STU MTF
SLOVENSKÁ TECHNICKÁ
UNIVERZITA V BRATISLAVE
MATERIÁLOVOTECHNOLOGICKÁ
FAKULTA SO SÍDLOM V TRNAVE

VEDECKÉ MONOGRAFIE

VYUŽITIE PROCESU ZÍSKAVANIA ZNALOSTÍ V OBLASTI PLÁNOVANIA A RIADENIA VÝROBNÝCH PROCESOV

Ing. Dominika JUROVATÁ, PhD.

Táto vedecká monografia bola schválená v súlade so Smernicou rektora č. 11/2012-N na základe rozhodnutia Vedenia MTF STU č. 126 z 26. 3. 2013.
Recenzenti: prof. Ing. Mária Franeková, PhD.
doc. Ing. Peter Bubeník, PhD.
Ing. Marián Urcikán, PhD.

SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
Materiálovotechnologická fakulta so sídlom v Trnave

Ing. Dominika Jurovatá, PhD.

VYUŽITIE PROCESU ZÍSKAVANIA ZNALOSTÍ V OBLASTI PLÁNOVANIA A RIADENIA VÝROBNÝCH PROCESOV

UTILIZATION OF KNOWLEDGE ACQUISITION PROCESS IN THE FIELD OF PRODUCTION PROCESSES PLANNING AND CONTROL

Ing. Dominika Jurovatá, PhD., Ústav aplikovanej informatiky, automatizácie a matematiky, Materiálovotechnologická fakulta STU, Paulínska 16, 917 24 Trnava

e-mail: dominika.jurovata@stuba.sk

Abstrakt

Monografia je zameraná na využitie procesu získavania znalostí v oblasti plánovania a riadenia výrobných procesov. Táto práca je orientovaná na predikciu správania sa systému z dát výrobného procesu. V monografii je uvedený opis používaných metód a techník v predikčných úlohách a zároveň prehľad metrík na ich vyhodnotenie. Ďalej monografia popisuje proces získavania znalostí realizovaný v programe STATISTICA Data Miner. Publikácia pozostáva zo štyroch častí. Prvá časť je venovaná teoretickým aspektom so zameraním na oblasti Business Inteligence, Knowledge Discovery a Data mining. Využité metódy, charakteristika objektu skúmania a postup realizácie experimentálnej štúdie tvoria druhú časť práce. Tretia kapitola poukazuje na potenciálne príležitosti využitia dolovania dát vo výrobných procesoch. V ďalšej časti práce je popísaná klasifikácia a numerická predikcia, zameraná na skúmanie správania sa výrobného procesu na základe zmeny vstupných parametrov. V závere sú zhrnuté a zhodnotené výsledky a ciele práce.

Kľúčové slová: riadenie procesov, získavanie znalostí z databáz, dolovanie dát, predikcia

Abstract

The monograph is focused on the use of the knowledge discovery process in the planning and control of production processes. This work is focused on the prediction of system behavior from the data manufacturing process. The description of the methods and techniques used for prediction tasks and also an overview of metrics for their evaluation are given in the monograph. Furthermore, monograph describes the process of knowledge discovery implemented in the program STATISTICA Data Miner. The publication consists of four main parts. The first part deals with theoretical aspects, focusing on the area of Business Intelligence, Knowledge Discovery and Data mining. Applied methods, characteristics of a research object and a realization method of the experimental study can be found in the second part of the work. The third part of monograph points to the potential opportunities of using data mining in manufacturing processes. The next part describes the classification and numerical prediction with focus on the behavior of the manufacturing process by varying the input parameters. In conclusion, there are summarized and evaluated the results and the objectives of work.

Key words: control processes, knowledge discovery in databases, data mining, prediction

VEDECKÝ PRÍNOS

Prínosy monografie spočívajú v analýze príležitostí a problémov vhodných na riešenie pomocou procesu získavania znalostí z databáz v priemysle, v identifikácii týchto príležitostí pre hierarchické riadiace systémy, vo vypracovaní konceptuálneho návrhu procesu získavania znalostí v hierarchických riadiacich systémoch, v aplikácii a overení navrhnutých postupov procesu získavania znalostí z databáz na riadenie konkrétneho výrobného procesu a v realizácii prognóz s nastavenými vstupnými parametrami. Bolo preukázané, že pomocou navrhnutého postupu riešenia je možné určiť, či zvolené nastavenie vstupných parametrov povedie k očakávanému splneniu cieľov procesu a zároveň predikovať konkrétne kvantitatívne ohodnotenie týchto cieľov, čo je možné považovať za ďalší vedecký prínos. Výsledok bádania overený v simulačnom prostredí Witness od firmy Lanner Group Ltd. možno taktiež považovať za prínos tejto monografie.

Monografia je konkrétnym príspevkom využitia metód získavania znalostí v riadení výrobných systémov. Východiskom pri návrhu prediktívnych modelov správania sa výrobného systému sa stal modul STATISTICA Data Miner od firmy StatSoft. Na podrobnejšiu analýzu dát boli využité matematické princípy korelácie medzi zvolenými premennými. Navrhnuté modely obsahovali deväť metód klasifikácie a deväť metód numerickej predikcie. Prínosom je taktiež porovnanie vybraných metód dolovania dát z hľadiska viacerých metrík, ako aj získanie nových poznatkov, resp. znalostí o analyzovanom procese a vytvorenie PMML súborov použiteľných v informačných a riadiacich systémoch. Vhodne zvolené metódy a algoritmy môžu priniesť nové znalosti, ktoré je možné použiť priamo v riadení výrobných systémov, a to viacerými spôsobmi. V plánovaní a riadení výrobných procesov môže pomôcť aplikácia procesu získavania znalostí z databáz pri identifikácii vplyvu výrobných parametrov na výrobný proces a následnej optimalizácii výrobného procesu, čo povedie k zvýšeniu produktivity a dosahovaniu väčšieho zisku.

Monografia sa zaoberá problematikou predikcie správania sa výrobného systému na základe zmeny vstupných parametrov. Význam monografie tkvie v originálnom prístupe k získavaniu znalostí na riadenie výrobných systémov. Riešený problém je aktuálnou témou pre výrobné spoločnosti, ktoré hľadajú rezervy aj na úrovni plánovania a riadenia procesov. Problematika plánovania a riadenia výroby sa stáva vhodným objektom na využívanie technológií dolovania dát a získavania znalostí.

ZOZNAM SKRATIEK A ZNAČIEK

BI	**B**usiness **I**ntelligence
BSP	**B**usiness **S**ystem **P**lanning
CandRT	**C**lassification and **R**egression **T**rees
CRISP-DM	**CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
CRM	**C**ustomer **R**elationship **M**anagement
CHAID	**Ch**i-squared **A**utomatic **I**nteraction **D**etector
DM	**D**ata **M**ining
DOLAP	**D**esktop **O**n-**L**ine **A**nalytical **P**rocessing
DSA	**D**ata **S**taging **A**rea
DSS	**D**ecision **S**upport **S**ystems
EAI	**E**nterprise **A**pplication **I**ntegration
EDI	**E**lectronic **D**ata **I**nterchange
EIS	**E**xecutive **I**nformation **S**ystems
ERP	**E**nterprise **R**esource **P**lanning
ETL	**E**xtract **T**ransform **L**oad
HOLAP	**H**ybrid **O**n-**L**ine **A**nalytical **P**rocessing
HTTP	**H**ypertext **T**ransfer **P**rotocol
Inc.	**I**ncorporated
IS/IT	**I**nformation **S**ystems / **I**nformation **T**echnology
ISO	**I**nternational **O**rganization for **S**tandardization
ISA 95	**I**nternational **S**tandard for the integration of enterprise and control systems
KDD	**K**nowledge **D**iscovery in **D**atabases
KNN	**K** – **N**earest **N**eighbour
Ltd.	**L**imited
MARS	**M**ultivariate **A**daptive **R**egression **S**plines
MES	**M**anufacturing **E**xecution **S**ystems
MSE	**M**ean **S**quared **E**rror
MIS	**M**anagement **I**nformation **S**ystems
MMABP	**M**ethodology for **M**odeling and **A**nalysis of **B**usiness **P**rocesess
MOLAP	**M**ultidimensional **O**n-**L**ine **A**nalytical **P**rocessing
MRP	**M**anufacturing **R**esource **P**lanning

ODBC	**O**pen **D**atabase **C**onnectivity
ODS	**O**perational **D**ata **S**tore
OIS	**O**ffice **I**nformation **S**ystems
OLAP	**O**n–**L**ine **A**nalytical **P**rocessing
RBF	**R**adial **B**asis **F**unction
ROLAP	**R**elation **O**n-**L**ine **A**nalytical **P**rocessing
SANN	**S**earch **A**utomatic **N**eural **N**etwork
SCADA	**S**upervisory **C**ontrol **A**nd **D**ata **A**cquisition
SVM	**S**upport **V**ector **M**achine
SQL	**S**tructured **Q**uery **L**anguage
TCP/IP	**T**ransmission **C**ontrol **P**rotocol/**I**nternet **P**rotocol
V1	**V**ýrobok **1**
V2	**V**ýrobok **2**
VD1	**V**eľkosť **D**ávky výrobku **1**
VD2	**V**eľkosť **D**ávky výrobku **2**
XML	**E**xtensible **M**arkup **L**anguage

ÚVOD

Znalosti a informácie sú v súčasnosti považované za jedny z najcennejších podnikových zdrojov. Konkurenčné trhové prostredie si vyžaduje, aby boli využívané všetky dôležité informácie potrebné pre činnosť podniku a bolo s nimi účelne narábané. Obrovské množstvo dát, ktorými podnik disponuje, potrebuje taký prístup, aby uložené dáta priniesli očakávanú pridanú hodnotu. Z tohto pohľadu je téma aktuálna a riešenie pre prax je významné. Riadiaci pracovníci potrebujú informácie dostatočné, presné a dynamické, aby mohli viesť podnik úspešne a vedeli reagovať na zmeny trhu. Preto sa proces dolovania dát javí ako vhodný zdroj informácií. Zavedením procesu získavania znalostí do oblasti plánovania a riadenia procesov je možné dosiahnuť lepšie pochopenie riadeného systému, získanie nových zaujímavých znalostí predikujúcich budúce správanie sa výrobného systému. Riadiacim pracovníkom pomôžu nové získané znalosti v ich rozhodovacom procese a môžu prispieť k dlhodobo udržateľnému rozvoju so zníženou citlivosťou na ekonomickú recesiu a trhové výkyvy.

Súčasný svet je charakteristický explóziou objemu dát zbieraných a ukladaných do databáz. Uchovávajú sa rôzne údaje, napríklad údaje o zákazníkoch, dodávateľoch, evidujú sa objednávky, faktúry a zvyšuje sa aj snaha uchovať čo najviac údajov o výrobnom procese. Podniky kumulujú svoje dáta v databázach, ale čo skutočne potrebujú, sú informácie, resp. poznatky alebo znalosti (68). Analytici musia získať informácie, aby boli schopní predvídať trendy a na základe analýzy umožnili kompetentným riadiacim pracovníkom robiť dynamické rozhodnutia. Plánovanie a riadenie výroby, podnikové analýzy, risk manažment, odhaľovanie porúch a nepodarkov, to sú príklady sfér riadiacich činností vyžadujúcich dokonalé znalosti od osôb, ktoré tieto činnosti vykonávajú. Hovorí sa o riadení znalostí alebo o systémoch riadenia znalostí. Zodpovedné osoby sa stávajú znalostnými pracovníkmi (58). Dôležitým krokom v systémoch riadenia znalostí je proces získavania znalostí z databáz, často nazývaný aj dolovanie dát. Riadiaci pracovníci a manažéri predstavujú koncových užívateľov týchto nástrojov a systémov na všetkých úrovniach riadenia. Ich nároky a potreby v dnešnej dobe narastajú s tým, ako sa zostruje konkurencia, zväčšujú sa objemy dát, ktoré je nutné analyzovať, zvyšujú sa nároky na rýchlosť analýzy a prijatia následného rozhodnutia.

V oblasti plánovania a riadenia výrobných procesov sa získavanie znalostí z výrobných databáz a z údajov o stave riadeného procesu doteraz minimálne používa (23, 56, 35).

Hlavným cieľom monografie je použitie procesu získavania znalostí v oblasti plánovania a riadenia výrobných procesov, a to predikovaním správania sa výrobného procesu na základe

zmeny vstupných parametrov. Správanie sa systému bude predikované pomocou klasifikácie a numerickej predikcie, čo vychádza z cieľov dolovania dát. Prvým cieľom dolovania dát je predikovať splnenie všetkých cieľových parametrov súčasne (klasifikácia) a druhým cieľom dolovania dát je predikovať konkrétne kvantitatívne hodnoty cieľových parametrov výroby (numerická predikcia). Táto práca je orientovaná na využitie metód dolovania dát a prostredníctvom nich získanie znalostí z dát výrobného procesu, ktoré sa využívajú pri plánovaní a riadení. Pre získanie potrebných údajov o riadenom výrobnom procese bol vytvorený simulačný model výrobného systému. Údaje získané zo simulačného modelu sú uchovávané v databáze. Pre konkrétny definovaný problém riadenia výroby je vytvorený model dolovania dát s použitím vybraných metód a techník. Nové znalosti o výrobnom systéme a možnosti dosahovania stanovených cieľov výroby sú získané zmenou vstupných parametrov a predikciou na základe týchto parametrov. Získané predikované správanie sa procesu sme overovali na simulačnom modeli výrobného systému.

1 TEORETICKÉ VÝCHODISKÁ

Táto kapitola je venovaná opisu súčasného stavu riešenej problematiky. Sú tu uvedené a rozobrané základné informácie a poznatky týkajúce sa témy získavania znalostí na plánovanie a riadenie výrobných procesov.

1.1 Business Intelligence

Využitie získavania znalostí v hierarchickom riadení procesov zahŕňa pojem Business Intelligence (skratka BI). Tento pojem zaviedol v roku 1989 Howard J. Dresner, ktorý definuje Business Intelligence ako „*súbor konceptov a metód určených na skvalitnenie rozhodnutí firmy*“ (54). Dresner taktiež uvádza nasledovnú definíciu BI: „Predstavuje poznatky získané pri prístupe a analýze obchodných informácií“ (11). BI môžeme chápať ako ucelený komplex procesov, úloh, činností, technológií a aplikácií, ktorých cieľom je účelne a účinne podporovať rozhodovacie procesy. Podporujú analytické a plánovacie činnosti v podniku a organizáciách. Sú postavené na princípoch multidimenzionálnych pohľadov na všetky dostupné dáta (50).

V rozsiahlej publikácii „The profit impact of business intelligence“ od Williamsových (81) je podrobne rozpísané ako BI zvyšuje zisk, aké chyby robia organizácie pri zavádzaní BI a podobne. Aktuálny stav BI uviedli Watson a Wisom (79) vo svojom príspevku The Current State of BI, ako aj Chaudhuri, Dayal, Narasayya (30) v článku An Overview of BI Technology.

Aktuálne sa používajú viaceré pomenovania pre BI a v princípe ide o to isté. Ide o pomenovania Business Intelligence a Business Analytics, resp. Business Discovery. Ak si porovnáme BI a BA, pri prvej „skratke“ sa kladie dôraz skôr na podporu rozhodovania na základe dodaných podkladov primárne od IT oddelení a v druhom prípade je to snaha podporiť rozhodovanie na základe vlastných analýz realizovaných v biznise s využitím technologických prostriedkov a dát zo systémov BA (3).

V súvislosti s BI sa aktuálne používa označenie BigData. Ide o tzv. nadstavbu, resp. rozšírenie BI. BigData sa zameriava na „veľké“ a v surovom stave „neužitočné“ dáta – t.j. dáta, ktoré používateľ do systému priamo nevložil, ako logy na webovom serveri, elektronické merače prietoku, dáta z bezpečnostných kamier, RFID snímače - teda milióny záznamov, ktoré sa po niekoľkých dňoch zahodia, aby nezaberali miesto na diskoch (39).

Účelom BI je konvertovať veľké objemy údajov na poznatky alebo znalosti, ktoré sú potrebné pre koncového užívateľa. Tieto poznatky, resp. znalosti potom môžu byť využité

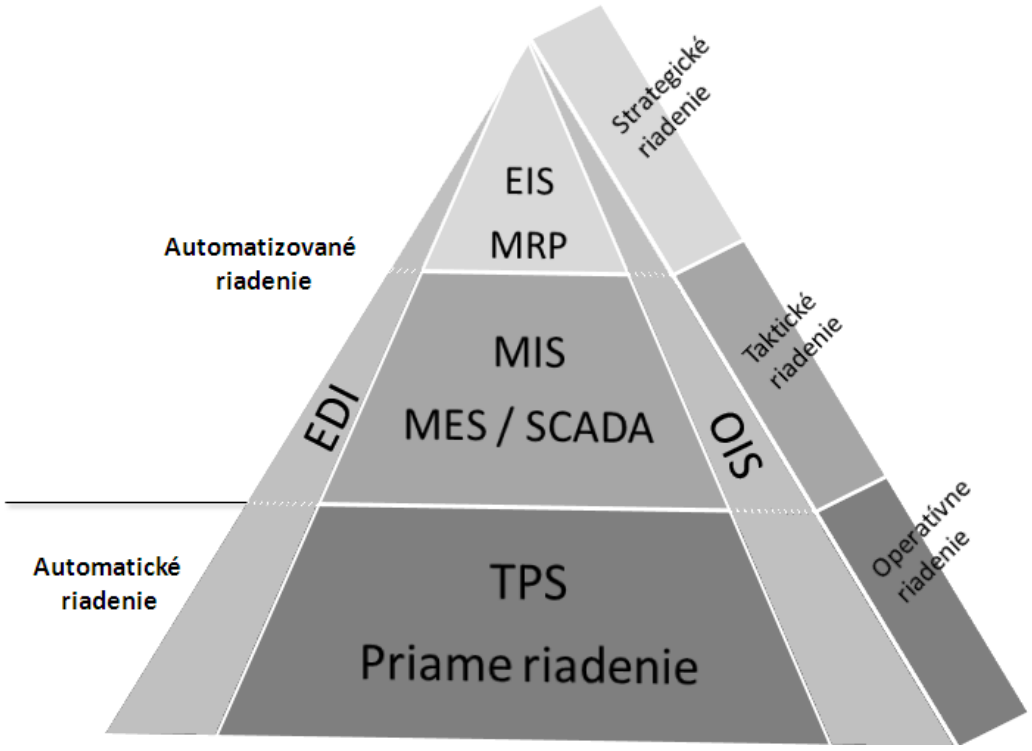
v procese riadenia na zvýšenie kvality či spoľahlivosti, ale aj bezpečnosti, efektívnosti a zvýšenie miery zisku. Poznatky a znalosti takto získané nie sú len konkrétne záznamy alebo množiny záznamov, môžu to byť znalosti určujúce závislosti medzi údajmi alebo poznatky získané z pozorovania istej veličiny. A preto moderné databázové servery obsahujú rozsiahlu podporu pre budovanie dátových skladov, analýzy OLAP a dolovania údajov. Okrem BI sa do skupiny systémov na podporu rozhodovania tiež zaraďujú:

- Management Information Systems (MIS),
- Decision Support Systems (DSS).

Podľa vzťahu k úrovni riadenia podniku je rozdiel medzi jednotlivými manažérskymi systémami v tom, či sú určené a využívané pre (77):

- operatívne,
- taktické,
- strategické riadenie.

Vyjadrenie vzťahu systémov k úrovni riadenia je zobrazené na obrázku 1.



Obr. 1 *Základné stavebné bloky architektúry IS/IT*

Systémy BI poskytujú hlboké analýzy nad veľkými objemami dát. Pomocou dataminingových modelov možno napríklad predpovedať z histórie údajov uložených v úložisku vyvíjanie sa daného procesu v budúcnosti (43). BI technológie sú používané pri

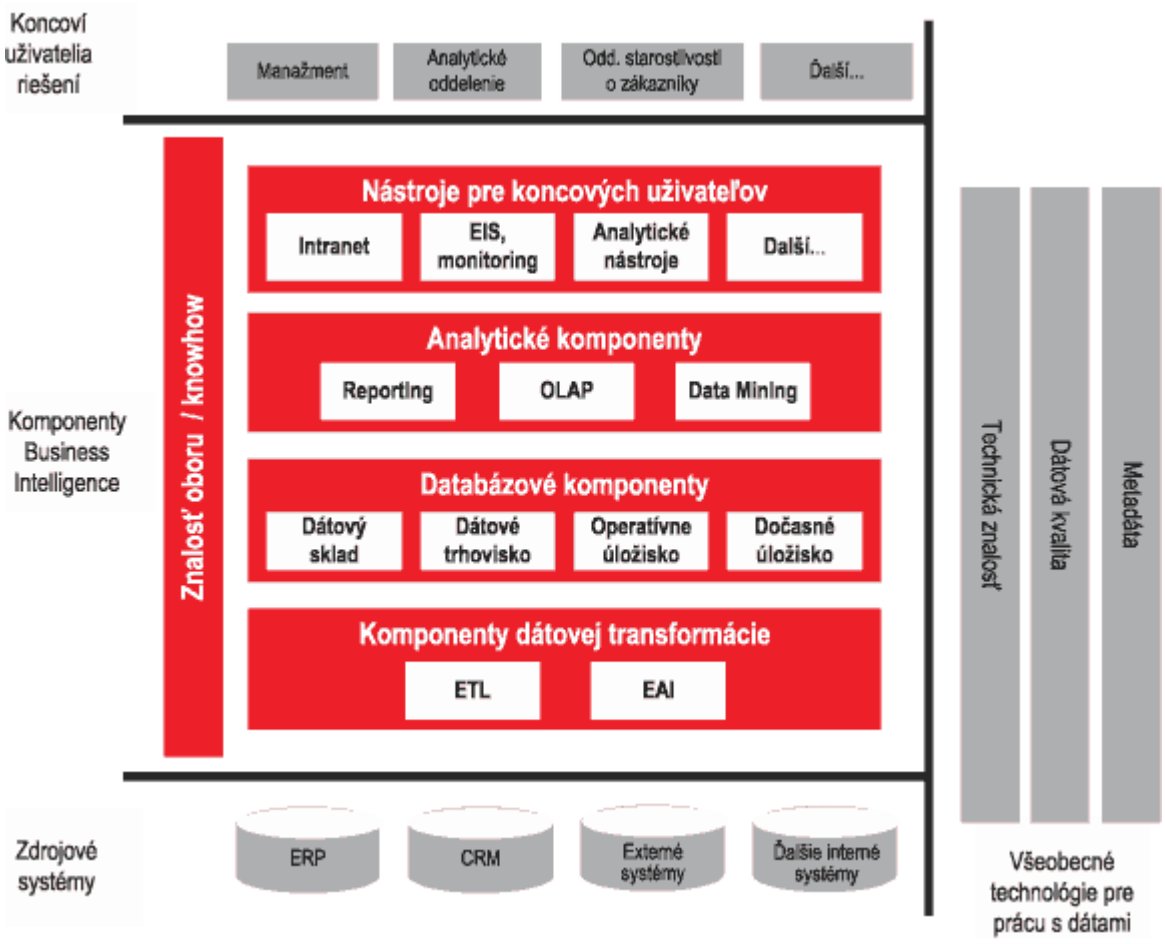
výrobe na objednávku a zákaznícku podporu, v obchode na vytváranie profilu užívateľa, vo finančných službách na analýzu likvidity a odhaľovanie podvodov, v doprave na správu vozového parku, v oblasti telekomunikácií na identifikáciu dôvodov straty zákazníkov, v oblasti technickej infraštruktúry na analýzu využitia energie, a pri zdravotnej starostlivosti na určenie výsledkov analýzy (30).

V rámci architektúry informačných systémov a úrovní riadenia je BI zaraďované k systémom orientovaným na strategické riadenie. Tieto strategické systémy sú tiež nazývané Executive Information Systems (EIS), ale podľa Lacka (41) sa častejšie stretávame s pojmom BI. „Aplikácie vytvorené pre riadenie podniku na tejto strategickej úrovni získavajú dáta z ostatných aplikácií podnikového informačného systému, ako sú Management Information Systems a z externých informačných zdrojov.“ Dáta sú tu agregované, vytvárajú časové rady a vzájomné väzby (77). Výstupy z týchto aplikácií sú následne určené pre rozhodovanie manažmentu či riadiacich pracovníkov. Jednotná štruktúra komponentov pri riešení BI neexistuje, a preto aj pri každom individuálnom riešení sa môžu využívať viaceré komponenty.

1.1.1 Komponenty BI

V tejto podkapitole sú opísané jednotlivé vrstvy a komponenty BI z hľadiska všeobecného prehľadu. Všeobecná koncepcia architektúry riešení BI obsahuje päť vrstiev (10, 22, 50). Typická architektúra pre podporu BI v rámci podniku je znázornená na obrázku 2.

Koncepcia BI ako zdroje dát využíva systémy ERP, CRM, rôzne externé a interné zdrojové a produkčné systémy, väčšinou transakčného charakteru. Vytvárajú z pohľadu BI zdrojové databázy, tieto systémy teda nie sú priamym komponentom riešení BI, ale iba zdrojom dát (22).



Obr. 2 *Všeobecná koncepcia architektúry riešenia BI (1)*

Komponenty dátovej transformácie – predstavujú vrstvu pre extrakciu, transformáciu, čistenie a nahrávanie dát. Zastrešujú oblasť zberu a prenosu dát zo zdrojových systémov do vrstvy na ukladanie dát v riešení BI. Patria sem tieto komponenty:

- Systémy ETL,
- Systémy EAI.

Skratkou **ETL** (Extraction, Transformation, Load) sú označované dátové pumpy, ktoré zabezpečujú extrakciu, transformáciu a ukladanie dát medzi produkčnými databázami a DSA, ODS, dátovým skladom a dátovými trhoviskami. Ide o riadenie procesu, pri ktorom sú jednotlivé údaje prenášané priamo medzi jednotlivými aplikáciami alebo databázami. Dáta sú najprv načítavané z interných a externých zdrojov podniku, potom sa čistia, a tým redukujú nepodstatné dáta.

EAI (Enterprise application integration) sú integračné nástroje pre on-line, okamžité aktualizácie dátového skladu, umožňujúce tzv. „dátové sklady v reálnom čase“ (10).

Databázové komponenty – predstavujú vrstvu na ukladanie dát. Zabezpečujú procesy ukladania, aktualizácie a správy dát pre riešenie BI. Zahŕňajú tieto komponenty:

- Dátové sklady (Data Warehouses),
- Dátové trhoviská (Data Marts),
- Operatívne dátové úložiská (Operational Data Stores),
- Dočasné úložiská dát (Data Staging Areas).

Dátový sklad – je základným databázovým komponentom. Podľa Inmona (32) je definovaný ako *subjektovo orientovaná, integrovaná, stála a časovo rozlíšiteľná kolekcia dát, určená na podporu procesu rozhodovania.*

Význam jednotlivých pojmov definície možno chápať takto:

Subjektovo orientovaný – dáta sú kategorizované podľa subjektu (napr. zákazník, dodávateľ, výrobok, atď.), nie podľa aplikácii, pre ktoré majú byť používané.

Integrovaný – dáta sú ukladané v rámci celej firmy, nie v rámci jednotlivých oddelení.

Stály – v dátových skladoch sa môžu realizovať len dva základné typy operácií, a to prvotné naplnenie dátového skladu a prístup k dátam (čítanie). Neexistujú v ňom žiadne iné operácie zmeny dát.

Časovo rozlíšený – načítané dáta musia so sebou niesť aj dimenziu času (história dát), aby bolo možné vykonať analýzu za určité časové obdobie.

Nevýhodou dátového skladu je zložitejšia realizácia, pomalšia implementácia a sekundárne procesy načítavania (z centrálneho dátového skladu do dátového trhoviska). Výhodami sú konzistentný obsah dátového skladu, menší počet procesov načítavania z prevádzkových systémov (primárne načítavacie procesy), jednoduchšia správa načítavacích procesov a jednoduchšie vytváranie nových dátových tržníc (detailné dáta sú k dispozícii už v dátovom sklade) (53).

Dátové trhoviská – sú to subjektovo orientované analytické databázy. Údaje pre dátové trhoviská sú vyberané s cieľom vyhovieť špecifickým požiadavkám, teda sú určené pre vymedzený okruh užívateľov. Sú to vlastne decentralizované dátové sklady, ktoré sa postupne integrujú do komplexného dátového skladu. V niektorých riešeniach tvoria medzistupeň pri transformácii dát z produkčných databáz aj po vytvorení už komplexného dátového skladu. Nezávislé dátové tržnice sú samostatné dátové úložiská pre jednotlivé aplikácie alebo útvary.

Operatívne dátové úložiská – reprezentujú podporné analytické databázy pre potreby taktického alebo aj strategického rozhodovania. V rámci spracovávania dát vykonáva

operatívne dátové úložisko operácie potrebné na zaistenie dátovej kvality. Zároveň integruje relevantné dáta z rôznych systémov. Na úrovni operatívneho dátového úložiska sú jasne a zreteľne identifikované a opísané jednotlivé dátové elementy.

Dočasné úložiská dát – predstavujú databázy na dočasné uloženie dát pred ich vlastným spracovaním do databázových komponentov riešenia BI.

Toto všetko sú finálne zdroje, odkiaľ je možné čerpať dáta na získavanie znalostí (50).

Analytické komponenty – predstavujú vrstvu na analýzu dát. Pre manažérov a riadiacich pracovníkov majú analytické komponenty veľký význam, pretože predstavujú činnosti spojené s vlastným sprístupnením a analýzou dát. Obsahujú tieto komponenty (10):

- Reporting,
- Systémy On Line Analytical Processing (OLAP),
- Dolovanie dát (Data Mining).

Reporting – je analytická vrstva zameraná na štandardný alebo ad hoc opytovací proces do databázových komponentov riešenia BI. Predstavuje činnosti spojené s otázkami do databáz pomocou štandardného rozhrania týchto databáz, napr. SQL príkazov. Pre koncových užívateľov to znamená možnosť získania analytických tabuliek a prehľadov na základe otázok do databáz dátových skladov, prípadne multidimenzionálnych databáz. V rámci reportingu sa rozlišuje:

- štandardný reporting (predpripravené otázky sú spúšťané v určitých časových periódach),
- ad hoc reporting (špecifické otázky - explicitne vytvorené používateľom, sú formulované na databázu jednorazovo).

Systémy OLAP, nazývané tiež ako manažérske aplikácie – exekutívne informačné systémy (EIS – Exekutívne Information Systems), predstavujú vyššiu flexibilitu vzhľadom na momentálne požiadavky koncových užívateľov. Umožňujú efektívne uloženie a interaktívne využívanie multidimenzionálnych dát. Základným princípom je niekoľkodimenzionálna tabuľka umožňujúca veľmi rýchlo a pružne meniť jednotlivé dimenzie (22).

Aplikácie BI využívajú tzv. OLAP nástroje, zaisťujúce vysoko efektívny mechanizmus viackriteriálnej analýzy. Ďalej umožňujú ad hoc analýzu s prijateľnou rýchlosťou a dostatočne malým obmedzením rozsahu vstupujúcich dát. Dôležité je, aby dáta na analýzu boli čisté

a konsolidované. OLAP technológia sa delí do štyroch skupín – podľa typu ukladania a spôsobu spracovania dát (22, 50):

- *ROLAP (Relation On-Line Analytical Processing)* rieši multidimenzionalitu uložením dát v relačnej databáze. Reprezentuje priamy prístup k dátam relačného primárneho systému, to znamená, že dáta prezentované v zobrazovacom nástroji sú získavané priamo z pôvodných dátových zdrojov, napr. z tabuliek databáz (prístup do týchto tabuliek je obvykle realizovaný prostredníctvom ODBC ovládačov v okamihu potreby). Pre uloženie dát sa teda používajú štandardné relačné databázy a dáta z nich sú vyberané pomocou SQL otázok.
- *MOLAP (Multidimensional On-Line Analytical Processing)* je charakteristický špeciálnym uložením dát v multidimenzionálnych – binárnych OLAP kockách. Informácie sú v tejto špeciálnej databáze navrhnuté ako množina multidimenzionálnych matíc a sú aktualizované a doplňované v určitom pravidelnom intervale.
- *HOLAP (Hybrid On-Line Analytical Processing)* predstavuje kombináciu predchádzajúcich prístupov, kedy detailné dáta sú uložené v relačnej databáze a agregované hodnoty sú uložené v binárnych OLAP kockách. Rozlíšenie správneho stupňa agregácie, ktorý bude predstavovať hranicu medzi podrobnejšími dátami ukladanými do relačnej dátovej štruktúry a agregovanými informáciami zavádzanými do multidimezionálnych databáz, je významnou úlohou pre implementátora EIS.
- *DOLAP (Desktop On-Line Analytical Processing)* umožňuje pripojenie k centrálnej dátovej kocke a stiahnutie potrebnej podmnožiny kocky na lokálny počítač. Všetky analytické operácie prebiehajú nad lokálnou kockou, užívateľ nemusí byť pripojený k serveru.

