TECHNIKY DOLOVÁNÍ DAT S VYUŽITÍM SAS ENTERPRISE MINER

Milena Macháčová
VŠB-TU Ostrava, milena.machacova@vsb.cz

ABSTRACT:
Rostoucí potřeba kvalitních informací pro rozhodování na vrcholové úrovni vede k rozvíjení oblasti zvané Business Intelligence, v jejímž rámci jsou vyvíjeny specializované techniky a nástroje pro kvalifikované analýzy dat, umožňující nalézat vztahy mezi nimi, které nejsou vždy na první pohled zcela patrné. Kromě tradičních statistických metod se zde tedy zvláště uplatňují dataminingové algoritmy a technologie OLAP.

KLÍČOVÁ SLOVA:
Dobývání znalostí z databází, SAS Enterprise Miner, metodologie SEMMA

1 ÚVOD

Vysoký stupeň úrovně informačních a komunikačních technologií umožňuje v současné době uchovávat velké objemy provozních dat, která jsou zpravidla skladována v rozsáhlých databázích se značně složitou strukturou. Síla těchto databází, pro něž se vžil název datové sklady, tkví v tom, že jsou v nich zaznamenaná data mapující časový úsek v řádu desítek let, to znamená, že obsahují takzvaná historická data.

Porozumění obsahu těchto dat, získání relevantních informací či dokonce znalostí je složitý proces, vyžadující značnou trpělivost a úsilí. Zhruba začátkem 90. let minulého století se v anglicky mluvících zemích začalo hovořit o Knowledge Discovery in Databases (KDD) nebo někdy též přímo o data miningu, v českém prostředí se později ustálil název Dobývání znalostí z databází. Jeho smyslem je získat co nejvíce nových, na první pohled skrytých, vztahů mezi daty, užitečných a srozumitelných pro jejich uživatele. Úlohy KDD se dnes realizují v mnoha aplikačních oblastech, především však v bankovnictví, pojišťovnictví, velkých výrobních podnicích a obchodních společnostech, telekomunikačních službách, ale také ve státní správě, vědě a výzkumu apod.

Dobývání znalostí z databází je oblastí, která je v současnosti podporována celou řadou softwarových produktů a metodik a to jak na komerční, tak i akademické bázi.

2 HLAVNÍ RYSY SAS ENTERPRISE MINER

Profesionální programový prostředek SAS Enterprise Miner je jedním z modulů rozsáhlého, původně statistického softwaru firmy SAS, který v sobě spojuje jak statistické metody, tak i postupy založené na neuronových sítích, rozhodovacích stromech, analýze asociací, či genetických algoritmech. Nabízí příjemné uživatelské prostředí, poučenější uživatelé mají možnost modifikovat implementované procedury eventuelně přímo použít vnitřní programovací jazyk.

Firma SAS vyvinula pro realizaci procesu dobývání znalostí vlastní metodologii SEMMA, která se skládá z několika základních kroků:

- **S**AMPLE – **příprava vstupních dat.**

  Jedná se především o výběr vzorků z rozsáhlých datových souborů, načtení a transformace dat pocházejících z různých zdrojů, náhodný výběr dat. Kompletní vstupní datový soubor je na závěr rozdělen na trénovací, validační a testovací množinu dat.

- **E**XPLORE – **průzkumná analýza dat.**

  V tomto kroku se provádějí prvotní analýzy datového souboru. Data podléhají statistickým šetřením, jejichž výsledky jsou graficky vizualizovány, dále jsou identifikovány důležité proměnné a je prováděna asociační analýza.

- **M**ODIFY– **příprava dat pro analýzu**

  Tento krok spočívá v transformaci dat, identifikaci a ošetření odlehlých a vlivných pozorování, imputaci chybějících hodnot či ve vytvoření dalších, dodatečných proměnných. V rámci přípravy dat pro analýzy je často prováděna jejich segmentace pomocí shlukové analýzy, dále může být použita analýza časových řad a nezřídka probíhá v tomto kroku i předběžný výběr prediktorů.

