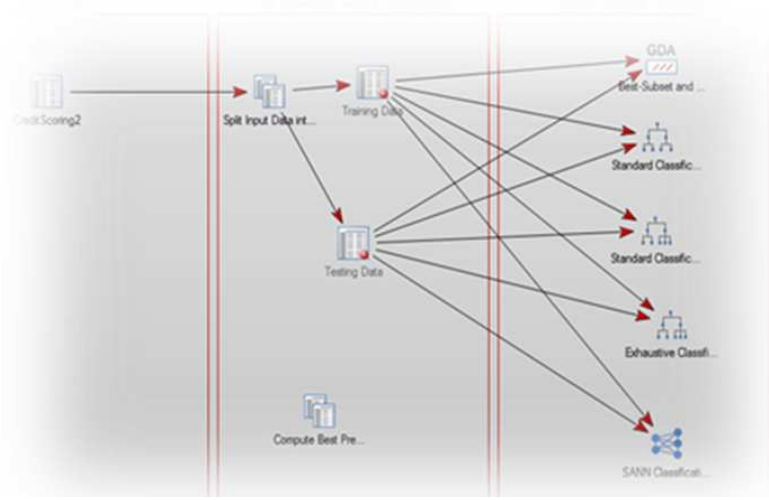
Data Miner Workspaces

Stručný průvodce
(ukázka práce v Data Miner Workspaces)



Data Miner Workspaces

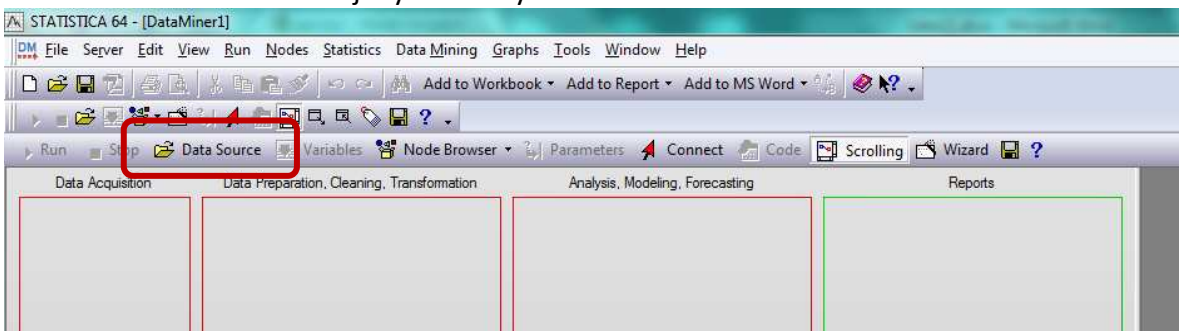
Data Miner workspaces - slouží k vytváření složitější DM struktury
Výhoda Data Miner Workspaces je zřejmá zejména při tvorbě složitějších data miningových úloh. Například mějme úlohu, kdy máme vytáhnout data, pak je rozdělit na testovací a trénovací, pak na těchto datech porovnat několik analýz a nakonec ze tří nejlepších modelů vytvořit predikci na základě majoritního hlasování. Toto by se přes klasické interaktivní rozhraní dělalo zdlouhavě a nevyvarovali bychom se chybám.

Grafické rozhraní Data Miner Workspaces se dělí do 4 logicky oddělených bloků:

- Datový zdroj
- Předzpracování dat
- Vlastní analýza
- Výstupy – report výsledku

Do těchto oblastí se jednoduchým způsobem vkládají uzly s požadovanými funkcionalitami. Jednak to mohou být datové zdroje, uzly pro transformaci, uzly analýz, uzly pro vytváření grafů, atd. Uzly lze na sebe napojovat a vytvářet tak složitější DM struktury. V grafickém rozhraní je pak přehledně viditelné, jak jdou jednotlivé analýzy za sebou a na jakých datech se provádějí.

Příklad vložení uzlu se zdrojovým datovým souborem:

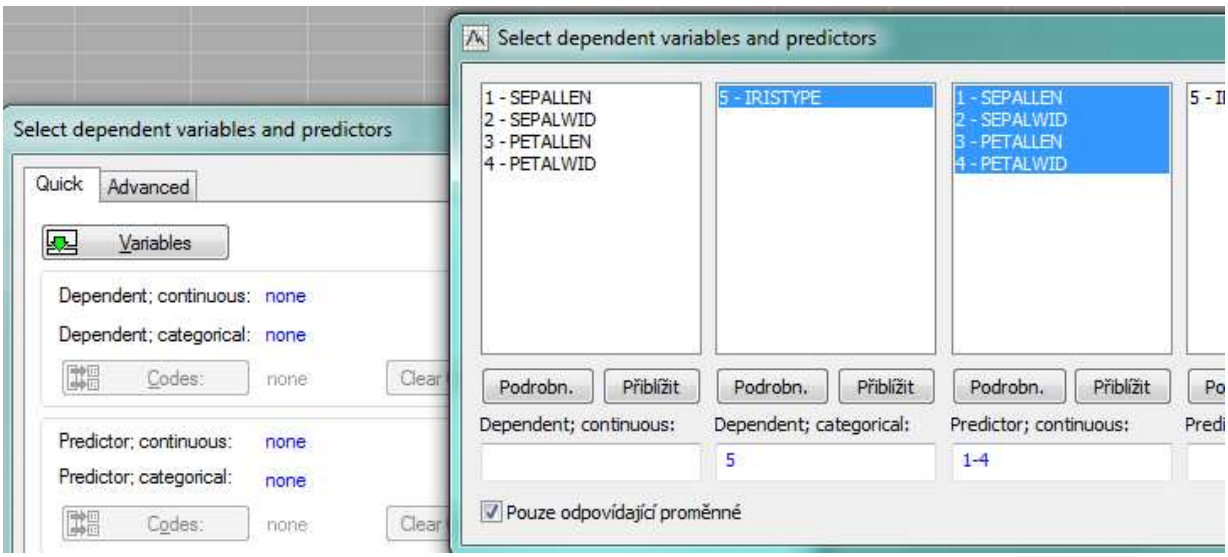


Vybrali jsme datový soubor v příkladech *STATISTICA*

Irisdat.sta *(Soubor -> Otevřít -> Datasets)*

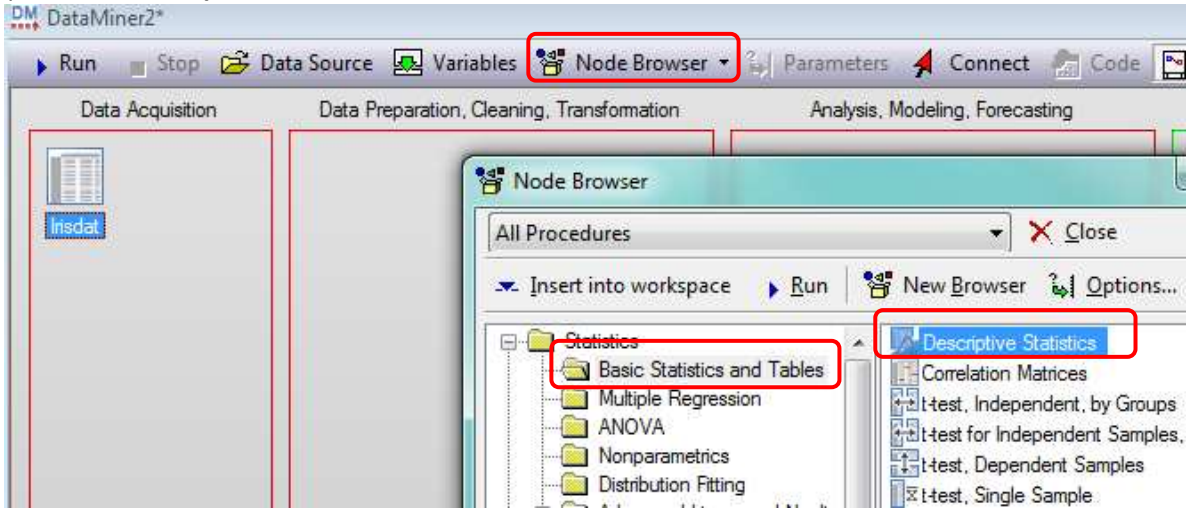
U každého datového souboru je potřeba vybrat proměnné, které půjdou do dalších analýz (uzlů).

- Zde volíme: Závislá proměnná je typ kosatce, nezávislé spojité prediktory máme 4:



Zobrazení základních statistických údajů

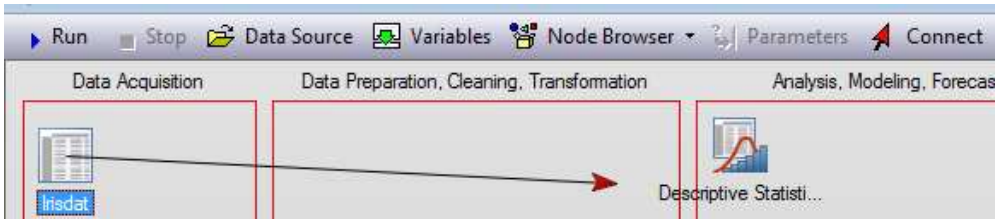
Jednotlivé uzly, které můžeme datům stavět do cesty, jsou k nalezení v prohlížeči uzlů (**Node Browser):**



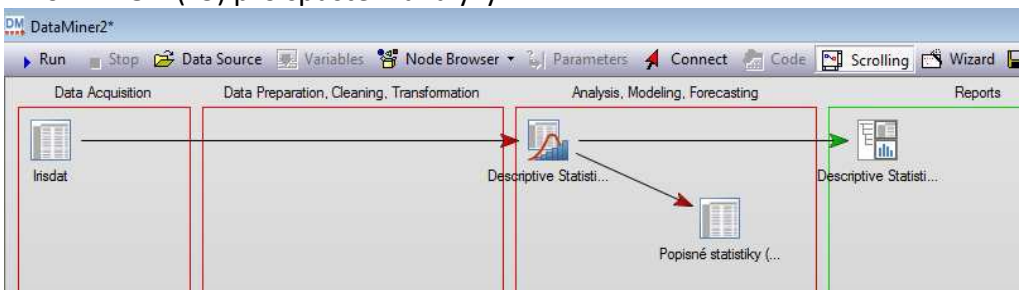
Varianty umístění do plochy:

- Přetáhnu do dialogu
- Dvojklik na vybranou metodu

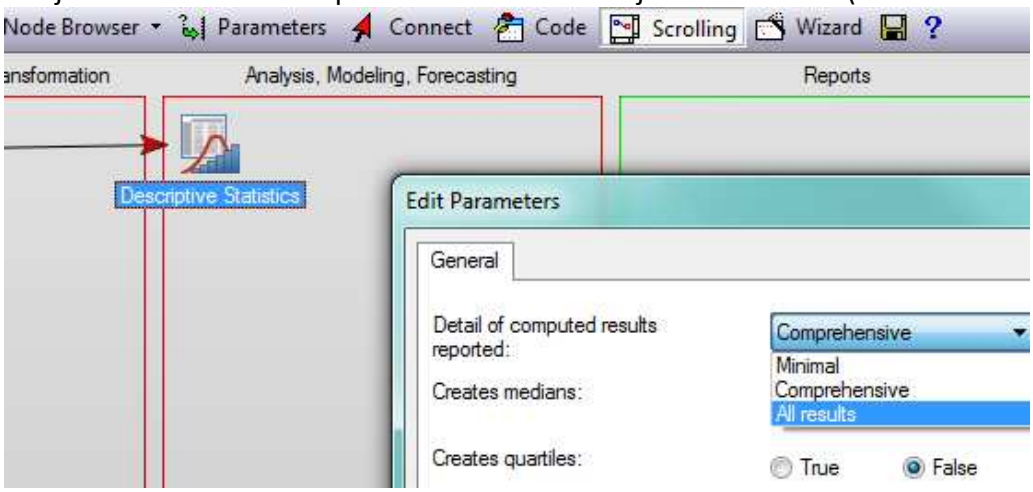
Přes ikonu **Connect** propojím data a popisné statistiky:



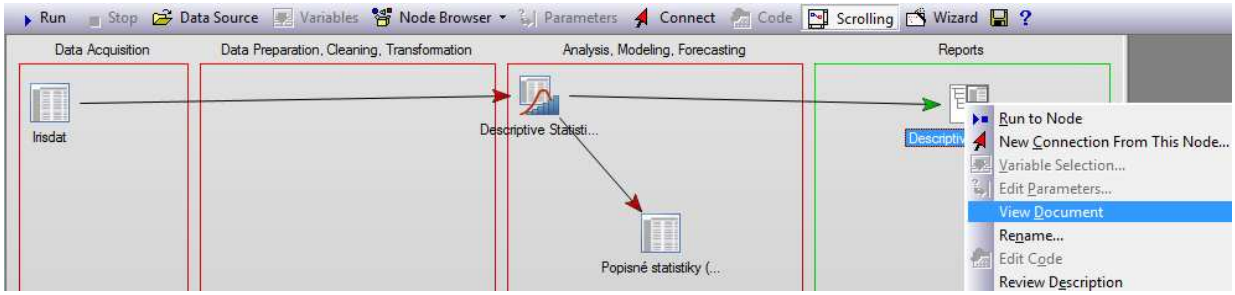
A volím **RUN** (F5) pro spuštění analýzy:



Dvojklikem na uzel Descriptive Statistics – definuji možnosti uzlu (volíme maximální výsledek):

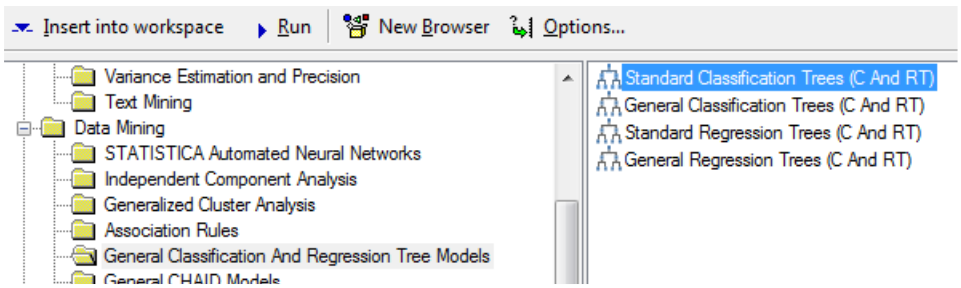


Pravým tlačítkem na list zobrazíme výsledek (u výsledných reportů ve formě sešitu *STATISTICA* je možné sešit výsledků zobrazit i dvojklikem na uzel):

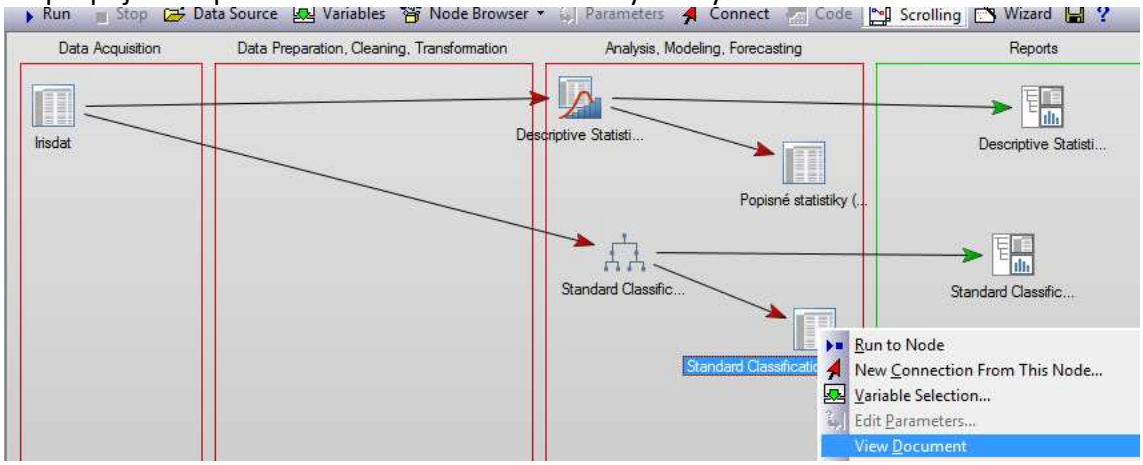


Přes záložku **Okno** – DataMiner1 se vrátím zpět na plochu **Workspaces**

Natrénovaní klasifikátoru:



Po propojení a spuštění sekvence si zobrazíme výsledky:

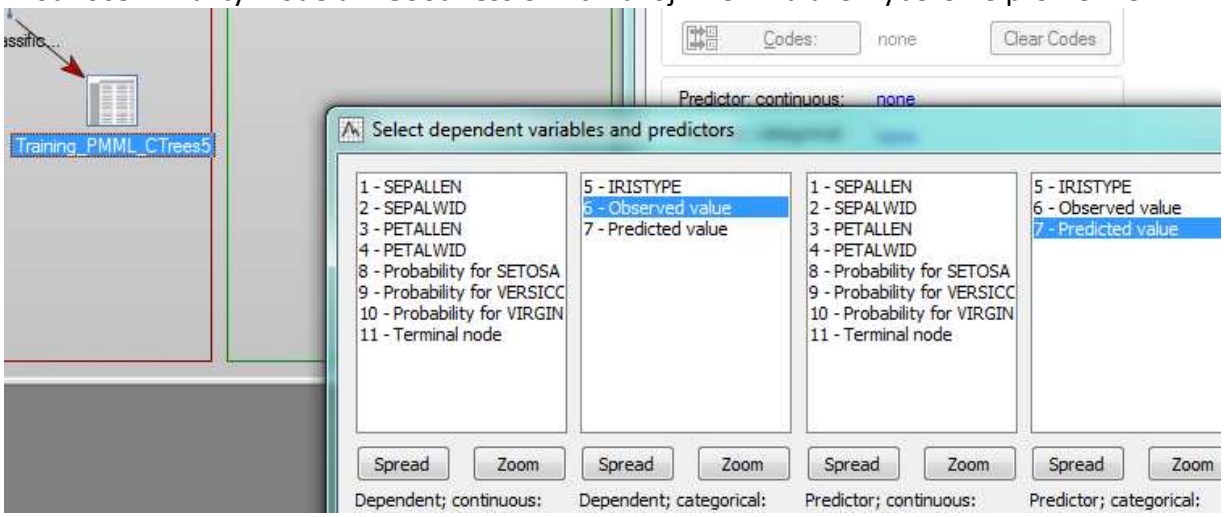


Pozorovaná vs. Predikovaná hodnota + jednotlivé pravděpodobnosti klasifikace:

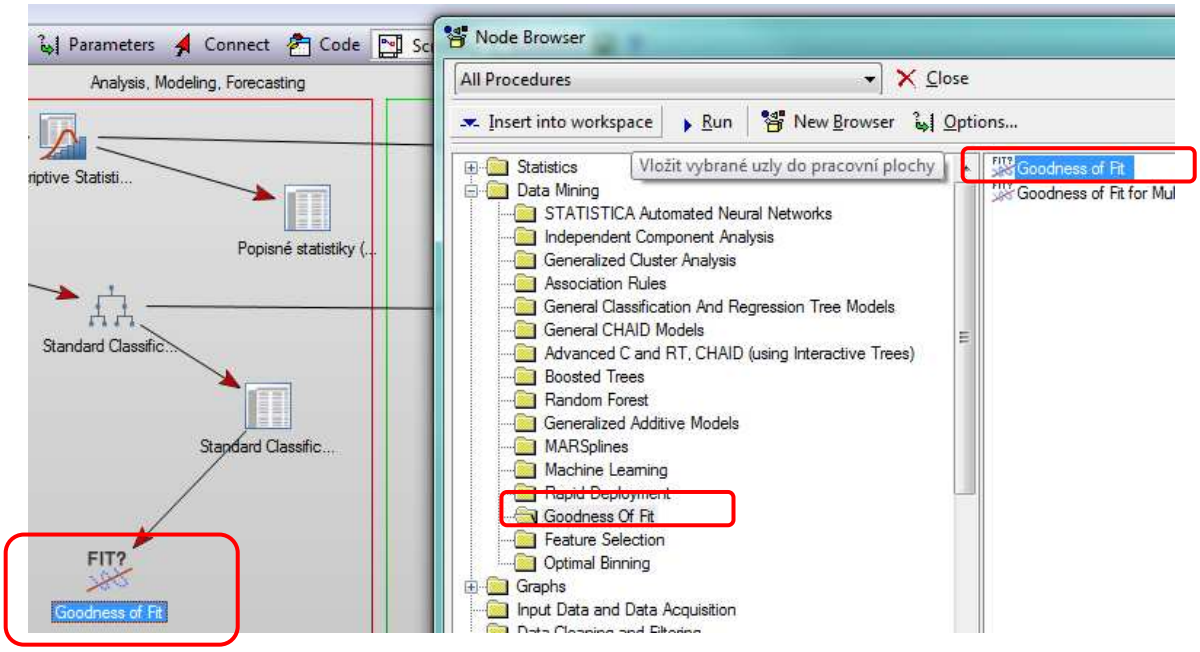
Predicted values 1
Dependent variable: IRISTYPE
Options: Categorical response

	1 SEPALLEN	2 SEPALWID	3 PETALLEN	4 PETALWID	5 IRISTYPE	6 Observed value	7 Predicted value	8 Probability for SETOSA	9 **Probability for VERSICOL**	10 Probability for VIRGINIC	11 Terminal node
19	5,7	2,8	4,5	1,3	VERSICOL	VERSICOL	VERSICOL	0,00000	1,00000	0,00000	8,00000
20	6,2	3,4	5,4	2,3	VIRGINIC	VIRGINIC	VIRGINIC	0,00000	0,02174	0,97826	5,00000
21	7,7	3,8	6,7	2,2	VIRGINIC	VIRGINIC	VIRGINIC	0,00000	0,02174	0,97826	5,00000
22	6,3	3,3	4,7	1,6	VERSICOL	VERSICOL	VERSICOL	0,00000	1,00000	0,00000	8,00000
23	6,7	3,3	5,7	2,5	VIRGINIC	VIRGINIC	VIRGINIC	0,00000	0,02174	0,97826	5,00000
24	7,6	3,0	6,6	2,1	VIRGINIC	VIRGINIC	VIRGINIC	0,00000	0,02174	0,97826	5,00000
25	4,9	2,5	4,5	1,7	VIRGINIC	VIRGINIC	VIRGINIC	0,00000	0,00000	1,00000	9,00000
26	5,5	3,5	1,3	0,2	SETOSA	SETOSA	SETOSA	1,00000	0,00000	0,00000	2,00000
27	6,7	3,0	5,2	2,3	VIRGINIC	VIRGINIC	VIRGINIC	0,00000	0,02174	0,97826	5,00000
28	7,0	3,2	4,7	1,4	VERSICOL	VERSICOL	VERSICOL	0,00000	1,00000	0,00000	8,00000
29	6,4	3,2	4,5	1,5	VERSICOL	VERSICOL	VERSICOL	0,00000	1,00000	0,00000	8,00000
30	6,1	2,8	4,0	1,3	VERSICOL	VERSICOL	VERSICOL	0,00000	1,00000	0,00000	8,00000

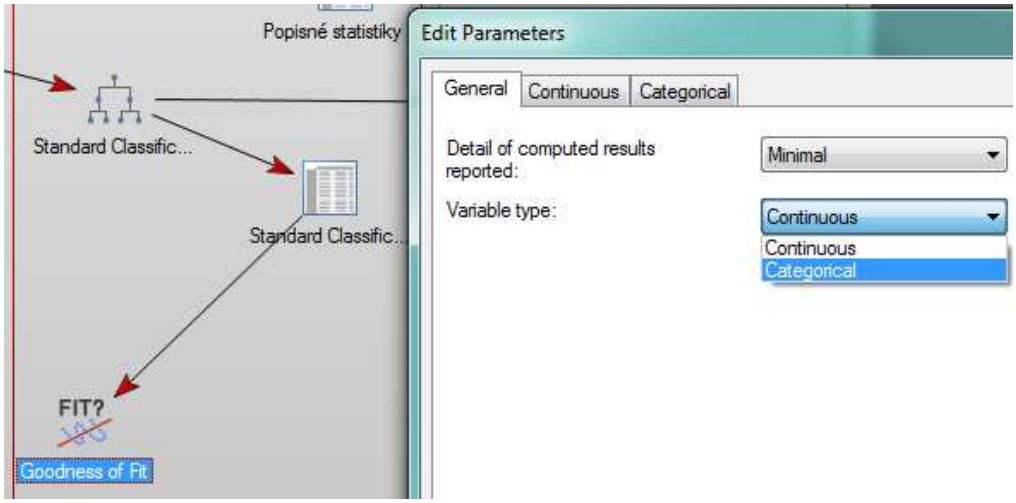
Na těchto datech chceme provádět další operace – tabulku použijeme jako vstup pro zhodnocení kvality modelu – **Goodness of Fit** – dvojklikem na uzel vybereme proměnné:



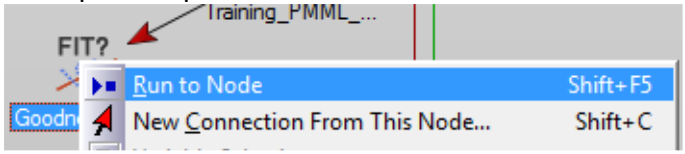
Přidáme samotnou metodu **Goodness of Fit:**



Dvojklikem na nový uzel – porovnáváme kategorické veličiny - nastavíme typ proměnných: ***Categorical:***