Data mining – je vysvetľovaný v samostatnej podkapitole 1.3.

<u>*Nástroje pre koncového používateľa*</u> – spadajú do prezentačnej vrstvy. Zabezpečujú komunikáciu koncových používateľov s ostatnými komponentmi riešenia BI, najmä zber požiadaviek na analytické operácie s následnou prezentáciou výsledkov. Patria sem tieto komponenty:

- Intranet,
- Systémy EIS - Executive Information Systems,
- rôzne analytické aplikácie.

Intranet - je počítačová sieť, ktorá používa rovnaké technológie (TCP/IP, HTTP) založené na Client/Server architektúre. Je určená pre malé skupiny používateľov, napríklad pre určitých pracovníkov daného podniku. Ide o tzv. interný Internet.

Systémy EIS - manažérske aplikácie EIS sú typom klientskych aplikácií BI, ktoré integrujú dôležité dátové zdroje. S tým sú spojené aj špecifické nároky na prezentáciu informácií a ich sprístupnenie manažérom či riadiacim pracovníkom.

Znalosť odboru / know-how - je nevyhnutné využívať odbornú znalosť a tzv. best-practices nasadzovania riešení BI pre konkrétnu situáciu v organizácii pri jednotlivých komponentoch dátovej transformácie, databázových a analytických komponentoch, ako aj pri nástrojoch pre koncových užívateľov (10, 50).

1.2 Získavanie znalostí z databáz

Proces získavania znalostí z databáz (z angl. KDD – Knowledge Discovery in Databases) je v literatúrach uvádzaný viacerými podobnými definíciami, ako napr. *KDD je netriviálny proces identifikácie platných, nových, doteraz neznámych, potenciálne použiteľných a dobre pochopiteľných znalostí o dátach* (14).

Pre správne chápanie definície je vhodné uviesť význam základných pojmov. Pod dátami sa rozumie množina usporiadaných faktov (napr. položiek v databáze) a pod znalosťou výraz v nejakom jazyku opisujúci podmnožinu dát alebo model aplikovateľný na túto podmnožinu. Preto v tomto ponímaní pod pojmom získavanie znalostí z databáz sa rozumie takisto návrh a upresňovanie modelu skúmanej reality tak, aby opisoval dáta čo najlepšie, hľadal opis štruktúry dát, a v neposlednom rade, aby tvoril opis dát na vyššej úrovni (metaznalosti).

Medzi základné vlastnosti procesu objavovania znalostí z databáz patrí, že KDD je v rámci svojich jednotlivých krokov nedeterministický, s mnohými možnými alternatívnymi rozhodnutiami v každom zo svojich krokov. Výber alternatívnej operácie, algoritmu alebo zmena jeho parametrov potom nutne vedú k iným výstupom, čo následne ovplyvňuje ďalšie kroky. Preto v rámci procesu KDD často dochádza k iteráciám v rámci jedného alebo viacerých jeho krokov s cieľom dosiahnuť čo najlepší výsledok. Z toho vyplýva, že KDD je iteratívny a interaktívny. Významnou charakteristikou procesu KDD je jeho multidisciplinárnosť, t.j. možnosť súčasne sledovať a analyzovať problémy z niekoľkých hľadísk podľa potrieb riadiacej úlohy (66).

1.2.1 Metodiky získavania znalostí z databáz

K najpoužívanejším metodikám získavania znalostí z databáz patria nasledujúce tri: SEMMA, 5A a CRISP-DM.

Metodika SEMMA

Metodika SEMMA od firmy SAS sa skladá z piatich krokov:

- Sampling – vzorkovanie,
- Exploration – skúmanie,
- Modification – modifikácia / úprava,
- Modeling – modelovanie,
- Assessment – vyhodnotenie.

Krok ***Sampling*** – výber vhodných vzoriek údajov. Databázy majú obvykle gigabajtové objemy, preto je potrebné uvážiť, či je pre analýzu potrebné použiť celú množinu údajov alebo bude postačujúca reprezentatívna vzorka údajov. Tento postup je štatisticky korektný. Vo všeobecnosti platí, že ak v celkových údajoch je obsiahnutý nejaký všeobecný vzťah/pravidlo, jeho vplyv musí byť viditeľný aj v reprezentatívnej vzorke údajov.

Krok ***Exploration*** – vizuálna explorácia a redukcia dát, diagnostika charakteristiky údajov. Zmyslom tohto kroku je ustáliť predstavu o otázkach, na ktoré môže analýza konkrétnych údajov poskytnúť odpovede. Prieskum údajov umožní zoznámiť sa s rozložením príslušných hodnôt v dátovom priestore a získať obraz o rozložení extrémnych hodnôt a rozpoznať existenciu sekvencií, asociácii alebo zoskupení.

Krok ***Modification*** – manipulácia a transformácia údajov. Príkladom možných modifikácií údajov je odstránenie nedefinovaných hodnôt, doplnenie opisov premenných, doplnenie nových informácií, vytvorenie zoskupení a pod.

Krok ***Modeling*** – konštrukcia abstraktných modelov. V podstate ide o hľadanie odpovede na otázku, čo je príčinou vzorov nájdených v údajoch. Odpoveď môže byť získaná napríklad konštrukciou štatistického modelu, ktorým je formulovaná a otestovaná explicitne vyjadrená hypotéza. Výber vhodnej metódy je závislý od charakteru vzorov, ktoré boli v údajoch rozpoznané pri ich prieskume.

Krok ***Assesment*** – porovnanie a posúdenie vytvorených modelov. V tomto kroku sa na základe porovnania získaných alternatív vyberie jeden model ako výsledné riešenie.

Intuitívna interpretácia abstraktného modelu je výsledkom, ktorý očakáva užívateľ od výsledku analýzy (60).

Metodika 5A

Metodika **5A** od firmy SPSS je podobná metodike SEMMA a zahŕňa taktiež päť krokov:

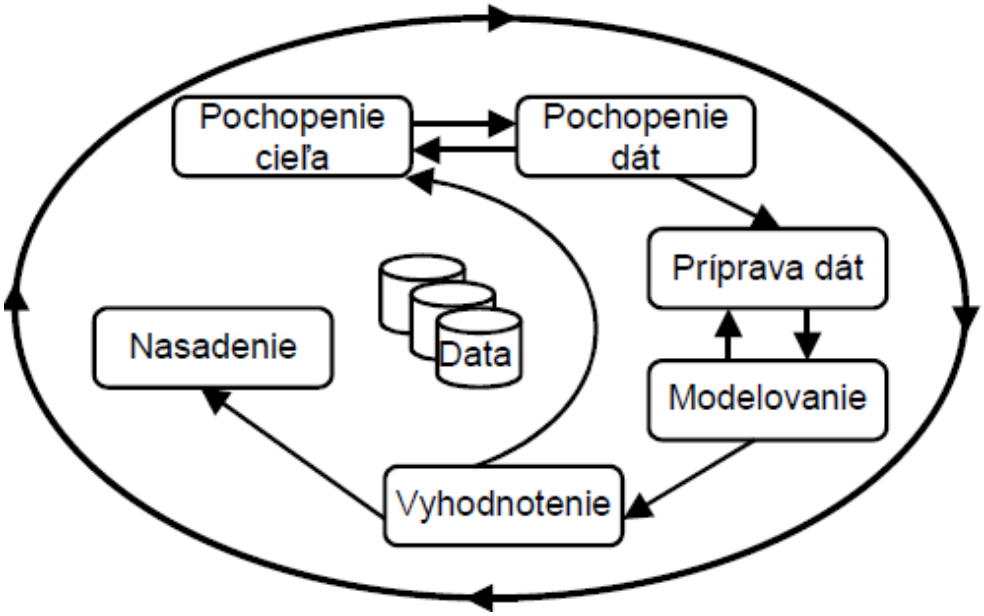
- Assess – posúdenie potrieb projektu,
- Access – zhromaždenie potrebných dát,
- Analyze - vykonanie analýz,
- Act - premena znalostí na akčné znalosti,
- Automate - aplikovanie výsledkov analýzy do praxe.

Metodika CRISP-DM

Metodika **CRISP-DM** (CRoss-Industry Standard Process for Data Mining), ktorá vznikla v rámci Európskeho výskumného projektu, predstavuje univerzálne použiteľný postup pre dolovanie dát. Táto metodika uvádza šesť fáz:

- Business Understanding – pochopenie cieľa (problematiky),
- Data Understanding - pochopenie dát,
- Data Preparation - príprava dát,
- Modeling – modelovanie,
- Evaluation - vyhodnotenie,
- Deployment - nasadenie.

V práci využijeme práve túto metodiku, pretože aj program STATISTICA, ktorý sme sa rozhodli použiť ako nástroj dolovania dát, je na tejto metóde založený. Nadväznosť a vzťahy medzi jednotlivými fázami vidieť na obrázku 3.



Obr. 3 *Procesný model pre KDD (53)*

Fáza ***pochopenia cieľa*** sa zameriava na pochopenie reálneho problému - obchodných alebo iných cieľov a požiadaviek, snaží sa ich pretransformovať na konkrétnu definíciu úlohy dolovania dát.

Fáza ***pochopenia dát*** začína počiatočným zberom dát, pokračuje aktivitami na podrobné oboznámenie sa s dátami, zistenie ich kvality a charakteru, prípadne zistením podmnožiny dát zaujímavých pre ďalší výskum.

Fáza ***prípravy dát*** zahŕňa všetky aktivity potrebné na vytvorenie množiny dát pre modelovanie. Operácie vykonávané v rámci prípravy dát väčšinou prebiehajú viackrát v nepredpísanom poradí. Patrí sem výber tabuliek, príkladov, atribútov, ako aj ich transformácia a čistenie.

Fáza ***modelovania*** je charakterizovaná výberom a aplikovaním nejakej metódy modelovania, nastavením a vyladením jej parametrov na optimálne hodnoty. Tu je nevyhnutná úzka interakcia s predchádzajúcou fázou prípravy dát pre špecifické požiadavky väčšiny metód na formu dát.

Fáza ***vyhodnotenia*** už disponuje veľmi kvalitnými modelmi z pohľadu dolovania v dátach, ale je potrebné ich vyhodnotiť z pohľadu stanovených obchodných alebo iných cieľov.

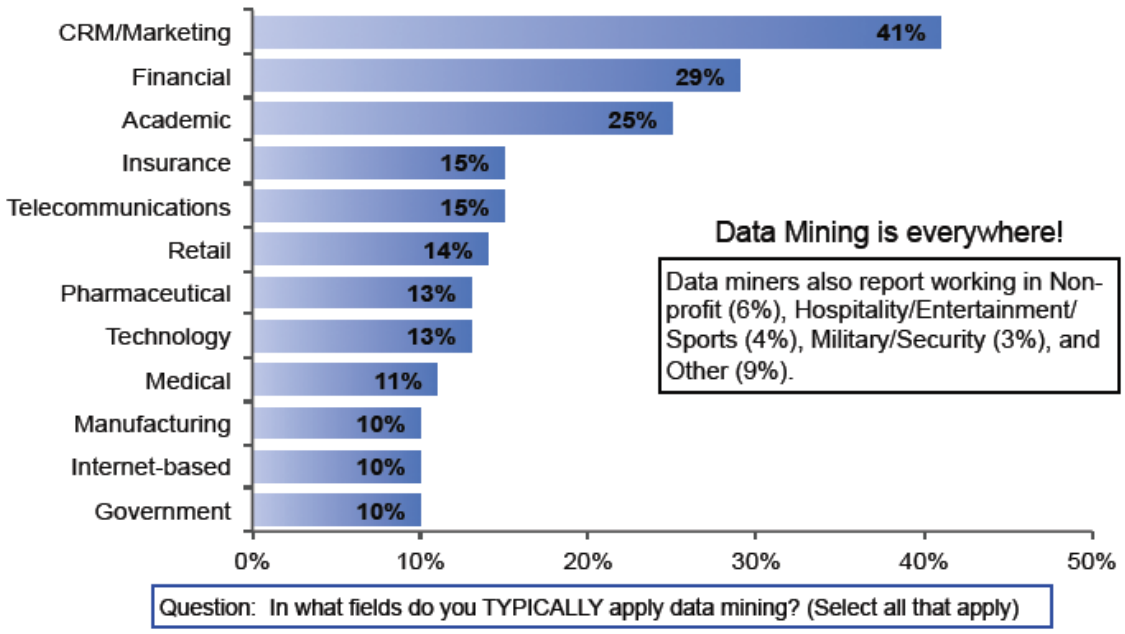
Fáza ***nasadenia*** získaných modelov môže byť pomerne jednoduchá, ak ide o vygenerovanie správy, ale môže byť aj zložitejšia, ako napr. implementácia opakovateľného procesu KDD pre danú aplikáciu (53, 80).

1.3 Dolovanie dát

V architektúre BI sa dolovanie dát nachádza vo vrstve analytických komponentov, kam zaraďujeme aj systémy OLAP a reporting. Z uvedenej skupiny je dolovanie dát najmladším komponentom. Medzi analýzami na báze OLAP a dolovania dát môžeme pozorovať podstatný rozdiel. Užívateľ pri použití OLAP vytvára určitú skupinu hypotéz, ktoré potom pomocou OLAP aplikácie potvrdzuje alebo vyvracia (toto sa nazýva deduktívny spôsob). Analýzy na báze dolovania dát naopak pomáhajú takéto hypotézy na základe analýzy skutočných dát vytvárať (50).

Dolovanie dát (z angl. DM – Data Mining) je kľúčovým krokom procesu KDD. Je to aplikácia inteligentných metód na získanie platných vzorov. V tejto fáze procesu teda ešte nehovoríme o znalostiach, ktoré vznikajú až vhodným výberom z vygenerovaných vzorov a ich aplikovaním v kontexte riešenej úlohy (53).

Súčasné odvetvia a oblasti využitia dolovania dát v praxi sú rôzne. V článku (52) je stručné zhodnotenie rôznych trendov aplikácií dolovania dát. Prieskum spoločnosti Rexer Analytics (56) ukazuje, že datamining je najviac využívaný v marketingu / CRM, ďalej vo financiách a vo výskume (obrázok 4). Tieto tri oblasti sú uvádzané vo všetkých nimi vykonaných prieskumoch (za roky 2007-2010).



Obr. 4 *Oblasti použitia dataminingu v praxi (56)*

Iný pohľad prináša prieskum spoločnosti KDnuggets (35) v tabuľke 1, ktorá uvádza porovnanie za roky 2010 a 2011. Oblasti s najväčším výrazným zvýšením využitia dataminingu boli cestovanie / pohostinstvo (o 429%), sociálne siete (100%), vzdelanie (65%) a biotechnológia / genomika (64%). Odvetvia s najväčším poklesom používania dataminingu boli výroba (-34%), reklama (-29%), e-Commerce (-25%), investície/akcie (-22%) a dolovanie s použitím internetu (-21%).

ODVETVIA / OBLASTI, V KTORÝCH SA VYUŽÍVA DATA MINING (35) Tabuľka 1

Industries / Fields where you applied Analytics / Data Mining in 2011? [228 voters]	2011 % of voters	2010 % of voters	
CRM/ consumer analytics (57)	25.0%	26.8%	
Banking (43)	18.9%	19.2%	
Health care/ HR (38)	16.7%	13.1%	
Education (37)	16.2%	9.9%	
Fraud Detection (32)	14.0%	12.7%	
Science (31)	13.6%	10.3%	
Social Networks (30)	13.2%	6.6%	
Credit Scoring (29)	12.7%	8.0%	
Direct Marketing/ Fundraising (28)	12.3%	11.3%	
Insurance (28)	12.3%	10.3%	
Finance (26)	11.4%	11.3%	
Telecom / Cable (25)	11.0%	10.8%	
Retail (24)	10.5%	8.0%	
Medical/ Pharma (22)	9.6%	8.0%	
Other (17)	11.7%	7.5%	
Biotech/Genomics (21)	9.2%	5.6%	
Government/Military (17)	7.5%	6.1%	
Travel / Hospitality (17)	7.5%	1.4%	
Advertising (16)	7.0%	9.9%	
Web usage mining (16)	7.0%	8.9%	
Software (16)	7.0%	0.0%	
e-Commerce (12)	5.3%	7.0%	
Manufacturing (12)	5.3%	8.0%	
Search / Web content mining (12)	5.3%	6.6%	
Investment / Stocks (10)	4.4%	5.6%	
Entertainment/ Music/ TV/Movies (8)	3.5%	3.3%	
Security / Anti-terrorism (4)	1.8%	1.9%	
Social Policy/Survey analysis (4)	1.8%	0.9%	
Junk email / Anti-spam (3)	1.3%	0.9%	

Typické úlohy pre datamining sú uvádzané vo viacerých literatúrach nasledovne (52, 65):

- Financie – detekcia podvodníkov
- Telekomunikácie – predpoveď odchodu zákazníkov
- Marketing – segmentácia, nákupný košík, stanovenie cieľového zákazníka
- Plánovanie – predpovede v časových radoch
- Zdravotníctvo – určenie diagnózy
- WebMining – typické vzory prezerania stránok a ďalšie.

Podľa SAS je možné akýkoľvek proces študovať, pochopiť a vylepšiť s použitím DataMiningu. DataMining je užitočný všade tam, kde je možné zhromažďovať údaje. V súčasnosti je s výhodou a úspešne aplikovaný v rezortoch, ktoré sú orientované na služby zákazníkom, poskytujú finančné služby alebo majú výrobný charakter (56).

V literatúre (16) sú uvedené teoretické zázemia pre aplikáciu dolovania dát v oblasti inžinierskeho navrhovania výroby a logistiky. Objavilo sa tiež niekoľko recenzií dolovania dát v priemyselnej výrobe. Napríklad v literatúre (2 , 7 , 8 , 11) je uvedených veľa možných oblastí použitia dolovania dát vo výrobe ako napr. riadenie kvality, plánovanie, diagnostika, analýza porúch, dodávateľské reťazce, systém na podporu rozhodovania. Niektoré ďalšie príklady aplikácií dolovania dát v priemyselných, lekárskych a farmaceutických odboroch sú prezentované v (16). V tejto monografii je aplikované využitie dataminingu v oblasti plánovania a riadenia výroby v priemyselnom odvetví.

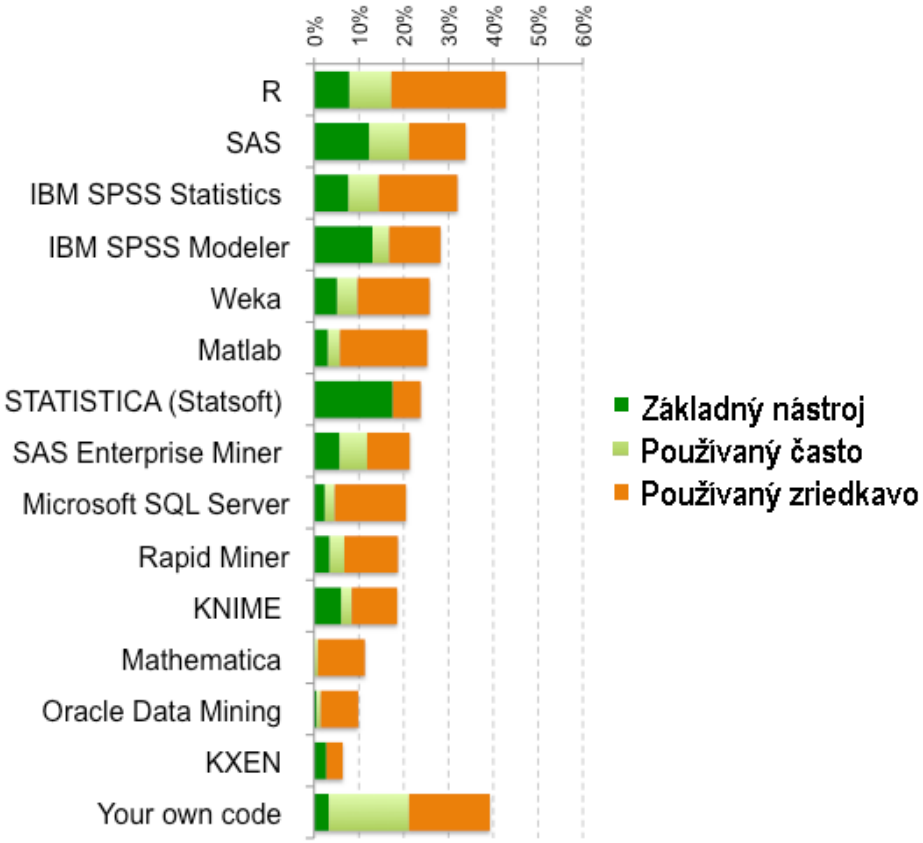
1.3.1 Nástroje dolovania dát

V súčasnosti je dostupných mnoho nástrojov na dolovanie dát či už komerčných alebo nekomerčných. Odlišujú sa počtom poskytovaných metód a techník na dolovanie dát, spôsobom prezentácie dosiahnutých výsledkov, možnosťou interaktivity počas procesu dolovania dát a pod. Všeobecný prehľad nástrojov je uvedený v nasledujúcej tabuľke 2 (32, 35, 44, 51, 55, 56).

PREHĽAD NÁSTROJOV NA DOLOVANIE DÁT Tabuľka 2

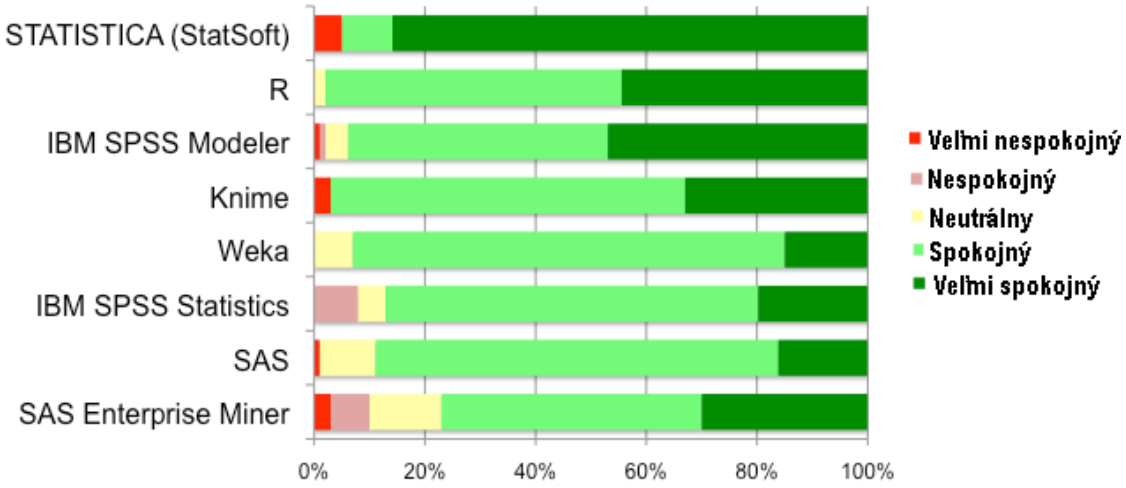
	Nástroj	Firma / Inštitúcia	Internetová stránka
KOMERČNÝ	***CART 6.0 ProEX***	Salford Systems	www.salford-systems.com/cart6.php
	Enterprise Miner	SAS Institute Inc.	www.sas.com
	Intelligent Miner	IBM Corp.	www.ibm.com
	KnowledgeStudio	Angoss Software Corp.	www.angoss.com
	KXEN	KXEN Inc.	www.kxen.com
	MineSet	Silicon Graphics	www.sgi.com
	SQL Server Data Mining	Microsoft	www.microsoft.com
	Oracle Data Mining	Oracle	www.oracle.com
	IBM SPSS Modeler	IBM Corp.	www.ibm.com
	STATISTICA	StatSoft Inc.	http://www.statsoft.com/products/statistica-data-miner/
	Spotfire Miner	TIBCO Software Inc.	spotfire.tibco.com/
NEKOMERČNÝ	***Ferda Data Miner***	Vysoká škola ekonomická v Prahe	http://ferda.wiki.sourceforge.net/
	KDD Package	Technická univerzita Košice	people.tuke.sk/jan.paralic/KDD/
	LISp-Miner	Vysoká škola ekonomická v Prahe	lispminer.vse.cz/
	Sumatra TT	České vysoké učení technické v Prahe, FEL	krizik.felk.cvut.cz/sumatra/
	TANAGRA	Univerzita Lumière Lyon, Francúzsko	http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html
	Weka	Univerzita Waikato, Nový Zéland	www.cs.waikato.ac.nz/ml/weka/

Štvrtý ročník nezávislého dataminingového prieskumu spoločnosti Rexer Analytics vyhodnotil najpoužívanejšie nástroje pre dolovanie dát a spokojnosť ich užívateľov. Prieskumu, ktorý potvrdil dôležitú úlohu dataminingu v procesoch spoločností, sa zúčastnilo 735 respondentov zo 60 krajín. Výsledok ankety na otázky: „Aké dataminingové / analytické nástroje ste použili? Ktorý softvérový balík používate najčastejšie?“, je zobrazený na obrázku 5.



Obr. 5 *Prieskum používaných softvérových balíkov pre dolovanie dát* (56)

Ďalej v tomto prieskume získala STATISTICA najvyššie hodnotenie spokojnosti (pozri obrázok 6). Užívatelia R a IBM SPSS Modeler sú tiež veľmi spokojní. Tieto tri nástroje dosiahli najvyššie hodnotenie spokojnosti aj v roku 2009. STATISTICA získala silné hodnotenie po všetkých stránkach (ako napr. kvalita a presnosť modelu; množstvo dostupných algoritmov; spoľahlivosť / stabilita softvéru; možnosti manipulácie s dátami; jednoduché používanie; schopnosť automatizovať opakujúce sa úlohy; kvalita užívateľského rozhrania; schopnosť spracovávať veľmi veľké súbory dát; dobré variabilné zistenia, profilovanie a výber; kvalita výstupu / ľahká interpretácia; rýchlosť; náklady na softvér a pod.), preto sme zvolili tento nástroj na riešenie cieľa monografie.



Obr. 6 *Prieskum spokojnosti užívateľov s nástrojmi pre dolovanie dát (56)*

1.3.1.1 STATISTICA

Produkt STATISTICA od firmy StatSoft Inc. je komplexný a efektívny systém, užívateľsky pohodlný nástroj pre celý proces získavania znalostí - od zberu dát z produkčných databáz, cez ich transformáciu a úpravu, dolovanie dát, až po vyhodnotenie získaných záverečných výstupov. Obsahuje rozsiahly výber analytických techník, široký výber algoritmov pre rôzne typy neurónových sietí, interaktívnych regresných a klasifikačných stromov, viacrozmerného modelovania, združovania a sekvenčnej analýzy. Ďalej veľký výber pripravených obsiahlych dataminingových projektov, ktoré sú kompletne prednastavené špecialistami spoločnosti StatSoft a externými expertmi pre časté dataminingové problémy.

Ľahko ovládateľný grafický interface je založený na metóde drag and drop, ktorá je ľahko použiteľná aj pre začiatočníkov, ale pritom umožňuje okamžitý prístup k používaným skriptom. Modely sú tvorené pomocou ikon, ktoré reprezentujú analytické uzly (moduly).

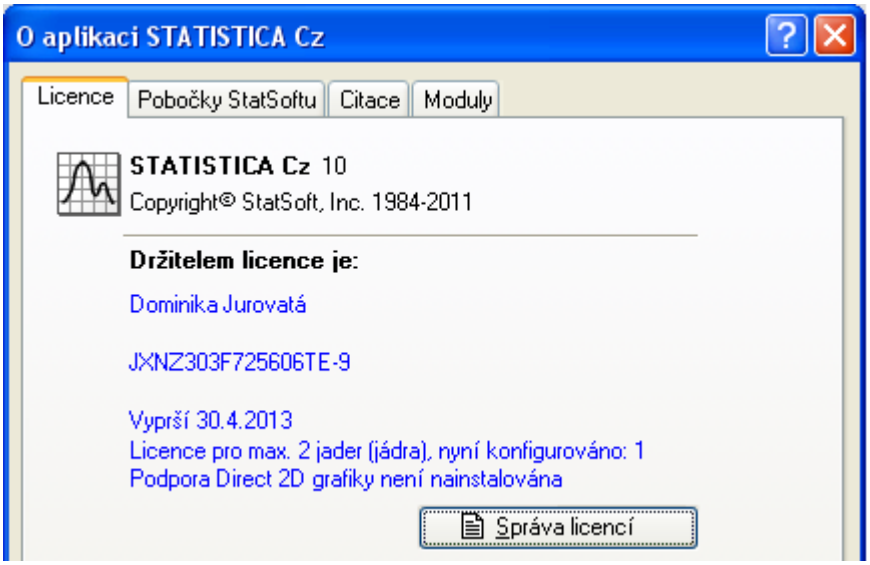
Poskytuje efektívne predspracovanie, čistenie a filtrovanie dát, nástroje na efektívnu funkciu výberu z tisícov kandidátov na prediktora, možnosti zlúčenia viacerých zdrojov údajov, zjednotenie dát na základe niekoľkých kritérií vrátane časových známok na rozdielnych intervaloch (dátová agregácia). Zaoberá sa odľahlými aj chýbajúcimi dátami, odstraňovaním duplicitných záznamov, atď. (63).

Modul STATISTICA Data Miner – datamining sa v mnohom podobá prieskumnej analýze dát (nemáme dopredu danú hypotézu, ktorú by sme prostredníctvom dát overovali, používajú sa podobné metódy a modely ako v prieskumnej analýze dát), avšak líši sa predovšetkým svojim účelom. Pri prieskumnej analýze dát je našim cieľom vytvoriť si istú predstavu

o zákonitostiach skrývajúcich sa v dátach a porozumieť im. V dataminingu je cieľom vytvoriť trefný model, ktorý bude poskytovať užitočné predikcie. Pritom nám už toľko nezáleží na tom, aby bol model zrozumiteľný človeku, a tak sa môžu uplatniť aj modely typu čierna skrinka (neurónové siete, support vector machine a pod.).

Spoločne so všetkými modulmi je k dispozícii aj množstvo techník, napr. na posúdenie kvality naučených modelov alebo na ľahké a rýchle použitie naučených modelov uložených v jazyku PMML (44).

Nástroj STATISTICA Data Miner verziu Cz 10 od firmy StatSoft použijeme na realizáciu procesu získavania znalostí (62), licencia je uvedená na obrázku 7. Vybrali sme si ho na základe vyššie uvedeného prieskumu spoločnosti Rexer Analytics.



Obr. 7 *Licencia STATISTICA Data miner*

1.3.2 Typy úloh dolovania dát

Úlohy v dolovaní dát sa rozčleňujú do niekoľkých typov (26):

- *Exploračné analýzy dát* – podstatou je preskúmať dáta bez predchádzajúcej znalosti, ktorá by určitým spôsobom naše hľadanie usmerňovala. Využívajú sa rôzne grafické metódy alebo špeciálne techniky.
- *Deskriptívne úlohy* – podstatou tohto typu úlohy je určitým spôsobom opísať celú dátovú množinu. Na riešenie deskriptívnych úloh sa používa napr. zhluková analýza (cluster analysis), prípadne modelovanie závislostí medzi premennými (dependency modeling).

- *Prediktívne úlohy* – cieľom je predpovedať hodnotu určitej veličiny na základe znalosti hodnôt ostatných veličín.
- *Hľadanie vzorov a pravidiel* – podstatou je hľadanie určitých vzťahov a vzorov správania sa v dátach.
- *Hľadanie podľa vzorov* – ide o rozpoznávanie vzorov v dátach na základe vopred danej šablóny. Využíva sa napr. metóda podobnosti vektorov.