- **M**ODEL – **výběr a odhad modelu.**

  Pomocí modelů probíhá vlastní analýza dat. Jako nástrojů se zde užívá neuronových sítí, statistických technik, rozhodovacích stromů nebo genetických algoritmů.

- **A**SSESS – **interpretace a vyhodnocení výsledků.**

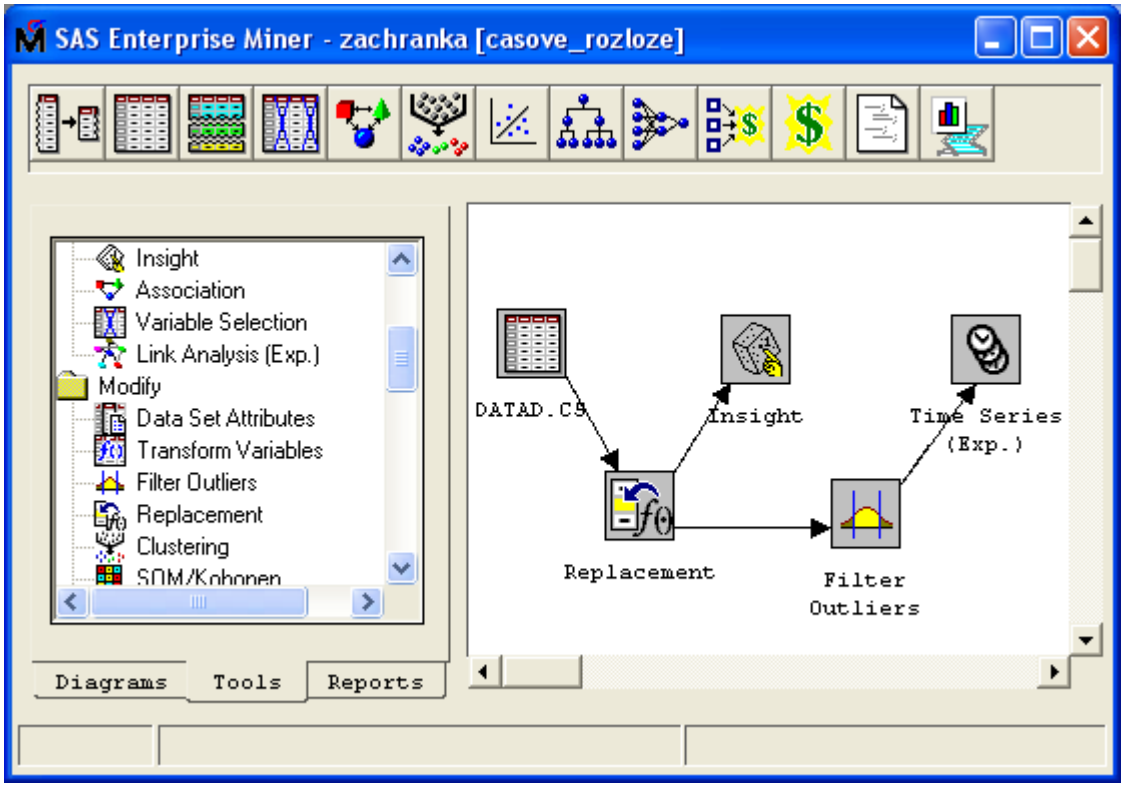
  Finální fáze dataminingového algoritmu srovnává výsledky získané jednotlivými typy modelů a předkládá je uživateli ve srozumitelné podobě zpravidla formou různých typů grafů a přehledných reportů s možností automatického generování ve tvaru webových stránek.

Jednotlivé kroky metodologie SEMMA se realizují v grafické podobě pomocí tzv. **procesních diagramů PFD** (Process Flow Diagram), které sestavuje uživatel vkládáním příslušných ikonek na pracovní plochu. Každá z ikonek reprezentuje určitý krok analýzy, spojnice definují jejich postupovou návaznost.

Na obrázku č.1 vidíme grafické prostředí pro zadávání PFD. V levé části obrazovky jsou zobrazeny kroky metodologie SEMMA spolu s úplným výčtem ikon reprezentujícím daný krok procesu. Nejpoužívanější z nich jsou pro uživatelovo pohodlí ještě zvláště vypíchnuty na horní liště.

