Univerzita Pardubice
Fakulta ekonomicko-správní

Modelování ekonomických ukazatelů na regionální úrovni

Bc. Alena Lukešová

Diplomová práce

2009

Univerzita Pardubice
Fakulta ekonomicko-správní
Ústav systémového inženýrství a informatiky
Akademický rok: **2008/2009**

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Alena LUKEŠOVÁ**
Studijní program:	**N6209 Systémové inženýrství a informatika**
Studijní obor:	**Informatika ve veřejné správě**
Název tématu:	**Modelování vybraných ekonomických ukazatelů na regionální úrovni**

Zásady pro vypracování:

Předpokládá se, že závěrečná práce bude obsahovat:
- popis ekonomických ukazatelů na regionální úrovni se zaměřením na podniky,
- výběr a sběr vhodných dat pro modelování,
- návrh modelu eknomických ukazatalů,
- sestavení a analýza modelu,
- popis prostředí, ve kterém bude model realizován.

Rozsah grafických prací:

Rozsah pracovní zprávy:

Forma zpracování diplomové práce: **tištěná/elektronická**

Seznam odborné literatury:

STUTELY R, DĚDEK O. Průvodce ekonomickými ukazateli: jak porozumět ekonomii. 2002.

LINGRAS P, JENSEN R. Survey of Rough and Fuzzy Hybridization, Fuzzy Systems Conference, 23-26 July, 2007, s. 1-6.

KAŠPAROVÁ M, JIRAVA P, KŘUPKA J. Hybrid Approach For Modelling Of Internal Human Population Migration Classifiers. Proc. of the 12th IASTED International Conference Artificial Intelligence and Soft Computing (ASC 2008), Ed. A.P. Del Pobil, September 1-3, 2008 Palma de Mallorca, Spain, ACTA Press: Anaheim, Calgary, Zurich, 2008, s. 50-54.

BLAŽEK J. Teorie regionálního rozvoje : nástin, kritika, klasifikace. 1. vyd. Praha: Karolinum, 2002. 211 s.

DUŠEK F. MATLAB a Simulink : Úvod do používání. Univerzita Pardubice, 2002.

ČESKÝ STATISTICKÝ ÚŘAD. Veřejná databáze ČSÚ [online]. URL: http://vdb.czso.cz/vdb/.

CZECH TRADE. Hlavní faktory regionálního rozvoje ČR [online]. 1997-2008. URL: http://www.businessinfo.cz/cz/clanek/ rozvoj-regionu/hlavni-faktory-regionalniho-rozvoje-cr/1001179/46055/.

Vedoucí diplomové práce: **doc. Ing. Jiří Křupka, Ph.D.**
Ústav systémového inženýrství a informatiky

Datum zadání diplomové práce: **6. října 2008**

Termín odevzdání diplomové práce: **1. května 2009**

doc. Ing. Renáta Myšková, Ph.D.
děkanka

L.S.

doc. Ing. Jiří Křupka, Ph.D.
vedoucí ústavu

V Pardubicích dne 6. října 2008

Prohlašuji:
Tuto práci jsem vypracovala samostatně. Veškeré literární prameny a informace, které jsem v práci využila, jsou uvedeny v seznamu použité literatury.

Byla jsem seznámena s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Souhlasím s prezenčním zpřístupněním své práce v Univerzitní knihovně.

V Pardubicích dne 27. 04. 2009

Bc. Alena Lukešová

SOUHRN

Diplomová práce se zabývá modelováním vybraných ekonomických ukazatelů na regionální úrovni. Je zaměřena na nezaměstnanost v okresech Pardubického kraje a další ekonomické ukazatele ji přímo či nepřímo ovlivňující. Modelování je provedeno v programu Clementine za pomoci rozhodovacích stromů a neuronové sítě.

KLÍČOVÁ SLOVA

ekonomické ukazatele, modelování, neuronové sítě, nezaměstnanost, regionální politika, rozhodovací stromy

TITLE

Modelling of selected economic indicators at the regional level

ABSTRACT

Diploma Thesis is focusing on modelling of selected economic indicators at the regional level. It is focused on unemployment in the districts of Pardubice Region, and other economic indicators to directly or indirectly affecting them. Modeling is done in the Clementine with the help of decision trees and neural networks.

KEYWORDS

economic indicators, modeling, neural network, unemployment, regional policy, decision trees

OBSAH:

SEZNAM OBRÁZKŮ

SEZNAM TABULEK:

SEZNAM ROVNIC:

SEZNAM PŘÍLOH:

ÚVOD

V práci se budu zabývat ekonomickými ukazateli pro jednotlivé okresy Pardubického kraje. Území Pardubického kraje je vymezeno územími okresů Pardubice, Chrudim, Ústí nad Orlicí a Svitavy. V kraji je 451 obcí, z toho 15 obcí s rozšířenou působností a 26 obcí s pověřeným obecním úřadem. Z celkového počtu obcí je 34 měst a 6 městysů. Sídelním městem kraje je statutární město Pardubice. Pardubický kraj má v současné době rozlohu 4.519 km^2, žije v něm 515 185[1] obyvatel a průměrná hustota je 114 obyvatel na 1km^2.

Z velkého výběru ekonomických ukazatelů jsem si vybrala nezaměstnanost a ty ukazatele, které s ní přímo či nepřímo souvisí a mají na ni určitý vliv. Toto odvětví jsem si vybrala, protože téma nezaměstnanost bylo, je a určitě vždy bude jedním z nejdiskutovanějších makroekonomických ukazatelů. Bez práce jsou lidé závislí na státu a to se odráží v jeho výdajích a celkovém rozpočtu. Téma nezaměstnanosti je diskutovanější o to více, že v současné době probíhá celosvětová hospodářská krize, která se podepsala na nejednom odvětví národního hospodářství.

V současné době Česká Republika zažívá nejvyšší růst nezaměstnanosti v celé její historii. Meziměsíční nárůst nezaměstnaných byl zaznamenán od začátku roku u všech 77 Úřadů práce. Vyšší míru nezaměstnanosti než je republikový průměr vykázalo 45 okresů. V Pardubickém kraji není situace, díky velkému počtu firem a investorů, vážná. V březnu 2009 se nezaměstnanost zvedla jen o 0,3%. Významným stimulem byla výroba dopravních prostředků, strojů a zařízení na elektrotechnický a elektronický průmysl.