Monografia je ďalej zameraná na predikciu, preto je podstatné uviesť len bližšiu špecifikáciu prediktívnej úlohy. Cieľom prediktívneho prístupu je na základe známej množiny dát, obsahujúcej kompletné atribúty, predpovedať výskyt hodnôt niektorých atribútov na dátach, ktorým budú tieto atribúty chýbať. Výstupom je najčastejšie tzv. model, čo môže byť súbor pravidiel, ktoré na základe hodnôt známych atribútov určujú hodnoty hľadaných atribútov (klasifikačné pravidlá, rozhodovacie stromy), prípadne dátová štruktúra naučená robiť predikciu týchto atribútov (neurónová sieť).

Prediktívne úlohy dolovania v dátach môžeme podľa typu predikovanej hodnoty rozdeliť na dve skupiny, klasifikáciu a predikciu.

Klasifikácia – nazýva sa aj kategorická predikcia (modeluje a predpovedá nominálne atribúty – triedy), predikuje kategorické označenia tried. Klasifikuje dáta (konštruuje model) na základe trénovacej množiny a daných zaradení do tried v klasifikačnom atribúte. Skonštruovaný model sa potom využíva na klasifikáciu nových príkladov. Najznámejšie používané techniky klasifikácie sú:

- Rozhodovacie stromy,
- K-najbližších susedov,
- Bayesovská klasifikácia,
- Klasifikačná analýza - klasifikačné pravidlá.

Proces klasifikácie sa skladá z dvoch krokov:

1. Učenie – v tomto kroku sa vytvorí klasifikačný model pomocou trénovacích dát, čo sú také vzorky dát, u ktorých poznáme výsledok klasifikácie, t. j. triedu, do ktorej patria. Z tohto dôvodu sa toto učenie nazýva učenie s učiteľom (supervised learning). Na prvý krok procesu klasifikácie sa môžeme pozerať aj ako na nájdenie takého mapovania alebo funkcie, pre ktorú bude platiť y = f(X), kde X je vektor známych hodnôt atribútov vzorky a y je predikovaný

atribút (label). Nájdené mapovanie alebo funkcia nám následne niekoľkodimenzionálny priestor rozdelí na jednotlivé triedy.

2. Vlastná klasifikácia – použitie vytvoreného modelu na klasifikáciu nových dát, pri ktorých label nie je známy. Podstatným krokom predchádzajúcim použitiu modelu je evaluácia modelu, t.j. určenie presnosti, s akou predikuje jednotlivé triedy. Táto evaluácia prebieha na testovacej množine, ktorej prvky sú obyčajne náhodne vybrané z celej množiny príkladov. Evaluácii a vyhodnocovaniu presnosti klasifikátora sa venuje samostatná podkapitola 1.3.4.

Na klasifikáciu príkladov do tried môžeme použiť rôzne techniky, nasledujúca podkapitola bližšie približuje nami vybrané metódy a techniky, ktoré sme sa rozhodli využiť v práci.

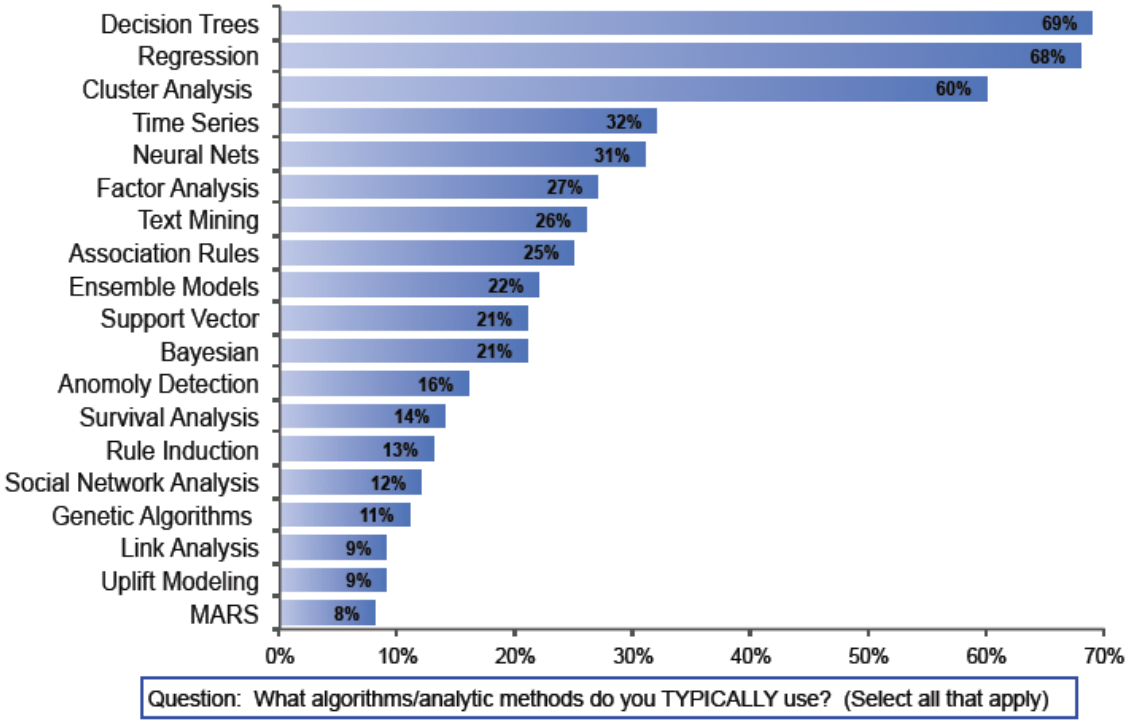
Predikcia – nazývaná numerická predikcia (modeluje a predpovedá numerické hodnoty) modeluje funkcie spojitých premenných, t .j. predikuje neznáme alebo chýbajúce hodnoty spojitého atribútu. Proces numerickej predikcie pozostáva z rovnakých fáz ako kategorická predikcia – z procesu učenia a z vlastnej predikcie. Na predikciu numerických hodnôt sa v oblasti štatistiky používa regresná analýza, na riešenie tohto problému sa dajú použiť aj iné metódy. Regresia sa často chápe ako synonymum k numerickej predikcii, ale Han a spol. (25) zdôrazňujú, že predikovanie numerickej hodnoty automaticky neznamená len použitie regresnej analýzy, ale napríklad aj metódy strojového učenia.

Najznámejšie používané techniky predikcie sú založené na regresii:

- Lineárna a viacnásobná regresia.
- Nelineárna regresia.
- M5´ - modelové stromy.
- Učenie založené na inštanciách (K-najbližších susedov).
- Neurónové siete.

V súčasnosti existuje množstvo nástrojov na vytváranie prediktívnych modelov. Niektoré využívajú štatistické techniky, ako sú napr. lineárna a logistická regresia. Iné sú založené na neštatistických či zmiešaných technikách, ako sú napr. neurónové siete, genetické algoritmy, klasifikačné a regresné stromy.

Podľa štvrtého ročníka nezávislého dataminingového prieskumu spoločnosti Rexer Analytics (56), naďalej tvoria trojicu základných algoritmov pre väčšinu špecialistov data miningu rozhodovacie stromy, regresia a zhluková analýza. Avšak používa sa široká škála algoritmov (obrázok 8).



Obr. 8 *Prieskum využívaných techník, algoritmov dolovania dát* (56)

Taktiež prieskumom vykonaným KDnuggets (tabuľka 3) boli rozhodovacie stromy, regresia a zhluková analýza označené za najpoužívanejšie metódy v roku 2011.

VYUŽÍVANIE METÓD / ALGORITMOV V ROKU 2011(35) Tabuľka 3

Which methods/algorithms did you use for data analysis in 2011? [311 voters]			
Decision Trees/Rules (186)	59.8 %	Neural Nets (84)	27.0 %
Regression (180)	57.9 %	Boosting (73)	23.5 %
Clustering (163)	52.4 %	Bayesian (68)	21.9 %
Statistics - descriptive (149)	47.9 %	Bagging (63)	20.3 %
Visualization (119)	38.3 %	Factor Analysis (58)	18.7 %
Time series/Sequence analysis (92)	29.6 %	Anomaly/Deviation detection (51)	16.4 %
Support Vector (SVM) (89)	28.6 %	Social Network Analysis (44)	14.2 %
Association rules (89)	28.6 %	Survival Analysis (29)	9.32 %

Ensemble methods (88)	28.3 %	Genetic algorithms (29)	9.32 %
Text Mining (86)	27.7 %	Uplift modeling (15)	4.82 %

1.3.3 Vybrané metódy a techniky dolovania dát

Znalosť jednotlivých techník a metód je nevyhnutná na ich správnu aplikáciu na riešený predikčný problém. V tejto časti sú preto podrobnejšie opísané vybrané metódy a techniky dolovania dát, ktoré sú použité v rámci programu STATISTICA pri riešení práce.

Pred definovaním jednotlivých metód je ešte potrebné uviesť niekoľko základných pojmov. Tento úvod stručne popisuje štatistické pojmy, ktoré poskytujú nevyhnutné základy pre ďalšie špecializované odborné znalosti v akejkoľvek oblasti štatistickej analýzy dát (36).

Množina skúmaných objektov sa v štatistike nazýva **súbor**. Predmetom skúmania sú prvky súboru – **štatistické jednotky**, nazývané aj pozorovania alebo objekty. Každá štatistická jednotka je nositeľom mnohých vlastností, ktoré možno skúmať. Tieto vlastnosti sa nazývajú **premenné** (Variables) atribúty alebo štatistické znaky. Premenné sú veci, ktoré meriame, sledujeme alebo s nimi manipulujeme počas výskumu. Líšia sa v tom, akú rolu zohrávajú v našom výskume a v spôsobe ich merateľnosti. Premenná u každej štatistickej jednotky nadobúda **hodnotu** (Value). V štatistike sa rozdiel medzi pozorovanou a predikovanou hodnotou nazýva **rezíduum** (zvyšok).

Dôležité je rozlišovať typ premenných, ako základné členenie sa uvádzajú:

1. číselné premenné (kvantitatívne) – v programe STATISTICA označované ako „continuous - spojité“,

2. kategorické premenné (kvalitatívne, slovné) – v programe tzv. „categorical“.

Toto jednoduché delenie často nepostačuje, preto sa používa delenie na štyri typy premenných podľa škály merania. Premenné rozdeľujeme podľa toho, ako dobre sú merateľné na nominálne, ordinálne, intervalové a pomerné. Pre naše účely postačí základné rozdelenie premenných.

Zo štatistického hľadiska ďalej premenné delíme na **závislú (vysvetľovanú, cieľovú) premennú** Y a **nezávislú (vysvetľujúcu) premennú** X, nazývanú tiež **prediktor**. Závislú premennú sa snažíme pomocou modelu vysvetliť na základe vysvetľujúcich premenných.

Na základe uvedených prieskumov spoločností Rexer Analytics a KDnuggets, samozrejme aj s ohľadom na úlohy dolovania dát a samotné dostupné dáta sme sa pre realizáciu predikcie

rozhodli použiť nasledovné najviac používané predikčné metódy a techniky: rozhodovacie stromy (CandRT, CHAID, Boosting Trees, Random Forest, MARSplines), neurónové siete (SANN), strojové učenie (Support Vector Machine, K-Nearest Neighbour a Naive Bayes) a viacnásobnú regresiu (Multiple Regression). Pri každej metóde je uvedený opis pre klasifikáciu a zároveň numerickú predikciu (regresiu), ak je možné riešiť obe úlohy danou metódou.

1.3.3.1 Rozhodovacie stromy

Rozhodovací strom sa skladá z koreňa, ktorý predstavuje celý súbor a postupne prebieha vetvenie do ďalších uzlov – strom rastie. Uzly, ktoré sa už ďalej nedelia, sa označujú ako terminálne uzly alebo listy. Stromy sú binárne alebo nebinárne podľa toho, či sa vetvia na dva alebo viacej vetiev. Rozhodovacie stromy môžeme rozdeliť podľa typu závislej premennej na klasifikačné a regresné (36).

I. CandRT (*Classification and Regression Trees*)

Je základným predstaviteľom klasických binárnych rozhodovacích stromov. Stromy typu CandRT sú vhodné pre kategorické aj regresné úlohy (8). Na začiatku tvorby stromu patria všetky pozorovania súboru do jedného uzla resp. koreňa. Následne sú tieto pozorovania rozdelené do dvoch dcérskych uzlov na základe hodnoty *a* prediktora *X*, ktoré sú ďalej delené opäť binárne na ďalšie uzly.

Priradenie hodnoty terminálnemu uzlu v strome CandRT

Pri klasifikačnom strome je každému uzlu, vrátane koreňového priradená výsledná kategória závislej premennej. Výslednou kategóriou je tá, ktorá má v danom uzle najväčšie zastúpenie. Nové pozorovanie je potom klasifikované podľa kategórie uzla, do ktorého je stromom zaradené. Môže sa stať, že po rozdelení do dvoch terminálnych uzlov bude obom uzlom priradená rovnaká kategória, najmä ak je podiel kategórií premennej *Y* nevyrovnaný. Výhodu teda môžu mať kategórie, ktoré sú v premennej *Y* viac zastúpené, pretože je väčšia pravdepodobnosť, že budú mať po rozdelení v uzle väčší počet hodnôt než menej početná kategória. V takom prípade je možné použiť váženie jednotlivých kategórií (36).

U regresných stromov je finálnemu uzlu priradený priemer hodnôt závislej premennej a ich variabilita. Z variability si môžeme urobiť predstavu, ako „presná“ je výsledná predikovaná hodnota, inými slovami, v akom intervale sa môže pohybovať. Pri predikcii stromom je novej vzorke priradená hodnota priemeru uzla, do ktorého je zaradený. Priradením obyčajného

priemeru finálnym uzlom prichádzame pri predikcii o pôvodný rozsah hodnôt závislej premennej. Počet výsledných predikovaných hodnôt bude totiž rovný počtu terminálnych uzlov. Výsledná nespojitá plocha je veľkou nevýhodou regresných stromov. Nespojitosť sa dá odstrániť napríklad preložením dát lineárnym regresným modelom v každom terminálnom uzle. Závislou premennou pri regresii sú hodnoty *yi* v danom uzle v závislosti od hodnôt *xi* prediktora *X*, ktorý bol použitý na rozdelenie do tohto dcérskeho uzla. Prediktor *X* však musí byť spojitý. Pre každý terminálny uzol tak dostaneme regresnú rovnicu, pomocou ktorej sa dá predikovať rozsah hodnôt v danom uzle. Tento postup je pomerne časovo náročný, terminálne uzly navyše nemusia obsahovať dostatočný počet vzoriek pre regresiu, poprípade nemusia byť nájdené žiadne vzťahy. Túto nevýhodu regresných stromov však riešia iné stromové metódy, napr. regresné lesy alebo metóda MARS. Regresné stromy sa preto používajú častejšie ako explanátorná technika na vysvetlenie vzťahu závislej premennej a prediktora než na predikciu. Ďalšie významné použitie je pre nájdenie medznej hodnoty, nazývanej diskriminačná hladina (*thresholdvalue)* pri rozdelení závislej premennej (36).

Algoritmus rastu stromu CandRT pre spojité prediktory

Najprv sa zoradia hodnoty každého prediktora od najmenšieho po najväčší. Prejdú sa všetky hodnoty prediktora *X* a spočíta sa kriteriálna štatistika všetkých možných rozdelení premennej *Y* na dva potenciálne dcérske uzly. Pokiaľ je deliaca hodnota *a* prediktora *X* väčšia alebo rovná hodnote *xi*, pozorovaná *yi* patrí do ľavého uzla, inak do pravého (poprípade naopak). Hodnota *a*, pre ktorú je kriteriálna štatistika minimálna, je vybraná ako najlepšie možné delenie závislej premennej *Y* pomocou daného prediktora. Pre každý prediktor tak získame jednu hodnotu (najlepšie potenciálne rozdelenie) kriteriálnej štatistiky. Následne je vybraný prediktor s najnižšou hodnotou kriteriálnej štatistiky a hodnota *a* je použitá na rozdelenie súboru (hodnôt *yi*) do dvoch dcérskych uzlov (36).

Pravidlá pre zastavenie rastu stromu CandRT (stopping rules)

Strom nemôže rásť donekonečna. Jeho maximálna veľkosť je daná veľkosťou súboru. Existujú určité pravidlá, kedy sa rast stromu zastaví. Strom sa zastaví sám v týchto prípadoch:

- terminálny uzol obsahuje len jedno pozorovanie;
- všetky pozorovania v uzle majú rovnakú hodnotu všetkých prediktorov;
- všetky pozorovania v uzle majú rovnakú hodnotu závislej premennej.

Strom môžeme v raste obmedziť nastavením niektorých parametrov a k ďalšiemu rozdeleniu nedochádza, pokiaľ sú dosiahnuté zadané hodnoty:

- maximálny počet vetvenia daného stromu;

• maximálny počet pozorovaní v koncovom uzle;

• frakcia pozorovania v uzle, ktorá už nemôže byť oddelená;

• veľkosti chyby v potenciálnych dcérskych uzloch - napríklad uzol sa nerozdelí, pokiaľ stredná kvadratická chyba (MSE) alebo percento nesprávne klasifikovaných vzoriek (percent disagreement) v dôsledku rozdelenia prekročí určitú hranicu (36).

II. CHAID (*Chi-squared Automatic Interaction Detector*)

Strom CHAID bol navrhnutý pre kategorické premenné (34). Je často využívaný v komerčných sférach, predovšetkým v marketingu. CHAID je strom nebinárneho typu, t.j. uzol môže byť rozdelený na väčší počet dcérskych uzlov než dva. Interpretácia väčšieho počtu uzlov je však zložitejšia než pri binárnych stromoch. Po prvom delení tiež nemusí zostať dostatok pozorovaní na vytvorenie ďalších „poschodí" stromu. Táto technika je preto vhodnejšia pre väčšie dátové súbory. Kriteriálnou štatistikou pre vetvenie je $\chi 2$ –test, čo vyplýva aj z názvu metódy. $\chi 2$ –test je použitý na zistenie nezávislosti v kontingenčnej tabuľke, ktorá je tvorená kombináciou kategórií závislej premennej a prediktora . Ak sú Y a X nezávislé, má testová štatistika približne Pearsonovo $\chi 2$ rozdelenie s $\upsilon = (r\text{-}1)(s\text{-}1)$ stupňami voľnosti, kde r je počet riadkov a s je počet stĺpcov kontingenčnej tabuľky. Nezávislosť v kontingenčnej tabuľke znamená, že sa obidve premenné navzájom neovplyvňujú v hodnotách, ktoré nadobúdajú. Hypotéza nezávislosti javov je tu nulovou hypotézou $H0$. Pearsonov $\chi 2$ –test je často označovaný ako test dobrej zhody.

Rast stromu sa zastaví, pokiaľ sú dosiahnuté nasledujúce pravidlá (36):

- Už nie je možné nájsť žiadne významné rozdelenie.
- Všetky pozorovania závislej premennej v uzle majú rovnakú hodnotu alebo identickú hodnotu pre každý prediktor.
- Pokiaľ je dosiahnuté užívateľom definovaných nastavení, ktoré sa týkajú:
  a) parametrov veľkosti stromu ako je nastavenie počtu terminálnych uzlov alebo vetiev;
  b) počtu pozorovaní v uzle, ktoré je menšie než minimum stanovené užívateľom alebo počtu pozorovaní, ktoré by po rozdelení viedlo k dcérskym uzlom s menším počtom pozorovaní, než je definované užívateľom.

Lesy

S myšlienkou použitia viacerých stromov na spresnenie klasifikácie a predikcie prišiel prvý krát Breiman (1996) a vytvoril tak les. Lesy sú teda nadstavbou nad rozhodovacími stromami. Môžu byť použité na klasifikáciu aj regresiu a odstraňujú niektoré problémy, ktoré nastávajú

pri použití stromov, najmä ich nestabilitu. Les je zložitejší a menej prehľadný než jeden strom a informácie jednotlivých stromov sa stratia v rámci celého lesa. Niekedy je táto metóda vďaka svojej „nepriehľadnosti“ označovaná ako čierna skrinka. Strata jednoduchosti však nie je dôvodom na nepochopenie pozadia metódy. Existuje niekoľko typov lesov. Najviac využívanou metódou je Random Forest (7).

III. Random Forest

Táto technika bola vyvinutá pre súbory obsahujúce veľké množstvo prediktorov a veľmi dobre funguje aj na malých dátových súboroch. Random Forest je určený na klasifikáciu aj regresiu. Náhodné lesy zvládnu kombinovať viac kategorických premenných a číselné premenné v jednej analýze (36).

Náhodný les sa skladá zo súboru jednoduchých stromov $T_1,...,T_N$, ktorých klasifikačná alebo regresná funkcia sa dá vyjadriť ako $h(X, \Theta_1),...,h(X, \Theta_N)$, kde h je funkcia, X je prediktor a $\Theta_1,...,\Theta_N$ sú nezávisle rovnako rozdelené náhodné vektory. Pre metódu Random Forest sa používajú binárne stromy typu CandRT. Podobne ako pri tvorbe jednotlivých stromov sa aj tu používa rozdelenie na testovací a trénovací súbor. Trénovacie súbory pre jednotlivé stromy *Ti* sú tzv. **bootstrapové výbery** z dátového súboru *L*. Bootstrapové výbery sú náhodnými **výbermi s opakovaním** o veľkosti *n*. Tvorbu výberu s opakovaním výberu si je možné jednoducho predstaviť ako losovanie čísiel. Množinou všetkých pozorovaní by bola množina všetkých čísel, z ktorých sa bude losovať. Pri losovaní sa náhodne vytiahne určitý počet čísiel. Každé číslo je po jeho vylosovaní vrátené späť, môže teda byť opäť vybrané. Dopredu stanovený počet vylosovaných čísel tvorí bootstrapový výber. Po losovaní sú všetky čísla vrátené späť a nový ťah (nový výber) začína opäť so všetkými číslami. Takto je možné pri každom ťahu vybrať rovnaký počet čísiel bez toho aby sa nám znižovala veľkosť množiny pôvodných pozorovaní. Týmto spôsobom sa dajú rozdeliť aj veľmi malé súbory na veľký počet trénovacích a testovacích súborov.

Pozorovania, ktoré sú v *i*-tom bootstrapovom výbere *Li,* sa použijú pri tvorbe stromu *Ti* (trénovací súbor), naopak pozorovania, ktoré sa do toho výberu nedostali (testovací súbor) sú použité k odhadu jeho chyby. Odhady chyby na testovacom súbore sa nazývajú *oob* (*out-of-bag, out of bootstrap sample)* odhady. Celkový počet *oob* pozorovaní tvorí 1/3 dátového súboru.

Pri použití lesov pre klasifikáciu získame z každého stromu informáciu o zaradení každého pozorovania do výslednej kategórie. Výsledok klasifikácie lesa je daný väčšinovým hlasovaním všetkých stromov. Pokiaľ je les použitý pre regresiu, predikcia každého pozorovania je jednoducho priemerom zo všetkých stromov.

Náhodný les zvyšuje presnosť (znižuje skreslenie) tým, že necháva narásť stromy dosť veľké, zároveň udržuje znesiteľnú variáciu kombinovaním výsledkov jednotlivých stromov (väčšinové hlasovanie/priemerovanie). Oproti ostatným lesom je tu však snaha zaistiť tiež nízku koreláciu medzi jednotlivými stromami. Pokiaľ totiž tvoríme výbery s opakovaním, ktoré nie sú navzájom nezávislé, budú výsledné stromy korelované. Táto „podobnosť" jednotlivých stromov môže viesť k nadhodnoteným výsledkom klasifikácie či predikcie. Zníženie korelácie medzi stromami sa dosiahne náhodným výberom len určitého počtu prediktorov. Pre každý strom sa tak najlepšie vetvenie pre daný uzol hľadá len z *m* prediktorov $X_1,..,X_M$. Náhodný les používa náhodný výber pozorovaní a zároveň náhodný výber prediktorov (36).

Algoritmus tvorby lesa

1. Vytvor bootstrapový podsúbor *Li* o veľkosti *N* - trénovací súbor.

2. Vyber náhodne *m* prediktorov.

3. Vytvor strom *Ti* na bootstrapovom súbore *Li* len s použitím *m* náhodne vybraných prediktorov (rovnako ako bolo popísané v metóde CandRT, hľadáme najlepšie rozdelenie daného uzla medzi prediktormi na dva dcérske uzly). Rast stromu sa zastaví, až strom dosiahne minimálne hodnoty veľkosti uzla.

4. Zaraď *oob* pozorovanie (testovací súbor) vytvorený stromom a urči výslednú klasifikačnú triedu (kategóriu) alebo predikciu všetkých *oob* pozorovaní. (Krok 1-4 sa opakuje do konečného počtu stromov v lese.)

5. Spočítaj celkový výsledok klasifikácie/predikcie celého lesa väčšinovým hlasovaním / priemerovaním.

Pre algoritmus lesa je potrebné vybrať správny počet premenných (*m*) pre náhodný výber a počet stromov (*ntree*) v lese. Určenie týchto parametrov je do istej miery experimentálne a vyžaduje skúsenosti. Klasickou cestou je vykonanie radu experimentov s rôznym nastavením týchto parametrov k získaniu lesa, ktorý má najmenšiu celkovú chybovosť. Vzhľadom na časovo náročné testovanie (zvlášť ak ide o súbory obsahujúce tisíce záznamov) je vhodné vybrať taký počet stromov, ktorý bude dostačujúci pre optimálnu klasifikáciu. Na začiatku teda nastavíme počet stromov v lese na vyššiu hodnotu (napr. 20 násobok počtu prediktorov). Po určitom čase začínajú stromy konvergovať k správnej hodnote *oob* odhadu. Minimálna veľkosť lesa sa dá určiť ako počet stromov, kde sa chyba *oob* odhadu s pribúdajúcimi stromami už nemení. Ďalším parametrom je počet náhodne vybraných prediktorov *p*. Pre náhodné lesy je odporučené nasledujúce nastavenie:

• pre klasifikáciu je hodnota $m = \sqrt{p}$ a minimálna veľkosť uzla je jedna;

• pre regresiu je hodnota $m = p/3$ a minimálna veľkosť koncového uzla je päť.

Vyššie uvedené hodnoty slúžia ako defaultné nastavenie vo väčšine softvérov. V praxi však určenie počtu prediktorov závisí od riešeného problému a parameter *m* je vhodné zvoliť podľa výsledkov testovania modelu s rôznym nastavením. Vyberieme také *m*, pri ktorom má výsledný les najmenšiu chybovosť. Vzhľadom na to, že stromy sa nedajú pretrénovať, je počet prediktorov najdôležitejšou hodnotou, ktorú musíme zvoliť, keďže počet stromov nás obmedzuje len časovo (36).

IV. Boosting Trees

Jeho základný princíp je v tom, že opakovanou zmenou váh jednotlivých pozorovaní vytvára aj zo slabých modelov modely veľmi silné. Výsledkom je potom skupina slabých modelov, z ktorých každý je expertom na jednotlivé časti vstupného priestoru. Na vytvorenie lesa používa, rovnako ako Random Forest, boostrapové výbery (13). Rozdiel oproti náhodným lesom v tvorbe jednotlivých stromov je ten, že nie sú náhodne vyberané prediktory, ale len pozorovania. **Boosting** les je tvorený veľkým množstvom stromov a výsledok, zaradenie pozorovania do kategórie, je daný väčšinovým hlasovaním rozhodovacích stromov, resp. priemernou hodnotou zo všetkých stromov. Každému pozorovaniu je priradená určitá váha podľa toho, ako dobre je možné ho klasifikovať. V prvom kroku je všetkým pozorovaniam priradená rovnaká váha, následne sa dátový súbor upraví podľa výsledku klasifikácie a väčšia váha sa priradí pozorovaniam, ktoré dopadli horšie. Týmto spôsobom sa dá výrazne zvýšiť presnosť a predikčná sila modelu. Dajú sa takto „natrénovať“ lesy aj na pozorovania, ktoré môžu byť odľahlé a stromy sa potom snažia vysvetliť aj šum v dátach. Toto nebezpečenstvo nastáva najmä pri súboroch s veľmi veľkou variabilitou. Boosting bol vytvorený pre klasifikačný problém, ale používa sa aj na regresiu (36).

V programe STATISTICA sa algoritmus Boosting stromu riadi postupom od Friedmana (19, 20).

Pre regresné problémy sa vykonávajú nasledujúce kroky:

1. Pre M stromov sa vyberie náhodná vzorka pozorovaní.
2. Vypočítajú sa predpokladané hodnoty pre pozorovania v tejto vzorke.
3. Nasadí sa regresný strom požadovanej zložitosti na zvyšky (residuals), STATISTICA používa funkciu chyby najmenších štvorcov (least squares error) pre rast stromu.
4. Aktualizuje priradené váhy.

Okrem toho, program počíta pri každom zvyšovaní kroku predikcie rezíduí pre nezávislé vzorky pozorovaní, priemernú štvorcovú chybu (zostatkovú) pre danú vzorku. Z tejto chybovej

funkcie pre testované dáta program automaticky zistí najlepšie číslo doplnkovej rozpínavosti, t.j. optimálny počet boosted stromov pre dosiahnutie dobrej prediktívnej platnosti. Program automaticky vyberie riešenie, ktoré dáva najmenšiu priemernú plošnú chybu v testovaných dátach (27).

Algoritmus pre klasifikačné problémy pracuje na základe algoritmu 6, ako je to popísané vo Friedman (20), a je v podstate rovnaký ako pre regresné problémy, s nasledujúcimi výnimkami. Namiesto jedného súboru po sebe idúcich regresných stromov program vytvára nezávislé sady boosted stromov pre každú kategóriu závislej premennej. Ďalej, namiesto jednoduchých súborov regresných rezíduí (vypočítaných v každom kroku) program uplatňuje logistickú transformáciu na predikované hodnoty. Pre výpočet konečnej klasifikácie, je logistická transformácia znovu aplikovaná na regresnú predikciu pre každú kategóriu nad za sebou idúcimi stromami.

V. MARSplines (*Multivariate Adaptive Regression Splines*)

Metóda MARS sa nachádza na rozhraní medzi stromovou technikou a parametrickou regresiou (21). Odstraňuje určité nedostatky binárnych regresných stromov, predovšetkým nespojitosti odhadnutých hodnôt závislej premennej (t. j. že finálny počet predikovaných hodnôt je rovný počtu terminálnych uzlov). Prediktory môžu byť spojité i kategorické. Výsledkom metódy je regresná rovnica, chýba teda klasická stromová štruktúra. Interpretácia výsledkov pri veľkom počte premenných môže tak byť zložitá. Zatiaľ čo u ostatných rozhodovacích stromoch pre rozdelenie uzla postačovala hodnota prediktora, ktorý bol určený pomocou kriteriálnej štatistiky, tu je potrebná polpriamka. Na rozdelenie pozorovaní závislej premennej sa teda nepoužíva konštanta, ale lineárna aproximácia.

Lineárna aproximácia je uskutočnená pomocou tzv. lineárnych splinov (*linear splines*). Spliny definované v metóde MARS sú po častiach lineárne funkcie $(x - t)_+$ a $(t - x)_+$ s uzlom v bode t, kde + označuje kladnú časť funkcie. Dvojice týchto funkcií sú označované ako zrkadlové páry a samotné funkcie $(x - t)_+$ a $(t - x)_+$ ako bázové funkcie (*basis functions*).

$$(x-t)_+ \begin{cases} (x-t) & ak\ x > t \\ 0, & inak \end{cases} \qquad (t-x)_+ \begin{cases} (t-x) & ak\ x < t \\ 0, & inak \end{cases} \qquad [1]$$

Často sa používa alternatívny zápis $max(0, x - t)$ a $max(0, t - x)$.

Najčastejšie použitie majú spliny v interpolačných úlohách. Pokiaľ prekladáme dáta polynómom vyššieho stupňa, dochádza často k osciláciám. Namiesto takého polynómu sa dá

použiť funkcia, ktorá bude polynómom nízkeho stupňa a bude prekladať dáta po častiach tak, aby na seba jednotlivé časti nadväzovali. Práve tieto funkcie sa nazývajú spliny (36).

