

IBM SPSS MODELER

Objevte vzory v historických datech, které budou sloužit k predikci budoucích událostí, dělejte lepší rozhodnutí a dosáhněte lepších výsledků.

IBM SPSS Modeler je komplexní platforma pro prediktivní analýzu, která dává do rukou jednotlivce, pracovních skupin nebo celé organizace nástroje pro nalezení skrytých vzorů v datech. S nástrojem SPSS Modeler nezáleží na velikosti vašeho řešení. Umožňuje pracovat na desktopovém počítači i používat řešení založené na rozsáhlém množství pokročilých algoritmů zpracovávaných na serverovém počítači. Použití těchto technik v rozhodování může vést k rychlému růstu ROI, k aktivnímu a opakovanému snižování nákladů při zvýšení produktivity.

Intuitivní ovládání

Software IBM SPSS Modeler využívá ověřenou a celosvětově uznávanou metodiku **CRISP-DM (Cross Industry Standard Process for Data Mining)**. Díky tomu Vám ponechává kontrolu nad daty a nad celým procesem, ve všech jeho fázích: porozumění úloze a jejímu cíli, porozumění datům, příprava dat, modelování, vyhodnocení, implementace.

Uživatelsky orientované, **intuitivní grafické rozhraní** umožňuje snadnou práci s programem a vytváření modelů i uživatelům bez hlubších technických znalostí. Analytici potom využijí pokročilejší funkce programu.

Ovládání softwaru je lehké, uživatel se může soustředit na konkrétní problém, nikoliv na jeho technické řešení.

Hlavní přínosy:

- datová nezávislost v přístupu k datovým skladům, databázím, Hadoop distribucím či ke klasickým textovým či excelovským souborům,
- sofistikované analytické prostředky pro hledání skrytých vzorů v datech,
- zpřístupnění statistických a dataminingových procedur širšímu spektru uživatelů díky uživatelsky jednoduché a intuitivní práci formou vizuálního programování,
- využitelnost dataminingové platformy pro řešení širokého okruhu obchodních úloh, od jednoduchých popisných statistik až po komplexní analytické modely,
- maximální využití stávající IT infrastruktury pro dosažení maximální rychlosti zpracování dat díky delegaci zpracování přímo do databází,
- rozšiřitelnost dataminingové platformy s možností její integrace s dalšími řešeními IBM,
- jednoduchá implementace a integrace řešení do stávající IT infrastruktury.

NOVÉ FUNKCIONALITY

- Podpora Azure Databricks jako datového zdroje s plnou podporou SQL pushback a integrací funkcí Modeleru.
- Nové Spark uzly – Logistic Regression a Bisecting K-Means Clustering využívající nativní algoritmy Spark.
- Podpora Intellisense a dynamického formátování pro SQL dotazy dostupné v databázovém zdrojovém uzlu.
- OLAP uzel umožňuje provádět pokročilou analýzu s okenními funkcemi pro řazení, číslování řádků a agregační výpočty bez nutnosti sbalování řádků.
- OFFSET uzel umožňuje načítat dřívější nebo pozdější hodnoty v rámci datového souboru. Můžete seskupovat, třídit a přizpůsobovat ovládání rozsahu.
- Podpora autentizace Teradata Wallet.
- Vylepšená integrace IBM SPSS Modeleru v CPLEX.
- Podpora čínského jazyka v Text Analytics.
- Optimalizace výkonu ve Spark uzlech.
- Podpora Apache Spark 3.5.4.
- Aktualizované kompilátory C++ pro zvýšení výkonu, stability a kompatibility s moderními platformami.
- Rozšířená podpora nových operačních systémů a databází.
- Integrace s IBM SPSS Statistics 31.
- Vylepšené zabezpečení opravou kritických zranitelností a aktualizací jádra a komponent s otevřeným zdrojovým kódem.

Bisecting-K-Means-AS

Fields | **Build Options** | Annotations

Regular

Model Name: Auto Custom

Number of Clusters: 4

Distance Measure

Distance Measure: Euclidean Cosine

Minimum Divisible Cluster Size: 1.0

Advanced

Advanced Settings

Max Iteration: 20

Tolerance: 1.0E-4

Set Random Seed

Generate

Random Seed: 12345678

OK Run Cancel Apply Reset

LogisticRegression-AS

Fields | **Build Options** | Annotations

Use predefined roles Use custom field assignments

Fields:

Sort: None

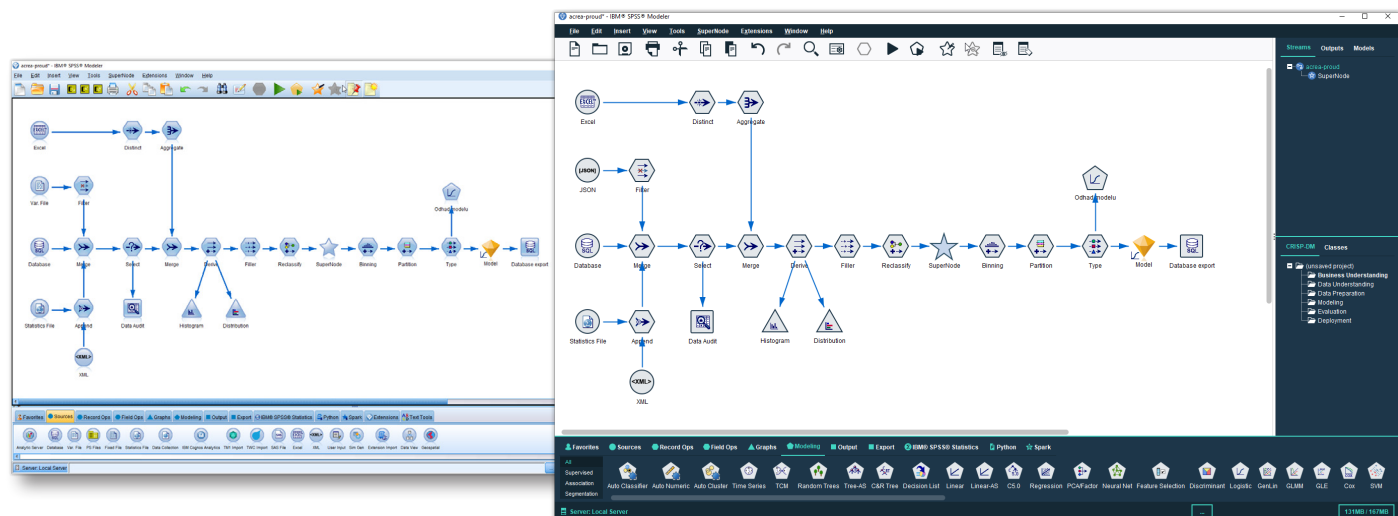
Target Field:

Input Fields:

All

OK Run Cancel Apply Reset

Nový, moderní a atraktivní vzhled. Uživatel však má možnost používat původní vzhled, na který byl zvyklý. Přepnutí je možné pomocí menu **Tools\User Options\Display**.



Přístup k různým typům dat

Z grafického rozhraní lze jednoduše přistupovat ke **všem typům databází** (IBM InfoSphere®, Microsoft SQL Server, Oracle a IBM Netezza), **tabulek, datových souborů** (jako jsou soubory IBM SPSS Statistics, SAS a Excel), **textových souborů, zdrojů z Web 2.0** (například RSS), IBM SPSS Data Collection, IBM Cognos® Business Intelligence, ze systémů s IBM Classic Federation server a zDB2 for z/OS.

Příprava dat a datové manipulace

Příprava dat je důležitý, ale časově náročný krok v dataminingovém procesu. Nástroj SPSS Modeler nabízí řadu způsobů jak manipulovat a připravovat data pro analýzy, ať už se jedná o přípravu záznamů v řádcích nebo přípravu proměnných ve sloupcích. Mezi metody používané k přípravě dat pro konkrétní typ analýzy se používají:

- **Operace se záznamy**

Uzly Select, Sample a Distinct umožňují vybrat si specifické řádky v tabulce. Uzly Merge a Append spojují dvě a více tabulek. Uzly Aggregate a RFM Aggregate vytvářejí agregace v podobě jednoho řádku. Balance uzel upravuje poměr záznamů v nevyvážené datové matici. Uzel Sort řadí záznamy na základě vybrané proměnné. Uzel Space Time Box transformuje geo-prostorová data a časové řady záznamů.

- **Operace s proměnnými**

Uzlem Type specifikujete metadata a vlastnosti datové matice. Uzel Filter odstraňuje záznamy z datové matice. Uzel Derive vytváří nové proměnné, zatímco uzel Filler přepisuje existující proměnné. Restrukturalizaci datové matice lze provádět za pomoci uzlů Set to Flag, Restructure nebo Transpose. K přeskupení kategorií v rámci jedné proměnné se používají uzly Reclassify nebo Binning. Před modelováním je vhodné rozdělit datovou matici na trénovací a testovací množinu prostřednictvím uzlu Partition. Uzly History a Time Intervals umožňují vytvářet nové proměnné při práci se sekvenčními daty (např. časové řady). Uzel Field Reorder upravuje pořadí proměnných v datové matici.