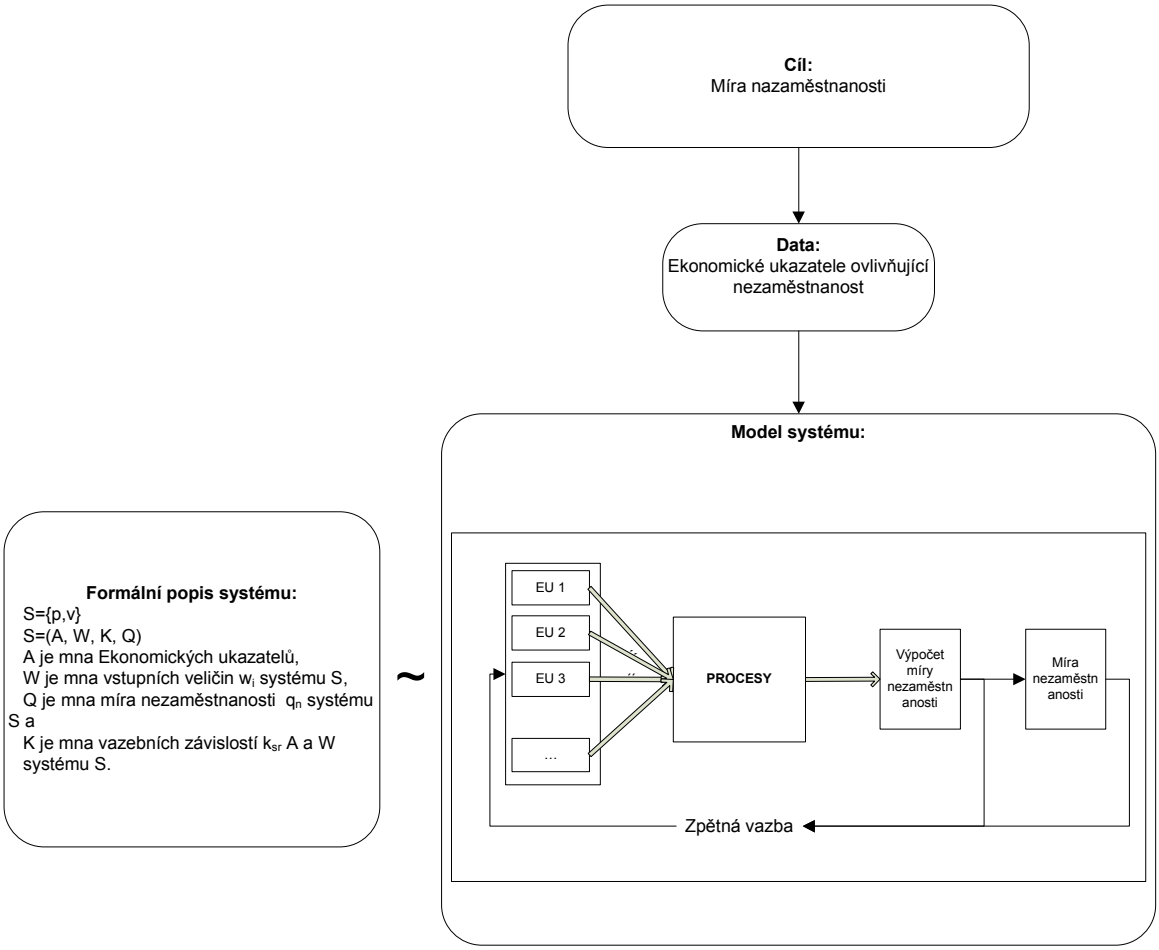
Nejsilnějším hospodářským odvětvím v tomto kraji je všeobecné strojírenství, dále pak průmysl textilní, oděvní, kožedělný, nejvyšší podíl na celostátní produkci má průmysl chemický. Významný je ale i zemědělský sektor. Vždyť z celkové rozlohy kraje zaujímá zemědělská půda 60,75 %, lesy 29 % a vodní plochy 1,35 %. Mezi nejsilnější průmyslové podniky v kraji patři např: Iveco Czech Republic Vysoké Mýto, Korado Česká Třebová, ETA Hlinsko, Paramo Pardubice (Česká rafinérská), Synthesia Semtín, Botas Skuteč, Orlické papírny Lanškroun,ZEZ SILKO Žamberk.

[1] Stav k 31. 12. 2008

Současná krize sice kartami poněkud zamíchala a Pardubický kraj jí utrpěl, potenciál ekonomického tygra republiky mu však podle Jičínského denníku zůstane.

Cílem práce je modelování nezaměstnanosti respektive míry nezaměstnanosti na základě ekonomických ukazatelů, přímo či nepřímo nezaměstnanost ovlivňujících, v okresech Pardubického kraje. Nezaměstnanost a přímo či nepřímo ji ovlivňující ukazatele jsem vybrala, protože nezaměstnanost je velmi důležitým ukazatelem ekonomického a sociálního rozvoje okresu, regionu i republiky.

Model systému pro odhad míry nezaměstnanosti jsem si před samotným začátkem práce navrhla takto:



Obrázek 1: Model systému [Zdroj: vlastní]

1 ZÁKLADNÍ POJMY

V této kapitole jsou definovány základní pojmy k tématu diplomové práce. Jsou zde popsány základní využité ekonomické ukazatele, dále pak pojmy z oblasti modelování.

1.1 Regionální politika

V teorii regionální politiky dosud nedošlo k jednotnému vyjádření jejího obsahového vymezení, které by se dalo universálně použít. Regionální politika představuje všechny veřejné intervence, vedoucí ke zlepšení geografického rozdělení ekonomických činností, respektive se pokouší napravit určité prostorové důsledky volné tržní ekonomiky ve smyslu dosažení dvou vzájemně závislých cílů: ekonomického růstu a zlepšení sociálního rozdělení ekonomických efektů. Jejím významným cílem je konvergence regionů v rámci určitého územního celku. Klíčovým znakem je její selektivnost, to znamená diferenciace zaměření intervencí a podporu vybraných problémových regionů, které výrazně zaostávají ve svém rozvoji za průměrem v míře, která jej společensky uznána za nežádoucí. [1], [2]

V souvislosti s koncepcí regionální politiky je nutno zdůraznit, že existují nejen významné rozdíly v citlivosti společnosti disparity v různých sférách, ale že existují i rozdíly v možnostech společnosti disparity v různých sférách ovlivnit. Regionální politika pak představuje konkrétní projev úsilí společnosti o snížení (změnu) velikosti regionálních rozdílů. [3]

Regionální politiku je nutno spíše chápat jako součást souboru ekonomických a sociálních politik, pomocí níže se státy snaží dosáhnout národních cílů, jako je ekonomický růst, sociální a politická stabilita, rovnost šancí obyvatel i rozdělován příjmů způsobem, který většina obyvatel považuje za spravedlivý a který je současně ekonomicky stimulující. [3]

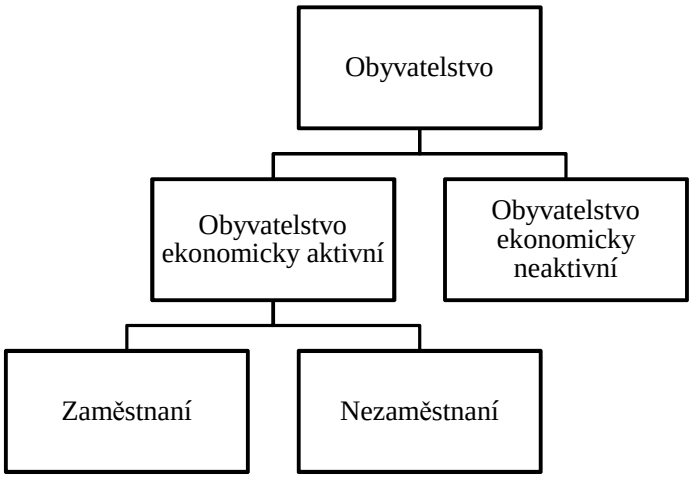
1.2 Ekonomické ukazatele

Ekonomické ukazatele [4] lze obecně rozdělit do tří skupin: podle načasování bodů zvratu v jejich časový řadách, ve vztahu k referenční řadě, která charakterizuje skutečný výskyt bodu zvratu. Nejsledovanější jsou předstihové ukazatele (mj. index vedoucích ekonomických indikátorů), které umožňují předvídání a prognózování bodů zvratu.

Dále jsou definovány některé základní ekonomické ukazatele použité pro modelování v této diplomové práci.