Viacrozmerné adaptívne regresné spliny (MARSplines) sa napriek názvu dajú využiť pre klasifikačné aj regresné problémy s kategoriálnymi aj spojitými vstupnými veličinami.

MARSpliny sú neparametrickou modelovacou procedúrou a nekladú žiadne nároky na vzťahy medzi vstupnými a výstupnými premennými. Konštrukcia modelu namiesto toho spočíva v určení koeficientov a počtu jednoduchých bázových funkcií. Ich konštrukcia je svojim spôsobom podobná konštrukcii stromov typu CandRT. Veľmi dobre sa uplatňujú pri riešení úloh s veľa vstupnými premennými, kde by iným metódam mohlo robiť problémy tzv. prekliatie rozmernosti, t.j. prudko narastajúca výpočtová zložitosť pri použití teoreticky optimálnych postupov (4).

Algoritmus je veľmi podobný predčasnému postupnému výberu (*forward stepwise selection*) vysvetľujúcich premenných v regresnom modeli, namiesto premenných sa ale vyberajú lineárne spliny. Začína sa s nulovým modelom (bez prediktorov). Postupne sa pridávajú jednotlivé členy do rovnice (bázové funkcie), len tie, ktorých príspevok k variabilite vysvetlenej modelom je štatisticky významný. Tento príspevok sa určuje na základe zníženia reziduálneho súčtu štvorcov modelu, resp. súčtu štvorcových odchýlok hodnôt y_i závislej premennej od hodnôt $\hat{y}_i$ odhadnutých modelom.

Algoritmus metódy MARS

1. Algoritmus začína s konštantnou funkciou $hm(X) = 1$.

2. Vytvoria sa spliny (zrkadlové páry) so svojim stredom (uzlom *t*) v každej hodnote *xij* pre každý prediktor *Xj*. Získame množinu všetkých „kandidátskych“ bázových funkcií *C* a model je tvorený prvkami z tejto množiny alebo ich kombináciou.

3. Z množiny *C* sú do modelu pridávané pomocou postupného výberu významné bázové funkcie, ktoré znižujú reziduálnu chybu modelu. Proces postupuje hierarchicky, významné interakcie sú pridávané do modelu len z kombinácie bázových funkcií, ktoré už boli do modelu vybrané (z kroku 1 - 3 sa získa rovnica s vybranými členmi).

4. Posledným krokom algoritmu je procedúra spätného odstraňovania. Z rovnice sú odstránené tie členy, u ktorých po ich odstránení dôjde k najmenšiemu zvýšeniu chyby modelu. Spätné odstraňovanie je vykonávané pomocou krosvalidácie. Hodnota *GCV* (*generalized cross-validation*) je spočítaná pre rôzne veľkosti modelu (s rôznym počtom členov v rovnici) a je vybraný model, pre ktorý je hodnota *GCV* minimálna. (36) GCV sa určí nasledovne:

$$GVC(\lambda) = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{(1 - M(\lambda)/N)^2}, \qquad [2]$$

kde N je počet pozorovaní, $\hat{y}_i$ je hodnota závislej premennej odhadnutá modelom a $M(\lambda)$ je parameter zložitosti modelu, ktorý má tvar: $M(\lambda) = r + cK$, kde r je počet nekonštantných bázových funkcií v modeli, c je konštanta a K je počet uzlov t v modeli, kde už prebehol výber parametrov pomocou vopred pripraveného výberu (36).

1.3.3.2 Strojové učenie

VI. Support Vector Machines

SVM patrí do konceptu v oblasti štatistiky a informatiky pre súbor metód strojového učenia. Tieto sa snažia využiť výhody poskytované efektívnymi algoritmami pre nájdenie lineárnej hranice, ktorá oddeľuje pozorovania vo vstupnom priestore. Zároveň sú schopné modelovať zložité nelineárne funkcie.

SVM sú založené na koncepte rozhodovacích rovín, ktoré definujú rozhodovacie hranice. Rozhodovacia rovina je tá, ktorá oddeľuje súbory objektov s rôznymi vlastnosťami triedy (27). Väčšina klasifikačných úloh však nie je taká jednoduchá a často sú potrebné zložitejšie štruktúry, aby bolo oddelenie optimálne, t. j. správne triedenie nových objektov na základe pozorovaní, ktoré sú k dispozícii. Niekedy objekty vyžadujú oddeľovaciu krivku (ktorá je zložitejšia než priamka). Hľadá sa tzv. hyperplocha, ktorá zabezpečí najväčšie oddelenie dvoch tried vstupných vzorov. Podstata separácie dát pomocou SVM sa dá opísať zjednodušene – SVM rozdeľuje p-rozmerné dáta pomocou p-1 rozmernej hyperroviny, a to tak, že maximalizuje vzdialenosť jednotlivých tried dát. Vzdialenosť rozdeľujúcej hyperroviny od najbližších výskytov dát jednotlivých tried má byť čo najväčšia. Za týmto účelom sú najbližšími výskytmi dát vedené tzv. Support Vektory. SVM hľadá takú separačnú hyperrovinu, ktorej vzdialenosť od oboch Support Vektorov je maximálna, a tým maximalizuje medzitriedny rozptyl (28).

Hlavnou myšlienkou metódy je previesť pomocou vhodnej transformácie úlohu klasifikácie do tried, ktoré nie sú lineárne separabilné (t.j. napr. nedajú sa oddeliť priamkou v dvojdimenzionálnom priestore), na úlohu klasifikácie do tried lineárne separabilných. Táto metóda teda v transformovanom priestore hľadá pomocou tzv. kernelových funkcií rozdeľujúcu nadrovinu (tzv. maximal margin hyperplane), ktorá má najväčší odstup od transformovaných príkladov z trénovacej množiny. Tým sa zabezpečí to, že nedôjde k preučeniu a vytvorený klasifikátor bude mať dostatočnú schopnosť generalizácie príkladov. Dobrá separácia je dosiahnutá nadrovinou, ktorá má najväčšiu vzdialenosť k najbližším trénovacím dátovým

bodom akejkoľvek triedy (tzv. funkčná hranica), pretože vo všeobecnosti čím je väčšia hranica, tým je nižšia zovšeobecnená chyba klasifikátora (70).

SVM sú priamo aplikovateľné len na problém klasifikácie do 2 tried. Ak majú byť použité na klasifikáciu do viacerých tried, musí byť použitý algoritmus, ktorý redukuje problém mnohotriedovej klasifikácie na niekoľko binárnych klasifikačných problémov. Existuje aj verzia SVM, ktorá umožňuje riešiť regresné úlohy – niekedy nazývaná aj SVR (Support Vector Regression). Namiesto predikcie tried sa SVR snaží naučiť vzťah medzi vstupnými príkladmi a príslušným výstupom, ktorý má charakter spojitej funkcie.

Pri metóde SVM sa robí najprv projekcia dát do vysokorozmerného priestoru. Táto metóda bola pôvodne vyvinutá ako lineárny klasifikátor. Neskôr bola modifikovaná použitím tzv. Kernelových metód. Jej modifikácia umožňuje aj nelineárne mapovanie dát do priestoru príznakov. Výhodou takého mapovania je, že berie do úvahy štatistické závislosti vyšších rádov. Existuje niekoľko kernelových rovníc, ktoré možno použiť pre modely Support Vector Machines. Patrí medzi ne lineárna, polynomiálna, radiálna bázová funkcia (RBF) a sigmoid (74).

VII. K-Nearest Neighbour

Metóda k-najbližších susedov je prototypová metóda, ktorá si zapamätá všetky vstupné prípady a nové hodnotí tak, že nechá hlasovať k najbližších susedov hodnoteného prípadu v trénovacích dátach (resp. spriemeruje ich výsledky).

Klasifikácia metódou K-najbližších susedov je klasifikácia na princípe učenia sa pomocou analógie. Vzorky z trénovacej množiny majú n číselných atribútov, každá vzorka teda reprezentuje bod v N-rozmernom priestore. Keď chce klasifikátor určiť cieľový atribút neznámej vzorky, hľadá v tomto priestore „*k*“ vzoriek z trénovacej množiny, ktoré sú najbližšie našej neznámej vzorke na základe miery, pomocou ktorej určuje vzdialenosť (napr. Euklidovská, Manhatanovská).

Majme vzorku X , určenú vektorom hodnôt jej n-atribútov $X= (x_1, x_2, \ldots, x_n)$, ktorú potrebujeme klasifikovať (pomocou euklidovskej miery). Euklidovská miera pre vzorku

X a vzorky $Y= (y_1, y_2, \ldots, y_n)$ patriace trénovacej množine je definovaná ako

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad . \qquad [3]$$

Proces predikcie určí zo vzoriek Y práve *k-najbližších* a neznámej vzorke X je priradená hodnota cieľového atribútu (cieľová trieda), ktorá je najčastejšia medzi k-najbližšími susedmi.

Ak je cieľový atribút spojitý, priradí sa neznámej vzorke priemerná hodnota cieľového atribútu medzi jej k-najbližšími susedmi.

Klasifikátory na princípe najbližších susedov sú tzv. inštančné, pretože ako model ukladajú celú trénovaciu množinu pre potreby hľadania susedov *k* neznámym vzorkám. Dôležitú úlohu tu preto zohráva rozumná implementácia a indexovacie techniky, keďže porovnávania v hustom priestore trénovacích vzoriek sú výpočtovo náročné. Na rozdiel od rozhodovacích stromov tieto klasifikátory defaultne uvažujú všetky atribúty s rovnakou váhou, čo môže viesť k chybným klasifikáciam v prípade, ak sa v trénovacej množine vyskytujú odľahlé hodnoty tzv. outliers (vzorky majúce chybné, prípadne nesprávne určené hodnoty atribútov), čím negatívne ovplyvňujú proces klasifikácie nových vzoriek.

Ďalším problémom môže byť tzv. „kliatba rozmerov“. Ide o situáciu, kedy majú vzorky veľký počet atribútov a proces klasifikácie určuje vzdialenosť na základe všetkých atribútov. Atribúty signifikantné pre cieľovú klasifikáciu môžu byť takto ”prebité” ostatnými atribútmi, keďže vzdialenosť dvoch vzoriek majúcich rovnaké hodnoty týchto signifikantných atribútov môže byť stále veľká kvôli hodnotám ostatných atribútov.

Problém tzv. outliers aj kliatby rozmerov je možné riešiť priradením rôznych váh jednotlivým atribútom, signifikantné atribúty sa prenásobia faktorom x >1 a menej podstatné atribúty faktorom x <1, faktor x = 0 plne eliminuje daný atribút z procesu klasifikácie (26, 47).

VIII. Naive Bayes

Je metóda na riešenie len klasifikačných úloh, dokáže predikovať pravdepodobnosť príslušnosti danej vzorky k cieľovej triede. Úlohy sú založené na Bayesovej teoréme, predpokladajú, že efekt atribútu na príslušnosť vzorky k istej cieľovej triede je nezávislý od ostatných atribútov. Často sa kvôli tomuto predpokladu používa termín Naivná Bayesovská klasifikácia.

Algoritmus Naivného Bayesovského klasifikátora pracuje nasledovne:

1. Každá vzorka je reprezentovaná n-dimenzionálnym vektorom vlastností $X = (x_1, x_2, \ldots, x_n)$, patriacim n-atribútom $A_1, A_2, \ldots, A_n$.

2. Majme *m* cieľových tried $C_1, C_2, \ldots, C_m$. Pre neznámu vzorku *X* (hodnota jej cieľového atribútu je neznáma) klasifikátor priradí triedu s najväčšou posteriórnou pravdepodobnosťou podmienenou na *X*. Teda klasifikátor priradí vzorke *X* cieľovú triedu $C_i : P(C_i|X) > P(C_j|X)$ *pre* $1 \leq j \leq m, j \neq i$

3. Hľadá sa maximálne $P(C_i|X)$. Podľa Bayesovej teorémy $P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$.

P(X) je konštanta pre všetky triedy, teda len *P(X|Ci)P(Ci)* je potrebné maximalizovať. Ak nie sú vopred určené apriórne pravdepodobnosti tried *C1, C2, . . . , Cm*, platí:

$P(C_i) = \frac{s_i}{s}$, kde s_i je počet vzoriek v trénovacej množine, prislúchajúcich triede C_i, s je počet všetkých vzoriek v trénovacej množine. Maximalizuje sa len člen $P(X|C_i)$.

4. Keďže pri veľkom počte atribútov by bolo výpočtovo náročné určovať *P(X|Ci)*, je urobený práve predpoklad nezávislosti atribútov pri príslušnosti k danej cieľovej triede, teda:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i). \quad [4]$$

Pravdepodobnosti $P(x_1|C_i), P(x_2|C_i), \ldots, P(x_n|C_i)$ sa dajú určiť z trénovacej množiny, kde: Ak A_k je diskrétny atribút, potom $P(x_k|C_i) = \frac{s_i}{s}$, kde s_{ik} je počet vzoriek z trénovacej množiny, patriacich triede C_i, pre ktoré $A_k=x_k.s_i$ je počet vzoriek z trénovacej množiny patriacich triede C_i.

Ak A_k je spojitý atribút, tak sa predpokladá Gausovo normálne rozdelenie hodnôt, teda:

$$P(x_k|C_i) = g(x_k, \mu_{Ci}, \sigma_{Ci}), \quad [5]$$

kde μ_{Ci} je stredná hodnota a σ_{Ci} rozptyl hodnôt atribútu A_k trénovacích vzoriek z triedy C_i.

5. Na klasifikáciu neznámej vzorky *X* je vypočítaný člen $P(X|C_i)P(C_i)$ pre každú triedu C_i. Vzorke *X* je priradená trieda C_i: $P(X|C_i)P(C_i) > P(X|C_j)P(C_j)$ *pre* $1 \leq j \leq m$, $j=i$, vzorka je teda priradená tej triede C_i, pre ktorú člen $P(X|C_i)P(C_i)$ je maximálny (53, 70).

1.3.3.3 Regresná analýza

Najčastejšou technikou, ktorá sa na riešenie problémov numerickej predikcie používa, je regresná analýza z oblasti štatistiky. Regresná analýza sa používa na modelovanie závislostí medzi nezávislými premennými (regresormi) a závislými premennými (regresantami). V kontexte dolovania dát sú regresormi známe atribúty popisujúce príklady a regresantom je atribút, ktorého hodnotu chceme predikovať. Regresia je metóda, ktorá odhaduje hodnoty jednej premennej na základe znalosti hodnôt druhej premennej. Najjednoduchší typ regresie predstavuje lineárna regresia, kde dáta sú aproximované pomocou regresnej priamky. Viacnásobná regresia je rozšírením lineárnej regresie na viac ako jednu nezávislú predikujúcu premennú (51).

IX. Multiple Regression

Viacnásobná regresia - (tento výraz ako prvý použil Pearson v roku 1908) skúma lineárnu závislosť medzi dvoma a viacerými premennými. Úlohou je odhadnúť β_j v rovnici:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \varepsilon_i \,, \qquad [6]$$

kde y_i – hodnota závislej premennej Y (kritéria) v i-tom pozorovaní;

x_{ij} – hodnota j-tej nezávislej premennej X (prediktora) v i-tom pozorovaní (i = 1, 2, ..., m);

β_j – neznámy regresný koeficient j-tej premennej X (j = 1, 2, ..., n);

ε_i – náhodná chyba i-teho pozorovania.

Parametre regresnej priamky sa najčastejšie získavajú metódou najmenších štvorcov, ktorá zvolí b (odhady neznámych parametrov β) tak, aby sa minimalizoval súčet štvorcov rezíduí. Výberová regresná rovnica sa zapisuje: $\hat{y} = b_0 + b_1 x_{i1}\, \hat{y} + b_2 x_{i2} + \ldots + b_n x_{in}$.

Rezídua e_i sú definované ako: $e_i = y_i - \hat{y}_i$.

Hlavným cieľom tejto metódy je minimalizácia chýb, ktoré predstavujú rozdiely medzi teoretickými (predikovanými) a pozorovanými (nameranými) hodnotami závislej premennej Y. Tieto rozdiely môžu nadobúdať ako kladné, tak aj záporné hodnoty, preto sa umocňujú na druhú a počítajú sa ich súčty (27, 51).

1.3.3.4 Neurónové siete

X. Neural Network

Neurónová sieť je množina spojených vstupno-výstupných jednotiek (umelých neurónov), pričom ku každému spojeniu prináleží určitá váha. Umelé neuróny sú založené na princípe biologických neurónov, ktoré tvoria nervovú sústavu človeka. Vstupné informácie sú vážené váhami. Od sumy vážených vstupných signálov sa odčíta prahová hodnota (threshold) a aktivačnou funkciou (activation function) sa signál transformuje na výstupný signál, ktorý je poslaný na vstup neurónom, s ktorými je daný neurón spojený.

Existuje niekoľko druhov neurónových sietí a algoritmov, ktoré sa používajú na ich učenie. Najpoužívanejší typ neurónovej siete je mnohovrstvová vopred realizovaná neurónová sieť, ktorá v učiacej fáze využíva algoritmus spätného šírenia chyby. Tento typ siete pozostáva z niekoľkých vrstiev neurónov – vstupnej vrstvy, niekoľkých skrytých vrstiev a výstupnej vrstvy (25).

Mnohovrstvové vopred realizované neurónové siete je možné použiť na klasifikáciu aj na predikciu spojitej funkcie (numerickú predikciu). Ak majú dostatok skrytých vrstiev a tréningových príkladov, dokážu aproximovať ľubovoľnú funkciu.

Nevýhodou neurónových sietí je pomerne dlhý čas učenia a slabá interpretovateľnosť. Neurónové siete sú vnímané ako čierna skrinka, pretože z hodnôt váh človek nedokáže extrahovať žiadnu znalosť. Navyše parametre neurónových sietí je nutné často určiť len empiricky. Výhodami neurónových sietí je ich tolerancia voči šumu v dátach a schopnosť modelovať komplexné vzťahy medzi vstupmi a výstupmi. Algoritmy neurónových sietí sa

dajú paralelizovať, čím sa zníži čas potrebný na výpočet (80).

1.3.4 Vyhodnocovanie modelov dolovania dát

Presnosť predikčného modelu vyjadruje mieru schopnosti modelu predikovať nad neznámymi dátami, ktoré neboli použité pri učení (25). K chybným, príliš optimistickým výsledkom by mohlo viesť použitie trénovacích dát na výpočet presnosti modelu. Preto existuje niekoľko techník, ktoré určujú, nad akými dátami trénovať a následne vyhodnocovať presnosť vzniknutého modelu. Medzi najpoužívanejšie patria nasledovné metódy:

1. Holdout - metóda spočíva v náhodnom rozdelení celkovej dátovej skupiny na dve nezávislé časti – trénovaciu a testovaciu. Na prvej z nich prebieha vlastná tvorba modelu (určenie). Potom sa jeho presnosť overuje na zostávajúcej časti.

2. Random sampling – (náhodné vzorkovanie) je rozšírením metódy holdout. Pri tomto spôsobe overovania presnosti sa realizuje n-krát postup predošlej metódy. Celkový výsledok je aritmetický priemer hodnôt získaných počas jednotlivých iterácií.

3. Cross-validation - (krížová validácia, resp. k-fold cross-validation) dáta sú tiež náhodne rozdelené, a to na „k“ disjunktných množín D_1 až D_k približne rovnakej veľkosti. Celý proces učenia a overovania presnosti prebieha v „k“ iteráciách. V každej iterácii je množina D_k použitá pre test a zostávajúce množiny slúžia na učenie. Celková presnosť sa vypočíta ako priemer dosiahnutých presností v jednotlivých iteráciách. Na rozdiel od holdout metódy každé pozorovanie je použité v rovnakom počte na učenie a raz na testovanie.

4. Leave-one-out - ide o extrémny prípad predchádzajúcej metódy, kedy „k“ (počet podskupín) je rovný počtu pozorovaní. To znamená, že v každom opakovaní slúži na tvorbu modelu celá dátová skupina okrem vždy jediného pozorovania. Na ňom je po skončení učenia overená presnosť.

Na testovacích dátach sa následne vyhodnocujú jednotlivé modely a porovnáva sa ich presnosť. Odhadnúť presnosť modelu nám pomáhajú rôzne techniky pre klasifikáciu a iné pre numerickú predikciu (25).

1.3.4.1 Vyhodnocovanie klasifikácie

Pri kategorickej predikcii sa presnosť modelov určuje rôznymi spôsobmi od jednoduchých po zložitejšie. Existuje viacero používaných metrík (37), ktoré sú určené pre klasifikáciu ako napríklad percento správne klasifikovaných, chybovosť, celková správnosť, krivka navýšenia, testy dobrej zhody a iné.

Najjednoduchšia metrika presnosti výsledného klasifikátora je počítaná ako pomer počtu správne klasifikovaných záznamov ku všetkým záznamom, vyjadrený v percentách. Často je uvádzaná aj chybovosť (error rate) klasifikátora, ktorú vyjadríme jednoducho ako *1- Acc(M),* kde *Acc* je presnosť klasifikátora *M*.

Ďalším nástrojom v porovnávaní presnosti jednotlivých tried je celková správnosť *OA* (***overall accuracy***), ktorá udáva pravdepodobnosť, že je pozorovanie správne klasifikované.

$$OA = \frac{n_p}{n}\,, \qquad [7]$$

kde *np* je počet správne klasifikovaných pozorovaní a *n* celkový počet pozorovaní.

Toto meranie však nezohľadňuje rôznu veľkosť skupín ani rozdielnosť oproti náhodnému výsledku, a preto môže ľahko prísť k nadhodnoteniu alebo naopak podhodnoteniu kvality modelu. Korekciu na veľkosť kategórií môžeme urobiť jednoduchou úpravou:

$$OA_{kateg} = \frac{1}{J}\sum_{c=1}^{J}\frac{n_{pc}}{n_c}\,, \qquad [8]$$

kde *J* je celkový počet kategórií, n_{pc} je počet správne klasifikovaných pozorovaní v kategórii *c* a n_c je počet všetkých pozorovaní v kategórii *c*.

Celková správnosť sa používa predovšetkým na porovnanie s ostatnými klasifikačnými metódami alebo pre výber vhodnej metódy. V praxi je však dôležitejšie percento správne klasifikovaných pozorovaní na každú kategóriu (36).

Krivka navýšenia (z angl. lift chart) vyjadruje mieru efektivity prediktívneho modelu. Do grafu sa vynášajú zisky bez použitia modelu (Base line) a zisky s použitím prediktívneho modelu. Poskytuje vizuálny prehľad o užitočnosti informácií poskytnutých jedným alebo

viacerými štatistickými modelmi pre predikciu kategorickej závislej premennej. Je použiteľná pre väčšinu štatistických metód, ktoré počítajú klasifikačnú predikciu pre dvojčlenné alebo viacčlenné kategórie. Klasifikácia je sprevádzaná numerickou hodnotou (nazývanou lift), ktorá vyjadruje, ako veľmi si klasifikátor dôveruje pri svojom rozhodnutí pre dané pozorovanie (37, 63).

Testy dobrej zhody (goodnes of fit) pre klasifikáciu, a teda diskrétne dáta sú štandardne využívané Chi-square statistic, G-square statistic, Percent disagreement (27).

Chi-square (Goodness of Fit) navrhol Karl Pearson v roku 1900 ako jednu z prvých induktívnych štatistických metód vôbec. Test vychádza z kontingenčnej (frekvenčnej) tabuľky a porovnáva rozdelenie početnosti v jednotlivých kategóriách s očakávanými početnosťami. Teda pri predikcii porovnáva pozorované a predikovavné hodnoty. Pearsonov Chi-kvadrát, test dobrej zhody testuje nulovú štatistickú hypotézu, ktorá tvrdí, že početnosti v jednotlivých kategóriách sa rovnajú očakávaným (predikovaným) početnostiam (57).

Pearson Chi-square

$$\chi^2{}_{N-1} = \sum_{i=1}^{N} (E_i - O_i)^2 / E_i \ , \qquad [9]$$

kde N – počet pozorovaných tried,

E_i - počet pozorovaní v pozorovanej triede i, ktoré sú predikované, že patria práve do triedy i (predikované alebo očakávané početnosti frekvencie výskytu) pre pozorovanú triedu i,

O_i - počet pozorovaní patriacich do triedy i (početnosť pozorovaní).

Táto hodnota je rovná 0 (nule), ak klasifikátor je perfektný (t.j., keď očakávané klasifikácie sú zhodné s pozorovanými klasifikáciami). To znamená, že je žiaduce, aby sa hodnota Chi-square blížila k nule (27).

G-square statistic (maximum likelihood Chi-square)

G- square pre test dobrej zhody je alternatíva k predchádzajúcej metrike chi- square. Na výpočet G^2 je využité chi-square rozdelenie. Tvar chi-kvadrátu závisí od počtu voľnosti. Termín maximálna vierohodnosť prvýkrát použil Fisher (1992), je to všeobecná metóda odhadu hodnôt parametrov populácie, ktoré maximalizujú funkciu vierohodnosti. Táto metóda nemá žiadne

požiadavky na nezávislé premenné a tie môžu byť nominálne, poradové (ordinálne) alebo intervalové (resp. spojité). Vzorec pre výpočet G^2 je nasledovný:

$$G^2 = 2\sum_{i=1}^{N} O_i * \ln(O_i/E_i), \qquad [10]$$

N – počet tried,

E_i - počet pozorovaní v pozorovanej triede i, ktoré sú predikované, že patria práve do triedy i (predikované alebo očakávané početnosti (frekvencie výskytu) pre pozorovanú triedu i.,

O_i - počet pozorovaní patriacich do triedy i (početnosť pozorovaní).

Väčšinou sa používa, keď máme jednu nominálnu premennú s dvoma alebo viacerými hodnotami. Nulová hypotéza je, že počet pozorovaní v každej kategórii je rovný počtu predikovaných. Čím väčší je rozdiel medzi pozorovanými a očakávanými, tým väčšia je hodnota G^2 (27, 57).

Percent disagreement

Percentuálny nesúhlas sa počíta ako percento všetkých pozorovaní, pre ktoré predikované klasifikácie nie sú zhodné (nesúhlasia) s pozorovanými klasifikáciami (27).

1.3.4.2 Vyhodnocovanie numerickej predikcie

Cieľom hodnotenia presnosti numerickej predikcie je analýza a výpočet chýb podľa určitej metriky. Chyby predikcie (t. j. rozdielu skutočnej a predikovanej hodnoty) sa posudzujú individuálne, ale pre porovnanie rôznych použitých postupov sa posudzujú skôr súhrnne pomocou rôznych mier.

Pri numerickej predikcii sa najčastejšie sledujú tieto metriky: mean square error, mean absolute error, mean relative squared error, mean relative absolute error a correlation coefficient (27).

Nech

N – počet pozorovaní,

E_i – predikovaná hodnota prípadu i,

O_i – pozorovaná hodnota prípadu i,

$\bar{E}$- stredná hodnota predikovanej premennej,

$\bar{O}$ – stredná hodnota pozorovanej premennej.

Potom (27):

Stredná kvadratická chyba (Least squares deviation - LSD, mean square error)

$$LSD = \sum_{i=1}^{N} \frac{(E_i - O_i)^2}{N} \ . \qquad [11]$$

Predstavuje aritmetický priemer druhých mocnín odchýlok. Je to najčastejšie používaná chyba, mnohé metódy minimalizujú strednú kvadratickú chybu, lebo sa s ňou dobre počíta.

Stredná absolútna chyba (Average deviation, mean absolute error)

$$AD = \sum_{i=1}^{N} |E_i - O_i|/N \ . \qquad [12]$$

Predstavuje priemernú vzdialenosť. Chyba pre odľahlé hodnoty je úmerná vzdialenosti – nie je zväčšovaná ako pri strednej kvadratickej chybe.

Niekedy nás môže zaujímať chyba relatívna voči predikcii.

Stredná relatívna kvadratická chyba (mean relative squared error)

$$RSE = \sum_{i=1}^{N} [(E_i - O_i)/E_i]^2/N. \qquad [13]$$

Stredná relatívna absolútna chyba (mean relative absolute error)

$$RAD = \sum_{i=1}^{N} |E_i - O_i|/\ E_i\ /N. \qquad [14]$$

Korelačný koeficient (Correlation coefficient - Pearson)

$$r = \sum_{i=1}^{N} (E_i - \bar{E}) * \frac{(O_i - \bar{O})}{\left[\sqrt{\sum_{i=1}^{N} (E_i - \bar{E})^2} * \sqrt{\sum_{i=1}^{N} (O_i - \bar{O})^2}\right]}. \qquad [15]$$

Korelačný koeficient meria silu štatistickej závislosti medzi dvoma kvantitatívnymi premennými. Premenná *Y* nezávisí od premennej *X*, ale dve náhodné premenné *X* a *Y* sa spoločne menia. Pearsonov korelačný koeficient (Pearson's product moment) z roku 1896 je mierou lineárnej závislosti dvoch premenných. Pearsonov korelačný koeficient *ρ (ró)* odhadnutý z náhodnej vzorky sa zapisuje r. Čitateľ rovnice sa nazýva kovariancia a vyjadruje, ako sa súčasne menia hodnoty dvoch premenných. Kladná hodnota znamená, že sa menia spoločne jedným smerom, záporná hodnota znamená, že sa menia opačným smerom a nula, že sa menia nezávisle. Vydelením kovariancie štandardnými odchýlkami sa vypočíta korelačný koeficient, ktorého hodnota sa nachádza v intervale od −1 do 1. Pearsonov korelačný koeficient sa rovná −1 v prípade, že všetky pozorovania ležia na klesajúcej priamke a 1, ak pozorovania

ležia na stúpajúcej priamke. Pearsonov korelačný koeficient je silne ovplyvniteľný extrémnymi hodnotami (outliers), a to v oboch smeroch. Môžu významne znížiť silnú závislosť, ale aj vyrobiť silnú závislosť tam, kde žiadna nie je. Dôležité závery by sa nemali robiť iba na základe hodnoty koeficientu. Ak r umocníme, získame koeficient determinácie, ktorý reprezentuje proporciu spoločného rozptylu, teda o koľko percent zmena jednej premennej ovplyvní druhú. Je to informácia o sile relácie medzi premennými (57).

Program STATISTICA poskytuje vyhodnotenie modelov pomocou error rate. Pre regresný typ problémov, keď existujú pozorované hodnoty pre výstupnú premennú, je štatisticky vypočítaná priemerná kvadratická chyba (zostatková) pre každý predikčný model (ERROR RATE = average squared error residual). Pre klasifikáciu to predstavuje celkovú mieru chýb predikcie každého modelu (27).

2 METODIKA A METÓDY SKÚMANIA

V prvej časti riešenia problémovej oblasti je využitá hlavne metóda analýzy zameraná na identifikáciu potenciálnych príležitostí a problémov v riadení výrobných procesov, ktoré je možné riešiť pomocou metód dolovania dát.

Ďalšia časť predstavuje proces získavania znalostí podľa metodiky CRISP-DM. Získanie údajov o reálnom výrobnom procese je problematické, pretože podniky si chránia svoje údaje. Využili sme metódu simulácie na získanie potrebných údajov o výrobnom procese. Pre tvorbu simulačného modelu výrobného procesu je dominantná udalosťami riadená simulácia (2). Simulačný model výrobného procesu ukladá údaje do pripravenej produkčnej databázy. Simulácia je realizovaná pomocou programu Witness, pretože s ním máme dlhoročné skúsenosti a aj pre jeho implementovanú podporu priamej spolupráce s databázovými systémami.

Witness

Program Witness od britskej spoločnosti Lanner Group Ltd. je jeden z najrozšírenejších softvérov na simuláciu a optimalizáciu výrobných, obslužných a logistických systémov. Tento simulátor využíva princípy simulácie s diskrétnymi udalosťami. Simulačný čas sa nemení spojito ani s konštantným inkrementom, ale nastavuje sa vždy na výskyt ďalšej udalosti. Narábanie so simulačným časom sa riadi podľa „kalendára udalostí“, ktorý zostavuje simulátor automaticky. Program má implementovanú podporu priamej spolupráce s vybranými databázovými systémami, tabuľkovými editormi a inými programami, ktoré využívajú ActiveX alebo ODBC (12, 42).