Uživatel může, ale nemusí, všechny výše vyjmenované kroky včlenit do svých analýz a naopak, pokud je to nezbytné, je možno jeden nebo více kroků opakovat tak dlouho, dokud není spokojen s výsledky. Spuštěním konkrétního nódu se automaticky spustí všichni jeho předchůdci.

V pravé části obrazovky je sestaven PFD, který v tomto případě umožňuje provádět průzkumnou analýzu prvotních, syrových dat a podle jejích předběžných výsledků data dále upravovat. Skládá se z ikon pro popis vstupního souboru dat **Input Data Source**, náhradu chybějících hodnot **Replacement**, interaktivní analýzu dat **Insight**, filtraci dat **Filter Outliers** a analýzu časových řad **Time Series**.



Obr.1 Uživatelské prostředí SAS Enterprise Miner

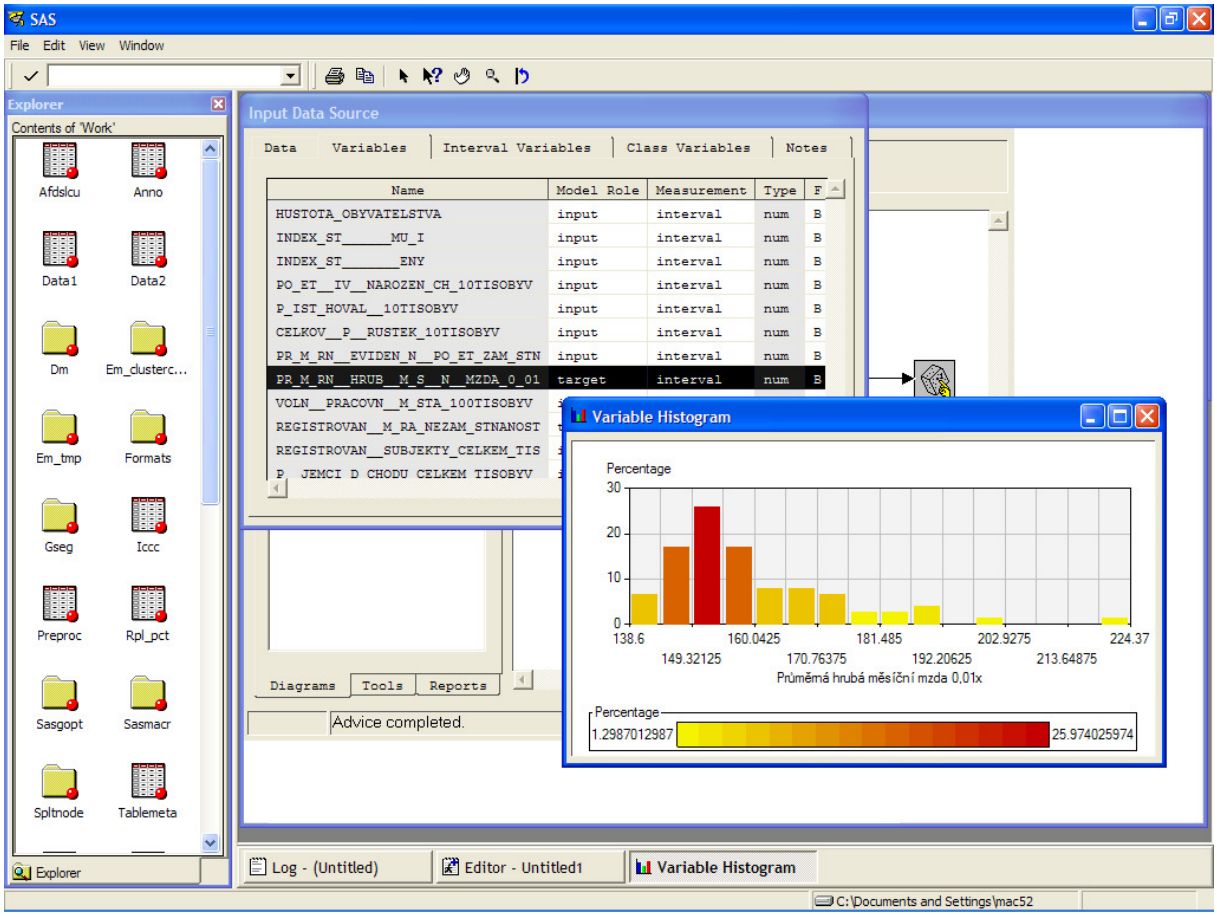
Příprava vstupních dat je z časového hlediska nejnáročnějším, ale současně nejdůležitějším a také nejobtížnějším krokem předcházejícím vlastní tvorbu modelu. Správné pochopení vlastností dat může rozhodnout o míře úspěšnosti celé dataminingové úlohy.