Definice nezaměstnanosti:

Populaci země lze rozčlenit na ekonomicky aktivní obyvatelstvo (pracovní sílu), tj. zaměstnané a nezaměstnané a obyvatelstvo ekonomicky neaktivní. [5]



Obrázek 2: Rozdělení obyvatelstva [5]

Za **nezaměstnané** jsou v České Republice (ČR), podle metodiky ILO, považovány všechny osoby, které dosáhly 15 let a více v daném období souběžně splňovaly tři podmínky:

- nebyly zaměstnané,
- aktivně hledaly práci,
- byly připraveny k nástupu do práce (nejpozději do 14 dnů).

Pokud osoba nesplňuje alespoň jednu ze tří výše uvedených podmínek, pak je buď zaměstnána, nebo ekonomicky neaktivní. [5]

Zaměstnané jsou všechny osoby 15leté a starší, které patří mezi placené zaměstnané nebo zaměstnané ve vlastním podniku. [5]

Celkový počet zaměstnaných a nezaměstnaných pracovníků představuje ekonomicky aktivní obyvatelstvo neboli **pracovní sílu**. [5]

Ekonomicky neaktivní obyvatelstvo (neboli osoby mimo pracovní sílu) jsou osoby, které nejsou zaměstnány a nejsou ekonomicky aktivní, např. děti předškolního věku, osoby navštěvující různé vzdělávací instituce, starobní důchodci, osoby dlouhodobě nemocné nebo invalidní apod. [5]

Výši nezaměstnanosti lze určit absolutně jako počet osob, nebo relativně jako míru nezaměstnanosti.

Míra nezaměstnanosti je procentuální podíl počtu nezaměstnaných na celkové pracovní síle neboli ekonomicky aktivním obyvatelstvu, tj. [5]

$$U = \frac{počet\ nezaměstnaných}{pracovní\ síla} \times 100$$

Rovnice 1: Výpočet míry nezaměstnanosti [5]

Kromě obecné míry nezaměstnanosti se zjišťují také specifické míry nezaměstnanosti, popisující nezaměstnanost podle věkové nebo jiné struktury obyvatelstva. [5]

Ministerstvo práce a sociálních věcí (MPSV) poprvé zveřejnilo míru nezaměstnanosti v ČR podle nové metodiky. Výsledky za měsíc červenec 2004 je tak možné lépe srovnávat s výsledky členských zemí Evropské unie (EU). [6]

MPSV dosud vykazovalo tzv. míru registrované nezaměstnanosti. Metodika, která byla zavedena 1. 1. 1997, vycházela z přesného počtu uchazečů o zaměstnání – občanů ČR, kteří jsou v evidencích úřadů práce v okrese jejich bydliště a z počtu zaměstnaných v národním hospodářství s jediným nebo hlavním pracovním poměrem. [6]

Druhým ukazatelem míry nezaměstnanosti, podle nové metodiky, v ČR je **obecná míra nezaměstnanosti,** která vychází z metodiky ILO (EUROSTATu). Ta je založena na bázi výběrových šetření pracovních sil. [6]

Stejnou metodiku pro výpočet míry nezaměstnanosti (kromě národní legislativy jednotlivých zemí) používají i členské státy EU. Pro výpočet registrované míry nezaměstnanosti bere MPSV v potaz, podle nově zaváděné metodiky, **tzv. dosažitelné uchazeče o zaměstnání**. Jedná se o lidi, kteří mohou bezprostředně nastoupit do zaměstnání. Tedy evidované uchazeče, kterým nebrání žádná objektivní překážka v nástupu do nového zaměstnání. Za dosažitelné uchazeče se považují lidé, kteří nejsou ve vazbě, ve výkonu trestu, nepobírají peněžitou pomoc v mateřství, hmotné zabezpečení po dobu mateřské dovolené, nejsou v pracovní neschopnosti, atd. Takto vymezená část uchazečů o zaměstnání lépe odpovídá definici nezaměstnaných používaných pro výpočet obecné míry nezaměstnanosti podle metodiky ILO. [6]

Důvodem ke změně metodiky je skutečnost, že je ČR od 1. května 2004 členem EU. Mezi uchazeče o zaměstnání se budou nově zahrnovat i případní uchazeči o zaměstnání ze zemí Evropského hospodářského prostoru – EHP a Švýcarska. [6]

Hrubý domácí produkt (HDP)

HDP je celková hodnota všech finálních výrobků a služeb, vyrobených v daném období na území daného státu. [5]

HDP se používá pro stanovení výkonnosti ekonomiky. Může být definován, resp. spočten třemi způsoby: [7]

- **produkční metodou**
- **výdajovou metodou**
- **důchodovou metodou**

Produkční metodou se HDP počítá jako součet hrubé přidané hodnoty jednotlivých institucionálních sektorů nebo odvětví a čistých daní na produkty (které nejsou rozvrženy do sektorů a odvětví). Je to také vyrovnávací položka účtu výroby za národní hospodářství celkem, kde se straně zdrojů zachycuje produkce a na straně užití mezispotřeba. Hrubá přidaná hodnota je rozdílem mezi produkcí a mezispotřebou. Vzhledem k tomu, že produkce se oceňuje v základních cenách a užití v kupních cenách, je strana zdrojů za národní hospodářství celkem doplněna o daně snížené o dotace na výrobky. [7]

Vztah při použití produkční metody lze zapsat takto:

$$HDP = P - M + d - D$$

Rovnice 2: Výpočet HDP: Produkční metoda [7]

- **P** je produkce výrobků a služeb, představující hodnotu zboží a služeb, které jsou výsledkem produkčních činnosti výrobních jednotek v daném časovém období na určitém území
- **M** je mezispotřeba, představující hodnotu zboží a služeb spotřebovaných v příslušném období místními producenty v procesu výroby jiného zboží a služeb,
- **d** jsou daně z produktu a
- **D** jsou dotace na produkty.

Výdajovou metodou se HDP počítá jako součet konečného užití výrobků a služeb rezidentskými jednotkami (skutečná konečná spotřeba a tvorba hrubého kapitálu) a salda vývozu a dovozu výrobků a služeb. Skutečná konečná spotřeba je odvozena prostřednictvím naturálních sociálních transferů od výdajů na konečnou spotřebu domácností, vlády a neziskových institucí sloužících domácnostem. Tvorba hrubého kapitálu se člení na tvorbu hrubého fixního kapitálu, změnu zásob a na čisté pořízení cenností. [7]

Při použiti výdajové metody, lze vztah pro výpočet HDP zapsat takto:

$$HDP = C + I + G + Nx$$

Rovnice 3: Výpočet HDP: Výdajová metoda [7]

- **C** jsou výdaje na konečnou spotřebu, představují výdaje domácností, vlády a neziskových institucí na konečnou spotřebu,
- **I** jsou soukromé hrubé domácí investice,
- **G** jsou výdaje státu na nákup výrobků a služeb, nezapočítávají se transferové platby a
- **Nx** je čistý vývoz, tedy rozdíl mezi exportem a importem.

Důchodovou metodou se HDP počítá jako součet prvotních důchodů za národní hospodářství celkem: náhrad zaměstnancům, daní z výroby a z dovozu snížených o dotace a hrubého provozního přebytku a smíšeného důchodu (resp. čistého provozního přebytku a smíšeného důchodu a spotřeby fixního kapitálu). [7]

Vztah pro měření HDP pomocí důchodové metody lze zapsat tímto způsobem:

$$HDP = w + p + r + i + a + T$$

Rovnice 4: Výpočet HDP: Důchodová metoda [7]

- **w** jsou náhrady zaměstnancům zahrnující mzdy, platy a sociální příspěvky zaměstnavatelů před zdaněním,
- **p** jsou důchody samozaměstnavatelů, kde jsou zahrnuty všechny formy plateb za výrobní faktory, jež používají osoby samostatně výdělečně činné, zisky firem v individuálním vlastnictví a zisky právnických osob,
- **r** jsou renty majitelů půdy,
- **i** jsou čisté úroky z kapitálu,

- **a** jsou odpisy a amortizace a
- **T** jsou nepřímé daně.