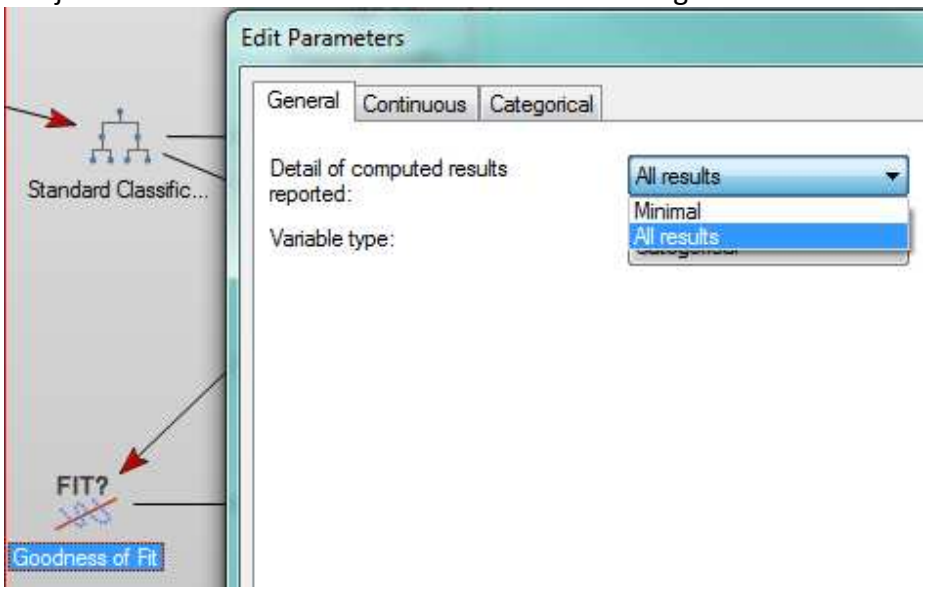


Chci spuštění pouze tohoto uzlu:



Při spuštění celého Workspaces se všechny uzly přepočítají, já chci analýzu nad právě současnými výsledky uzlu => **Run to Node** (možnost přepočítat pouze část uzlů).

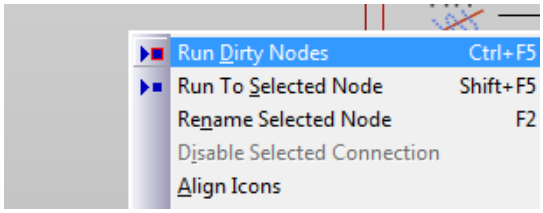
Dvojklikem na uzel **Goodness of Fit** zobrazíme dialog **Edit Parameters** a volíme All results:



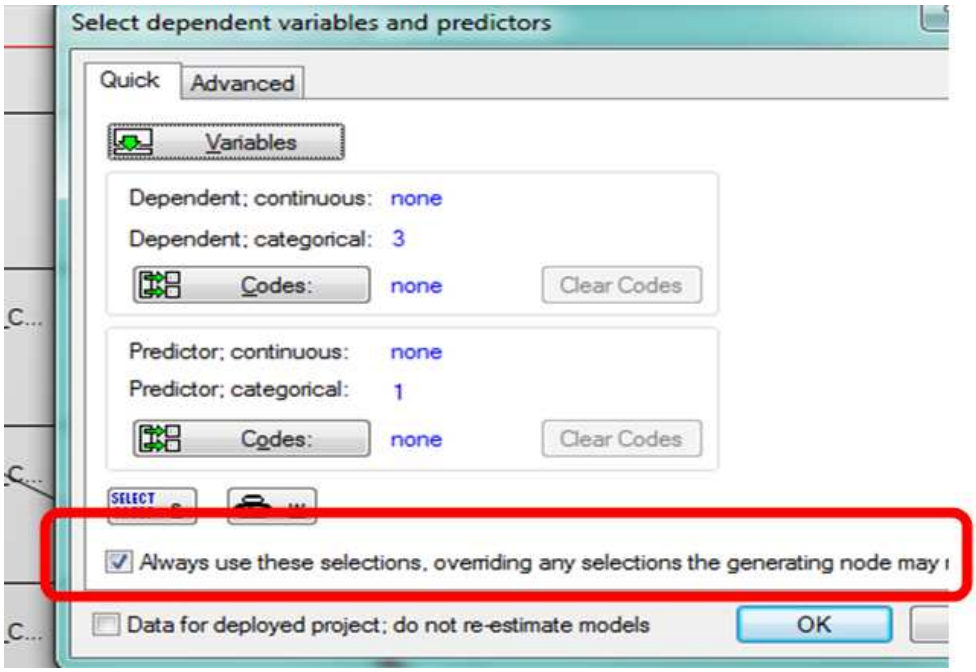
Po potvrzení je uzel výstupu této analýzy označen červeně:

Výsledky v uzlu nejsou aktuální (uzly s červenou kuličkou) - uzel je třeba přepočítat:

Příkazem (Ctrl+F5) **Run Dirty Nodes** přepočteme všechny neaktuální uzly:



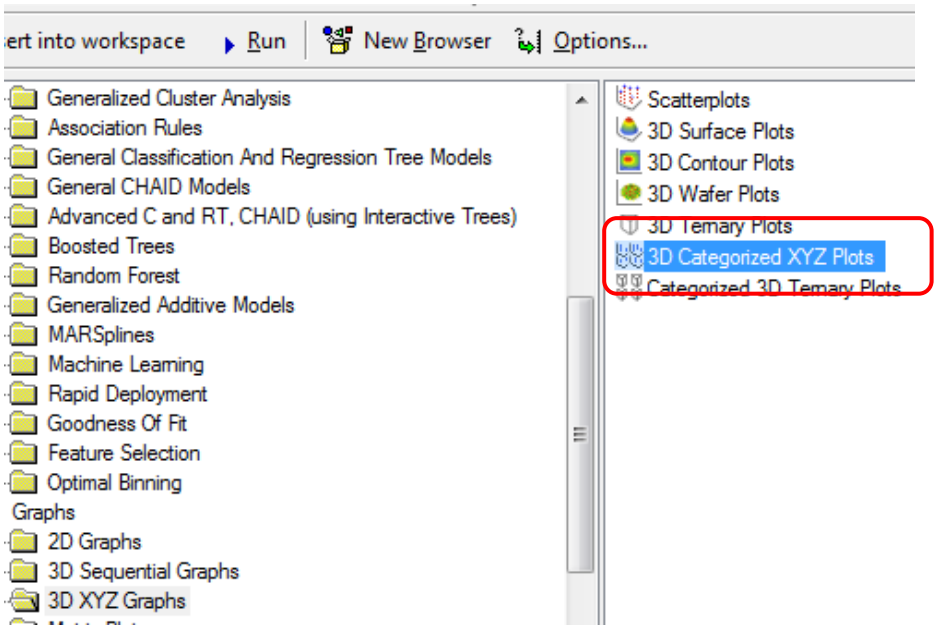
Pokud dáme pouze RUN (F5) – přepočítat úplně všechno – bude v navazujících uzlech vymazáno nastavení proměnných! Při volbě proměnných lze však zaškrtnout volbu: *Always use these selections...*



a model lze přepočítávat celý (F5), ke zrušení nastavení již nedojde.