Rozsáhlá nabídka modelovacích technik

IBM SPSS Modeler disponuje širokou škálou pokročilých dataminingových modelovacích nástrojů pro řešení úloh, na které může analytik narazit.

- **Klasifikační algoritmy**

Vytváří předpovědi budoucího chování na základě historických dat pomocí modelovacích technik, jako jsou klasifikační stromy, neuronové sítě, logistická regrese, podpůrné vektory, Coxova regrese, zobecněné smíšené lineární modely (GLMM) a další. Využijete automatické modelování jak pro kategorizované, tak pro spojité cílové proměnné a vytvořte samo-učící se model (SLRM), který můžete opakovaně aktualizovat nebo znovu odhadovat bez nutnosti předělat váš model.

- **Segmentační algoritmy**

Seskupuje zákazníky a hledá neobvyklé vzory pomocí automatického nebo uživatelského seskupování, detekce anomálií a neuronových sítí. Automatické seskupování současně vyzkouší více segmentačních algoritmů a vyberete vhodnou seskupovací metodu.

- **Asociační algoritmy**

Nacházejí typické kombinace jevů a vztahy v posloupnostech událostí (např. nákupů) pomocí **algoritmů Apriori, CARMA** a sekvenčních asociací.

- **Časové řady**

Vytvářejí předpovědi pro jednu nebo více časových řad pomocí statistických modelovacích technik.

- **Rozšíření o jazyk R**

SPSS Modeler umožňuje používat transformace, modely a výstupy v **jazyce R**. Pro kolegy, kteří nechtějí přímo psát zdrojový kód, lze vytvořit vlastní, standardně ovládaný uzel využívající nástroje jazyka R.

- **Monte Carlo simulace**

Umožňuje vložit prvek nejistoty do predikčních modelů. Vstupy do modelu obsahující prvek nejistoty jsou nasimulovány podle historických dat nebo pravděpodobnostního rozdělení. Tento proces může být mnohokrát opakován. Výsledkem je pak rozdělení výsledků, které mohou přispět k zodpovězení komplexních otázek.

Proč IBM SPSS Modeler?

Prostřednictvím nezávislých studií a díky našim zkušenostem z projektů jsme identifikovali následující hlavní kategorie přínosů, které Vám **software IBM SPSS** přinese:

- jednoduchá a rychlá implementace (do ½ dne),
- nízké náklady na zaškolení v důsledku nenáročného ovládnání software,
- rychlé předání znalostí a know how užívání softwaru v rámci organizace,
- zvýšení produktivity jako důsledek práce s jednoduchým grafickým uživatelským prostředím bez nutnosti programování,
- snadné sdílení výsledků práce díky dokumentaci datového procesu přímo v prostředí programu,
- přístup k datům v datových skladech, databázích, Hadoop distribucích či samotných souborech,

- silný nástroj nejen pro tvorbu modelů, ale také pro (často zásadní) přípravu vstupních dat,
- komplexní analýza velkého množství dat v kratším čase s použitím stávajícího IT vybavení, výpočty prováděné v databázi, minimální přesun dat,
- neustálý technologický vývoj v rámci portfolia produktů IBM a pravidelné upgrady.

Specifikace IBM SPSS Modeler Professional

Pochopení dat:

- **komplexní přehled proměnných v souboru včetně kontroly jejich kvality,**
- **široká škála interaktivních grafů,**
- **pavučinový graf pro analýzu vztahů v datech,**
- **interaktivní výběr dat z grafu pro vizualizaci nebo modelování,**
- **přístup k procedurám a grafům z programu IBM SPSS Statistics přímo v programu Modeler.**

Příprava dat:

- přístup k datům z IBM SPSS Collaboration and Deployment Services Repository, Cognos Business Intelligence, Cognos TM1, IBM DB2®, Oracle®, Microsoft SQLServer™, IBM Informix®, IBM Netezza, MySQL (Oracle), Hadoop Distributed File System, datovým zdrojům Teradata, stejně tak jako k databázím zDB2 a IBM Classic Federation Server Support,
- import textových souborů pevné délky nebo s oddělovači, import datových souborů **IBM SPSS Statistics, SAS, IBM SPSS Data Collecton nebo XML,**
- paleta nástrojů pro čištění dat od odebrání či nahrazení chybných údajů až po automatické vkládání chybějících hodnot a zmírnění vlivu odlehlých pozorování a extrémních hodnot,
- automatické ověření kvality dat a jejich příprava k modelování,
- výběr proměnných, přejmenování, odvození nových proměnných, kategorizace, nahrazení hodnot a změna pořadí proměnných,
- výběr případů, náhodné výběry, spojení dat a textových řetězců, třídění, agregace a vážení,
- restrukturalizace dat, rozdělení na tréninkovou a testovací množinu a transpozice,