Minimální mzda

Minimální mzda je nejnižší přípustná výše odměny za práci v pracovněprávním vztahu. Její základní právní úprava je stanovena zákoníkem práce (zákon č. 262/2006 Sb., ve znění pozdějších předpisů). [8]

Minimální mzda se vztahuje na všechny zaměstnance v pracovním poměru nebo právním vztahu založeném dohodami o pracích konaných mimo pracovní poměr (dohoda o provedení práce a dohoda o pracovní činnosti). Nárok na minimální mzdu vzniká v každém pracovním poměru nebo právním vztahu založeném dohodami o pracích konaných mimo pracovní poměr samostatně. Minimální mzda platí jako jediná mzdová veličina pro zaměstnance v organizacích podnikatelské sféry, v nichž se uplatňuje kolektivní vyjednávání o mzdách. [8]

1.3 Modelování

Modelování představuje objektivní a progresivní přístup k rozhodování. Pomocí matematického modelu, který je zjednodušeným obrazem reality, se provádí analýza a hledají se nejlepší varianty řešení. V této DP jde o modelování ekonomických ukazatelů.

Modelování vždy sleduje snahu usnadnit dosažení cíle, kterýs i člověk vytyčil, a to nepřímo, oklikou. Má tedy charakter cílevědomých činností, sledujících předem určený cíl, ale také sledujících hledání nejlepší cesty zmožných, které tomu cíly vedou. [9]

Jedním z možných nástrojů pro modelování ekonomických ukazatelů je Data Mining.

Data Mining je pojem zastřešující širokou škálu technik používaných v řadě odvětví. DM umožňuje pomocí speciálnímu algoritmu automaticky objevovat v datech strategické informace. Je to analytická technika pevně spjatá s datovými sklady, jako velmi kvalitním zdrojem pro tyto speciální analýzy. [10]

Definice: [11]

- DM je proces hledání jistých závislostí, vzoru a trendu na základě vlastností dat uložených v databázích prostřednictvím technik pro rozpoznávání vzoru nebo jiných matematických a statistických technik.
- DM je netriviální dobývání skrytých předem neznámých a potenciálně užitečných informací z dat. Při jejich objevování se využívají expertní systémy a grafické a statistické techniky a prezentují se způsobem srozumitelným lidem
- DM je netriviální **proces** zjišťování **platných**, **neznámých**, potenciálně **užitečných** a snadno **pochopitelných** závislostí v datech.

V DP je modelování prováděno pomocí rozhodovacích stromů a neuronové sítě, obě tyto metody jsou podrobně popsány v následující kapitole.

2 VYBRANÉ METODY MODELOVÁNÍ

Tato kapitola popisuje použité metody modelování, rozhodovací stromy a neuronové sítě, a také program Clementine, ve kterém bylo modelování provedeno.

2.1 Rozhodovací stromy (RS)

Základní princip, z kterého vychází RS, je, že výchozí uzel se dále dělí na další uzly, až je dosaženo zadaného cíle. RS jsou velmi častou technikou zejména díky své snadné interpretaci ve formě rozhodovacích pravidel. [10]

RS lze definovat jako strom (stromový graf), kde každý nelistový uzel stromu představuje test hodnoty atributu a větve vedoucí z tohoto uzlu možné výsledky testu. V každém kroku je záznam otestován podle testu v aktuálním uzlu rozhodovacího stromu a dále pokračuje po větvi shodné s konkrétním výsledkem testu. Pokud takto záznam dojde až do listového uzlu, je oklasifikován třídou identifikovanou hodnotou příslušného listu RS. [14]

Při tvorbě RS se postupuje metodou „rozděl a panuj“. Trénovací data se postupně rozdělují na menší a menší podmnožiny tak, aby v těch podmnožínách převládaly příklady jedné třídy. Atribut vhodný pro větvení stromu vybíráme na základě jeho charakteristik převzatých z teorie informace a pravděpodobnosti: entropie, informačního zisku, poměrného informačního zisku, χ^2 – testu, Giniho indexu a dalších. [14]

Stromové grafy dovolují vizuálně prozkoumat výsledky a posoudit vhodnost modelu. RS lze poměrně snadno převést na rozhodovací pravidla. Každé cestě stromem od kořene k listu odpovídá jedno pravidlo. [14]

RS mohou být založeny na množství algoritmů pro přímou tvorbu pravidel. Příkladem některých RS založených na samostatných algoritmech jsou C5.0 – vytváří RS na základě tzv. metody TDIDT (top down induction of decision trees), data se trénovací data se postupně rozděluji na menší podmnožiny (uzly stromu), tak aby v těchto podmnožinoách převládaly příklady jedné třídy (tzv. metoda rozděl a panuj), C&RT (Classification and Regression Trees), který vytváří binární stromy tzn., že skupina je štěpena vždy na dvě části, CHAID (Chisquared Automatic Interaction Detector), což je statistický altoritmus založený na optimální hodnotě χ^2 – testu závislosti nebo F-testu, který štěpí skupiny vždy na vhodný počet statisticky homogenních podskupin a QUEST (The Quick, Unbiased, Efficient Statistical Tree), který je statistic-

ký algoritmus vybírající proměnné nevychýleně a rychle a efektivně vytváří přesné binární stromy. [15], [16]

RS jsou vhodné pro úlohy, ve kterých má být provedena klasifikace nebo předpověď. Užitečné jsou v oblastech, ve kterých lze hodnoty proměnných rozdělit do relativně malého poctu skupin. [17]

Silné stránky rozhodovacích stromu:

- jsou schopny vytvářet srozumitelná pravidla;
- provádějí klasifikaci bez velkých početních požadavku;
- jsou vhodné k zpracování jak spojitých tak i kategorických veličin, výsledkem stromu je však veličina kategorická.

Nevýhody rozhodovacích stromu:

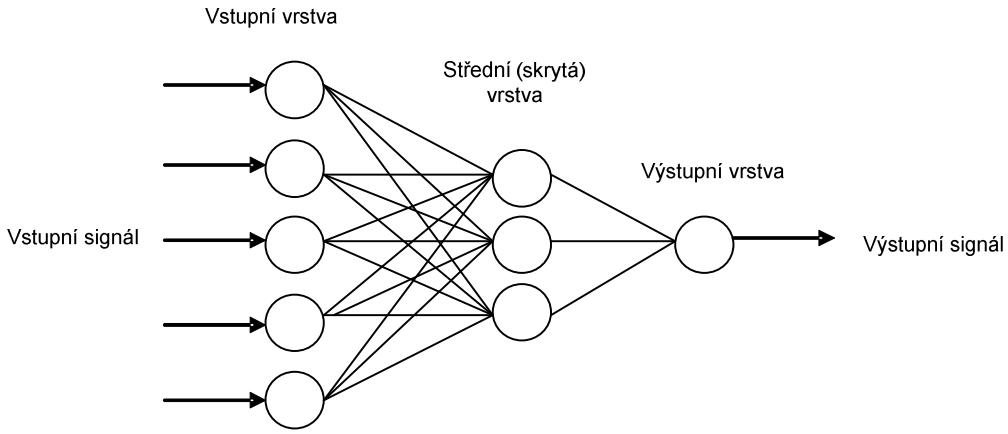
- jsou méně vhodné pro odhadnutí úloh, kde je cílem predikce hodnoty nastávající proměnné,
- jsou také problematické pro časové řady.