V ďalších krokoch riešenia boli pre aplikáciu navrhnuté vybrané metódy a techniky dolovania dát na základe prieskumu najviac používaných metód a zároveň s ohľadom na úlohy dolovania a dostupných dát v databáze. Pôjde o predikčnú úlohu, kde bude využitá klasifikácia a numerická predikcia. V rámci jednotlivých metód dolovania dát budú použité nasledujúce predikčné metódy a techniky, opísané v predchádzajúcej kapitole:

1. Rozhodovacie stromy – Decision Trees:
   I. Klasifikačné a regresné stromy – C&RT
      (Classification and Regression Trees)
   II. Chi-kvadrátová automatická interakčná detekcia – CHAID
      (Chi-squared Automatic Interaction Detector)

III. Náhodné lesy – Random Forrest
IV. Zosilnené stromy – Boosting Trees
V. Viacrozmerné adaptívne regresné spliny – MARSplines (Multivariate Adaptive Regression Splines)

2. Strojové učenie – Machine learning:
VI. Stroj s podpornými vektormi – Support Vectore Machine
VII. K-najbližších susedov – K-Nearest Neighbour
VIII. Naivný Bayes – Naive Bayes

3. Viacnásobná regresia – Multiple Regression.
4. Neurónové siete – Neural Network.

Na klasifikáciu nebude využitá viacnásobná regresia a naopak pre numerickú predikciu nebude môcť byť použitá technika Naive Bayes – je to z dôvodu obmedzenia týchto techník na spracovanie vstupných dát. Viacnásobná regresia požaduje len numerické (spojité) závislé premenné a Naive Bayes zase len kategorické (continuous) závislé premenné. Spolu je použitých 9 techník pre regresiu a 9 techník pre klasifikáciu. Podrobnejšie sú techniky popísané v podkapitole 1.3.3 Vybrané metódy a techniky dolovania dát.

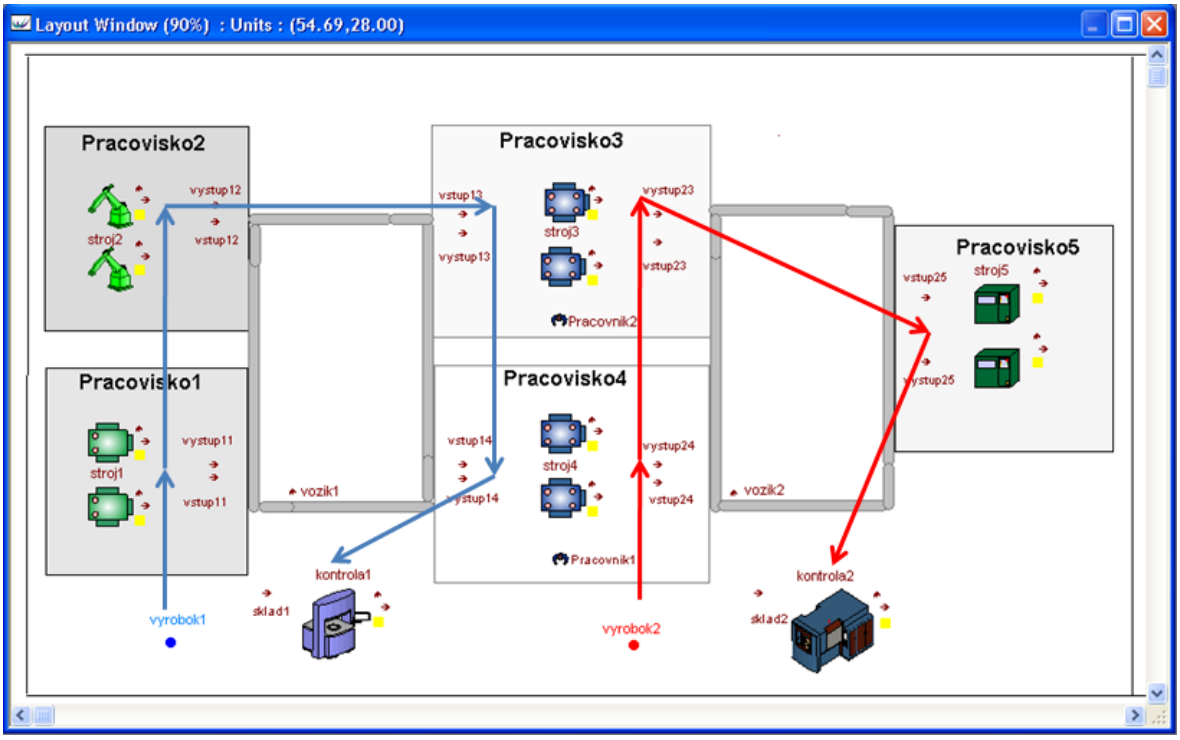
Pre dolovanie dát sme zvolili nástroj STATISTICA Data Miner taktiež na základe spomínaného nezávislého dataminingového prieskumu spoločnosti Rexer Analytics. Dosiahnuté výsledky z predikcie budú verifikované na simulačnom modeli.

2.1 Charakteristika objektu skúmania

Objektom skúmania tejto monografie je správanie sa výrobného procesu za účelom jeho správneho riadenia. Riadenie výroby je jeden z najpodstatnejších druhov riadenia konkrétnych procesov. Výrobu je možné definovať ako časť transformačného procesu, tzn. že je to konkrétna premena výrobných faktorov (vstupov, resp. riadiacej veličiny) na výrobky (výstupy, resp. riadené veličiny). Táto premena prebieha ako výrobný proces, ktorý pozostáva z celého radu pracovných, automatických aj prírodných procesov a je ohraničený časovým intervalom, v ktorom sa východiskové vstupy premieňajú na výstupy. Usporiadanie a štruktúra konkrétnych výrob a ich riadenia závisí od charakteru výrobku, resp. služby, trhu, objemu výroby, charakteru zákaziek, použitých technológií a niektorých ďalších faktorov. Výrobný proces môže byť chápaný len ako proces priebehu výroby výrobku (napr. od rozrezania materiálu, cez opracovanie a kontrolu, až po vyexpedovanie hotového výrobku) alebo ako súhrn činností,

ktoré tento priebeh zaisťujú a ktoré sa na ňom priamo podieľajú. Bezprostredný výrobný proces je možné charakterizovať ako proces, pri ktorom sa vplyvom pôsobenia človeka stávajú z pracovných predmetov hotové výrobky, ktoré spoločnosť potrebuje ku svojmu životu a svojmu rozvoju. Tento proces prebieha na výrobných pracoviskách priemyselných podnikov a jeho podstatným rysom je zmena mechanických, fyzikálnych, chemických a iných vlastností pracovného predmetu (38).

Simulačný model výrobného procesu, zobrazený v simulátore na obrázku 9, slúži na generovanie výrobných údajov pre riadenie daného procesu a tie sú uložené v databáze. Dáta zahŕňajú kumulované údaje po uplynutí jedného pracovného dňa. Súčasne, model slúži na overovanie získaných výsledkov predikcie.



Obr. 9 *Simulačný model výrobného procesu*

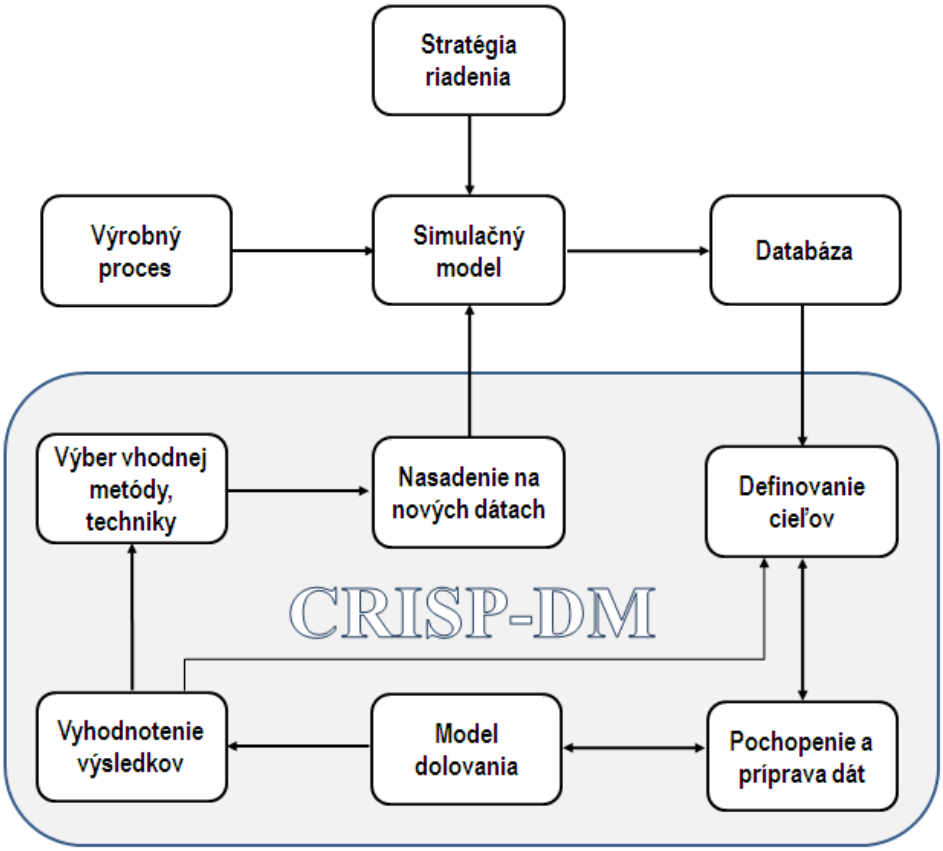
Formulácia výrobného procesu:

Skladá sa z 5 pracovísk na vykonávanie technologických operácií a dvoch výstupných kontrolných pracovísk pre dva finálne výrobky. Výrobky vchádzajú do pracovísk po dávkach. Každé pracovisko má priradené vstupné a výstupné sklady, ktorých funkciou je medzioperačné skladovanie. Transport výrobných dávok sa realizuje pomocou vozíkov po presne určených cestách. Cesty spájajú príslušné pracoviská, čím je umožnená preprava výrobkov medzi operáciami. Pracoviská 1 a 2 vykonávajú operácie pre prvý výrobok, ktorého postup

pracoviskami závisí od vyťaženosti strojov. V pracovisku 5 sa realizujú technologické operácie len pre druhý výrobok. Oba typy výrobkov prechádzajú pracoviskom 3 a 4, kde zoraďovanie operácií je vykonávané prostredníctvom robotníkov. Kontrola 1 a 2 sú výstupnými pracoviskami, odkiaľ sú výrobky odosielané na export.

2.2 Postup realizácie práce

Postup realizácie práce je zobrazený na obrázku 10. Podľa uvedeného výrobného procesu bol vytvorený simulačný model. Pre simulačný model je navrhnutá stratégia riadenia na dosiahnutie konkrétnych cieľov výroby. Prostredníctvom simulačných experimentov sú generované údaje o priebehu výroby a dosiahnutých výsledkoch, pričom budeme sledovať, ktoré vstupné veličiny ovplyvňujú deklarované výstupy a ako. Na základe týchto údajov bude vytvorená databáza. Ďalej budeme simulačný model považovať za čiernu skrinku a jeho štruktúra sa nebude meniť. Získané údaje z jednotlivých behov reprezentujú hodnoty vybraných cieľových parametrov a vstupné hodnoty riadiacich (regulovateľných) parametrov za sledované obdobie práce výrobného systému. Každá takto získaná množina údajov reprezentuje jedno rovnaké časové obdobie histórie práce výrobného systému.

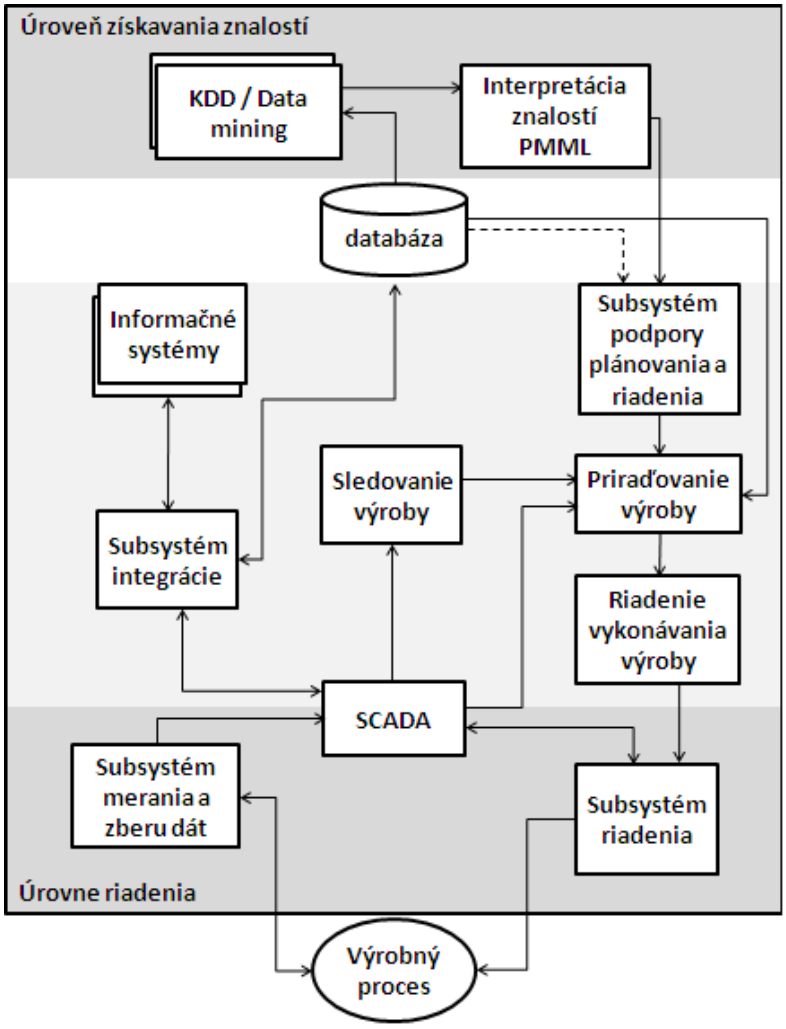


Obr. 10 *Postup realizácie práce*

V rámci ďalšieho postupu sme uplatňovali metodológiu CRISP-DM, na obrázku. 10 sú zvýraznené jej fázy. Proces získavania znalostí potom pozostáva z nasledujúcich kľúčových fáz:

- definovanie cieľov,
- pochopenie a príprava dát – zahŕňa extrakciu množiny dát, zoskupenie, vyčistenie a transformáciu dát s ohľadom na vybrané metódy a ciele dolovania dát,
- vytvorenie modelu dolovania dát a aplikácia metód na transformovanú množinu dát,
- vyhodnotenie výsledkov a porovnanie jednotlivých modelov
- výber vhodného modelu pre aplikovanie na nových dátach,
- nasadenie vybraného modelu na nových dátach.

Celý postup je iteratívny, teda je možné sa vrátiť do ľubovoľného kroku v procese. Pre overenie predikovaných hodnôt je použitá simulácia.



Obr. 11 *Konceptuálna schéma návrhu využitia získavania znalostí*

Na obrázku 11 je znázornená schéma návrhu využitia procesu získavania znalostí pre výrobný proces v priemyselnom podniku so zavedenými podpornými informačnými a riadiacimi systémami. V konceptuálnej schéme sú úrovne riadenia informačne prepojené na samotný výrobný proces (vychádza sa zo štandardného pyramídového modelu). Úroveň získavania znalostí obsahuje subsystém KDD a subsystém na interpretáciu znalostí. Správne získané a interpretované znalosti môžu byť použité na zvýšenie kvality riadenia procesov. Je potrebné zapezpečiť spätný transport získaných znalostí do výrobného procesu, to sa realizuje prostredníctvom subsystému riadenia. Tento modul je možné chápať ako interfejs medzi úrovňou získavania znalostí a úrovňou riadenia. Je napojený na SCADA systém a prostredníctvom neho sa priamo realizujú intervencie do výrobného procesu. Moduly „Priraďovanie výroby“ a „Riadenie vykonávania výroby“ zabezpečujú rozhodovanie o spôsobe zadávania vstupných údajov do výroby.

3 IDENTIFIKÁCIA POTENCIÁLNYCH PRÍLEŽITOSTÍ VYUŽITIA DOLOVANIA DÁT PRE VÝROBNÉ PROCESY

Výrobný proces v priemyselnom podniku zahŕňa mnohé príležitosti (možné prípady) a je priam vhodným prostredím pre aplikovanie dolovania dát. Ideálne je, ak daný podnik má vybudovaný integrovaný informačný systém, často nazývaný aj informačný a riadiaci systém. Štandardná štruktúra takéhoto systému je uvedená na obrázku 1 v kapitole 1 a detailnejší pohľad je ďalej v kapitole 4 – obrázok 12. Keď prejdeme jednotlivé úrovne riadenia, zistíme, že na každej je možné aplikovať proces získavania znalostí.

Na podnikovej úrovni (strategické riadenie) – dolovanie z dát na tejto úrovni je možné použiť na vyhodnocovanie a analýzu dopytu, predaja, nákladov či správania sa zákazníkov vzhľadom na spotrebu a produkciu. Získané znalosti sú určené pre manažment a ekonómov.

Na prevádzkovej úrovni (taktické riadenie) – na základe dát z tejto úrovne (zbieraných a spracovaných procesných informácií meraním, či reguláciou) je pomocou dolovania dát možné opísať aktuálny stav alebo predikovať budúce správanie sa výrobného systému, predikovať anomálie či havarijné stavy vo výrobe a riadiť kvalitu (SCADA). Ďalej je možné dolovanie dát využiť pri návrhu a riadení interakcie človeka a stroja, pri podpore rozhodovania sa v kritických situáciách a podobne (MES).

Na procesnej úrovni (operatívne riadenie) – dáta na tejto úrovni sú zbierané pri priamom pôsobení na procesy, sú to informácie z procesov. Znalosti získané z dát na tejto úrovni je možné použiť na predikciu správania sa jednotlivých výrobných zariadení a na ich riadenie.

Dolovanie dát sa javí ako vhodný nástroj na podporu rozhodovania (48). Rozhodovanie je tá časť procesu, ktorá predchádza v riadiacom cykle skutočným zásahom a pripravuje ho. V automatizovaných systémoch riadenia vykonáva rozhodovanie operátor, dispečer (tzv. účastník rozhodovacej situácie) alebo automatizovaný riadiaci člen (tzv. formátor) (59). Zo systémového pohľadu môžeme pojem rozhodnutia definovať ako výsledok spoločných znakov rôznych analýz (17): „Rozhodnutie je cieľavedomou voľbou medzi verziami činností v danom prostredí, kde sa rôzne verzie činností v procese rozhodovania v čase pred rozhodnutím, prejavujú ako možnosti činností“. Stratégia je alternatívny plán, predkladaný účastníkovi rozhodovacej situácie na zváženie, prijatie a uskutočnenie. Stratégie väčšinou vystupujú ako vzájomne sa vylučujúce alternatívy, aj keď niekedy je možné prekrývanie niektorých vlastností viacerých alternatív. Samotný výber alternatívy je ukončením (výsledkom) rozhodovacieho procesu. Voľba stratégie sa môže uskutočniť len z množiny takých prípustných stratégií, o ktorých sa účastník na základe dostupných informácií domnieva, že sa môžu realizovať.

Aplikovanie procesu získavania znalostí o rôznych problémoch môže pomôcť v rozhodovacom procese a umožniť lepšie riadiť daný proces, ako aj kvalitu výroby. Medzi typické príležitosti, resp. problémy vyskytujúce sa v oblasti plánovania a riadenia výroby, na ktoré môže byť aplikované dolovanie dát, patria (33, 68, 72):

- Identifikácia vplyvu výrobných parametrov na výrobný proces a predikcia jeho správania na základe zmeny týchto parametrov – tu môže byť využitá numerická aj klasifikačná predikcia.
- Predikcia havarijných stavov riadeného procesu – tu môžu byť použité prediktívne úlohy: numerická i klasifikačná predikcia. Na základe údajov z histórie riadeného procesu je možné určiť, kedy a za akých podmienok daný havarijný stav vzniká.
- Detekcia chybových stavov výrobných zariadení, ako aj jednotlivých výrobkov (odhaľovanie výskytu nepodarkov) – podobne ako pri predikcii havarijných stavov aj tu je možné použiť predikciu alebo hľadanie vzorov a pravidiel.
- Identifikácia rôznych neštandardných stavov, ktoré majú vplyv na výrobný proces a ktoré musí riešiť operátor výroby najčastejšie neplánovanou odstávkou stroja alebo časti technológie – taktiež je vhodné využiť klasifikáciu.
- Predikcia preventívnych kontrol výrobných zariadení (súvisí s údržbou) – bolo by vhodné použiť napríklad klasifikačné stromy, aby bolo jasne viditeľné, kedy je odporúčané vykonať preventívnu kontrolu.
- Identifikácia a optimalizácia relevantných parametrov riadenia, ktoré majú vplyv na zvyšovanie bezpečnosti riadenia procesov alebo kvalitu produkcie a podobne – tu je opäť možné využiť prediktívne úlohy.
- Diagnostika výrobných systémov s ohľadom na celkovú životnosť týchto systémov – vhodné je použiť deskriptívne úlohy a zároveň hľadanie vzorov a pravidiel.
- Priebežné sledovanie kvality procesu riadenia na základe hodnotenia kvality z on-line získaných údajov – vhodné je využiť klasifikáciu.
- Predikcie pre potreby manažmentu podniku, rôzne ad hoc reporty – vhodná je numerická aj klasifikačná predikcia.
- Predikcia spotreby elektrickej alebo tepelnej energie – tu je potrebné využiť numerickú predikciu.
- Predikcia spotreby výrobného materiálu – vhodné je použiť obe predikčné úlohy.

Uvedené body môžu predstavovať perspektívy ďalšieho výskumu. Ďalej bude práca orientovaná na predikciu správania sa systému na základe zmeny vstupných parametrov. Toto je cieľ, ktorý chceme dolovaním dát dosiahnuť.

Vo výrobnom procese vzniká veľké množstvo dát, ktoré sa dajú zbierať a spracúvať na prevádzkovej úrovni (SCADA) a potom je možné ich ďalej využívať vo vyšších úrovniach riadenia. Na tieto dáta je vhodné aplikovať proces dolovania dát, a to je predmetom tejto práce. Získané znalosti z takýchto dát môžu predikovať budúce správanie sa výrobného systému pri zmene niektorých parametrov či predikovať možné poruchy a stavy ohrozujúce plynulosť výrobného procesu na základe predchádzajúcich informácií o výrobnom procese, priebeh výrobných dávok, nastavenie výrobných strojov a podobne. Riadenie výroby na prevádzkovej úrovni je zamerané na dosiahnutie konkrétnych cieľov, medzi ktoré patria:

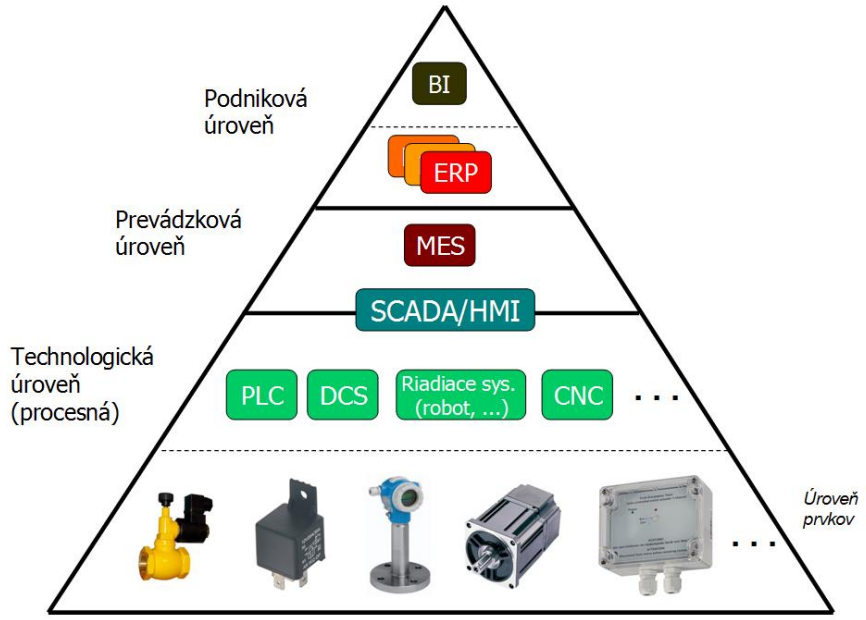
- plnenie požadovaných počtov odvádzaných výrobkov,
- dodržovanie termínov odvádzania (minimalizácia priebežných časov výroby),
- minimalizácia časových sklzov,
- minimalizácia nákladov,
- minimalizácia nepodarkov, dodržiavanie stanovenej kvality výroby,
- minimalizácia rozpracovanej výroby,
- maximálne využitie kapacít,
- minimalizácia skladových zásob (73).

Tieto ciele sú ovplyvňované rôznymi faktormi a riadenými vstupmi.

Výrobný proces obsahuje veľa príležitostí a problémov, pre ktoré je vhodné využiť rozsiahle nástroje dolovania dát. Dôležité je mať vhodné dáta a správne pochopiť problém, na základe problému potom určiť úlohy, resp. stanoviť ciele a nájsť zodpovedajúci spôsob riešenia konkrétnej úlohy dolovania dát. V tejto monografii sa zameriame na predikciu správania sa výrobného procesu na základe zmeny vstupných parametrov určenú pre podporu riadenia na prevádzkovej úrovni – úroveň riadenia výroby.

4 PREDIKCIA SPRÁVANIA SA VÝROBNÉHO SYSTÉMU

Táto kapitola je nosnou časťou celej práce. Zameriava sa na použitie predikcie na konkrétny problém, ktorého riešenie prispeje ku kvalitnejšiemu a efektívnejšiemu rozhodovaniu v riadení procesov. Riešenie sa zameriava na prevádzkovú úroveň riadenia, kde na úrovni riadenia výroby sa rozhoduje o spôsobe zadávania vstupných údajov do výroby. V rámci MES systémov to zabezpečujú funkcie Priraďovanie výroby a Riadenie vykonávania výrobných úloh (podľa klasifikácie ISA 95). Výstupné hodnoty cieľových parametrov sú zbierané priebežne a do vyhodnocovania sú prijaté hodnoty po skončení danej plánovacej periódy - jeden deň (zmena). Tieto údaje slúžia na sledovanie výroby, ako aj na vyhodnocovanie výkonnosti výrobného systému. Aj tieto funkcie patria do množiny funkcií MES systémov podľa klasifikácie ISA 95. Hodnotenia sú potom odovzdávané vyššej úrovni riadenia, zvyčajne ERP systému. Použijeme predikciu na podporu rozhodovania na danej riadiacej úrovni. Tieto rozhodovacie procesy ovplyvnia finálne procesnú úroveň riadenia.



Obr. 12 *Pyramídový model distribuovaného riadenia (68)*

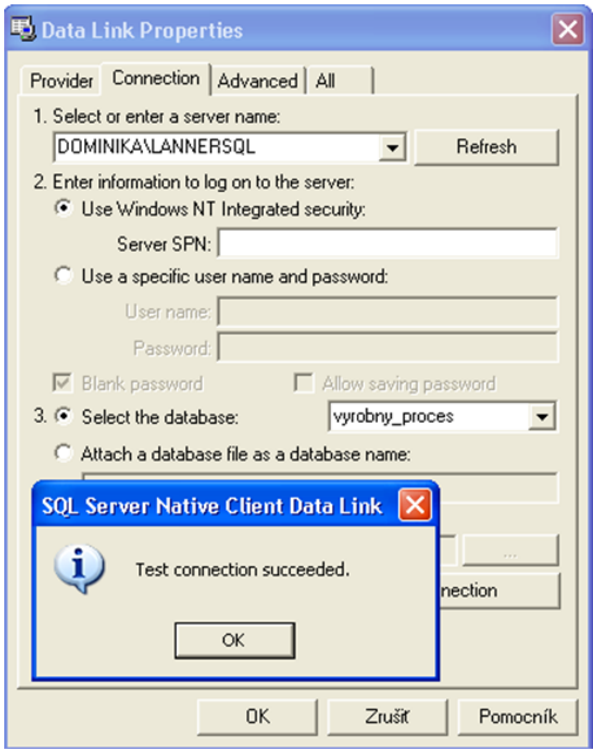
Komplexné riadenie procesov nie je dnes možné bez využitia informačných systémov. Ide o informačné a riadiace systémy na jednotlivých úrovniach riadenia od najvyššej úrovne – podnikovej úrovne riadenia (manažérske informačné systémy, ERP systémy a pod.) až po procesnú úroveň riadenia so systémami nachádzajúcimi sa priamo vo výrobnom procese. Aplikácia moderných informačných a riadiacich systémov výrazne zvýšila kvalitu riadiacich

procesov a priniesla podnikom efektívnu prevádzku a zisk (68). Pyramídový model (na obrázku 12) je stručne opísaný v kapitole 1, predstavuje všeobecne uznávaný model komplexného riadenia procesov.

V kapitole 2 je stručne opísaný sledovaný výrobný proces, následne je v súlade s metodikou CRISP-DM definovaný problém, vykonaná analýza a príprava dát, modelovanie pomocou rôznych predikčných techník, ich vyhodnotenie a vybraný vhodný model PMML pripravený na integráciu do rozhrania informačných a riadiacich systémov pre podporu rozhodovania.

4.1 Výber dát z databázy

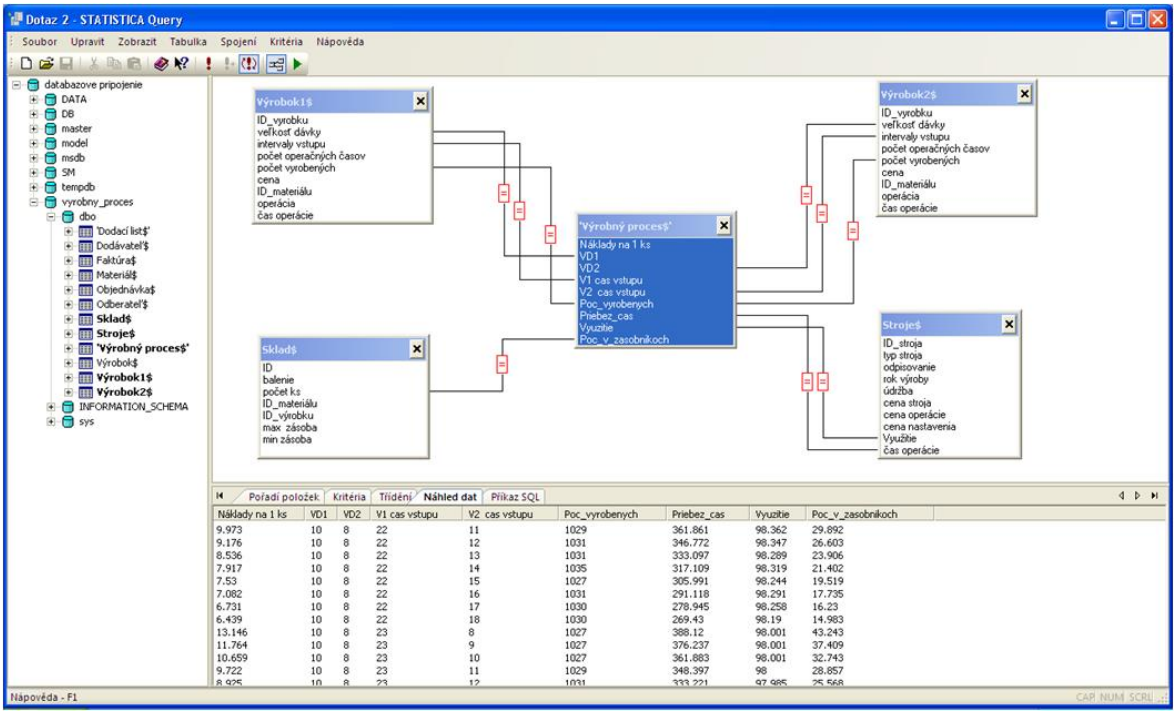
V databáze sú zaznamenané údaje o výrobnom procese získané simulátorom. Každý riadok v databáze predstavuje súhrnné údaje po uplynutí sledovanej doby. Pripojenie programu STATISTICA na databázový server sme realizovali pomocou nástroja Query, ktorý umožňuje definovanie prepojenia na niekoľko podporovaných databázových serverov. Jedným z nich je aj databázový server Microsoft SQL, ktorý sme využili. Na pripojenie bol použitý SQL Server Native Client 10.0 (obrázok 13). Prepojenie databázy z Witness cez LANNER SQL.



