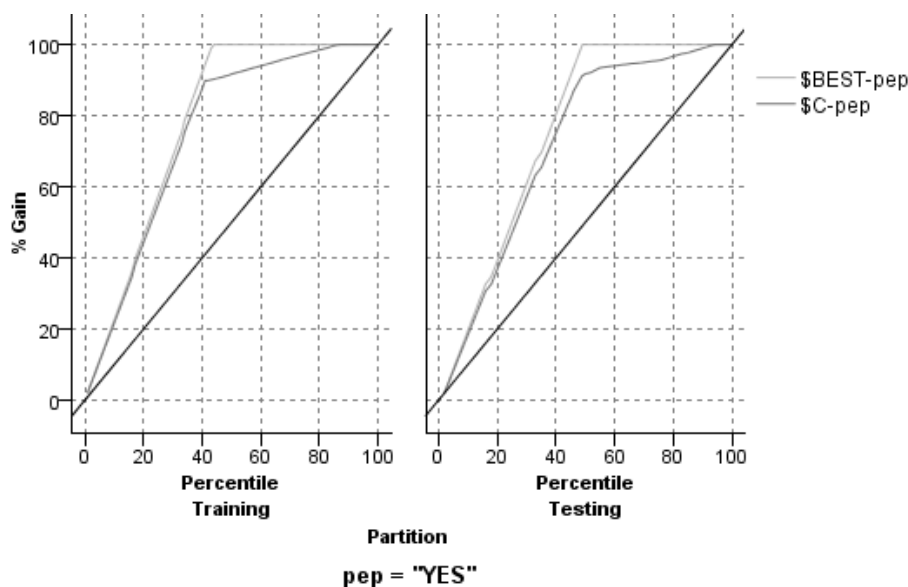
Obr. 30: Model C5.0 – Záložky Model a Viewer

Kvalita modelu se hodnotí pomocí evaluačních grafů v uzlu **Evaluation** na záložce **Graphs**. Ukažme si graf nazvaný *Gains chart*.

Za uzel **Type** zapojte do proudu hotový model (ikonka žlutého diamantu) ze správce výstupů, za model pak zapojte uzel evaluačního grafu. V uzlu **Evaluation** nastavte na záložce **Plot** v horní části dialogu typ evaluačního grafu *Chart type: Gains*, dále zaškrtněte možnost *Include best line* a pro přehlednost jej v záložce **Annotation** nazvěte *Gains*. Pracovní proud a výstup vypadají následovně:



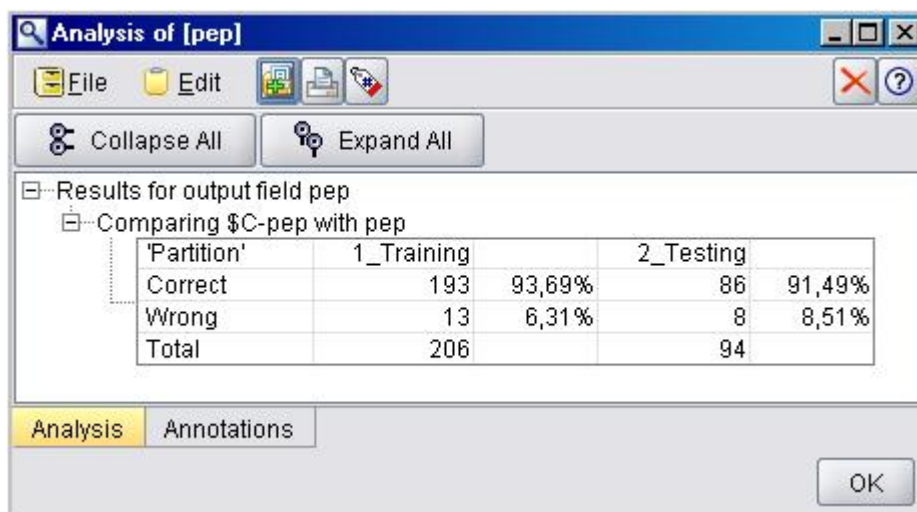




Obr. 31: Evaluační graf Gains

Vrchní křivka zobrazuje ideální situaci 100% kvalitního modelu bez chyb, dolní přímka pak situaci náhodného přiřazování kategorií cílové proměnné. Křivka uprostřed popisuje situaci vytvořeného modelu. Čím kvalitnější je model, tím těsněji se k sobě blíží křivka modelu a křivka ideální situace.

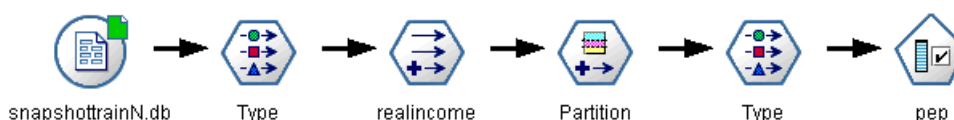
Pomocí uzlu **Analysis** (paleta *Output*) lze kvalitu modelu vyčíslit. Uzel není nutné nijak nastavovat, stačí jej napojit do proudu a spustit.



Obr. 32: Výstup z uzlu Analysis

V trénovací množině dat bylo 93,69 % případů klasifikováno správně. V testovací množině je tento poměr 91,49 %. Výsledek lze proto považovat za uspokojivý.

Modifikací uvedeného postupu může být automatická detekce vhodných vysvětlujících proměnných uzlem **Feature Selection** (paleta *Modeling*), který na bázi statistických testů ověří statistickou významnost každého potenciálního prediktoru a proměnné seřadí podle důležitosti.



Obr. 33: Proud s uzlem Feature Selection

Výstupem je interaktivní seznam proměnných, seřazených podle statistické významnosti pro predikci zvolené cílové proměnné – *PEP*.

	Rank	Field	Type	Importance	Value
<input checked="" type="checkbox"/>	1	children	Ordered Set	★ Important	1,0
<input checked="" type="checkbox"/>	2	realincome	Range	★ Important	1,0
<input checked="" type="checkbox"/>	3	income	Range	★ Important	0,998
<input checked="" type="checkbox"/>	4	age	Range	★ Important	0,998
<input type="checkbox"/>	5	sex	Flag	□ Unimportant	0,896
<input type="checkbox"/>	6	mortgage	Flag	□ Unimportant	0,764
<input type="checkbox"/>	7	current_act	Flag	□ Unimportant	0,705
<input type="checkbox"/>	8	married	Flag	□ Unimportant	0,372
<input type="checkbox"/>	9	save_act	Flag	□ Unimportant	0,277
<input type="checkbox"/>	10	region	Set	□ Unimportant	0,156
<input type="checkbox"/>	11	car	Flag	□ Unimportant	0,082

Selected fields: 4    Total fields available: 11

☒ ★ > 0,95  
 ☐ + <= 0,95  
 ☐ □ < 0,9

0 Screened Fields

Field	Type	Reason
-------	------	--------

Model   Summary   Annotations

OK

**Obr. 34: Výstup z uzlu Feature Selection**