2.2 Neuronové sítě (NS)

Neuronové sítě lze nazvat jako masivní paralelní systém (procesor) vytvořený z jednoduše zpracovatelných neuronů, který má přirozený sklon k ukládání či uchovávání empirických znalostí a jeho dalšího využívání. NS se podobá mozku ve dvou aspektech: [18]

- poznatky získává síť ze svého okolí při učení,
- mezineuronová spojení nazývaná též synaptické váhy jsou použita pro uložení získaných znalostí.

NS je orientovaný ohodnocený graf složený z velkého počtu neuronů, kde každý uzel je označen číslem vrstvy a pořadím uzlu ve vrstvě. Neurony jsou mezi sebou vzájemně propojeny synapsemi s ohodnocenými váhami, tyto synapse se nazývají hrany. Hrany spojují jen uzly sousedních vrstev. Vstup dat do NS probíhá přes vstupní vrstvu, kde je vstupem každého uzlu hodnota části analyzovaného záznamu. Na Obrázku 3 je znázorněna struktura neuronové sítě.



Obrázek 3: Struktura NS [Zdroj: vlastní]

Použití neuronové sítě

- NS jsou dobrou volbou pro většinu klasifikačních a predikčních úloh, když výsledek modelu je důležitější než porozumění tomu jak model pracuje.

- Ve skutečnosti reprezentují komplex matematických rovnic s počítáním sumačních, exponenciálních funkcí. Tyto rovnice jsou pravidla sítí.

- NS nepracují dobře, pokud je mnoho vstupních hodnot. NS dělá problémy najít vzor velkému počtu hodnot. To může mít za následek, že doba trénování bude dlouhá a nikdy nenajde dobré řešení. V takových případech mohou NS dobře spolupracovat s RS, které jsou dobré pro hledání nejdůležitějších proměnných a ty pak mohou být NS využity jako trénovací data. [19]

Vstupy NS jsou v podstatě číselné, i když dovedou rozeznávat i složité strukturální (nečíselné) obrazce. Typy výstupu NS jsou stejné jako u vstupu. Zpravidla se používají stejné typy vstupu a výstupu v celé síti a modifikující váhy se připisují na vstupy neuronu, na které je výstup připojen. [20]

2.3 Systém Clementine

Systém Clementine od britské firmy SPSS patří mezi přední komerční systémy pro dolování dat a znalostí. Systém nabízí řadu metod pro klasifikační, predikční i deskriptivní úlohy. [12]

Firma SPSS, která tento program, vyrábí je celosvětově působící společnost vyvíjející software a řešení pro prediktivní analýzy. Nabízí široké portfolio programů zaměřených na statistické analýzy, business inteligence, Data Mining, na analýzy v marketingu, sběr dat a prezentaci výsledků. Společnost byla založena roku 1968 a dnes působí ve více než 60 zemích světa. Společnost SPSS CR spol. s r. o. je v České a Slovenské republice výhradním dodavatelem analytických a statistických služeb. [13]

Mezi specialisty je Clementine celosvětově oblíben, protože umožňuje: [12]

- snadno se připojit, upravit a spojovat různé zdroje strukturovaných dat i ostatních typů dat, jako jsou volné texty, webové logy nebo data z výzkumů
- rychle vytvářet a validizovat modely s využitím sofistikovaných statistických postupů a metod strojového učení
- snadno sdílet závěry s mnoha uživateli a účinně do praxe nasadit predikční modely, které poskytují výsledky dávkově nebo v reálném čase, aby tak byly k dispozici všem, kdo je potřebují pro řízení a rozhodování ve vhodný čas a na vhodném místě.

3 VÝBĚR DAT PRO MODELOVÁNÍ

V této kapitole jsou popsána vstupní data a jejich výběr, důvod proč byla zvolena zrovna tato data a zdroje z kterých byla data získána. Dále je zde uvedena analýza dat a úvodní příprava dat.

3.1 Vstupní data

Jako vstupní data byla použita data zveřejněná na stránkách MPSV ČR, ČSÚ Pardubického kraje a z veřejné databáze ČSÚ ČR. Zdrojové stránky jsou uvedeny v použité literatuře. [21], [22], [23] Všechny použité ukazatele přímo či nepřímo určitým způsobem ovlivňují míru nezaměstnanosti kraje respektive okresu. Vybraná vstupní data mají 21 atributů. Atributy představují jednotlivé ekonomické ukazatele.

Ekonomické ukazatele přímo ovlivňující nezaměstnanost:

- Pracovní síla,
- Nově hlášení uchazeči o zaměstnání,
- Vyřazení uchazeči ve sledovaném měsíci (celkem),
- Vyřazení uchazeči ve sledovaném měsíci a umístěni,
- Počet uchazečů o zaměstnání,
- Počet uchazečů na rekvalifikaci,
- Volná pracovní místa,
- Přírůstek/úbytek uchazečů o zaměstnání,
- Veřejně prospěšné práce (vytvořená místa),
- Veřejně prospěšné práce (umístění uchazeči),
- Absolventské praxe (vytvořená místa),
- Absolventské praxe (umístění uchazeči).

Ekonomické ukazatele nepřímo ovlivňující míru nezaměstnanosti:

- Přírůstek/úbytek subjektů,
- Počet subjektů,
- Počet obyvatel v jednotlivých okresech,
- Míra nezaměstnanosti,
- Minimální mzda (měsíční),
- Minimální mzda (hodinová) a
- HDP (důchodová metoda).

Tyto ukazatele byly vybrány po odborné konzultaci s odborníkem z Ústavu ekonomie UPa. Většina použitých ukazatelů je sledována měsíčně, v DP jsou využita i data, která jsou sledována za delší časová období (čtvrtletně, ročně), tyto ukazatele byly vhodně upraveny.

Ekonomické ukazatele „Počet obyvatel“ je sledován čtvrtletně, ve všech třech měsících jednotlivých čtvrtletí jsou udány stejné hodnoty, které byly určeny na konci čtvrtletí. „Minimální mzda (měsíční)“ a Minimální mzda (hodinová)“ jsou dány zákonem většinou jednou ročně, je tedy v každém roce 12 stejných hodnot. Jen v roce 2006 byla zákonem stanovená, na začátku roku, měsíční minimální mzda 7.570 Kč změněna od července na 7.955 Kč. Ekonomické ukazatele „Počet subjektů“, „Vzniklé/zaniklé subjekty“ a „HDP (důchodová metoda)“, které jsou také sledovány čtvrtletně, byly určeny pro jednotlivé měsíce jednoduchým aritmetickým průměrem v daném čtvrtletí.

Analýza vstupních dat

Všechna sebraná data byla vložena do jediné tabulky v souboru typu xls. Ukázka z této tabulky je v Příloze č. 1. Takto uložená data, v programu MS Excel, byla rozdělena na data kategorizovaná a spojitá a byl vytvořen datový slovník. Pro zpracování v programu Clementine byl soubor převeden do formátu csv.