Obr. 13 *Vytvorenie otázky na databázu „vyrobny_proces“*

V programe STATISTICA a jeho grafickom móde Query je možnosť vybrať požadované databázové tabuľky a označiť v nich požadované atribúty hneď po pripojení na zvolený databázový server. Ďalej STATISTICA Query umožňuje definovať základné kritéria pre spojenie jednotlivých databázových tabuliek a ďalších podmienok týkajúcich sa tohto spojenia. Tabuľka výrobný proces, vytvorená pomocou simulácie, môže obsahovať spojenia na iné tabuľky v databáze. S ohľadom na riešenú problematiku urobíme výber len nami požadovanej tabuľky výrobný proces a všetkých atribútov v tejto tabuľke.

Pomocou grafického módu sme vytvorili výberovú otázku (obrázok 14). Keďže sme vedeli aké dáta budeme potrebovať, už pri simulácii boli pripravené všetky potrebné matematické operácie na získanie výsledných dát. Preto nebolo potrebné definovať žiadny zložitejší výber vnorenej otázky, náhradu hodnôt pre atribúty s hodnotou „null“ či použitie agregovaných funkcií a podobne. Konkrétny výber dát na učenie modelov môžeme realizovať až po detailnejšej analýze dát.



Obr. 14 *STATISTICA Query – vytvorená otázka na databázu „vyrobny_proces“*

Ukážka SQL otázky vytvorenej súčasne s označením požadovaných tabuliek a atribútov v grafickom móde:

SELECT TABLE1."Náklady na 1 ks", TABLE1."VD1", TABLE1."VD2", TABLE1."V1#cas vstupu", TABLE1."V2 #cas vstupu", TABLE1."Poc_vyrobenych", TABLE1."Priebez_cas",

TABLE1."Vyuzitie", TABLE1."Poc_v_zasobnikoch" FROM "DB"."dbo"."'Výrobný proces$'" TABLE1

Následne po spustení výberovej otázky sa prepne STATISTICA späť do nástroja Data Miner a načítajú sa dáta z databázy do internej tabuľky. Riadky tejto tabuľky zodpovedajú jednotlivým riadkom získaným z vytvorenej otázky a stĺpce reprezentujú jednotlivé atribúty otázky. Takto načítané dáta je možné ďalej upravovať alebo vykonať analýzu, transformáciu, čistenie či priamo dolovanie. Súbor s načítanými dátami (obrázok 15) obsahuje 9 stĺpcov (premenné, resp. atribúty) a 10 000 riadkov (pozorovania, resp. prípady). Formát dát uložených v databáze je v požadovanom tvare, a tak zatiaľ nie je potrebné vykonať ich ďalšiu úpravu.

Data: výrobný proces (9s krát 10000ř)

	1 Náklady_na_1_ks	2 VD1	3 VD2	4 V1_cas vstupu	5 V2_cas vstupu	6 Poc_vyrobenych	7 Priebez_cas	8 Vyuzitie	9 Poc_v_zasobnikoch
1	5.293	2	3	15	8	658	127.234	59.471	5.103
2	4.339	2	3	15	9	658	90.872	56.389	2.943
3	3.585	2	3	15	10	656	55.828	53.396	1.304
4	3.278	2	3	15	11	636	37.988	50.7	0.558
5	3.319	2	3	15	12	601	37.434	48.012	0.508
6	3.355	2	3	15	13	571	37.104	45.78	0.466
7	3.39	2	3	15	14	545	36.796	43.883	0.431
8	3.445	2	3	15	15	522	36.282	42.164	0.394
9	3.429	2	3	15	16	504	36.82	40.725	0.389
10	3.442	2	3	15	17	486	36.725	39.399	0.372
11	3.441	2	3	15	18	471	36.971	38.259	0.357
12	5.317	2	3	16	8	645	126.377	58.308	5.084
13	4.343	2	3	16	9	645	90.788	55.198	2.937
14	3.587	2	3	16	10	642	55.669	52.238	1.297
15	3.268	2	3	16	11	623	37.692	49.567	0.549
16	3.304	2	3	16	12	586	37.158	46.806	0.497
17	3.354	2	3	16	13	557	36.968	44.608	0.46
18	3.388	2	3	16	14	533	36.861	42.739	0.428
19	3.428	2	3	16	15	508	36.885	40.936	0.404
20	3.458	2	3	16	16	491	36.368	39.553	0.374
21	3.458	2	3	16	17	473	36.543	38.264	0.363

Obr. 15 *Načítané dáta z databázy*

4.2 Definovanie cieľov dolovania dát

Cieľ, ktorý chceme dosiahnuť aplikáciou procesu získavania znalostí z databáz, je predikcia správania sa systému na základe zmeny vstupných parametrov. Tento cieľ sme rozdelili na dve úlohy dolovania dát:

- Kategorická predikcia splnenia všetkých výrobných cieľov súčasne - celkovej kvality produkcie.
- Numerická predikcia jednotlivých ukazovateľov výrobných cieľov.

Vstupné regulovateľné parametre je potrebné nastaviť tak, aby bol proces efektívny a boli dosiahnuté všetky požadované ciele procesu podľa určitej stratégie riadenia. Regulovateľné

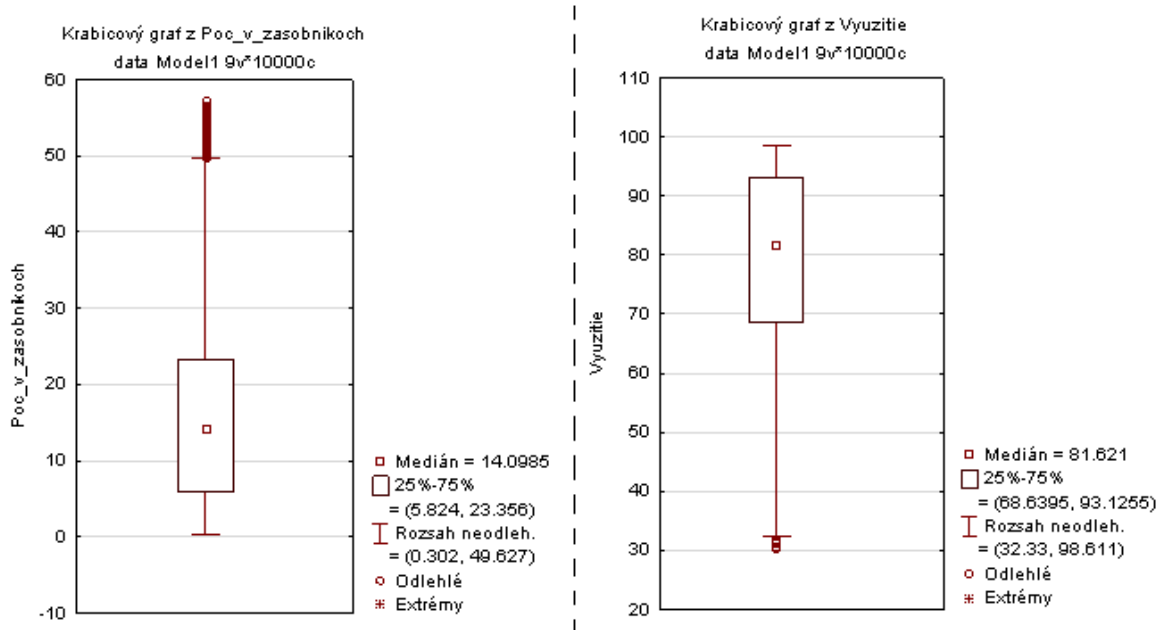
výrobné parametre použijeme ako prediktory v úlohách dolovania dát - klasifikácii a numerickej predikcii.

4.3 Príprava - Analýza dát

Vstupné dáta boli získané vykonaním testov s rôznym nastavením vstupných parametrov procesu. Nastavenie sa týkalo veľkosti výrobnej dávky a intervalu zadávania materiálu do výroby (vstupného času). Celkovo v jednotlivých testoch boli regulované 4 parametre v rozsahoch: VD1 2-10, VD2 3-7, casV1 15-27, casV2 8-18. Rozsahy parametrov boli nastavené cielene tak, aby ich okrajové hodnoty reprezentovali rôzne zaťažený systém. Sledujeme vplyv týchto nastavení a hľadáme najdôležitejšie faktory, ktoré najlepšie rozlišujú medzi hodnotami ukazovateľa. Pomocou týchto faktorov potom dokážeme vysvetliť príčinu variability ukazovateľa a efektívne ovplyvňovať kvalitu procesu.

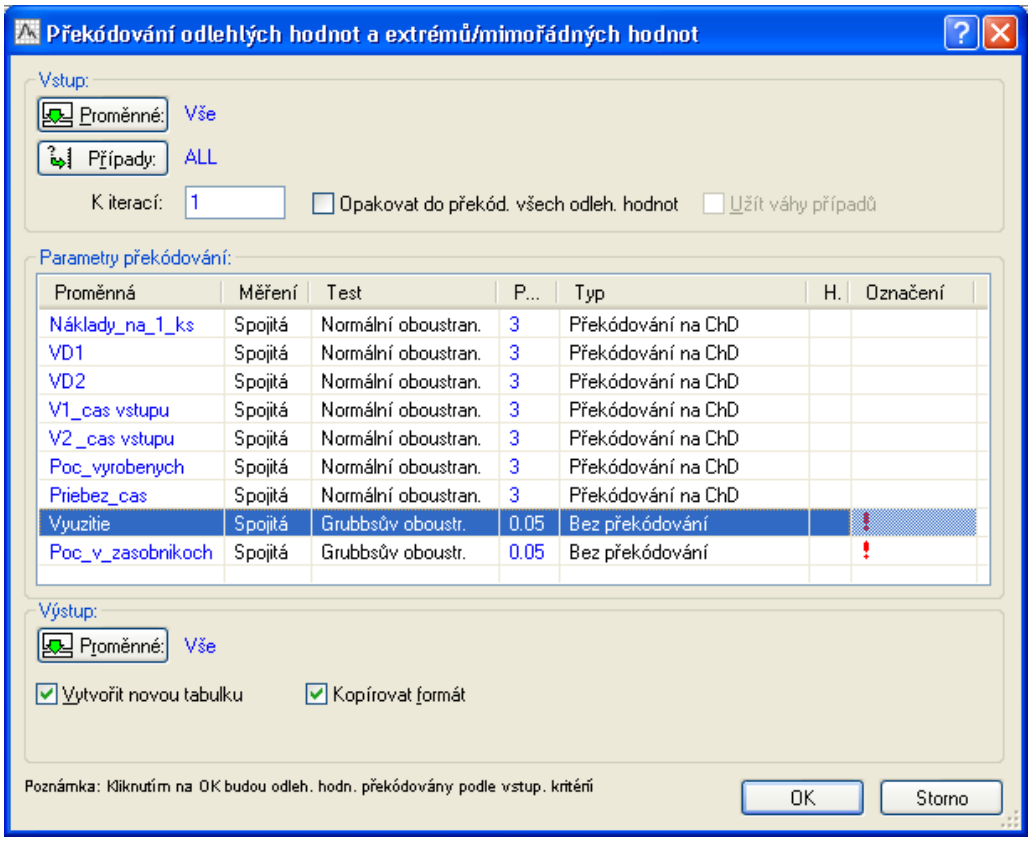
Vykonáme deskriptívnu, prieskumovú a korelačnú analýzu viacrozmerných dát, ktorých cieľom je urobiť akýsi prvý pohľad na dátový súbor, za účelom odhalenia odľahlých pozorovaní, identifikácie rôznych štruktúr, ako je závislosť v dátach a podobne. Je asi zrejmé, že na rozdiel od jednorozmerných dát, sú prieskumové analýzy vo viacrozmerných dátových súboroch násobne náročnejšie a dajú sa v podstate robiť len s využitím vhodného nástroja. Na základe týchto výsledkov môžeme urobiť určité korekcie v dátach, už pred vlastnými štatistickými analýzami, cieľom ktorých je urobiť určité závery.

Deskriptívnou analýzou dát sme zistili, že premenné Poc_v_zasobnikoch a Vyuzitie obsahujú odľahlé hodnoty tzv. OUTLIERS (obrázok 16). Pred ďalšou analýzou je potrebné tieto odľahlé hodnoty odstrániť, pretože počet v zásobníkoch predstavuje vysokú rozpracovanosť výroby. Obmedzíme rozpracovanosť v priemere na maximálne 10 kusov v medziskladoch. V učení metód dolovania dát, t .j. v trénovacom súbore budeme uvažovať s využitím vyšším ako 50%.



Obr. 16 *Outliers v premenných Poc_v_zasobnikoch a Vyuzitie*

Označené prípady (obrázok 17), ktoré obsahujú odľahlé hodnoty, nebudeme brať do úvahy pre nasledujúcu analýzu. Takto vznikol súbor s počtom pozorovaní 1981, ktorý sme ďalej použili v procese získavania znalostí.



Obr. 17 *Označenie outliers v súbore dát*

POPISNÉ ŠTATISTIKY Tabuľka 4

	N platných	Priemer	Minimum	Maximum	Sm.odch.
Náklady_na_1_ks	1981	3.9087	3.1470	4.999	0.5782
VD1	1981	6.8985	3.0000	10.000	1.8655
VD2	1981	4.1161	3.0000	7.000	1.0414
V1_cas vstupu	1981	21.1464	15.0000	27.000	3.6171
V2 _cas vstupu	1981	14.2468	8.0000	18.000	2.8627
Poc_vyrobenych	1981	897.8410	671.0000	1058.000	102.1542
Priebez_cas	1981	103.0428	44.7020	187.611	37.9129
Vyuzitie	1981	78.6595	60.0390	96.964	10.2910
Poc_v_zasobnikoch	1981	3.9207	0.8010	9.326	2.1754

Tento súbor dát, pripravený na ďalší krok v dolovaní dát, podrobíme ďalšej analýze. V tabuľke 4 sú zobrazené popisné štatistiky jednotlivých premenných. Pre daný počet pozorovaní je vypočítaný priemer, minimálna a maximálna hodnota, smerodajná odchýlka premenných. V rámci popisnej štatistiky nástroj STATISTICA umožňuje zobraziť početnosti výskytu, krabicové grafy, normálne pravdepodobnostné grafy jednotlivých premenných a podobne. Pre základnú analýzu dát sme ďalej využili aj korelačnú maticu pre všetky premenné v súbore (tabuľka 5).

Označenia jednotlivých premenných sme zvolili podľa abecedy, aby bola tabuľka prehľadná a zmestila sa na šírku strany.

A – Náklady na 1 ks

B – VD1 (veľkosť dávky 1)

C – VD2 (veľkosť dávky 2)

D – V1 čas vstupu

E – V2 čas vstupu

F – Počet vyrobených výrobkov

G – Priebežný čas

H – Využitie

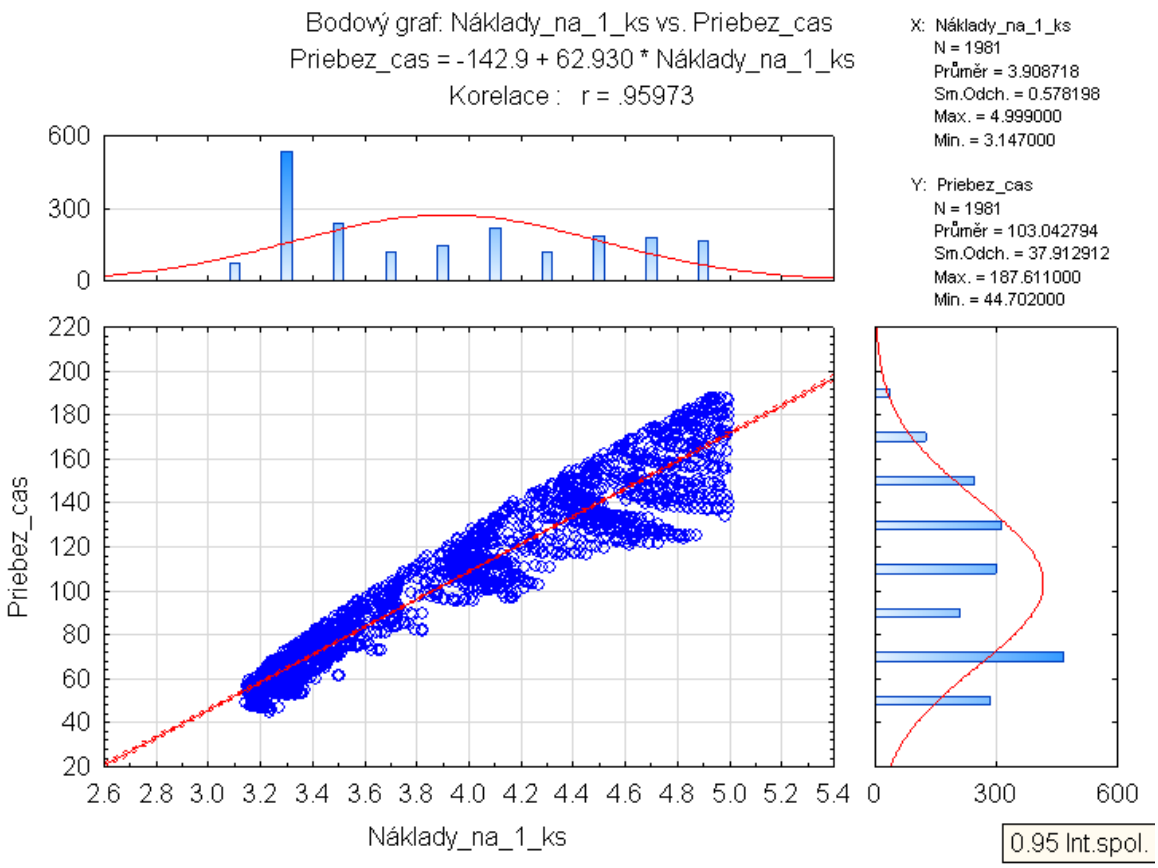
I – Počet výrobkov v zásobníkoch (rozpracovaná výroba)

KORELAČNÁ MATICA VŠETKÝCH PREMENNÝCH Tabuľka 5

	A	B	C	D	E	F	G	H	I
A	1	0.094	0.390	-0.085	-0.192	0.280	0.960	0.462	0.938
B	0.094	1	-0.130	0.425	0.075	0.670	0.246	0.652	0.347
C	0.390	-0.130	1	0.017	0.549	0.114	0.448	0.191	0.364
D	-0.085	0.425	0.017	1	-0.018	-0.168	-0.024	-0.150	-0.091
E	-0.192	0.075	0.549	-0.018	1	-0.097	-0.188	-0.115	-0.160
F	0.280	0.670	0.114	-0.168	-0.097	1	0.478	0.976	0.562
G	0.960	0.246	0.448	-0.024	-0.188	0.478	1	0.641	0.977
H	0.462	0.652	0.191	-0.150	-0.115	0.976	0.641	1	0.706
I	0.938	0.347	0.364	-0.091	-0.160	0.562	0.977	0.706	1

Každá bunka v korelačnej matici má hodnotu v rozsahu od –1,00 do 1,00, ktorá udáva tesnosť štatistickej závislosti medzi premennými. Čím vyššia je absolútna hodnota korelačného koeficientu, tým užší je lineárny vzťah. Ak je hodnota kladná, vzťah je „pozitívny“. Teda vysoké hodnoty jednej premennej odpovedajú vysokým hodnotám druhej premennej. Podobne nízke hodnoty jednej premennej odpovedajú nízkym hodnotám druhej. Pre zápornú hodnotu je to opačne (nízke hodnoty jednej premennej odpovedajú vysokým hodnotám druhej). V tabuľke sú zobrazené zvýraznenou farbou všetky korelačné koeficienty, ktoré sú významné pri $p < 0,05$ (62).

Môžeme zdôrazniť napríklad vzťah medzi priebežným časom a nákladmi na 1 kus (zobrazený na obrázku 18). Na obrázku 18 je zostava grafov, ktoré obsahujú bodový diagram bodov každej korelácie, graf rozdelenia (histogram) každej premennej, ako aj výpočet príslušného korelačného koeficientu a regresnú rovnicu. Z tohto vzťahu vyplýva, že čím sú vyššie priebežné časy, tým vyššie sú náklady na 1 ks. Tento vzťah zodpovedá závislosti medzi nákladmi na jednotku produkcie a priebežným časom pre výrobné systémy z malosériovej až sériovej výroby (Job-shop až Batch-shop production), kde sa súčasne vyrába viac finálnych výrobkov (78, 46).



Obr. 18 *Korelácia medzi priebežným časom a nákladmi na 1 ks*

Ďalej môžeme zdôrazniť vzťahy medzi nákladmi a rozpracovanou výrobou, počtom vyrobených výrobkov a využitím strojov, priebežným časom výroby a rozpracovanou výrobou. Všetky tieto vzťahy majú vysoký kladný korelačný koeficient nad hodnotu 0,9. Naopak premenná čas vstupu V2 záporne ovplyvňuje náklady na 1 kus, priebežný čas, aj rozpracovanú výrobu (počet v zásobníkoch).

Premenná čas vstupu V1 zase ovplyvňuje celkový počet vyrobených výrobkov. Tieto vzťahy majú záporný koeficient nad hodnotu - 0,16. Z toho vyplýva, že pre výrobný proces je žiaduce, aby čas vstupu V2 bol vyšší, tým sa znížia náklady, priebežný čas a zároveň rozpracovaná výroba. A pre zvýšenie celkového počtu vyrobených výrobkov je potrebné, aby čas vstupu V1 bol nižší. Ďalej si môžeme všimnúť, že samotné veľkosti dávok jednotlivých výrobkov sa navzájom záporne ovplyvňujú s korelačným koeficientom -0,130. Teda ak sa zvýši veľkosť jednej dávky, veľkosť druhej sa zníži.

4.4 Klasifikácia

Je veľa spôsobov ako sa dá posudzovať vplyv na výrobný proces. Ak je žiaduce dosiahnuť komplexný pohľad na situáciu a sledovať aj rôzne zložitejšie vzťahy v dátach, ponúka sa práve možnosť použitia dataminingových metód. Klasifikácia je jedna z radu možných prístupov využívajúcich metódy dataminingu.

Aby bola výroba efektívna a boli dosiahnuté požadované ciele procesu, je treba vhodne nastaviť výrobné parametre. Je potrebné zistiť, čo všetko má podstatný vplyv na výsledné ciele a čoho je treba sa vyvarovať, aby ciele výroby neklesli pod určité medze. Musí byť zostavený model, ktorý dokáže komplexne popísať vzťahy medzi rôznymi nastaveniami jednotlivých parametrov a cieľmi výrobného procesu.

V uvedenej analýze sa zameriame výhradne na regulovateľné faktory, aj keď, samozrejme, výsledné ciele procesu závisia taktiež od faktorov neregulovateľných. Ciele výrobného procesu, a teda jeho kvalitu sme posudzovali na základe kvantitatívnej charakteristiky jednotlivých cieľov výroby. Bolo vybraných päť cieľových charakteristík a ich kvantitatívne ohodnotenia, ktoré platia pre konkrétny výrobný systém a konkrétne sledované obdobie jeho činnosti.

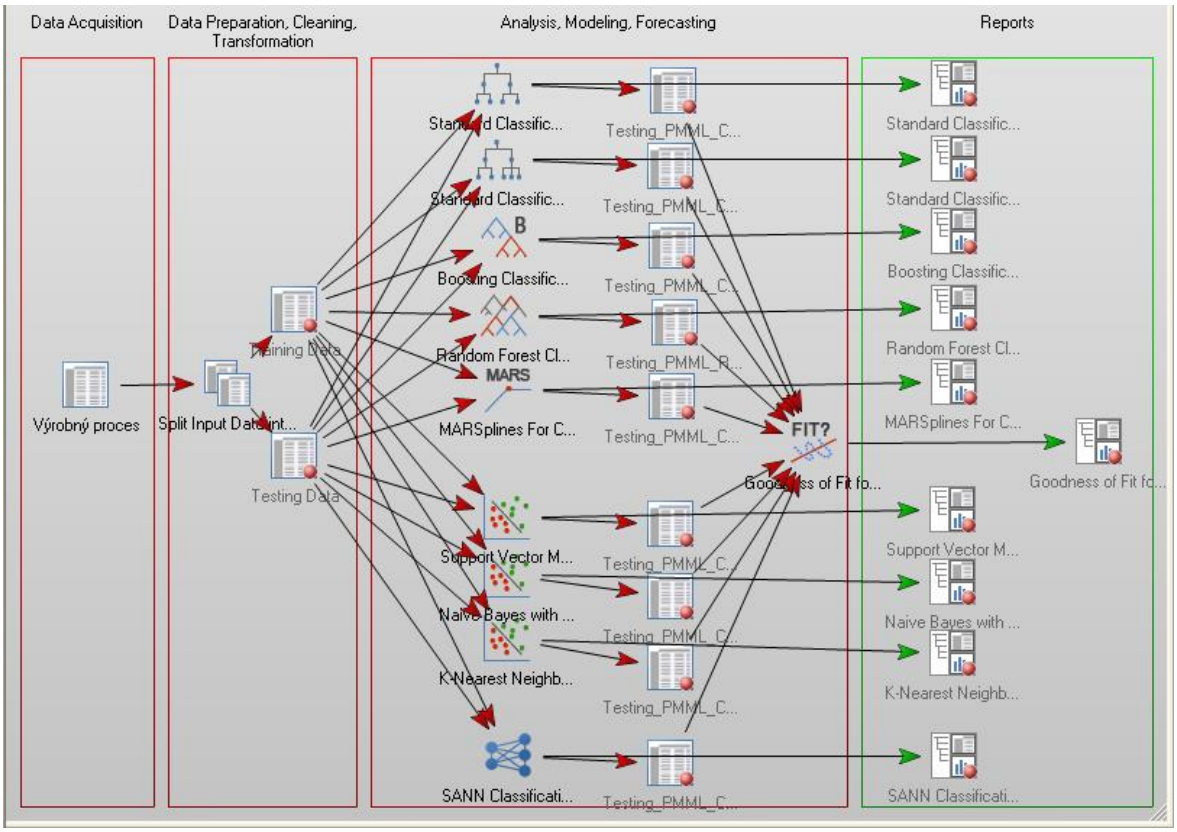
Data: Data_Výrobný_proces* (10s krát 1981ř)

	1 Náklady_na_1_ks	2 VD1	3 VD2	4 V1_cas vstupu	5 V2 _cas vstupu	6 Poc_vyrobenych	7 Priebez_cas	8 Vyuzitie	9 Poc_v_zasobnikoch	10 ciele spolu
1	4.985	3	3	15	8	760	133.796	68.929	5.319	ZLE
2	4.141	3	3	15	9	760	96.731	65.765	3.149	DOBRE
3	3.502	3	3	15	10	759	61.517	62.865	1.511	DOBRE
4	3.236	3	3	15	11	738	44.702	60.083	0.801	DOBRE
5	4.157	3	3	16	9	740	96.682	64.028	3.13	DOBRE
6	3.498	3	3	16	10	739	61.497	61.087	1.493	DOBRE
7	4.173	3	3	17	9	722	96.609	62.501	3.11	DOBRE
8	4.166	3	3	18	9	709	96.573	61.077	3.092	DOBRE
9	4.726	3	4	15	11	761	125.837	68.412	4.759	ZLE
10	4.13	3	4	15	12	759	99.134	65.83	3.253	DOBRE

Obr. 19 *Kategorická premenná v dátach*

Ak proces spĺňa požadované podmienky, tak sú ciele výroby považované za splnené. Tieto podmienky sú definované nasledovne: „Náklady_na_1_ks < 5 and Poc_vyrobenych > 600 and Priebez_cas < 110 and Vyuzitie > 60 and Poc_v_zasobnikoch < 4“. Na základe uvedenej podmienky sme vytvorili kategorickú premennú „ciele spolu“. Táto kategorická premenná, uvedená v poslednom stĺpci (obrázok 19), nadobúda hodnoty DOBRE alebo ZLE. To znamená, že ak premenná „ciele spolu“ má hodnotu DOBRE - proces spĺňa všetky požadované ciele a naopak, ak nadobúda hodnotu ZLE - proces nespĺňa všetky ciele súčasne.

Nasledujúci obrázok zobrazuje návrh dataminingového modelu s použitím vybraných metód pre klasifikáciu. Návrh modelu obsahuje závislú kategorickú premennú – ciele spolu a prediktormi sú regulovateľné spojité premenné - veľkosti dávok VD1, VD2 a ich časy zadávania do výroby V1_cas_vstupu, V2_cas_vstupu.



Obr. 20 *Návrh modelu klasifikácie*

Funkčné celky sme vytvorili pospájaním jednotlivých častí dolovacieho modelu. Základný dátový súbor upravený podľa vyššie uvedenej prípravy dát sme rozdelili pomocou uzla Split Input Data na trénovacie a testovacie dáta s využitím Random sampling a nastavením približne 30% prípadov pre testovacie dáta (approximate percent of cases for testing). Týmto sme vytvorili trénovací súbor o veľkosti 1395 pozorovaní a testovací súbor s 586 pozorovaniami. Aby sme vytvorili funkčný model, bolo potrebné prepojiť zdroj dát s jednotlivými dolovacími metódami. Takto vytvorený dolovací model bol pripravený na svoju ďalšiu činnosť. Ako je vidieť na obrázku 20, v návrhu dataminingového modelu klasifikácie sme použili deväť metód, a to zhora nadol rozhodovacie klasifikačné stromy (Standard CandRT, CHAID, Boosting Trees, Random Forest, MARSplines), Strojové učenie (Support Vector Machine, Naive Bayes a K-Nearest Neighbour), Neurónové siete (SANN).

Na transformovanú množinu dát bol aplikovaný vytvorený model dolovania dát a následne z výsledkov jednotlivých modelov na testovacích dátach sme vyhodnotili presnosť ich klasifikácie. Po vytvorení dolovacieho modelu sme pristúpili k samotnému procesu dolovania dát, kedy sa postupne vytvárali výstupné správy (Reports) pre každú použitú metódu a techniku dolovania dát.

Výstupom jednotlivých metód je okrem iného aj tzv. PMML (Predictive Model Markup Language) dokument, ktorý je pripravený pre použitie na nových dátach vo forme XML súboru. PMML je jazyk založený na XML, ktorý umožňuje efektívnu výmenu (naučených) prediktívnych modelov a spoločné využívanie modelov medzi rôznymi aplikáciami. PMML dokument zvyčajne obsahuje informácie opisujúce plne naučené alebo parametrické analytické modely tak, aby mohli byť ľahko nasadené na nových dátach inou aplikáciou. PMML dokumenty môžu byť uložené prakticky zo všetkých dostupných metód v programe STATISTICA pre predikciu a klasifikáciu (27).