Input Data Source (na obrázku 1 reprezentovaný souborem DATAD.CS) čte zdrojová data a definuje jejich atributy pro následné zpracování dolovacími algoritmy. Současně pro každou proměnnou automaticky vytváří metadata, kde nastavuje inicializační hodnoty a určuje roli proměnné v modelu. Pokud uživatel není spokojen s tímto automatickým nastavením parametrů, má možnost jejich hodnoty změnit. Každá proměnná však musí mít bezpodmínečně nastaveno své postavení v rámci budoucího modelu. V nódu je zobrazena celková statistika týkající se intervalových a kategoriálních proměnných. Má-li zájem, může uživatel získat i vizuální představu o jednotlivých proměnných. Obrázek č.2 ukazuje část karty zachycující základní charakteristiky proměnných spolu s vizualizací jedné z nich.

Přestože současné počítače disponují velkou paměťovou kapacitou, ukazuje se, že pro urychlení zpracování je vhodné využívat pouze určitý reprezentativní vzorek dat, který by měl co nejlépe vystihovat vlastnosti celého datového souboru. V SAS Enterprise Mineru zastupuje tuto možnost ikona **Sampling**.

Vybranou datovou množinu je třeba dále zkontrolovat, zda neobsahuje **chybné hodnoty, hodnoty mimo povolený rozsah** a **chybějící hodnoty**.

Chybějící hodnoty se vyskytují téměř v každé sadě dat. Řada softwarových nástrojů záznamy s chybějícími hodnotami ignoruje a tak se dopouští zkreslení, protože i skutečnost, že hodnota chybí, může mít prediktivní vlastnosti, takže tuto informaci je nutné zachytit. Imputace chybějících hodnot se provádí pomocí ikony **Replacement**, která nabízí několik způsobů náhrady. Chybějící hodnotu můžeme nahradit některou z existujících hodnot atributu a to buď nejčetnější hodnotou, proporcionálním podílem všech hodnot nebo libovolnou hodnotou.



Obr.2 Karta vstupních proměnných v nódu Input Data Source

Jinou možností je nahradit chybějící hodnotu novou hodnotou „nevím“. Pro doplnění chybějící hodnoty se může použít i model. Atribut s chybějícími hodnotami se považuje za cílový, pro trénování a testování se použijí záznamy se známými hodnotami tohoto atributu. Chybějící hodnoty se doplní na základě výsledků modelu.