Analýza vstupních dat je v programu Clementine provedena pomocí uzlu STATISTICS, který zjistí u každé proměnné počet hodnot (Count), průměr (Mean), minimální hodnotu (Min), maximální hodnotu (Max), medián (Median) a modus (Mode). Tento uzel je schopný vygenerovat i další statistické hodnoty, pro základní analýzu ale stačí těchto 6.

Ne pro všechny atributy je možné zjistit medián, je tedy vypisována nejmenší hodnota z těch, které se nejčastěji opakují. Pokud je v atributu každá hodnota jen jednou, vypisuje program nejmenší hodnotu daného atributu. Na Obrázku 4 je ukázka analýzy některých atributů. Celý náhled je přiložen v Příloze č. 2.

Pracovní síla

Statistics

Count	336
Mean	65060.408
Sum	21860297.000
Min	49449
Max	96786
Range	47337
Median	61949
Mode	49449*

*Multiple modes exist. The smallest value is shown.

Nově hlášení uchazeči o zaměstnání

Statistics

Count	336
Mean	665.869
Sum	223732.000
Min	350
Max	1219
Range	869
Median	624.500
Mode	450*

*Multiple modes exist. The smallest value is shown.

Obrázek 4: Statistika vstupních dat [Zdroj: vlastní]

3.2 Datový slovník:

Datový slovník byl vytvořen v programu MS Excel, podle statistiky vytvořené programem Clementine. V datovém slovníku je název atributu, typ proměnné daného atributu, typ proměnné v Clementine, Rozsah daného atributu a stručná definice atributu.

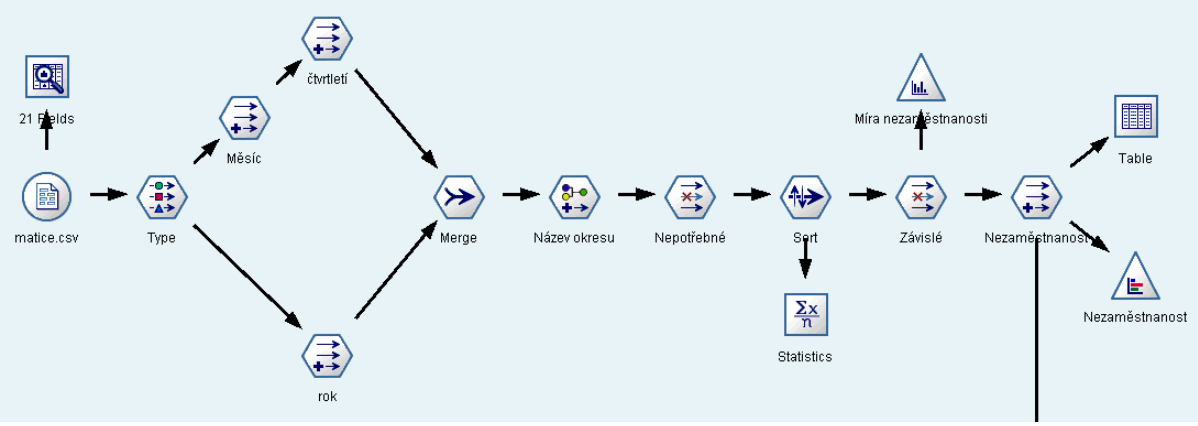
Sloupec	Atributy	Typ proměnné	Typ v Clementinu	Rozsah	Definice
1	Označení	kategorizovaná	Set	1/2002, 2/2002, …	označení období (měsíc/rok)
2	Okres	kategorizovaná	Set	1, 2, 3, 4	rozdělení okresů Pardubického kraje
3	Pracovní síla	spojitá	Range	<49449;96786>	počet ekonomicky aktivních lidí
4	Nově hlášení uchazeči o zaměstnání	spojitá	Range	<350;1219>	nový uchazeči o zaměstnání nahlášení na ÚP
5	Vyřazení ve sledovaném měsíci	spojitá	Range	<257;1164>	uchazeči vyřazeni v daném měsíci z ÚP
6	Vyřezení ve sledovaném měsíci a umístěni	spojitá	Range	<141;959>	uchazeči vyřazeni v daném měsíci z ÚP, kteří byli umístěni
7	Počet uchazečů o zaměstnání	spojitá	Range	<2529;7258>	celkový počet uchazečů hlášených na ÚP
8	Rekvalifikace	spojitá	Range	<0;252>	uchazeči přihlášeni na ÚP v rekvalifikaci
9	Volná pracovní místa	spojitá	Range	<275;6588>	hlášená volná místa zaměstnavateli
10	Přírustek uchazečů o zaměstnání	spojitá	Range	<-1052;1612>	rozdíl počtu uchazečů mezi minulým a současným měsícem
11	Veřejně prospěšné práce (vytvořená místa)	spojitá	Range	<0;491>	vytvořená místa jako veřejně prospěšné práce
12	Veřejně prospěšné práce (umístění uchazeči)	spojitá	Range	<0;440>	počet umístěných uchazečů na veřejně prospěšné práce
13	Absolvenstské praxe (vytvořená místa)	spojitá	Range	<0;268>	vytvořená místa jako absolventské praxe
14	Absolventské praxe (umístění uchazeči)	spojitá	Range	<0;268>	počet umístěných uchazečů na absolventské praxe
15	Vzniklé/zaniklé subjekty	spojitá	Range	<-46;333>	rozdíl počtu subjektů mezi minulým a současným měsícem
16	Počet subjektů	spojitá	Range	<16706;39209>	počet subjektů, zaměstnávajících více než 20 zaměstnanců
17	Počet obyvatel	spojitá	Range	<101728;166039>	počet obyvatel daného okresu
18	Míra nezaměstnanosti	spojitá	Range	<3.12;14.40>	vypočítaná míra nezaměstnanosti v daném měsíci
19	Minimální mzda (měsíční)	spojítá	Range	<5700;8000>	zákonem stanovená minimální měsíční mzda
20	Minimální mzda (hodinová)	spojitá	Range	<33.90;48.10>	zákonem stanovená minimální hodinová mzda
21	HDP (důchodová metoda)	spojitá	Range	<568896938203>	HDP celé ČR vypočítaný důchodovou metodou

Tabulka 1: Datový slovník [Zdroj: vlastní]

4 PŘÍPRAVA DAT PRO MODELOVÁNÍ

Tato kapitola popisuje přípravu dat, jejich úpravu, korelaci, vytvoření a odstranění atributů. Tyto úpravy vstupní matice jsou nutné, aby data byla použitelná k další práci a k modelování. Příprava dat stejně jako návrh modelu byla provedena v programu Clementine vytvořený firmou SPSS.

Na Obrázku 5 je průběh (Stream) celé přípravy dat v programu Clementine, která je dále v této kapitole podrobně popsána.



Obrázek 5: Příprava dat v Clementine [Zdroj: vlastní]

Prvním krokem je načtení vybraných dat, která byla předpřipravena v programu MS Excel. Do programu jsou exportována data ze souboru typu csv pomocí uzlu VAR.FILE.

4.1 Posouzení kvality dat

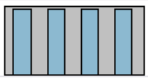
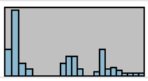
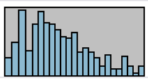
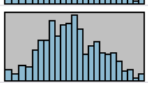
Obrázek 6 ukazuje datové typy atributů vstupního souboru. Podle zaškrtnutého pole „Override“ je poznat, že datový typ atributu Okres byl změněn (byl přetypován). Protože Okresy jsou ve vstupní matici označeny čísly, zařadil Clementine tento atribut mezi Integer, v tomto případě se s nimi nepracuje jako s čísly ale jako s označením, jsou to tedy kategorizovaná data, proto jsou přetypována na String. Ostatní atributy zůstali tak jak je Clementine nadefinoval.