Druhá volba: *Data for deployed project: do not re-estimate models* slouží například pro deployement, mohou být tedy použity jako vstup do uzlů analýz, ale v těchto uzlech nebude již podle těchto dat přepočítán model, přepočítají se tedy jen předpovědi (viz příklad s testovací a trénovací množinou níže).

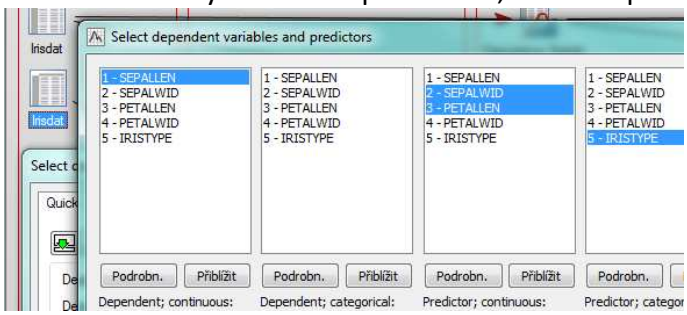
Přidání grafu



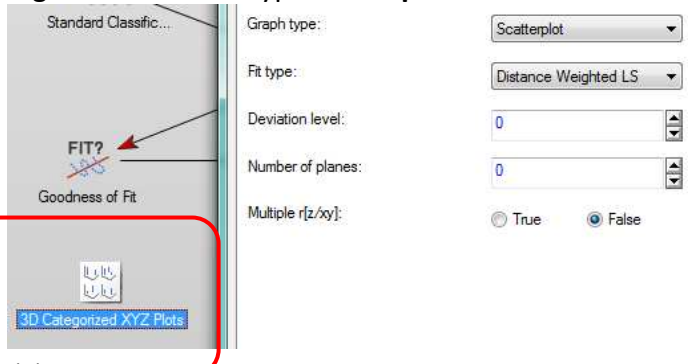
Přes Data Source načteme vstupní tabulku ještě jednou (nemůžeme použít stávající tabulku, protože chceme využívat jiné proměnné, je tedy potřeba vytvořit nový datový zdroj s novým nastavením proměnných – buď můžeme znovu načíst další soubor nebo využijeme uzel *Multiple Copies of Data Sources* – příklad tohoto postupu bude níže)

<u>Nastavíme 3 souřadnice:</u>

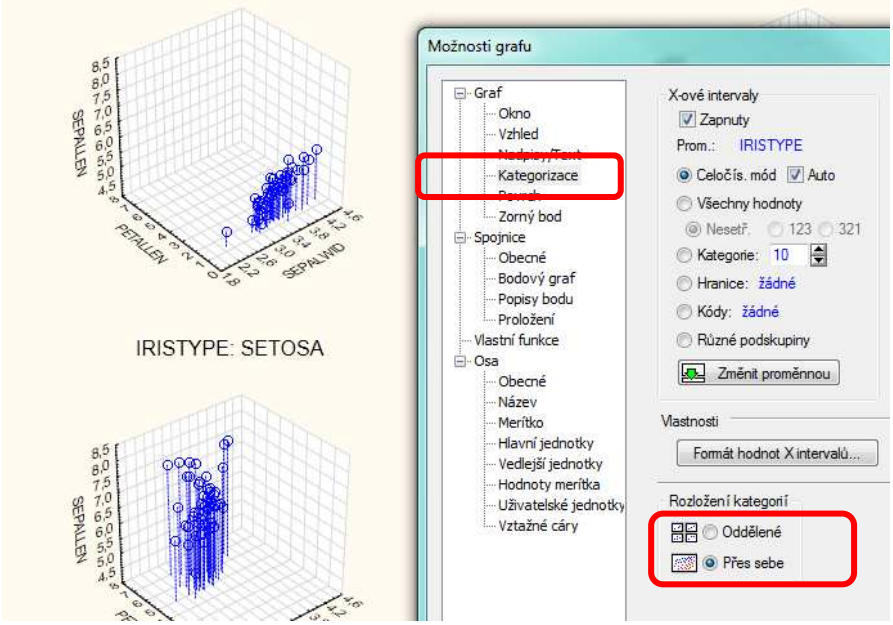
2 vodorovné osy- nezávislé proměnné, barvíme podle typu kosatce…



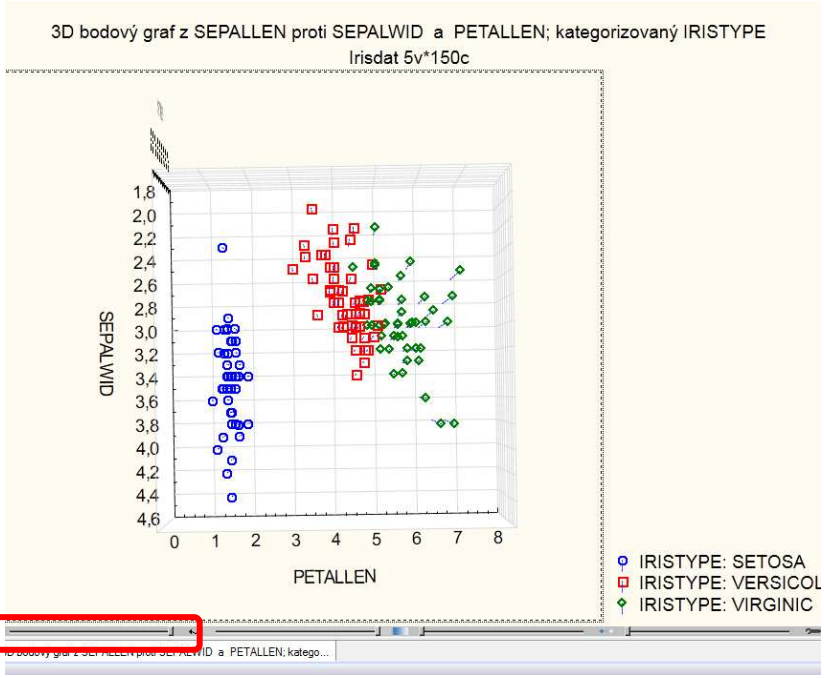
V grafu nastavíme typ: **Scatterplot**



Spustíme uzel a zobrazíme výstup – v grafu vyvoláme dialog **Možnosti grafu** a v záložce Kategorizace volíme **Přes sebe**:

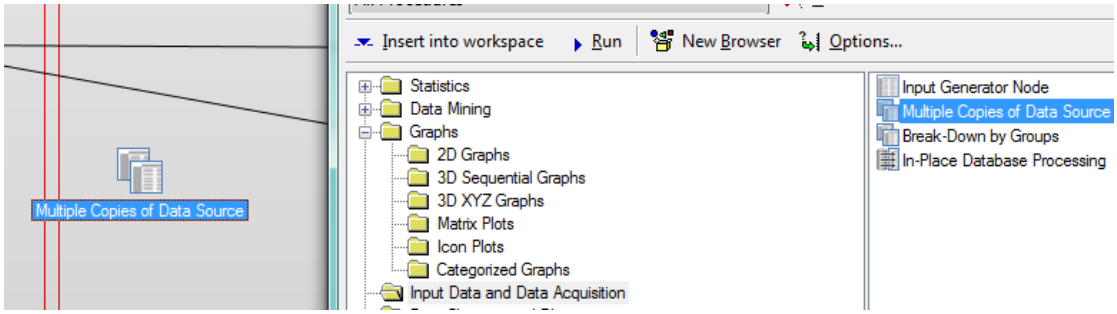


Výsledné rozložení:

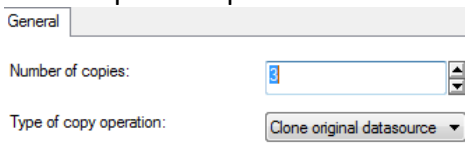


Zkopírování datového souboru

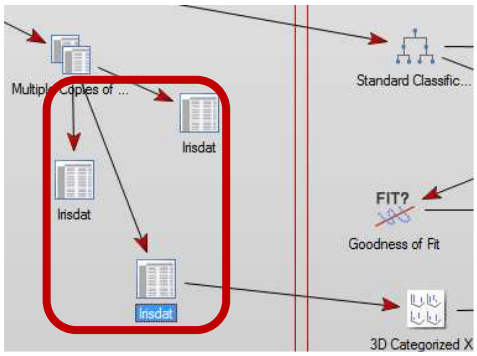
V **Node Browser:**



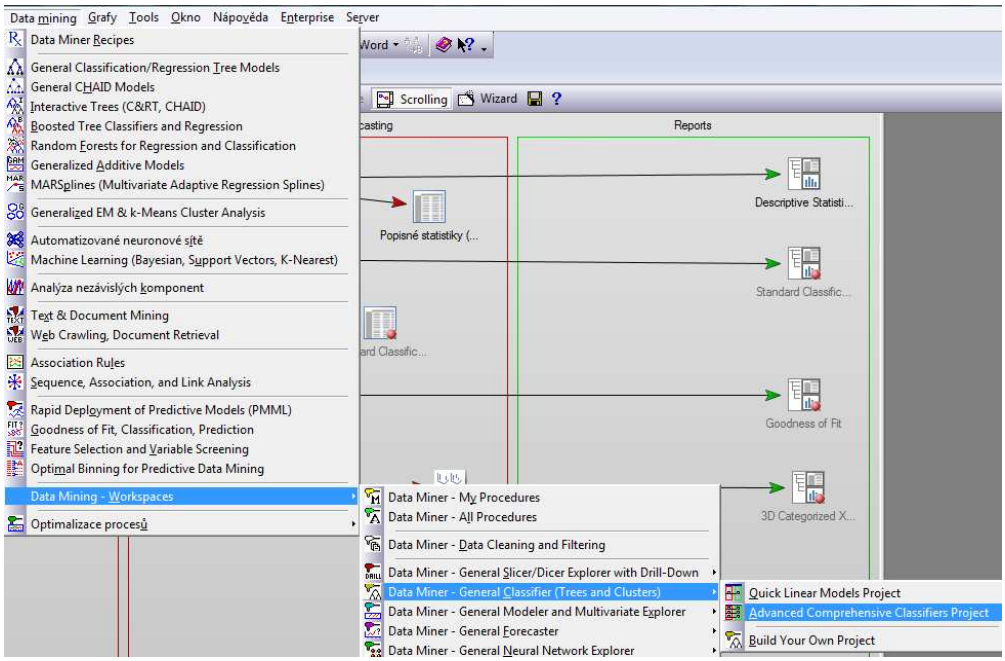
Zvolíme počet kopií:



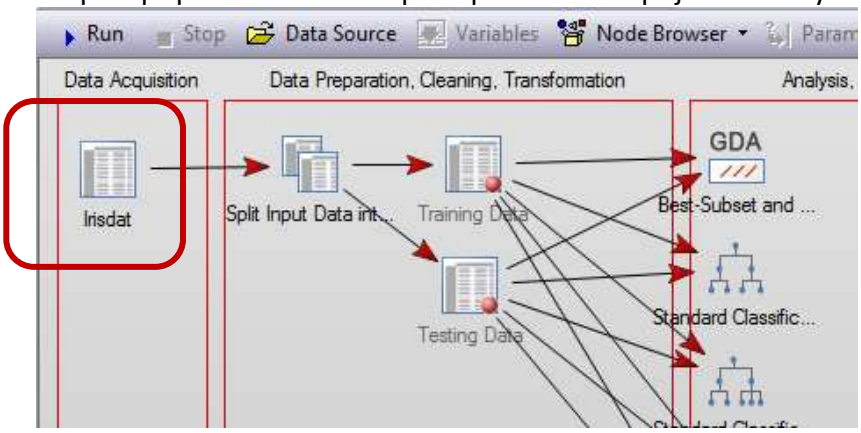
Spustím tento uzel a dostanu 3 virtuální kopie, které mohu dále využívat:



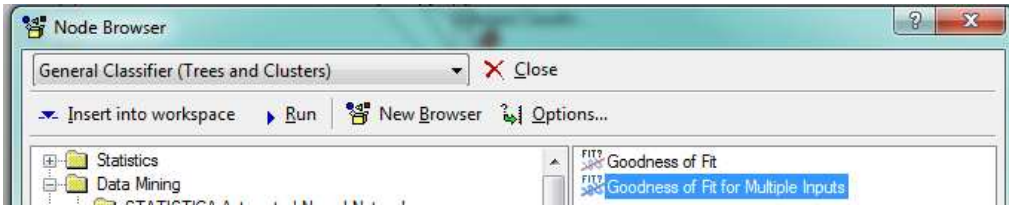
Předpřipravené analýzy pro klasifikaci:



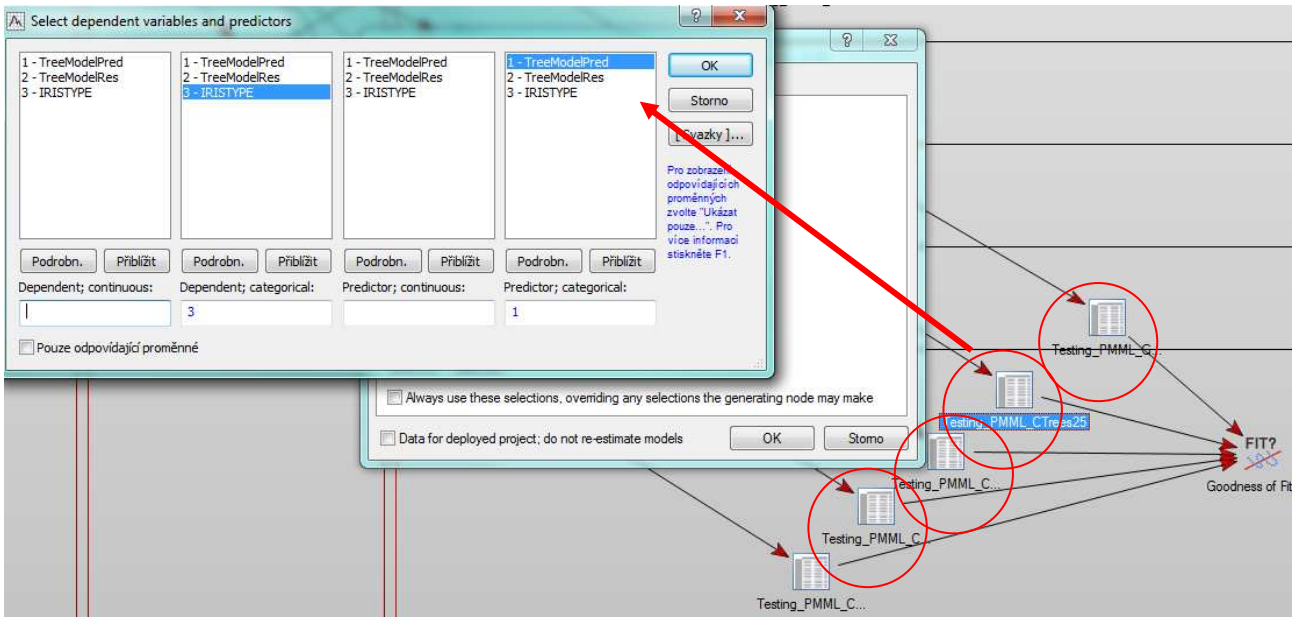
Do předpřipraveného workspace přidáme a napojíme datový soubor:



Spustíme analýzu – RUN:
Pro hodnocení všech modelů vybereme **Goodness of Fit Multiple**:

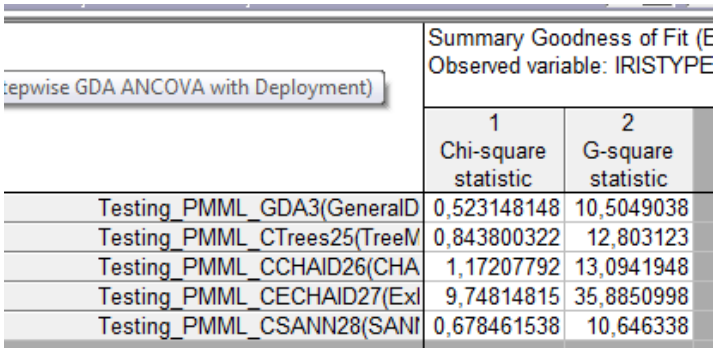


Výsledky jednotlivých klasifikátorů (Testing sample) připojíme na uzel **Goodness of Fit** a na každém uzlu Testing vybereme proměnné:



Na uzlu **Goodness of Fit** nastavíme typ vstupu:

a kombinací Ctrl+F5 přepočteme a zobrazíme tabulku, která porovnává jednotlivé klasifikátory:



	Summary Goodness of Fit (E Observed variable: IRISTYPE	
tepwise GDA ANCOVA with Deployment)	1 Chi-square statistic	2 G-square statistic
Testing_PMML_GDA3(GeneralD	0,523148148	10,5049038
Testing_PMML_CTrees25(TreeM	0,843800322	12,803123
Testing_PMML_CCHAID26(CHA	1,17207792	13,0941948
Testing_PMML_CECHAID27(Exl	9,74814815	35,8850998
Testing_PMML_CSANN28(SANI	0,678461538	10,646338

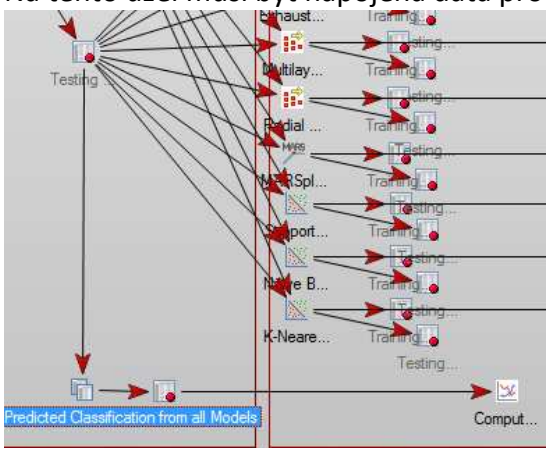
Speciální analytické uzly

Kromě klasických uzlů obsahujících analytické metody, metody pro práci s daty, existují i uzly, které byly navrženy speciálně proto, aby uživatelům Data Miner Workspaces ulehčily práci. Následuje výběr nejdůležitějších uzlů tohoto typu:

Multiple copies of Data Source – vytváří kopii datového zdroje tak, aby ho mohlo být využito v různých analýzách s různým nastavením proměnných. Tento uzel ale nevytváří fyzickou kopii souboru a dochází k úspoře místa oproti vložení mnoha stejných datových souborů

Split Input Data Into Training and Testing Samples – rozdělí data na testovací a trénovací množinu, buď náhodně podle daného poměru nebo využije proměnnou *Learning/testing indicator* v záložce *Advanced* datového zdroje.

Compute Best prediction From All models – Tento uzel automaticky vyhledá v projektu všechny analytické uzly, které slouží k predikci, jejich výstupy předepsaným způsobem zkombinuje a vytvoří predikce, které bývají často přesnější než predikce kteréhokoli z modelů. Na tento uzel musí být napojena data pro testování, jako například na dalším příkladě:



Compute Overlaid Lift Charts From All Models - Pokud potřebujete vytvořit grafy liftu pro všechny modely, použijte tento uzel. Poskytne vám jednoduché vizuální vodítko pro rozhodnutí, který z modelů přinese nejvyšší užitek