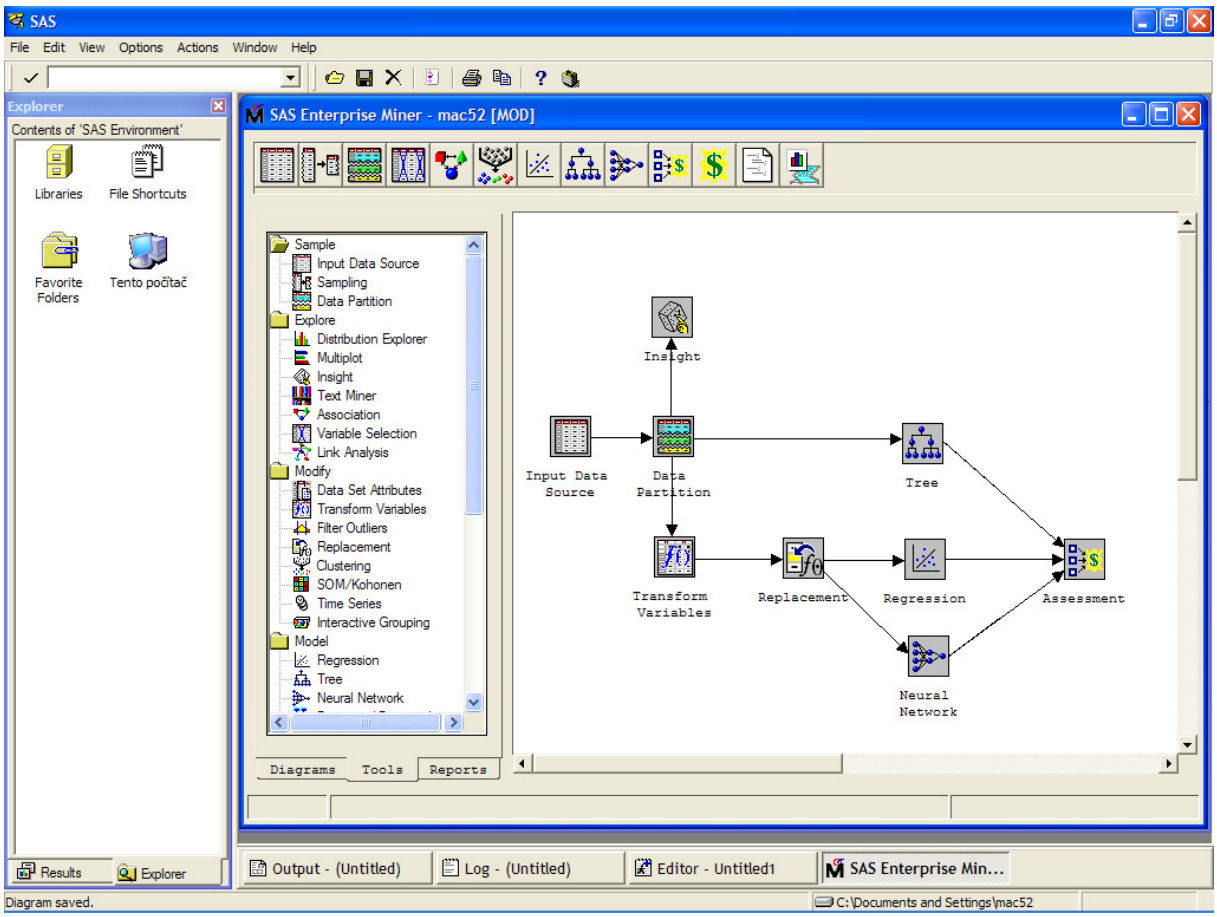
Ikona **Insight** otevírá zvláštní režim práce umožňující interaktivně prozkoumávat a analyzovat data. Za pomocí vestavěných nástrojů pro statistické analýzy dovoluje vytvářet průzkumné vysvětlující modely dat.

V případě, že množství vybraných dat začíná neúměrně narůstat, roste i **potřeba redukovat počet proměnných**. Provádět hloubkovou analýzu s každou proměnnou není efektivní, navíc řada proměnných spolu navzájem koreluje. Dále platí, že ne všechny proměnné musejí být podstatné pro zamýšlenou analýzu. Obecně je tento problém možno vyřešit za pomocí experta, který většinou odhadne, které proměnné jsou pro danou úlohu podstatné, anebo je možno využít automatických metod.

Takovouto automatickou možnost selekce, kdy z existujících proměnných jsou vybírány pouze ty nejdůležitější, které nejlépe přispějí ke klasifikaci záznamů do jednotlivých tříd, může představovat ikona **Filter Outliers**, kdy se ke každé proměnné spočítává charakteristika vyjadřující její vhodnost pro klasifikaci. Obdobně je možno řešit odstranění tzv. **odlehlých hodnot**, tj. případů, kdy se hodnota proměnné vyskytuje jednou nebo daleko od střední hodnoty i od většiny ostatních hodnot této proměnné.