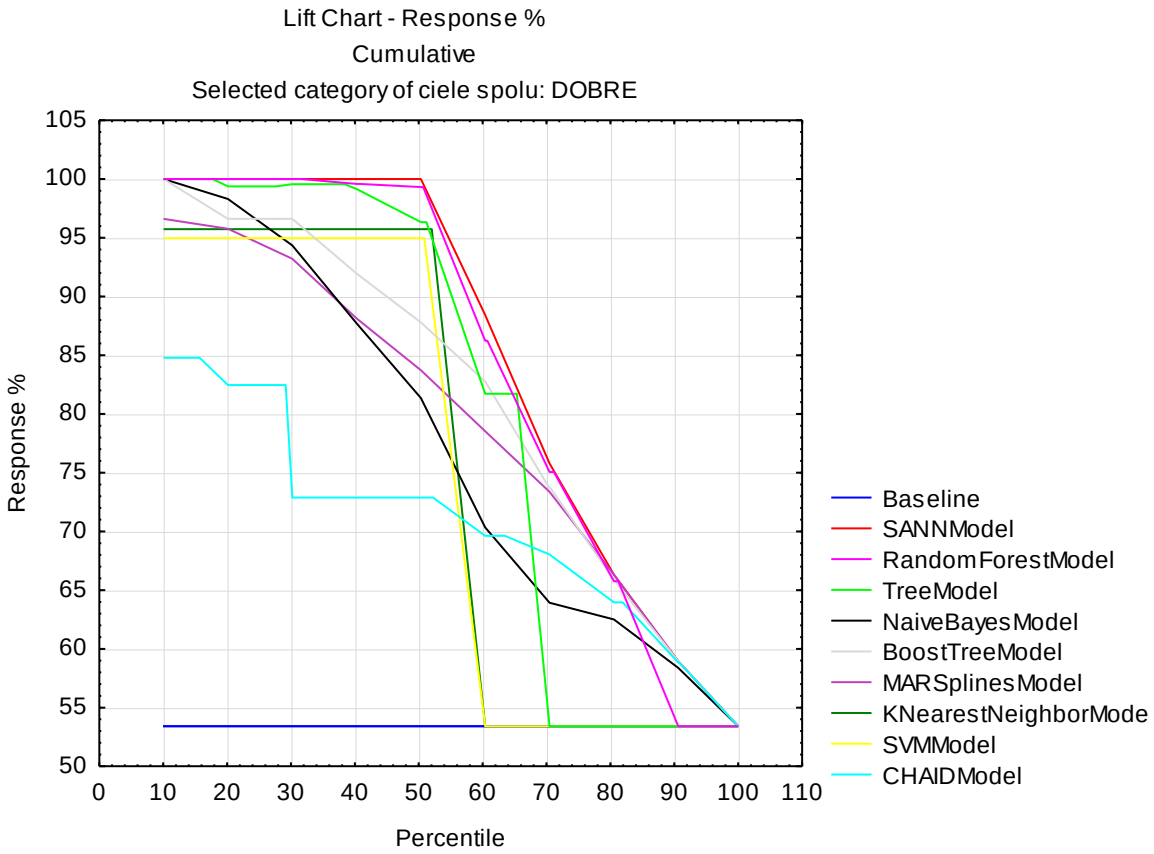
Testing PMML súbory sme spojili s uzlom Goodness of Fit pre automatické vyhodnotenie jednotlivých použitých metód samotným programom. Za účelom ďalších podrobnejších analýz sme uložili všetky získané výstupy do pracovného zošita. Forma jednotlivých výstupných správ je rôzna a závisí od konkrétnej použitej metódy alebo techniky dolovania dát. Ide o výstupy textové, grafické alebo numerické vo forme sumárnych informácií o vykonanom procese danou metódou alebo parciálnych informácií o jednotlivých častiach modelu a podobne.

Pomocou modulu Rapid Deployment sme najprv vyhodnotili zostatkovú chybovosť - Error rate a Lift chart. Načítali sme všetky použité metódy a aplikovali ich na testovací súbor. Najnižšia chybovosť je dosiahnutá pri použití metódy SANN, ďalšie sú uvedené vzostupne v nasledujúcej tabuľke 6.

ERROR RATE JEDNOTLIVÝCH MODELOV
PRE KLASIFIKAČNÚ PREDIKCIU Tabuľka 6

	Error rate
SANN	0.013652
Random Forest	0.032423
CandRT	0.040956
K-Nearest Neighbour	0.058020
SVM	0.076792
Boosting Tree	0.141638
MARSplines	0.189420
Naive Bayes	0.216724
CHAID	0.284983

Na obrázku 21 je znázornený % Response Lift Chart. Graf posudzuje úspešnosť klasifikácie zo skupiny DOBRE (spĺňa podmienky pre efektívnu výrobu). Modrá krivka (Baseline) zodpovedá situácii, kedy nepoužijeme ku klasifikácii žiadny model. Môžeme z nej vyčítať, že zastúpenie skupiny DOBRE v danom súbore je 53,41%.



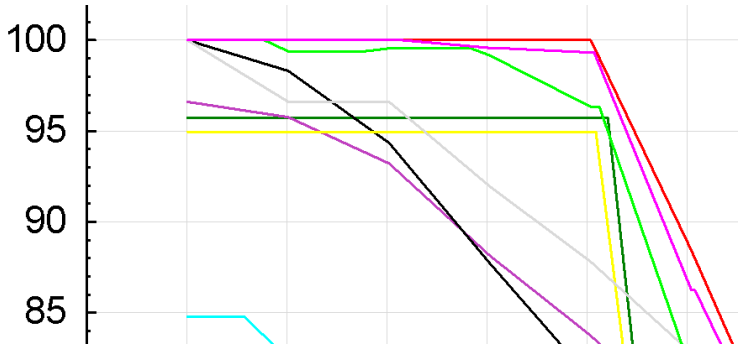
Obr. 21 *Lift chart všetkých použitých metód*

Ostatné krivky ukazujú, aký efekt prinesie použitie jednotlivých modelov. Čím viac sa krivka blíži pravému hornému rohu, tým lepšie model klasifikuje. Lift chart znázorňuje percentuálne zastúpenie vzorky zo skupiny DOBRE v určitej časti dátovej série, v ktorej sú prvky zoradené podľa pravdepodobnosti príslušnosti ku skupine DOBRE. Táto pravdepodobnosť je vypočítaná klasifikačným modelom. Zoberme z celého súboru 50%, ktorým modely predpovedali najvyššiu pravdepodobnosť, že patria do skupiny DOBRE (t. j. prvých päť decilov).

Vďaka použitému modelu SANN je v tomto výbere 100% prvkov klasifikovaných do skupiny DOBRE. Je to teda výrazne viac, než keby sme nepoužili žiadny model. Pokiaľ vypočítame podiel 100% / 53,41%, dostaneme hodnotu navýšenia (liftu). Lift vyjadruje, ako je model efektívny. V našom prípade dostávame hodnotu liftu 1,872. Teda inými slovami

pomocou metódy SANN dokážeme vybrať 50% pozorovaní tak, že výber bude obsahovať 1,872krát viac pozorovaní zo skupiny DOBRE, než by tomu bolo bez použitia modelu (Baseline).

Podobne môžeme interpretovať výsledky, ktoré sú pre kategóriu ZLE. Väčšinou záleží na odborných znalostiach a heuristike, koľko prvých decilov bude uvažovaných pre výber najlepšieho modelu.



Obr. 22 *Priblíženie Lift chartu pre prvých päť decilov*

Priblíženie Lift Chartu kategória DOBRE je na obrázku 22. Prvých päť decilov najlepšie klasifikovali modely SANN (červená), Random Forest (ružová), CandRT (svetlozelená, označená ako TreeModel). V nasledujúcej tabuľke 7 sú jednotlivé modely zoradené zostupne podľa hodnoty navýšenia v prvých piatich deciloch.

HODNOTA NAÝŠENIA – LIFT Tabuľka 7

	hodnota Lift
SANN	1.872204
Random Forest	1.859597
CandRT	1.803557
K-Nearest Neighbour	1.792406
SVM	1.777966
Boosting Tree	1.643732
MARSplines	1.567575
Naive Bayes	1.523149
CHAID	1.364384

Presnosť, s akou modely dokážu klasifikovať, preskúmame aj detailnejšie. Využijeme frekvenčné tabuľky (obrázok 23) pre pozorované, verzus predikované hodnoty závislej premennej „ciele spolu“.

	Frequency table for predicted variable SANNModelPred Observed variable: ciele spolu		
	OBSERVED ciele spolu	PREDICTED SANNModelPred	COUNT
1	ZLE	ZLE	268
2	ZLE	DOBRE	5
3	DOBRE	ZLE	3
4	DOBRE	DOBRE	310

Obr. 23 *Frequency table pre určenie presnosti modelu SANN*

Podľa týchto tabuliek prepočítame percentuálne presnosti klasifikácie jednotlivých modelov ako celkovú klasifikačnú silu, správne klasifikované pozorovania a nesprávne klasifikované pozorovania.

Výsledky sú zobrazené v tabuľke 8, kde celková klasifikačná sila zahŕňa všetky správne klasifikované pozorovania pre obe kategórie DOBRE aj ZLE. Správne klasifikované pozorovania kategórie DOBRE predstavujú správne predikované kategórie DOBRE na DOBRE. Nesprávne klasifikované pozorovania kategórie DOBRE predstavujú predikovanie pozorovaní DOBRE na hodnotu kategórie ZLE.

Aj v tomto prípade dosahuje najlepšie výsledky metóda SANN. Pri metóde CandRT je síce celková klasifikačná sila nižšia ako pri metóde Random Forest, ale z hľadiska správne a nesprávne klasifikovaného pozorovania DOBRE je presnejšia.

VYHODNOTENIE Z HĽADISKA PERCENTUÁLNEJ PRESNOSTI KLASIFIKÁCIE Tabuľka 8

	celková klasifikačná sila v %	**správne klasifikované pozorovania DOBRE (%)**	**nesprávne klasifikované pozorovania DOBRE (%)**
SANN	98.63	52.9	0.51
Random Forest	96.76	51.37	2.05
CandRT	95.91	51.54	1.88
K-Nearest Neighbour	94.2	49.83	3.58
SVM	92.32	48.29	5.12
Boosting Tree	85.83	48.12	5.29
MARSplines	81.06	44.03	9.39
CHAID	71.5	44.2	9.22
Naive Bayes	69.45	42.66	10.75

Ďalej uvádzame porovnanie modelov pomocou testov dobrej zhody – „Goodness of fit“ (tabuľka 9). Nulová hypotéza H_0: Medzi pozorovanými a predikovanými hodnotami nie je signifikantný rozdiel (početnosti v jednotlivých kategóriách sa rovnajú predikovaným početnostiam).

Testujeme na hladine signifikancie 5%. Počet stupňov voľnosti je 1, pretože máme dve kategórie DOBRE a ZLE. Hraničná hodnota Chi-kvadrátu pri 1 stupni voľnosti a hladine signifikancie 5% je 3,841 (podľa tabuľky pre nájdenie kritickej hodnoty uvádzanej v (64)).

GOODNESS OF FIT METRIKY Tabuľka 9

	Chi-kvadrát	**G-kvadrát**	**Percentuálny nesúhlas (%)**
SANN	0.12231584	16.1216479	1.36518771
Random Forest	0.662615842	38.6547871	3.24232082
CandRT	1.05066225	49.035314	4.09556314
K-Nearest Neighbour	2.16027397	70.1147467	5.80204778
SVM	4.05230504	93.9291062	7.67918089
Boosting Tree	15.6430955	180.664136	14.1638225
MARSplines	26.1764191	246.318142	18.9419795
CHAID	91.0649373	410.275528	28.4982935
Naive Bayes	101.583006	442.750053	30.5460751

Naša vypočítaná hodnota Chi pre SANN model je 0.12231584, teda je omnoho nižšia než tabuľková hodnota. Leží v obore potvrdenia nulovej hypotézy, teda predpokladáme, že nulová hypotéza platí. Znamená to, že rozdiel medzi početnosťami zistenými v pozorovaných a predikovaných kategóriách môže byť dôsledkom náhodného výberu. To znamená, že rozdiel nie je štatisticky významný. Rovnako môžeme potvrdiť nulovú hypotézu na hladine signifikancie 5% pre ďalšie tri modely Random, CandRT a KNN.

V tabuľke Goodness of fit sú jednotlivé modely zoradené podľa hodnôt metrík. Čím väčší je rozdiel medzi počtami pozorovaných a predikovaných tried, tým väčšia je hodnota G-kvadrátu, a taktiež hodnota percentuálneho nesúhlasu stúpa.

Vyhodnotenie modelov klasifikácie

Na základe všetkých výsledkov jednotlivých použitých metód bola vytvorená sumárna tabuľka 10, v ktorej sú zaznamenané poradia úspešnosti týchto modelov v daných metrikách. Bola využitá štatistická bodovacia metóda. Poradie a zároveň body od 1 po 9 sú prideľované

od najúspešnejšej metódy po najmenej úspešnú v danej metrike. Počet použitých modelov je deväť a metrík je osem, t. j. celkový najlepší možný počet bodov je 8 a najhorší možný počet bodov je 72. Metóda s najnižším prideleným počtom bodov je najlepšia a naopak s najvyšším počtom bodov je najhoršia.

SUMÁRNE VYHODNOTENIE MODELOV KLASIFIKÁCIE Tabuľka 10

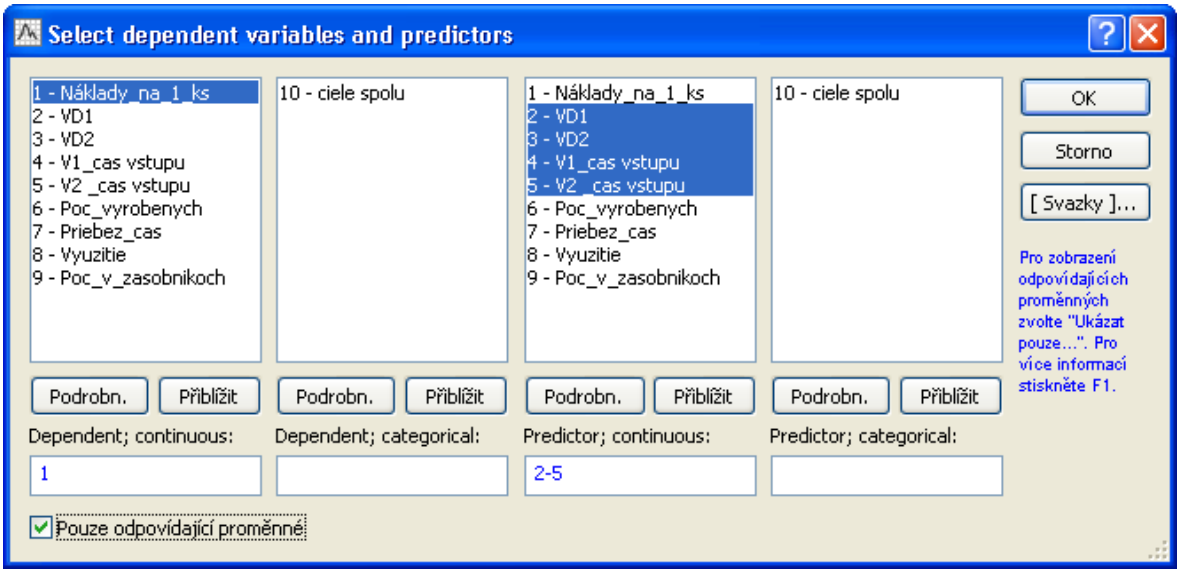
	SANN	Random Forest	CandRT	KNN	SVM	Boosting Tree	MAR Splines	CHAID	Naive Bayes
celková klasifikačná sila	1	2	3	4	5	6	7	8	9
správne klas. DOBRE	1	3	2	4	5	6	8	7	9
nesprávne klas. DOBRE	1	3	2	4	5	6	8	7	9
Error rate	1	2	3	4	5	6	7	8	9
Lift chart	1	2	3	4	5	6	7	9	8
Chi-kvadrát	1	2	3	4	5	6	7	8	9
G-kvadrát	1	2	3	4	5	6	7	8	9
Percentuálny nesúhlas	1	2	3	4	5	6	7	8	9
súčet bodov	8	18	22	32	40	48	58	63	71

Na základe porovnania a vyhodnotenia všetkých metód použijeme pre klasifikáciu na nových dátach metódu SANN, ktorá dosahuje vo všetkých metrikách najlepšie výsledky. Metódy CHAID a Naive Bayes dosiahli najhoršie výsledky, čo sme očakávali z dôvodu nesplnenia podmienky normálneho rozdelenia premenných a zároveň nesplnenia podmienky nezávislosti premenných.

4.5 Numerická predikcia

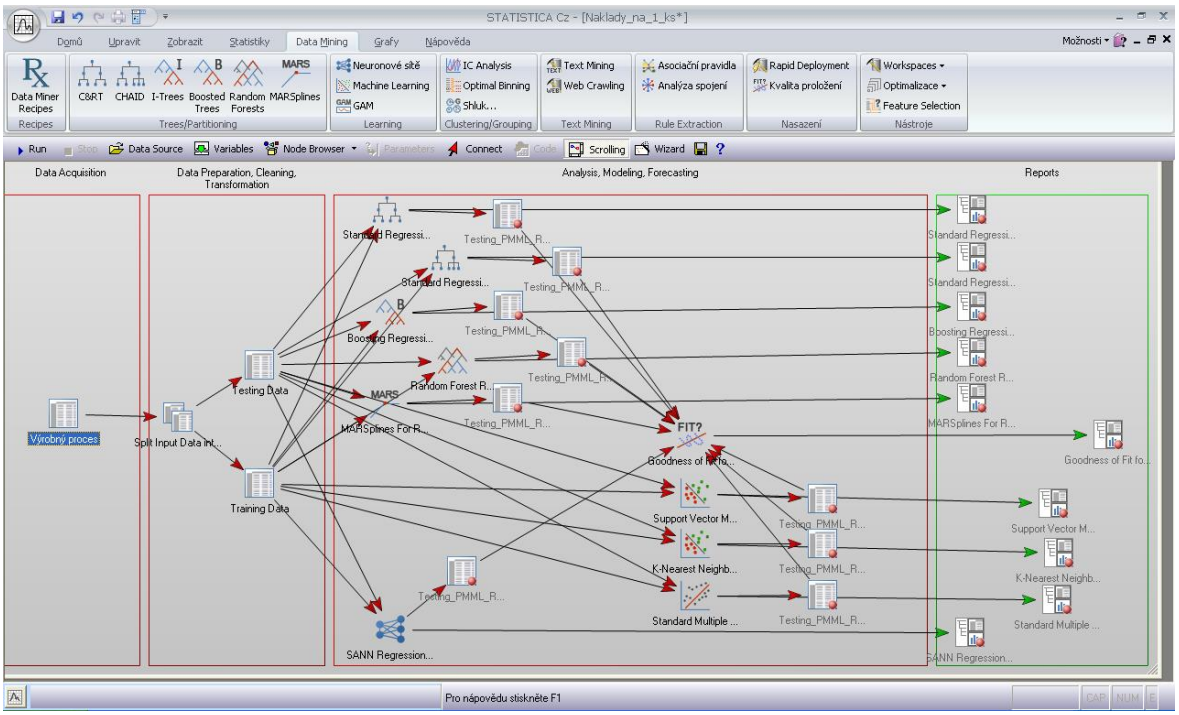
Zaujíma nás správanie sa systému pri zmene vstupných parametrov. Toto správanie budeme teraz sledovať pomocou numerickej predikcie konkrétnych ukazovateľov výrobných cieľov. Zostavíme modely, ktoré nám dokážu predikovať numerické hodnoty na základe

vzťahov medzi rôznymi nastaveniami jednotlivých vstupných regulovateľných parametrov a cieľmi výrobného procesu.



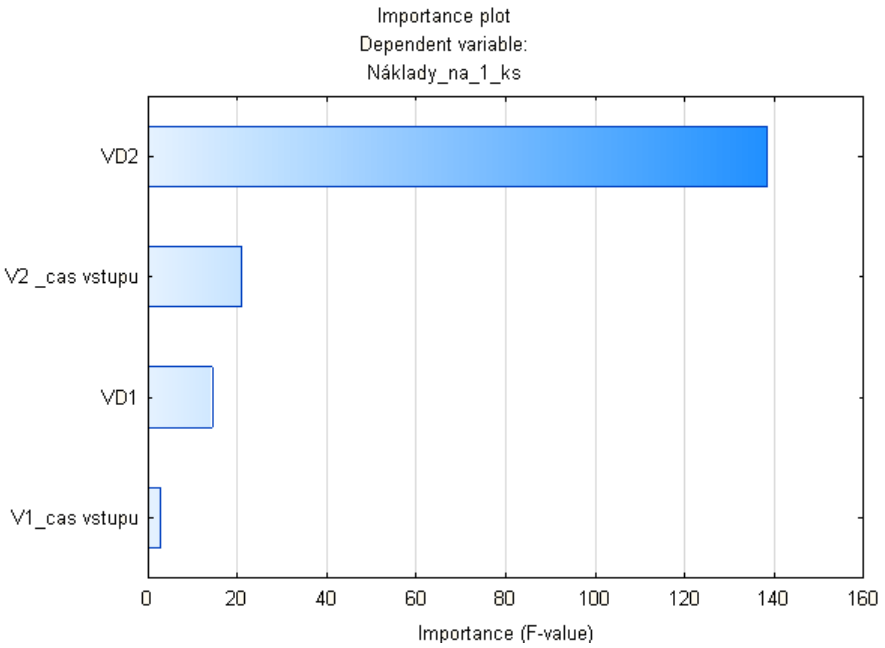
Obr. 24 *Výber závislej premennej a prediktorov*

Numerickú predikciu vykonáme pre každý výrobný cieľ samostatne. Každý dataminingový model obsahuje závislú premennú – spojitú, ktorá reprezentuje ukazovateľa výrobného cieľa a prediktory veľkosti výrobných dávok VD1, VD2, spolu s intervalmi zadávania (obrázok 24). Návrh dataminingového modelu pre numerickú predikciu je zobrazený na obrázku 25. Vytvorili sme päť dataminingových modelov pre jednotlivé výrobné ciele, a to náklady na 1 ks, počet vyrobených výrobkov, priebežný čas výroby, využitie strojov a počet výrobkov v medziskladoch – teda rozpracovanosť výroby. Na uvedenú transformovanú množinu dát bol postupne aplikovaný každý vytvorený model dolovania dát a následne z výsledkov jednotlivých modelov na testovacích dátach sme vyhodnotili presnosť ich predikcie. V návrhu dataminingového modelu numerickej predikcie sme použili deväť metód, a to rozhodovacie regresné stromy (Standard CandRT, CHAID, Boosting Trees, Random Forest, MARSplines), Strojové učenie (Support Vector Machine, K-Nearest Neighbour), Viacnásobnú regresiu (Multiple Regression) a Neurónové siete (SANN). Návrh je podobný dataminingovému modelu pre klasifikáciu, líši sa použitím metódy Multiple Regression namiesto Naive Bayes.



Obr. 25 *Návrh dataminingového modelu pre numerickú predikciu výrobného cieľa „Náklady na 1ks“*

Pre výrobný cieľ „náklady na 1 ks“ sme určili významnosť prediktorov pomocou nástroja Feature selection (obrázok 26).



Obr. 26 *Vplyv vstupných parametrov na cieľ „náklady na 1 ks“*

Najväčší vplyv z regulovateľných vstupných parametrov na výslednú kvalitu predikcie výrobného cieľa „náklady na 1 ks“ má výrobná dávka VD2 a čas jej vstupu, následne až VD1

a čas vstupu V1. Čas vstupu V1 má síce len okrajový vplyv na predikciu nákladov, ale rozhodli sme sa ho v modeli ponechať.

Presnosť numerickej predikcie analyzujeme pomocou výpočtu chýb jednotlivých metód podľa viacerých metrík. Na vyhodnotenie použitých modelov sme najprv využili modul Rapid Deployment. Pomocou neho sme zistili priemernú kvadratickú chybu rezíduí (zostatkov = average squared error residual). V tabuľke 11 sú jednotlivé metódy zoradené vzostupne od najnižšej chybovosti.

ERROR RATE PRE „NÁKLADY NA 1 KS“ Tabuľka 11

	Error rate
SANN	0.001600
K-Nearest Neighbour	0.018314
CandRT	0.021583
Random Forest	0.026241
SVM	0.052129
Boosting Tree	0.144016
CHAID	0.164278
MARSplines	0.166651
Multiple Regression	0.166850

Tabuľka 11 obsahuje chyby predikcie vybraných modelov v testovacom súbore dát. Menšie hodnoty chýb ukazujú lepšie a výkonnejšie modely. V tomto prípade dosiahol SANN model najmenšiu zostatkovú chybovosť.

Ďalej podľa metrík dobrej zhody pre numerickú predikciu sme mohli porovnať presnosť jednotlivých modelov. V nasledujúcej tabuľke 12 sú metódy zoradené vzostupne podľa dosiahnutých hodnôt v jednotlivých metrikách. Najmenšie chyby v predikcii sú práve pri metóde SANN, ktorá sa javí ako najpresnejšia.

GOODNESS OF FIT PRE PREMENNÚ „NÁKLADY NA 1 KS“ Tabuľka 12

	Mean square error	**Mean absolute error**	**Mean relative squared error**	**Mean relative absolute error**	**Correlation coefficient**
SANN	0.00302435977	0.0412196534	0.000173186886	0.0102213569	0.995525969
CandRT	0.0196653699	0.0816930626	0.00120265334	0.0201803379	0.970602989
K-Nearest Neighbour	0.0364194664	0.125211435	0.00212260008	0.0304546601	0.947707591

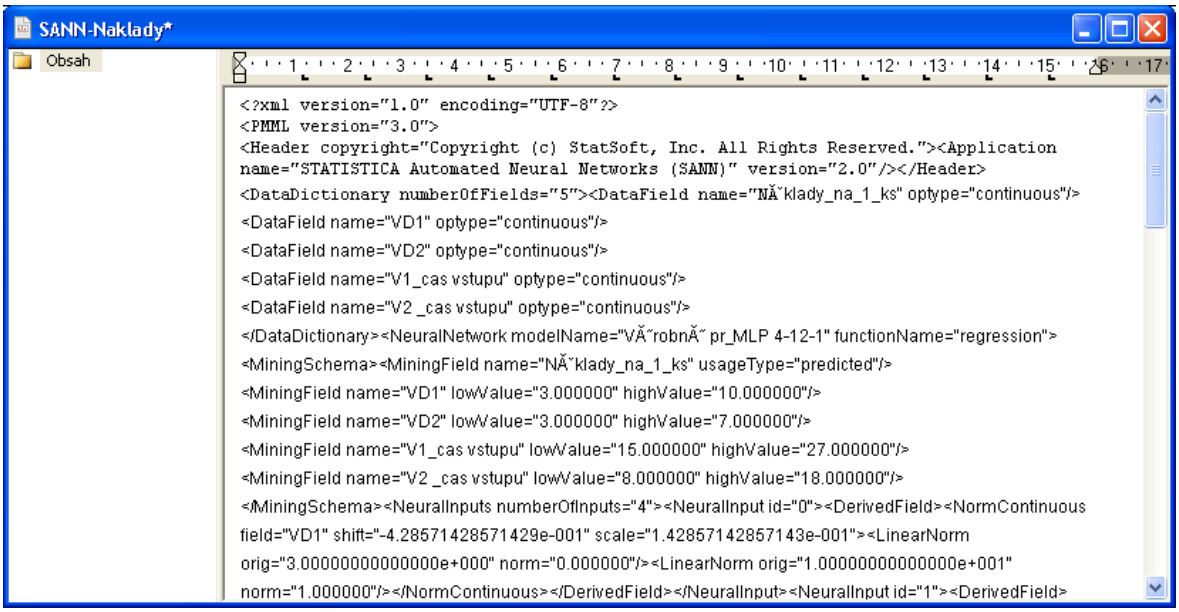
SVM	0.0506216211	0.181379106	0.00328550644	0.0463830423	0.925350905
Random Forest	0.0685247262	0.201773538	0.00427675699	0.0511670527	0.923186498
CHAID	0.133628755	0.268972113	0.00978380826	0.0716485586	0.778829863
Boosting Tree	0.163222033	0.31928279	0.0112108909	0.0830061994	0.725067214
MARSplines	0.163970652	0.326338163	0.0115438836	0.0852848012	0.719605363
Multiple Regression	0.174044743	0.335287136	0.0122151193	0.0876591252	0.699072046

Podobne boli vykonané dataminingové modely pre každý cieľ výroby – Počet vyrobených výrobkov, Využitie, Počet v zásobníkoch a Priebežný čas. Metóda SANN dosiahla najlepšie výsledky aj vo vyhodnotení Goodness of Fit pre ostatné ciele.

Metóda SANN pre výrobný cieľ využitie predikuje so strednou kvadratickou chybou 0.0427, počet výrobkov v zásobníku s chybou 0.0388, priebežný čas s chybou 5.301. Pre výrobný cieľ počet vyrobených výrobkov vznikla väčšia chyba predikcie 38.156, ktorú sa nám síce podarilo zmenšiť na 0.004539 pomocou štandardizácie. Štandardizácia premennej bola vykonaná pomocou vzorca $\acute{x}_{ij} = \frac{x_{ij}-\bar{x}}{s}$, kde x_{ij} je pozorovaná hodnota, $\bar{x}$ je priemer a s smerodajná odchýlka. Po spätnej transformácii premennej sme však dostali horšie výsledky rezíduí. Môže to byť spôsobené veľkým rozsahom hodnôt tejto premennej, ktoré sú uvedené v tabuľke 2 - Popisné štatistiky.

4.6 Nasadenie vybraných metód na nové dáta

Použitie už natrénovaných modelov na nových dátach či na vzorke dát umožňuje modul Rapid Deployment. Pomocou tohto modulu je možné predikované hodnoty zapísať späť do vstupného súboru dát, rovnako aj do externých databáz alebo dátového skladu. Pre správne nasadenie je nevyhnutné, aby premenné v súbore s novými dátami mali rovnaký názov a boli rovnakého typu ako v trénovacom súbore. Postupné nasadenie prediktívnych modelov na nové hodnoty sme teda realizovali pomocou tohto modulu. Na obrázku 27 je ukážka PMML súboru metódy SANN určená na predikciu nákladov na 1 kus.



Obr. 27 *Ukážka PMML súboru metódy SANN na predikciu nákladov*

Z analýzy dát vyplýva, že je lepšie, ak čas vstupu VD1 je nižší a naopak, čas vstupu VD2 je vyšší, zároveň sa ovplyvňujú aj veľkosti výrobných dávok – ak je jedna vyššia druhá je nižšia. S ohľadom na túto analýzu dát sme zvolili nové hodnoty vstupných parametrov. Boli zvolené aj hodnoty parametrov, ktoré sa nenachádzali v trénovacom súbore dát. Nové nastavenia vstupných parametrov výrobného procesu sú nasledovné:

I. VD1=3 , VD2=7, V1_cas vstupu=10, V2_cas vstupu=21

II. VD1=5 , VD2=4, V1_cas vstupu=15, V2_cas vstupu=19

III. VD1=6 , VD2=3, V1_cas vstupu=16, V2_cas vstupu=18

IV. VD1=7 , VD2=5, V1_cas vstupu=20, V2_cas vstupu=11

V. VD1=8 , VD2=6, V1_cas vstupu=18, V2_cas vstupu=20

Nasadenie na nové dáta sme najprv urobili pre všetky dostupné metódy pomocou natrénovaných súborov PMML a výsledky klasifikácie jednotlivých metód sú v tabuľke 13, ďalej pre numerickú predikciu cieľa náklady na 1 kus sú výsledky v tabuľke 14.

KLASIFIKÁCIA NA NOVÝCH DÁTACH SO VŠETKÝMI METÓDAMI Tabuľka 13

	Boosting Tree	CandRT	CHAID	KNN	MAR Splines	Naive Bayes	Random Forest	SANN	SVM	Voted prediction
I.	ZLE	DOBRE	ZLE	DOBRE	ZLE	ZLE	DOBRE	DOBRE	ZLE	ZLE
II.	DOBRE	DOBRE	DOBRE	DOBRE	DOBRE	DOBRE	DOBRE	DOBRE	DOBRE	DOBRE
III.	DOBRE	DOBRE	DOBRE	DOBRE	DOBRE	DOBRE	DOBRE	DOBRE	DOBRE	DOBRE
IV.	ZLE	ZLE	ZLE	ZLE	ZLE	ZLE	ZLE	ZLE	ZLE	ZLE
V.	ZLE	ZLE	ZLE	ZLE	ZLE	ZLE	ZLE	ZLE	ZLE	ZLE

Metódy Boosting Tree, CHAID, MARSplines, Naive Bayes a SVM nesprávne klasifikovali I. nastavenie vstupných parametrov. STATISTIKA označila prvý riadok ako nezhodu v klasifikácii (tabuľka 13). Niekedy je možné využiť aj Voted prediction – teda predikciu hlasovaním všetkých metód (používa sa súčet hlasov). Podobne pre numerickú predikciu je možné využiť priemernú predikovanú hodnotu všetkých metód – Avg. prediction.

NUMERICKÁ PREDIKCIA NÁKLADOV NA 1 KS S POUŽITÍM VŠETKÝCH METÓD

Tabuľka 14

	SVM	Boosting Tree	CandRT	CHAID	KNN	MAR Splines	Random Forest	Multiple Regression	SANN	Avg. prediction
I.	4.317	4.541	4.955	4.955	4.269	4.527	4.457	4.271	3.930	4.469
II.	3.286	3.316	3.274	3.548	3.291	3.323	3.425	3.254	3.275	3.332
III.	3.567	3.127	3.274	3.548	3.340	2.937	3.473	3.022	3.389	3.297
IV.	5.485	4.612	4.867	4.867	4.868	4.743	4.505	4.805	5.671	4.936
V.	4.134	4.649	4.103	4.146	4.490	4.399	4.394	4.203	4.296	4.312

Pre klasifikáciu aj numerickú predikciu sme vybrali metódu SANN, a to na základe predchádzajúceho vyhodnotenia všetkých modelov. Postupne sme aplikovali jednotlivé súbory PMML pri klasifikácii a potom pri predikcii jednotlivých ukazovateľov cieľov procesu.