- funkce pro práci s textovými řetězci: tvorba řetězců, nahrazování znaků, vyhledávání, ořezávání a odebrání mezer,
- **RFM skórování**, agregace transakčních dat pro kompletní RFM analýzu,
- export dat do databází, **IBM Cognos Business Intelligence**, **IBM SPSS Statistics**, **IBM SPSS Data Collection**, **textových dokumentů**, **Excel**, **SAS**, **XML**.

Modelování a ověřování modelů:

- pokročilé data miningové algoritmy pro získání informací z dat,
- automatická klasifikace a seskupování pro rychlé nalezení vhodných modelů,
- interaktivní prohlížeč modelů a přehledné statistické výstupy,
- vizualizace analytických výsledků na geografických mapách,
- grafické zobrazení relativní důležitosti prediktorů pro závislou proměnnou,
- kombinace několika modelů (metamodelování), nebo analýza jednoho modelu pomocí druhého,
- **Component-Level Extension Framework (CLEF)** pro tvorbu vlastních aplikací,
- přístup k nástrojům **jazyka R** včetně hladké implementace procedur v jazyce R do prostředí programu,
- možnost práce v jazyku **Python a Python for Spark**,
- propojení s **IBM SPSS Statistics**,
- simulování dat metodou **Monte Carlo**.

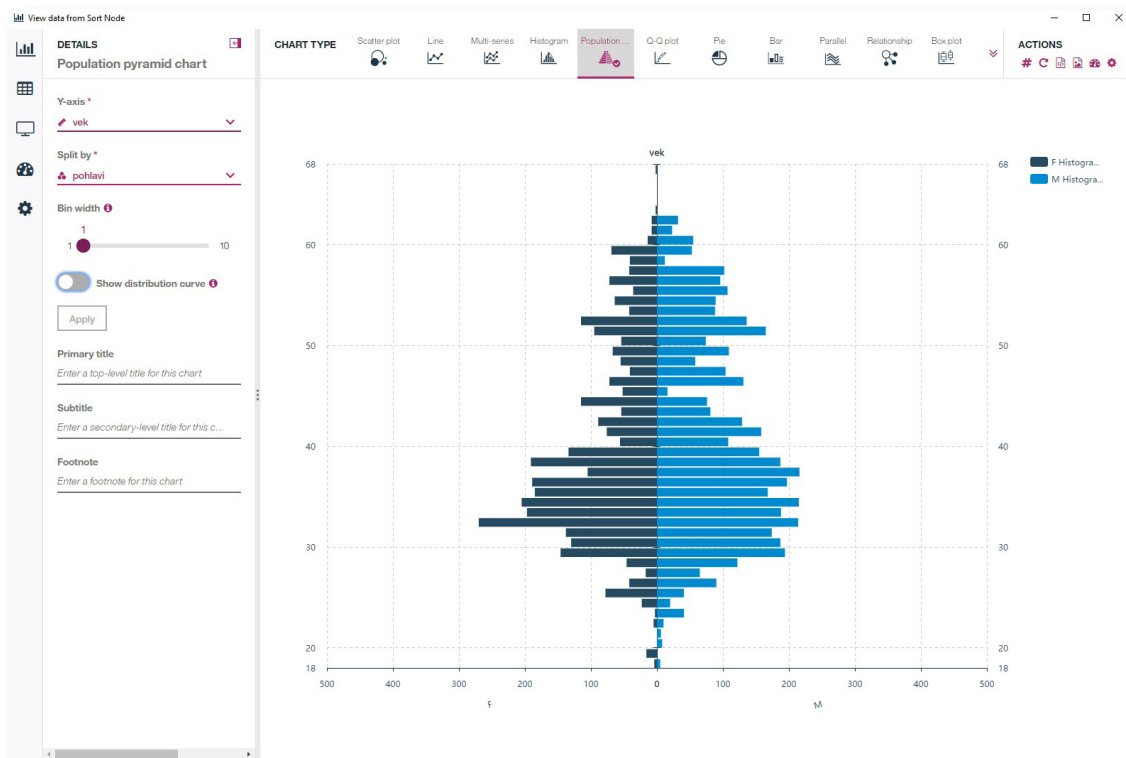
Modelovací algoritmy:

- **C&RT, C5.0, CHAID & QUEST** – rozhodovací a klasifikační stromy s možností interaktivního růstu, náhodné lesy,
- **Decision List** – interaktivní algoritmus pro vytváření pravidel,
- **Kohonenovy sítě**, metody K-Means a Two Step, diskriminační analýza a metoda podpůrných vektorů (SVM) – seskupovací a segmentační algoritmy,
- **Faktorová analýza**, analýza hlavních komponent – algoritmy pro redukci dimenzionality,
- **Lineární regrese**, zobecněná lineární regrese (GLM) a zobecněné lineární smíšené modely (GLMM) – odhady parametrů v lineárních modelech,

- **Logistická regrese** - modelování kategorizovaných proměnných,
- **SLRM** – bayesovský model s postupným učením,
- **Analýza časových řad** – automatické generování a odhady parametrů časových řad, metody Temporal Causal Modelling,
- **Neuronové sítě** – vícevrstvá síť se zpětnou propagací, síť s radiální bazickou funkcí,
- **Podpůrné vektory (SVM)** a lineární podpůrné vektory (LSVM) – pokročilé algoritmy vhodné pro rozsáhlé datové soubory,
- **Bayesovské sítě** – modely založené na podmíněné pravděpodobnosti,
- **Coxova regrese** – odhad času do konkrétní události,
- **Detekce anomálií** – nalezne neobvyklé záznamy pomocí seskupovacích algoritmů,
- **KNN** – klasifikace metodou nejbližších sousedů,
- **Apriori** – oblíbený asociační algoritmus s pokročilými funkcemi pro vyhodnocení výsledků,
- **CARMA** – asociační algoritmus s možností vícenásobných důsledků,
- **Sequence** – nalezení asociací v záznamech uspořádaných podle času,
- **Spatio-Temporal Prediction** – modely pro hledání časových a prostorových vztahů v datech,
- **XGBoost Tree** - učení klasifikačních lesů,
- **XGBoost Linear** - hledání koeficientů lineárního modelu,
- **One-Class SVM** - hledá anomálie pomocí nesupervizovaného seskupování,
- **Gaussian Mixture** - nesupervizovaný klasifikátor,
- **Kernel Density Estimation (KDE)** - pro simulaci a modelování jádrového odhadu hustoty,
- **Hierarchical Density-Based Spatial Clustering (HDBSCAN)** - nesupervizovaný klasifikátor.

Vyžaduje IBM SPSS Modeler Server Professional

- **modelovací algoritmy v databázi IBM InfoSphere:** Apriori, seskupování, rozhodovací stromy, logistická regrese, naivní bayesovské klasifikátory, regresní modely, hledání asociací v sekvencích a časové řady,
- **modelovací algoritmy v databázi IBM Netezza:** bayesovské sítě, naivní bayesovské klasifikátory, rozhodovací a regresní stromy, hierarchické seskupování, seskupování metodou K-Means, zobecněné lineární modely, metoda hlavních komponent a časové řady,
- **modelovací algoritmy pro databázi Microsoft SQL Server:** Apriori, seskupování, rozhodovací stromy, lineární regrese, naivní bayesovské klasifikátory, neuronové sítě, sekvenční seskupování a časové řady,
- **modelovací algoritmy pro databázi Oracle:** adaptivní bayesovské sítě, naivní bayesovské klasifikátory, Apriori, umělá inteligence (AI), rozhodovací stromy, zobecněný lineární model (GLM), metoda K-Means, minimální popisná vzdálenost (MDL), faktorizace pozitivně semidefinitních matic, O-Cluster (ortogonální seskupování), podpůrné vektory (SVM),
- přístup k modelovacím nástrojům přímo v databázích,
- paralelní spuštění proudů a modelů,
- bezpečný přenos citlivých dat mezi klientem a serverem pomocí kódování Secure Sockets Layer (SSL),
- převedení transformací a výběrů do SQL a jejich provedení přímo v databázi přes SQL pushback.



Náhled View Data - populační pyramida věku dle pohlaví.

Závěr

SPSS Modeler je platforma pro prediktivní analýzy vyznačující se širokou škálovatelností. Umožňuje nasazení na desktopovou stanici jednoho zaměstnance až po integraci se systémy v organizaci, která přinese relevantní výsledky analýz jednotlivcům, skupinám nebo celé organizaci.

Vaše organizace může SPSS Modeler využít k provedení analýz bez ohledu na to, kde jsou data uložena nebo zda jsou data strukturovaná nebo nestruturovaná. Architektura client – server vám umožňuje delegovat datové manipulace a výpočty do datového zdroje, tak aby se minimalizoval pohyb dat a zvýšila se efektivita výpočtu.