Pro **hodnoty mimo rozsah** je také možno vytvořit překrývací pravidlo tj. takové pravidlo, které umožní nahrazení extrémní hodnoty jinou hodnotou, která se nachází v akceptovatelných mezích.

Na obrázku č.3 můžeme vidět diagram, který pokrývá celý životní cyklus metodologie SEMMA:



Obr.3 Životní cyklus metodologie SEMMA

Kromě již známých ikon se zde objevují další nódy.

Data Partition nód umožňuje uživateli rozdělit původní vstupní soubor dat na trénovaní, testovací a validační skupinu. Trénovaní soubor je používán pro předběžný průzkum dat, validační soubor slouží pro monitorování a vyladění odhadu vah jednotlivých proměnných v modelu a také pro závěrečné vyhodnocení. Testovací soubor dat umožňuje verifikovat a zhodnotit výsledky postaveného modelu. Ke své práci nód využívá jednoduchého náhodného vzorkování, stratifikovaného náhodného vzorkování eventuelně mohou být data rozdělena do skupin přímo uživatelem.

Možnost dalšího předzpracování dat nabízí nód **Transform Variables**. Kromě jiného je zde možno provádět další redukce, kdy z existujících proměnných je vytvořen menší počet proměnných nových. Uplatňuje se zde například faktorová analýza, analýza hlavních komponent a jiné. Tyto metody předpokládají použití výhradně numerických proměnných.

Jádrem každé dataminingové úlohy je výstavba vlastního modelu, který může mít charakter deskriptivní nebo prediktivní v závislost na povaze řešeného problému. Na obrázku č.3 jsou znázorněny tři typy modelů.

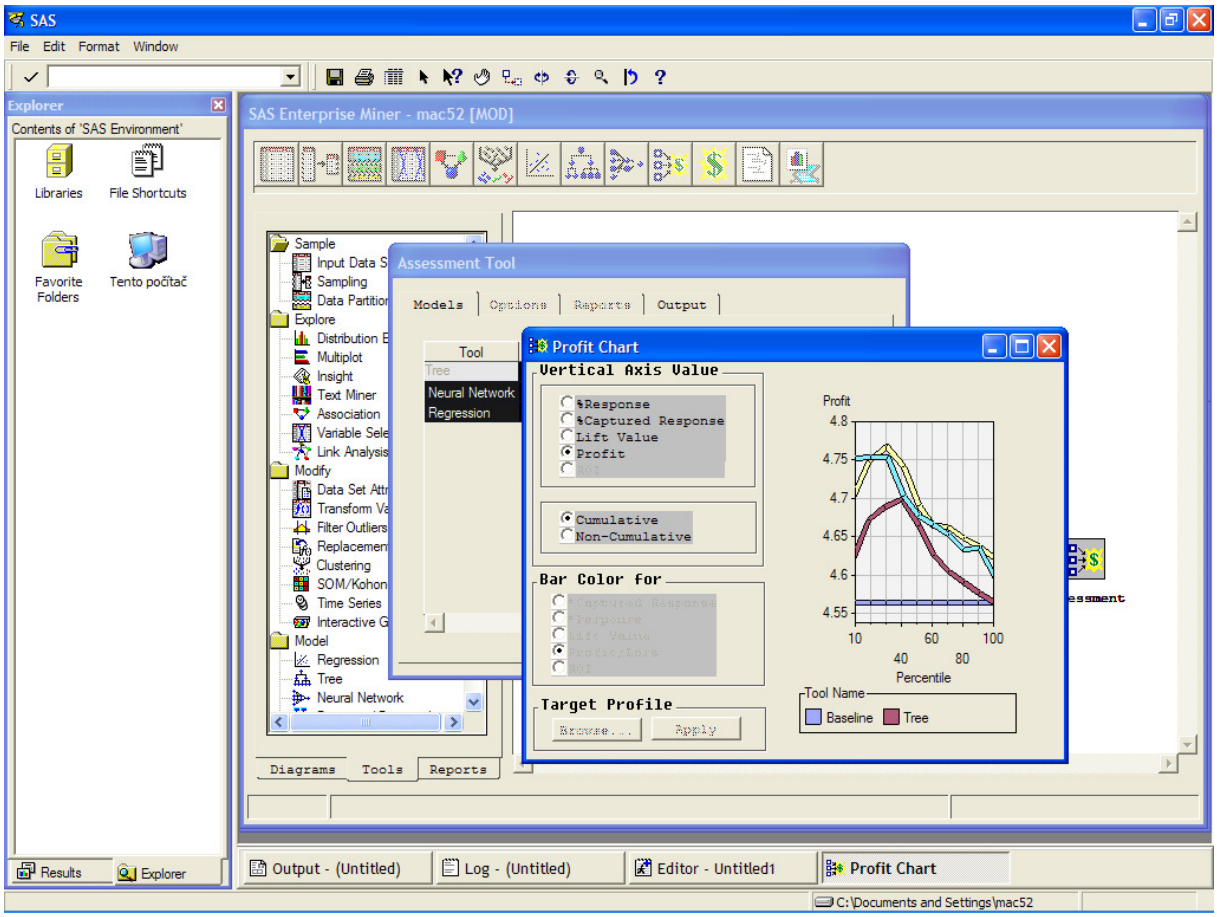
První z nich používá technologii speciálního rozhodovacího stromu vyvinutou firmou SAS a spojující v sobě vlastnosti CHAID, CART a C4.5 algoritmů. **Tree** nód umožňuje uživateli realizovat multidimenzionální rozdělení dat databáze na základě zadaných kriterií. Podporuje jak automatické, tak interaktivní rozdělení. Pokud modelování probíhá v automatickém módu,