Field	Override	Storage	Input Format
Označení	☐	String	
Okres	☑	String	
Pracovní síla	☐	Integer	
Nově hlášení uchazeči o zaměstn...	☐	Integer	
Vyřazení ve sledovaném měsíci	☐	Integer	
Vyřezení ve sledovaném měsíci a ...	☐	Integer	
Počet uchazečů o zaměstnání	☐	Integer	
Rekvalifikace	☐	Integer	
Volná pracovní místa	☐	Integer	
Přírustek uchazečů o zaměstnání	☐	Integer	
Veřejně prospěšné práce (vytvoře...	☐	Integer	
Veřejně prospěšné práce (umístě...	☐	Integer	
Absolvenstské praxe (vytvořená m...	☐	Integer	
Absolventské praxe (umístění uch...	☐	Integer	
Vzniklé/zaniklé subjekty	☐	Integer	
Počet subjektů	☐	Integer	
Počet obyvatel	☐	Integer	
Míra nezaměstnanosti	☐	Real	
Minimální mzda (měsíční)	☐	Integer	
Minimální mzda (hodinová)	☐	Real	
HDP (důchodová metoda)	☐	Integer	

Obrázek 6: Datové typy atributů vstupního souboru [Zdroj: vlastní]

Pro vstupní analýzu dat je použit uzel DATA AUDIT, který umožňuje analyzovat vstupní data. Tento uzel zobrazuje některé charakteristiky dat: název, grafické zobrazení dat ve formě grafu odpovídajícího datovému typu, datový typ, a nakonec počet platných hodnot. Uzel může zobrazovat dále i některé statistické charakteristiky jako je minimu maximu, rozpět atd., tyto údaje již byly využity při základní analýze dat, není proto nutné je zde znovu zobrazovat. Základní statistické charakteristiky jsou vypsány v Příloze č. 2.

Na Obrázku 7 je jen výběr atributů, celý audit je přiložen k diplomové práci v Příloze č. 3. V posledním sloupci „Valid“ je vidět, že v matici nechybí v žádném atributu ani jedna hodnota. V každém sloupci je platných 336 hodnot. To znamená, že vstupní matice je „validní“ tedy v pořádku a lze ji použít pro další práci.

Field	Graph	Type	Valid
Označení		Set	336
Okres		Set	336
Pracovní síla		Range	336
Nově hlášení uchazeči o zaměstnání		Range	336
Vyřazení ve sledovaném měsíci		Range	336

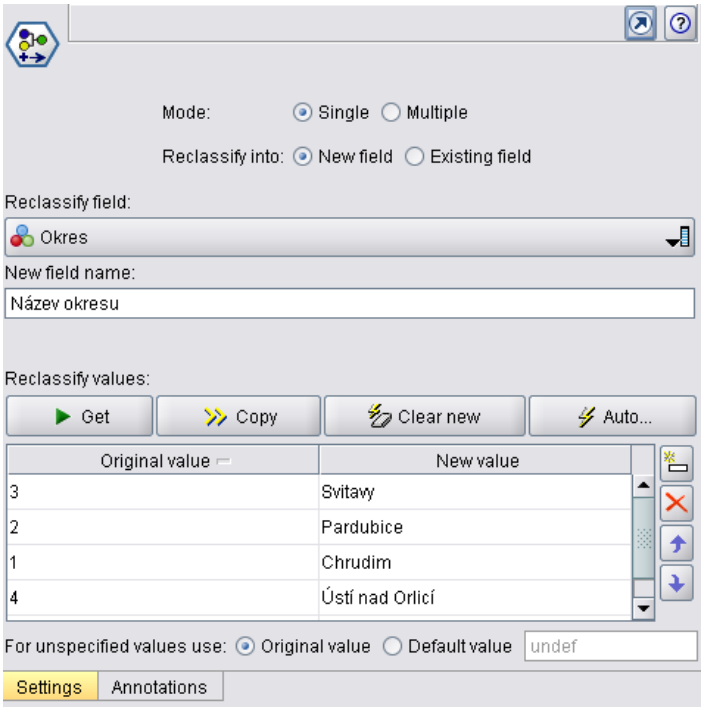
Obrázek 7: Kvalitativní analýza vstupních dat [Zdroj: vlastní]

Pro další práci s těmito daty je potřeba je vhodně připravit a upravit. Některé atributy budou rozděleny, odstraněny nebo upraveny za pomoci uzlů programu Clementine.

4.2 Úprava atributů

Prvním krokem úprav je úprava prvního sloupce „Označení“ a to tak, že bude rozdělen do tří sloupců na roky, čtvrtletí a měsíce, protože každý zvlášť působí na míru nezaměstnanost více než toto označení, které je pro každý údaj unikátní. Budou tedy přidány tři nové atributy a jeden bude odstraněn. Rozdělení a vytvoření nových atributů je provedeno pomocí uzlů DERIVE.

Jako další je provedena úprava atributu Okres, označení okresu je převedeno z čísel na jméno okresu uzlem RECLASSIFY. Tato úprava má jenom kosmetický charakter, pro přehlednost, na výsledky modelování to vliv nemá. Reklasifikace je vidět na Obrázku 8.



Obrázek 8: Reklasifikace označení okresu [Zdroj: vlastní]

Na výstupu z uzlu DATA AUDIT je vidět, že atributy „Absolventské praxe (vytvořená místa)“ a „Absolventské praxe (umístění uchazeči)“ mají poměrnou část hodnot rovno nule. Je to dáno tím, že od března roku 2003 nebyly přímo absolventské praxe vytvářeny. Tato skutečnost by další analýzu a modelování mohla ovlivnit, proto jsou tyto dva atributy pomocí uzlu FILTER z další analýzy vyloučeny.

Protože vybrané ekonomické ukazatele, které tvoří atributy, se všechny týkají přímo či nepřímo nezaměstnanosti je možné, že některé jsou korelované (závislé), je proto provedena pro vybrané ukazatele základní statistika včetně korelace a následně odstraněny závislé ukazatele.

Základní statistické včetně korelace údaje spočítá uzel STATISTICS. Provedení korelace ukázalo, že mezi atributy je několik, které jsou na sobě závislé. Jeden ze závislých je vždy odstraněn z další analýzy.

Závislé atributy:

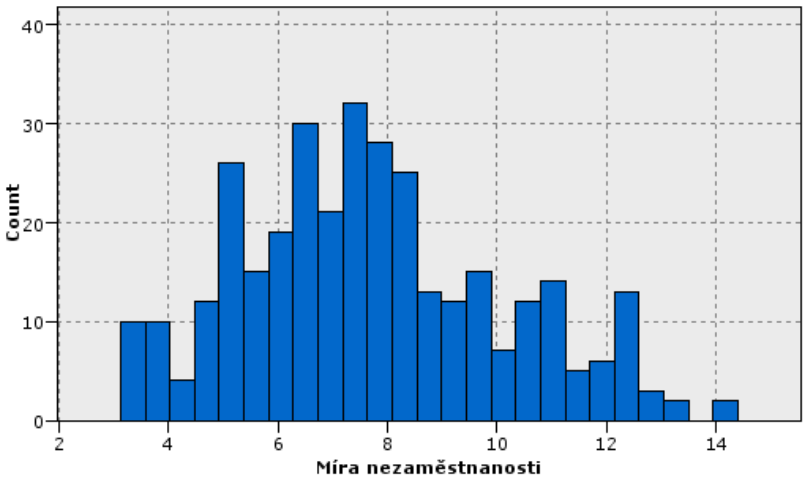
Z analýzy vyšly jako závislé tyto atributy: Vyřazení uchazeči ve sledovaném měsíci (celkem) a Vyřazení uchazeči ve sledovaném měsíci a umístěni, Veřejně prospěšné práce (umístění uchazeči) a Veřejně prospěšné práce (vytvořená místa) a Minimální mzda (měsíční) a Minimální mzda (hodinová). Vždy jeden z dvojice závislých atributů byl odstraněn. Následující tabulka ukazuje závislé a odstraněné ukazatele.