Najprv sme klasifikáciou zistili správanie sa systému, a teda či pri daných vstupných parametroch bude súčasne spĺňať požadované kvantitatívne charakteristiky. V tabuľke 15 je uvedená výsledná klasifikácia pri zvolených nastaveniach vstupných parametrov. Šedou zvýraznené časti tabuliek sú hodnoty získané predikciou. Pri metóde SANN sú prvé tri nastavenia klasifikované na hodnotu DOBRE, to znamená, že je predpovedané splnenie všetkých charakteristík súčasne, a teda proces s týmito nastaveniami by mal byť úspešný. Nastavenia parametrov IV a V podľa klasifikácie nesplnia charakteristiky pre všetky výrobné ciele súčasne. Overenie klasifikácie na simulácii je uvedené v tabuľke 17, kde na základe simuláciou získaných hodnôt jednotlivých charakteristík je možné určiť, či nastavenia spĺňajú alebo nespĺňajú, požadované podmienky stratégie riadenia.

SANN – KATEGORICKÁ PREDIKCIA NA NOVÝCH DÁTACH Tabuľka 15

Nastavenia parametrov	VD1	VD2	V1_cas vstupu	V2 _cas vstupu	ciele spolu
I.	3	7	10	21	DOBRE
II.	5	4	15	19	DOBRE
III.	6	3	16	18	DOBRE
IV.	7	5	20	11	ZLE
V.	8	6	18	20	ZLE

Následne sme zisťovali hodnoty konkrétnych ukazovateľov kvantitatívnych charakteristík nasadením numerickej predikcie na nové dáta. Tieto predikované hodnoty potvrdili aj predchádzajúcu klasifikáciu, že IV. a V. nastavenie parametrov nesplní charakteristiky pre všetky výrobné ciele súčasne. A teda nespĺňajú požiadavku na rozpracovanú výrobu, kedy má byť počet v zásobníkoch menší ako 4 kusy a súčasne požiadavku na priebežný čas, ktorý má byť pod 110 minút.

PREDIKOVANÉ HODNOTY NA ZÁKLADE NASTAVENÝCH PARAMATETROV Tabuľka 16

Nastavenia parametrov:	I.	II.	III.	IV.	V.
VD1	3	5	6	7	9
VD2	7	4	3	5	6
V1 čas vstupu	10	15	16	20	18
V2 čas vstupu	21	19	18	11	20
Náklady na 1 ks	3.929976	3.274970	3.388896	5.671034	4.295689
Počet výrobkov	912.538	822.005	844.494	977.078	1048.968
Priebežný čas	107.7801	54.23636	60.77595	212.1174	145.236
Využitie	80.54391	72.25269	73.16904	92.83138	93.16227
Počet v zásobníkoch	4.000555	1.10235	1.682129	9.524493	6.154498

Za účelom overenia získaných predikovaných hodnôt sme nové nastavenia vstupných parametrov implementovali na navrhnutý simulačný model v programe Witness, ktorého dáta slúžili ako vstupné dáta do procesu získavania znalostí z databáz. Zámerom bolo porovnať simulované a predikované hodnoty. A zároveň určiť, či by vybrané modely a konkrétne algoritmy v PMML súboroch mohli byť použité pre rozhodovanie v riadení daného procesu.

SIMULÁCIOU ZÍSKANÉ HODNOTY PRE NOVÉ NASTAVENIE VSTUPNÝCH PARAMETROV Tabuľka 17

Nastavenie parametrov	Náklady na 1 ks	Počet výrobkov	Priebežný čas	Využitie	Počet v zásobníkoch	Ciele spolu
I.	4.006	907	110.379	80.052	3.999	DOBRE
II.	3.266	841	55.68	71.707	1.399	DOBRE
III.	3.345	835	57.072	72.468	1.518	DOBRE
IV.	5.774	973	212.101	90.963	11.155	ZLE
V.	4.189	1043	134.249	92.634	6.031	ZLE

Ako je vidieť v poslednom stĺpci „ciele spolu“ v tabuľke 17, simulácia potvrdila klasifikáciu jednotlivých nastavení. Hodnoty v tomto stĺpci sa zhodujú s rovnakým stĺpcom v tabuľke 15. Týmto môžeme klasifikáciu pomocou metódy SANN považovať za 100% úspešnú.

Teraz porovnáme numerické hodnoty simulácie a predikcie, teda tabuľky 16 a 17. Najvyššie rozdiely vznikali pri predikcii počtu vyrobených výrobkov. Počet vyrobených výrobkov predikovaný modelom SANN sa od reálnej hodnoty líši približne od 4 do 19 kusov, konkrétne rezíduá jednotlivých predikcií počtu vyrobených výrobkov sú v nasledujúcej tabuľke 18.

REZÍDUÁ PRE PREMENNÚ „POČET VYROBENÝCH“ Tabuľka 18

	Simulácia	Predikcia	Rezíduá
I.	907	912.538	-5.53836
II.	841	822.005	18.99537
III.	835	844.494	-9.49427
IV.	973	977.078	-4.07815
V.	1043	1048.968	-5.96836

Podľa popisnej štatistiky, ktorú sme vykonali pri analýze dát, sa počet vyrobených výrobkov v dátovom súbore pohybuje v rozmedzí 671 – 1058. Chybu predikcie 19 kusov preto môžeme považovať za pomerne dobrý výsledok.

Simulácia potvrdila, že predikované metódy sú správne určené pre všetky sledované ciele výroby. Predikované hodnoty nákladov na 1 kus boli veľmi presné a maximálna odchýlka bola približne 10 centov pri IV. a V. nastavení vstupných parametrov. Predikcia priebežného času bola tiež veľmi presná, najväčšia odchýlka 10.9870, t. j. približne 11 minút, bola pri nastavení

číslo VI. Veľmi presne boli predikované aj hodnoty využitia strojov, kde maximálna odchýlka bola 1.86838 % pri IV nastavení parametrov. V numerickej predikcii rozpracovanej výroby (počet v zásobníkoch) bola najvyššia odchýlka 1.630507 pri nastavení číslo IV., teda rozdiel približne 2 kusy. Všetky uvedené odchýlky môžeme považovať za prijateľné.

Na základe týchto výsledkov môžeme konštatovať, že zvolené algoritmy (PMML) predikcie – klasifikácie aj numerickej predikcie – sú vhodné na implementáciu do informačných a podporných systémov riadenia. Vieme potvrdiť, že zvolené vstupné parametre povedú k splneniu deklarovaných cieľov procesu a zároveň dokážeme predikovať konkrétne hodnoty cieľov s tolerovanou presnosťou. Pre každý vstup vieme predikovať očakávané výstupy.

Formulácia znalostí

Na základe vykonaného procesu dolovania dát môžeme formulovať získané znalosti. Znalosti by sa dali reprezentovať formou produkčných pravidiel, ktoré majú tvar *AK podmienka POTOM dôsledok*. Rozhodli sme sa použiť formuláciu znalostí jednoduchou reprezentáciou komplexných faktov, pretože cieľom práce nebolo venovať sa formulácii znalostí.

- Vstupné parametre nemajú rovnaký vplyv na sledované ciele. Ich vplyv sa mení podľa zvoleného cieľového kritéria. Pre daný výrobný systém vieme deklarovať, že pre dosiahnutie požadovaných cieľov náklady na jednotku výroby, priebežný čas a rozpracovanosť výroby v konkrétnom kvantitatívnom vyjadrení má najväčší vplyv zmena veľkosti výrobnej dávky výrobku dva.
- Pre dosiahnutie konkrétne definovaných cieľov využite strojov a celkový počet vyrobených kusov má najväčší vplyv zmena veľkosti výrobnej dávky prvého výrobku a následne druhého.
- Zvyšovanie nákladov na jeden kus spôsobuje zvyšovanie priebežného času, pričom pre daný systém je táto závislosť lineárna.
- V danom výrobnom systéme existuje vysoká závislosť medzi nákladmi na jednotku výroby a rozpracovanou výrobou (korelačný koeficient je 0,9).
- V danom výrobnom systéme existuje vysoká závislosť medzi počtom vyrobených výrobkov a využitím strojov.
- V danom výrobnom systéme existuje vysoká závislosť medzi priebežným časom a rozpracovanou výrobou.

- Nastavenia veľkosti výrobných dávok sa navzájom ovplyvňujú. V riadení výrobného systému treba uvažovať, že ak sa zvýši veľkosť jednej dávky, je potrebné znížiť veľkosť druhej dávky. Závislosť je definovaná konkrétnym korelačným koeficientom.
- Vieme prognózovať, či zvolené vstupné parametre povedú k dosiahnutiu stanovených cieľov výroby priradením kategórie splniteľnosti.
- Vieme prognózovať, k akým numericky vyjadreným cieľovým hodnotám povedú zvolené vstupné parametre.

Pre tieto znalosti môžeme vytvoriť produkčné pravidlá nasledovne:

AK veľkosť výrobnej dávky druhého výrobku VD2 je vyššia,
POTOM náklady na jednotku výroby sú zvýšené;
priebežný čas je zvýšený;
rozpracovanosť výroby je zvýšená.

AK veľkosť výrobnej dávky prvého výrobku VD1 je vyššia,
POTOM využitie strojov je zvýšené;
celkový počet vyrobených výrobkov je zvýšený.

AK náklady na jednotku výroby sú vyššie,
POTOM priebežný čas je zvýšený;
rozpracovanosť výroby je zvýšená.

AK veľkosti výrobných dávok oboch výrobkov a ich intervaly zadávania sú známe,
POTOM kategória splnenia výrobných cieľov môže byť určená;
numerická hodnota cieľových parametrov môže byť určená.

ZÁVER

Monografia sa zaoberá využitím procesu získavania znalostí z databáz v oblasti plánovania a riadenia procesov. Je zameraná na predikciu správania sa systému z dát výrobného procesu, ktoré sa využívajú pri plánovaní a riadení. Realizovali sme klasifikáciu a numerickú predikciu. Vytvorili sme niekoľko predikčných modelov a pomocou vybraných metód a techník dolovania dát sme predikovali správanie sa výrobného procesu na základe zmeny vstupných parametrov. Vieme zistiť, či zvolené parametre povedú k splneniu všetkých cieľov procesu a zároveň s určitou tolerovanou presnosťou dokážeme predpovedať konkrétne hodnoty kvantitatívnych ukazovateľov cieľov výroby. Dôležitým výsledkom práce bolo vytvorenie PMML súboru vhodného na nasadenie do informačných a podporných systémov pre riadenie konkrétneho procesu. Tento súbor vo formáte XML je plne integrovateľný s väčšinou systémov pre oblasť dataminingu.

Analýzou problémov riadenia výrobných procesov a definovaných cieľov výroby, ktoré sme sledovali, boli identifikované možné príležitosti, problémy a závislosti, ktoré sa vyskytujú vo výrobnom procese a bolo by možné riešiť ich metódami dolovania dát. V práci sú identifikované príležitosti uplatnenia procesu KDD na rôznych hierarchických úrovniach riadenia podniku od procesnej až po podnikovú úroveň. Plánovanie a riadenie výroby musí zabezpečiť dosiahnutie rôznych cieľov výroby v danom časovom horizonte. Tieto ciele sú často protichodné a ich dosiahnutie závisí od mnohých faktorov.

Keďže dáta z reálneho výrobného procesu sa nám nepodarilo získať, jedným z čiastkových cieľov práce muselo byť vytvorenie simulačného modelu. Tento model reprezentoval typického predstaviteľa pre malosériovú – dávkovú výrobu, kde sa súčasne vyrába viac typov výrobkov. Vytvorený simulačný model slúžil na generovanie výrobných údajov pre riadenie tohto procesu. Generované dáta boli uložené v databáze. Súčasne tento model bol použitý na overenie získaných znalostí, čo reprezentuje validáciu prognóz cieľových parametrov.

Pri návrhu modelu dolovania dát bolo potrebné rozhodnúť, ktoré metódy budú použité pre vytváraný model z pohľadu dát, ktoré boli získané z výrobného systému a zároveň tieto metódy museli rešpektovať zvolené úlohy dolovania dát. Toto bolo realizované vo viacerých krokoch. V prvom kroku bola vykonaná analýza dát získaných z daného výrobného systému s využitím vybraného nástroja dolovania dát. Dôležité bolo získanie základného prehľadu o analyzovaných dátach popisnou štatistikou pred ich ďalším použitím v procese dolovania dát. Po vhodnom výbere, transformácii a úprave dát mohol byť vykonaný ďalší krok, a to vytvorenie

návrhu dolovacích modelov s použitím vhodných metód. Pre vytváraný model museli byť vybrané predikované a predikujúce premenné. V rámci tohto kroku bolo vykonané dolovanie dát za účelom získania predikcie pre analyzovaný výrobný proces. Tieto predikcie pomôžu dosiahnuť väčšiu istotu v plánovaní a riadení výrobného systému. Posledným krokom bolo vyhodnotenie použitých metód pre obe úlohy dolovania dát (klasifikačnú a numerickú predikciu) na základe viacerých metrík určujúcich presnosť modelov. Na základe porovnania metód mohla byť vybraná vhodná metóda pre nasadenie klasifikácie aj numerickej predikcie na nových dátach, a to konkrétne metóda SANN – neurónové siete.

Správanie sa systému bolo predikované pomocou klasifikácie a numerickej predikcie, vychádzajúc z úloh dolovania dát. Prvou úlohou dolovania dát bolo predikovať splnenie všetkých cieľových parametrov súčasne a druhou úlohou dolovania dát bolo predikovať konkrétne numerické hodnoty cieľových parametrov výroby. Na základe zmeny vstupných parametrov vieme určiť, či zvolené nastavenie vstupných parametrov povedie k očakávanému splneniu cieľov procesu a zároveň dokážeme predikovať konkrétne kvantitatívne ohodnotenie týchto cieľov.

Pomocou vytvorených PMML súborov bolo možné aplikovať naučené metódy na nové dáta. Predikcia správania sa systému s novým nastavením vstupných parametrov bola následne overená pomocou simulácie. Na simulačnom modeli boli verifikované všetky predikované hodnoty a simulácia potvrdila predikciu.

ZOZNAM BIBLIOGRAFICKÝCH ODKAZOV

1. ADASTRA: Medzinárodná konzultačná spoločnosť. 2012. Business Intelligence & Data Warehouse [cit. 2012-01-17]. Dostupné na internete: <http://www.sk.adastragrp.com/dokument.aspx?id=9>
2. ALBRECHT M.C. 2010. Introduction to Discrete Event Simulation. [cit. 2012-01-13]. Dostupné na internete: <http://www.albrechts.com/mike/DES/Introduction%20to%20DES.pdf>
3. BALAŽOVIČ, Igor. 2013. BI, BA alebo BusinessDiscovery?. In Infoware IX. (6-7), pp. 20-21. ISSN: 1335−4787.
4. BERANOVÁ, Petra. 2008. Příklad efektivního využití dataminingových metód v oblasti kontroly kvality výroby. © Publikační činnost Statsoft CR s.r.o. [cit. 2012-12-10]. Dostupné na internete: <www.statsoft.cz>
5. BERSON, A., SMITH, S., THEARLING, K. 1999. *Building data mining applications for CRM*. New York: McGraw-Hill. ISBN 978-0071344449
6. BREIMAN, L. 1996. Bagging Predictors. *Machine Learning*. **24**,123-140.
7. BREIMAN, L. 2001. Random Forests. *Machine Learning*. **45**, 5-32.
8. BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., STONE, C.J. 1984. Classification and Regression Trees. New York: Chapman & Hall. ISBN 0-412-04841-8
9. CASSANDRAS G.G., STRICKLAND, S. G. 1989. Sample path properties of time discrete event systems. *Proceedings of the IEEE*. **77**(1), 59-71. ISSN 0018-9219
10. ČARNICKÝ, Š. 2006. Business Intelligence v riadení podnikov a prístupy k jeho riešeniu. In: *Podniková revue,* **5**(9), 36 – 52. ISSN 1335-9746
11. DRESNER, Howard, 2007. The Performance Management Revolution: Business Results Through Insight and Action, John Wiley&Sons.
12. DYNAMIC FUTURE: Organizácia poskytujúca logistické poradenstvo. Witness power with easy, 2012 [cit. 2012-01-07]. Dostupné na internete: <http://www.dynamicfuture.cz/witness/>
13. EFRON, B., TIBSHIRANI, R.J. 1993. *An Introduction to the Bootstrap*. London: Chapman and Hall. ISBN 0-412-04231-2
14. FAYYAD, U. M., PIATETSKI-SHAPIRO, G., SMYTH, P. 1996a. From Data Mining to Knowledge Discovery: An Overview. In: *Advances in Knowledge Discovery and Data Mining*. California: AAAI Press/The MIT Press, s. 1-34. ISBN 02-62560-97-6

15. FAYYAD, U.M., PIATETSKI-SHAPIRO, G., SMYTH, P. 1996b. The KDD Process for Extracting Useful Knowledge from Volumes of Data. *COMMUNICATIONS OF THE ACM*, **39**(11), 27-34. ISSN 0001-0782
16. FENG, J.C.-X., KUSIAK, A. (2006). Data mining applications in engineering design, manufacturing and logistics. *International Journal of Production Research*, 44:14, 2689-2694, DOI:10.1080/00207540600681072.
17. FIKES, R.E. , NILSSON, N.J. 1971. Hierarchical Control, STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence (IJCAI'71)*, 608-620. [cit. 2012-12-10]. Dostupné na internete: <http://ijcai.org/Past% 20Proceedings/IJCAI-1971/CONTENT/content.htm>
18. FISHMAN, G. S. 2001. *Discrete-event simulation: modeling, programming, and analysis.* New York: Springer. ISBN 0-387-95160-1
19. FRIEDMAN, H. Jerome. 1999b. Stochastic Gradient Boosting. [cit. 2012-10-10]. Dostupné na internete: <http://www-stat.stanford.edu/~jhf/>
20. FRIEDMAN, H. Jerome. 1999a. Greedy Function Approximation: A Gradient Boosting Machine. [cit. 2012-10-10]. Dostupné na internete: <http://www-stat.stanford.edu/~jhf/>
21. FRIEDMAN, H. Jerome. 1991. Multivariate Adaptive Regression Splines. In: *Annals of Statistics **19*** (1-141),
22. GÁLA, L., POUR, J., ŠEDIVÁ, Z. 2009. *Podniková informatika: 2., přepracované a aktualizované vydání.* Praha: Grada. ISBN 978-80-247-2615-1
23. GIUDICI, P., FIGINI, S. 2009. *Applied Data Mining for Business and Industry*. 2nd Edition. United Kingdom: Wiley Computer Publishing. ISBN 978-0-470-05887-9
24. GRAUER, M., KARADGI, S., MILLER, U., METZ, D., and SCHEFER, W. (2010). Proactive Control of Manufacturing Processes Using Historical Data. *Knowledge-Based and Intelligent Information and Engineering Systems*, 6277, 399-408.
25. HAN, J., KAMBER, M., PEI, J. 2006. *Data Mining: Concepts and Techniques.* Canada: Morgan Kaufmann Publishers. ISBN 0080475582
26. HAND, D., MANNILA, H., SMYTH, P. 2001. *Principles of Data Mining.* MESTO: MIT Press. ISBN 0-262-08290-X.
27. HILL,T., LEWICKI, P. 2005. *STATISTICS: Methods and Applications Book.* Tulsa: Statsoft, Inc. ISBN-10: 1884233597
28. HSU, Ch.-W., CHANG, Ch.-Ch., LIN, C.-J. 2003. A Practical Guide to Support Vector Classification. [cit. 2012-12-10]. Dostupné na internete: <www.csie.ntu.edu.tw/~cjlin/ talks/freiburg.pdf>

29. HUŠEK, R., LAUBER, J. 1987. *Simulační modely*. Praha: SNTL/ALFA. ISBN 04-326-87
30. CHAUDHURI Surajit, DAYAL Umeshwar, NARASAYYA Vivek: An Overview of Business Intelligence Technology, In Communications of the ACM, Vol. 54 No. 8, Pages 88-98. August 2011. Dostupné na internete: <http://cacm.acm.org/ magazines/2011/8/114953-an-overview-of-business-intelligence -technology/fulltext>
31. CHOUDHARY, A. K., HARDING, J. A., and TIWARI, M. K. (2009). Data mining in manufacturing: a review based on the kind of knowledge. Journal of Intelligent Manufacturing, 20, 501-521.
32. INMON, W. H. 2005. *Building the Data Warehouse.* United States: Wiley Publishing. ISBN: 0-7645-9944-5
33. JUROVATÁ, D. KEBÍSEK, M., KURNÁTOVÁ, J. 2012. Application of Knowledge Discovery in Databases Process in the Production Systems. In: *CECIIS 2012*. Varaždin: University of Zagreb, s. 37-42. ISSN 1847-2001
34. KASS, G.V. 1980. An Exploratory Technique for Investigating Large Quantities of Categorical Data. In: *Applied Statistics.* ***29*** (2), 119-127. Publisher: JSTOR. ISSN 00359254
35. KDnuggets: Data Mining Community‘s Top Resource. © 2011 [cit. 2012-10-10]. Dostupné na internete: <http://www.kdnuggets.com>
36. KOMPRDOVÁ, Klára. 2012. *Rozhodovací stromy a lesy.* Brno: IBA. ISBN 978-80-7204-785-7
37. KOSKOVÁ, Gabriela. 2007. Vyhodnocovanie metód dolovania znalostí. Prednášky. Bratislava: FIIT STU
38. KUPLOVIČ, Milan. 2002. *Podnikové hospodárstvo: komplexný pohľad na podnik.* Bratislava: Sprint. ISBN 8088848938
39. KUZÁR, Tomáš. 2012. BigData – keď Business Intelligence nie je len sluha. Blog, DWH/BI. [cit. 2013-11-02]. Dostupné na internete: <http://robime.it/bigdata-ked-business-intelligence-nie-je-len-sluha/>
40. KÖKSAL, G., BATMAZ, I., TESTIK, M. C. 2011. A review of data mining applications for quality improvement in manufacturing industry. *Expert Systems with Applications*, 38, 13448-13467.
41. LACKO, Ľ. 2005. *Business Intelligence v SQL Serveru 2005*. Praha: Computer Press. ISBN: 9788025111109
42. LANNER: Provider of simulation software. WITNESS. 2012. [cit. 2012-01-07]. Dostupné na internete: <http://www.lanner.com/en/witness.cfm>

43. LOUDA, P. 2008. Business intelligence v novém hávu. *Computerworld*. ISSN 1210-9924
44. MELOUN, M., MILITKÝ ,J., HILL, M. 2005. *Počítačová analýza vícerozměrných dat v příkladech*. Praha: Akademia. ISBN 80-200-1335-0
45. MIČIETA, B., KRÁL, J. 1998. *Plánovanie a riadenie výroby.* Žilina: Žilinská univerzita v Žiline. ISBN 80-7100-430-8
46. MILTENBURG, John. 2005. *Manufacturing strategy: How to formulate an implement a winning plan.* New York: Productivity Press. ISBN 1-56327-317-9
47. MITCHELL, M. Tom. 1997. Machine learning. United States: MIT Press and McGraw-Hill. ISBN 0-07-042807-7
48. MLADENIC, D., LAVRAC, N., BOHANEC, M., MOYLE, S. 2003. *Data mining and decision support: Integration and collaboration.* United States: Kluwer Academic Publisher. ISBN 1-4020-7388-7
49. NELSON, B.L., NICOL, D. M., BANKS, J., CARSON, J.S. 2005. *Discrete-event system simulation*, 4. vyd. Upper Saddle River: Pearson Prentice Hall. ISBN 0-13-129342-7
50. NOVOTNÝ, O., POUR, J., SLÁNSKÝ, D. 2005. *Business intelligence : Jak využít bohatství ve vašich datech.* Praha: Grada. ISBN 80-247-1094-3.
51. ODED, M., LIOR, R. 2005. *Data Mining and Knowledge Discovery Handbook*. USA: Springer. ISBN 0-387-24435-2
52. PAIDI, Annan Naidu. 2012. *DataMining: FutureTrends and Applications*. In International Journal of Modern Engineering Research (IJMER),Vol.2, Issue.6, Nov-Dec. 2012 pp. 4657-4663. ISSN: 2249-6645
53. PARALIČ, Ján. 2003. *Objavovanie znalostí v databázach*. Košice: Elfa. ISBN 80-89066-60-7
54. POWER, D.J. 2007. *A Brief History of Decision Support Systems*. DSSResources.COM. [cit. 2012-12-10]. Dostupné na internete: <http://dssresources. com/history/dsshistory.html>
55. R&D: Informatics Research and Development Laboratory. *Open Source Data Mining Software Evaluation:* US Centers for Disease Control and Prevention, 2010. [cit. 2012-01-07]. Dostupné na internete: <http://www.phiresearchlab.org/downloads/ OpenSourceDataMining.pdf>
56. Rexer Analytics: *4th Annual Data Miner Survey – 2010 Survey Summary Report.* 2010. [cit. 2012-01-13]. Dostupné na internete: <http://www.rexeranalytics.com/Data-Miner-Survey-2010-Intro2.html>

57. RIMARČÍK, Marián. 2007. Štatistika pre prax. Košice: Marián Rimančík. ISBN 978-80-969813-1-1
58. ŘEPA, V. 2007. *Podnikové procesy: Procesní řízení a modelování.* 2-vyd. Praha: Grada. ISBN 978-80-247-2252-8
59. SARNOVSKY, J., MADARASZ, L., BIZÍK, J., CSONTÓ, J. 1992. *Riadenie zložitých systémov*, 1.vyd. Bratislava: Alfa. ISBN 80-05-00945-3
60. SAS: vendor in the business intelligence. 2012. *SAS Enterprise Miner - SEMMA*. [cit. 2012-01-13]. Dostupné na internete: <http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html>
61. SHANNON, R.E. 1975. *Systems Simulation – the art and science.* New Jersey: Prentice-Hall. ISBN 9780138818395
62. STATSOFT, Inc. 2011. STATISTICA (data analysis software system), version 10. [cit. 2012-01-13]. Dostupné na internete: <www.statsoft.com>.
63. STATSOFT, Inc.: Global provider of software for statistics, data analysis. 2012. *STATISTICA Data Miner*. [cit. 2012-01-13]. Dostupné na internete: <http://www.statsoft.com/products/statistica-data-miner/>
64. SULLIVAN, Michael. 2010. *Statistics: Informed Decisions Using Data III*. Upper Saddle River, N.J. : Prentice Hall/Pearson. ISBN-10: 0321568028
65. ŠMEHÝL, M., Kremeňová, I. 2003. *Možnosti využitia dataminingu v podnikovej praxi.* Katedra spoj, F-PEDaS, Žilinská univerzita v Žiline, ININFOS 9. Medzinárodná konferencia združenia EUNIS Sk, Nitra, Dostupné na internete: <http://www.fem.uniag.sk/uninfos2003/zbornik/smehyl_kremenova.pdf>
66. ŠTOFKO, M., OCELÍKOVÁ, E. 2005. *Získavanie znalostí z databáz a ich využívanie vo vizualizácii informačných a riadiacich systémov* (1). *AT&P Journal*, Bratislava: HMH s.r.o. **12**(6), 53-54. ISSN 1335-2237
67. TANIAR, D. 2008. *Data Mining and Knowledge Discovery Technologies.* Hershey, PA: IGI Publishing. ISBN 978-1-59904-960-1
68. TANUŠKA, P. 2013. *Získavanie znalostí z databáz v hierarchickom riadení technologických a výrobných procesov.* Inauguračná prednáška. Trnava: MTF STU.
69. TOMEK, G., VÁVROVÁ, V. 1999. *Řízení výroby*. Praha: Grada Publishing. ISBN 80-7169-578-5
70. TU v Košiciach: Fakulta elektrotechniky a informatiky. *Objavovanie znalostí – Support Verctore Machine. 2012.* [cit. 2012-12-10]. Dostupné na internete:

<https://mining.fei.tuke.sk/oz/index.php/operatory-rapidminer/164-support-vector-machine>

71. TVRDÍKOVÁ, M. 2008. *Aplikace moderních informačních technologií v řízení firmy:* Nástroje ke zvyšování kvality informačních systémů. 1-vyd. Praha: Grada. ISBN 978-80-247-2728-8
72. VAŽAN, P., KEBÍSEK, M., TANUŠKA, P., JUROVATÁ, D. 2011. The data warehouse suggestion for production system. In: *Annals of DAAAM and Proceedings of DAAAM Symposium*. Vienna: DAAAM International, s. 0017-0018. ISBN 978-3-901509-83-4
73. VAŽAN, P., TANUŠKA, P., JUROVATÁ, D., KEBÍSEK, M. 2012. Analysis of Production Process Parameters by Using Data Mining Methods. In: *CECOL 2012*. Trnava : AlumniPress. s. 342-349. ISBN 978-80-8096-179-4
74. VAPNIK, N. Vladimir. 1995. *The Nature of Statistical Learning Theory*. New York: Springer. ISBN:0-387-94559-8
75. VERCELLIS, C., GIOVANNI, F. 2007. *Mathematical Methods for Knowledge Discovery and Data Mining*. United Kingdom: Idea Group Reference. ISBN 1599045281
76. VERCELLIS, C. 2009. *Business Intelligence: Data Mining and Optimization for Decision Making*. United Kingdom: TJ International. ISBN 978-0-470-51138-1
77. VOŘÍŠEK, Jiří. 1997. *Strategické řízení informačního systému a systémová integrace.* Praha: Management Press, 1. vyd. ISBN 80-85943-40-9
78. WACKER G., J. 1987. *The complementary nature of manufacturing goals by their relationship to throughput time: A theory of internal variability of production systems. ELSEVIER*, **7**(1-2) 91-106. ISSN 0272-6963
79. WATSONn, J. Hugh, WISOM, H. Barbara. 2007. *The Current State of Business Intelligence*. In Computer, Volume: 40 (9), pp.96-99. ISSN: 0018-9162.
80. WILLIAMS, G. J., SIMOFF, S. J. 2006. *Data Mining – Theory, Methodology, Techniques, and Applications*. Berlin/Heidelberg: Springer. ISBN 978-3-540-32547-5
81. WILLIAMS, Steve, WILLIAMS, Nancy. 2007.*The profit impact of business intelligence*. San Francisco: Morgan Kaufmann. ISBN-10: 0-12-372499-6.
82. WITTEN, I., EIBE, F. 1999. *Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations.* San Francisco: Morgan Kaufmann. ISBN 1558605525
83. WOJCIK, W., GROMASZEK, K. (2011). *Knowledge-Oriented Applications in Data Mining: Data Mining Industrial Applications*. Shanghai: InTech, (Chapter 26).

84. YLIMÄKI, T. 2004. *Tools and Techniques for Enterprise Architecture.* [cit. 2012-01-10]. Dostupné na internete: <http://haddock.titu.jyu.fi/larkkidata/eatools.html>

OBSAH

Autor:	Ing. Dominika Jurovatá, PhD.
Názov:	VYUŽITIE PROCESU ZÍSKAVANIA ZNALOSTÍ V OBLASTI PLÁNOVANIA A RIADENIA VÝROBNÝCH PROCESOV UTILIZATION OF KNOWLEDGE ACQUISITION PROCESS IN THE FIELD OF PRODUCTION PROCESSES PLANNING AND CONTROL
Miesto vydania:	Trnava
Vydavateľ:	AlumniPress
Rok vydania:	2014
Vydanie:	prvé
Rozsah :	95 strán (27 obrázkov, 18 tabuliek, 5,94 AH)
Edícia:	Vedecké monografie
Edičné číslo:	1/AP-VM/2014

Publikácie vydávané v tejto edícii prešli jazykovou korektúrou.

ISBN 978-80-8096-199-2
EAN 9788080961992

zverejnené na http://www.mtf.stuba.sk