vstupní proměnné jsou automaticky zařazeny na určitou hierarchickou úroveň stromové struktury. Uvedená klasifikace proměnných může být použita pro jejich následný výběr za účelem dalšího modelování.

Druhý typ modelu využívá klasickou statistickou metodu regrese. **Regression** nód dovoluje uživateli použít k modelování dat jak lineární, tak logistickou regresi. Jednotlivé cílové proměnné mohou být ordinální, spojité nebo binární, jako vstupy je možno mít jak spojité tak diskrétní proměnné. Nód podporuje vícenásobnou regresi, dopředný i zpětný mechanismus.

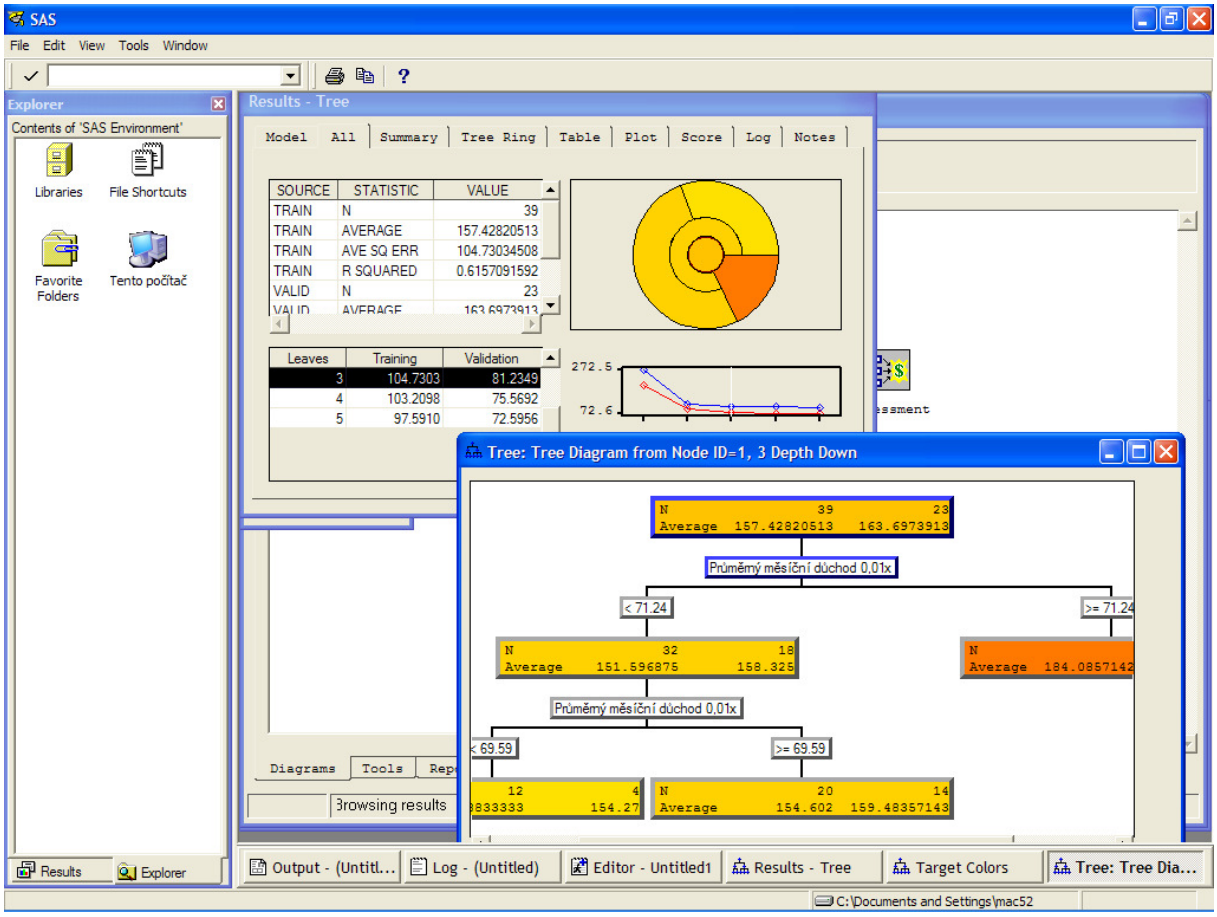
Poslední model je postaven na základě neuronových sítí a reprezentuje ho ikona **Neural Network.** Tento nód dovoluje uživateli konstruovat modely pro predikci budoucích hodnot sledovaných proměnných. Má možnost výběru z devíti typů základních architektur neuronových sítí, které je možno za pomocí parametrů dále modifikovat a získat tak další možné varianty původních struktur.

Závěrečnou fázi metodologie SEMMA představuje v diagramu ikona **Assessment,** pomocí které jsou vytvářeny podmínky pro vzájemné porovnání výsledků získaných modelováním v jednotlivých typech modelů. Očekávané a aktuálně vypočtené hodnoty proměnných jsou zobrazovány formou různých typů grafů, jeden z nich je možno vidět na obrázku č.4.



Obr.4 Porovnání výstupů vytvořených modelů

S každým z vytvořených modelů je možno samostatně experimentovat a získat tak další velkou sadu nezávislých výstupů. Jak vypadají výsledky modelování pomocí modelu sestaveného na základě rozhodovacího stromu ukazuje obrázek č.5.



Obr.5 Výstupy modelu Rozhodovací strom

3 ZÁVĚR

Dobývání znalostí z databází, potažmo data mining, představuje v současné době jednu ze sofistikovaných možností, jak lépe zhodnotit data uložená v rozsáhlých, složitě strukturovaných databázích a zhusta zachycující časový úsek několika desítek let. V současné době se tento proces již uplatňuje v mnoha aplikačních oblastech a je podporován celou řadou profesionálních softwarových prostředků.

LITERATURA

[1] Berka, P.: Dobývání znalostí z databází, Praha, ACADEMIA, 2003, ISBN 80-200-1062-9, s.366

[2] Cook book Data Mining Using Enterprise Miner software: A Case Study Approach published by company SAS Institute , February 2000

[3] Cook book Getting Started with Enterprise Miner Software, Version 4.0 published by company SAS Institute , July 2000

[4] Macháčová, M.: Využití inteligentních nástrojů pro analýzu technologických dat, Sborník vědeckých prací Vysoké školy báňské – Technické univerzity Ostrava, řada hornicko-geologická, ročník XLIX, 2/2003, Ostrava, ISBN 80-248-0562-6, ISSN 0474-8476, s.43-52