Tabulka 2: Závislé ukazatele [Zdroj: vlastní]

Vybrané ukazatele		**Odstraněné ukazatele**
Vyřazení uchazeči ve sledovaném měsíci (celkem)	×	Vyřazení uchazeči ve sledovaném měsíci a umístěni
Veřejně prospěšné práce (vytvořená místa)	×	Veřejně prospěšné práce (umístění uchazeči)
Minimální mzda (měsíční)	×	Minimální mzda (hodinová).

Analýza ukázala, že kromě výše uvedených atributů jsou závislé ještě další atribut, jsou to například Počet subjektů a Pracovní síla, a další. Po úvaze bylo rozhodnuto, že u těchto atributů je závislost náhodná, proto tyto atributy byly ponechány pro další analýzu a modelování.

Poslední úprava, která je provedena, je vytvoření nového atributu „Nezaměstnanost“ z atributu „Míra nezaměstnanosti“. Tento krok je zde zařazen proto, že rozhodovací strom typu C5.0, který je pro modelování také využit, nepracuje na výstupu se spojitými proměnnými. Míra nezaměstnanosti je rozdělena do pěti skupin podle četnosti, kterou určuje HISTOGRAM. Pomocí uzlu DERIVE je vytvořen nový sloupec s těmito hodnotami nahrazujícími původní míry nezaměstnanosti. Na Obrázku 9 je vidět graf, podle kterého bylo provedeno rozhodnutí o rozdělení míry nezaměstnanosti do skupin.



Obrázek 9: Histogram - Míra nezaměstnanosti [Zdroj: vlastní]

Rozdělení bylo provedeno takto:

1 = <0;5,5>, velmi nízká nezaměstnanost

2 = (5,5;6,9>, nízká nezaměstnanost

3 = (6,9;8>, střední nezaměstnanost

4 = (8;9,7>, vysoká nezaměstnanost

5 = (9,7;∞). Velmi vysoká nezaměstnanost

Na Obrázku 10 je zobrazen graf rozdělení, v každé skupině je od 66 do 69 údajů.

Value	Proportion	%	Count
1		19,64	66
2		20,24	68
3		19,64	66
4		19,94	67
5		20,54	69

Table Graph Annotations

Obrázek 10: Rozdělení míry nezaměstnanosti do skupin [Zdroj: vlastní]

V této fázi jsou data připravena pro modelování, které je popsané v následující kapitole.

4.3 Rozdělení dat na trénovací a testovací

Po úpravě vstupních dat je potřeba pro modelování data rozdělit na trénovací a testovací. Rozdělení dat bude náhodné, ale ne zcela. Protože trénovací data by měla obsahovat dvě třetiny původního počtu dat a data testovací jednu třetinu. Obvykle se přidávají ještě data validační, v tomto případě je velmi malý počet vstupních dat proto nejsou v tomto modelování brány

v úvahu. Protože je dále prováděna analýza a srovnání všech použitých metod, mělo by být rozdělení náhodné, ale stejné pro všechny metody.

Rozdělení provede uzel PARTITION, v kterém nastavíme počet trénovacích a testovacích dat v poměru 65% ku 35 %. To aby byl náhodný výběr pokaždé stejný je zajištěno zaškrtnutím políčka „Set random seed“ a nastavením prvního výchozího vzorku (hodnoty) od které bude rozdělení provedeno. Protože uzel PARTITION využívá pro rozdělení stále stejný algoritmus, nastavením výchozího vzorku je zajištěn stejný výběr při každém opakování.

Na Obrázku 11 je vidět rozdělení na data trénovací a testovací podle rozdělení nezaměstnanosti. Z obrázku je jasné, že není potřeba upravovat výchozí vzorek, protože data jsou rozdělena víceméně rovnoměrně a všechny skupiny nezaměstnanosti jsou v obou typech dat.

Nezaměstnanost

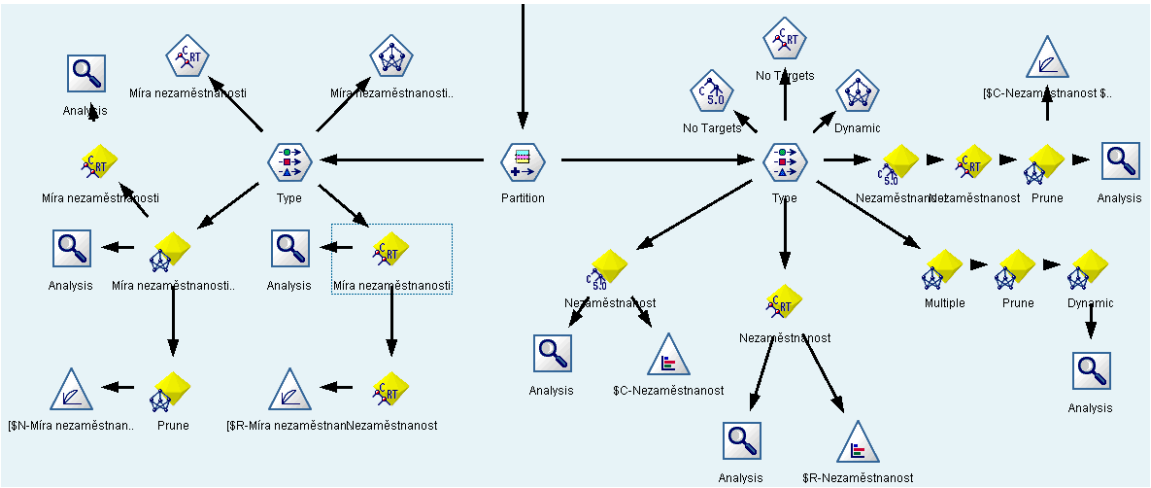
Partition		1	2	3	4	5
1_Training	Count	38	53	41	40	44
	Column %	57.576	77.941	62.121	59.701	63.768
	Total %	11.310	15.774	12.202	11.905	13.095
2_Testing	Count	28	15	25	27	25
	Column %	42.424	22.059	37.879	40.299	36.232
	Total %	8.333	4.464	7.440	8.036	7.440

Cells contain: cross-tabulation of fields (including missing values)
Chi-square = 7,57, df = 4, probability = 0,11

Obrázek 11: Rozdělení dat na trénovací a testovací - MATRIX [Zdroj: vlastní]

5 NÁVRH MODELŮ V PROSTŘEDÍ CLEMENTINE

V této kapitole je popsáno modelování v Clementine. Pro modelování budou použity rozhodovací stromy (C5.0 a C&RT) a neuronová síť. Pro všechny tyto metody budou natypovány stejné vstupy a výstupy, aby bylo možné výsledky všech metod porovnat. Návrh modelu (Stream) v programu Clementine je zobrazen na Obrázku12.



Obrázek 12: Návrh modelu v Clementine [Zdroj: vlastní]

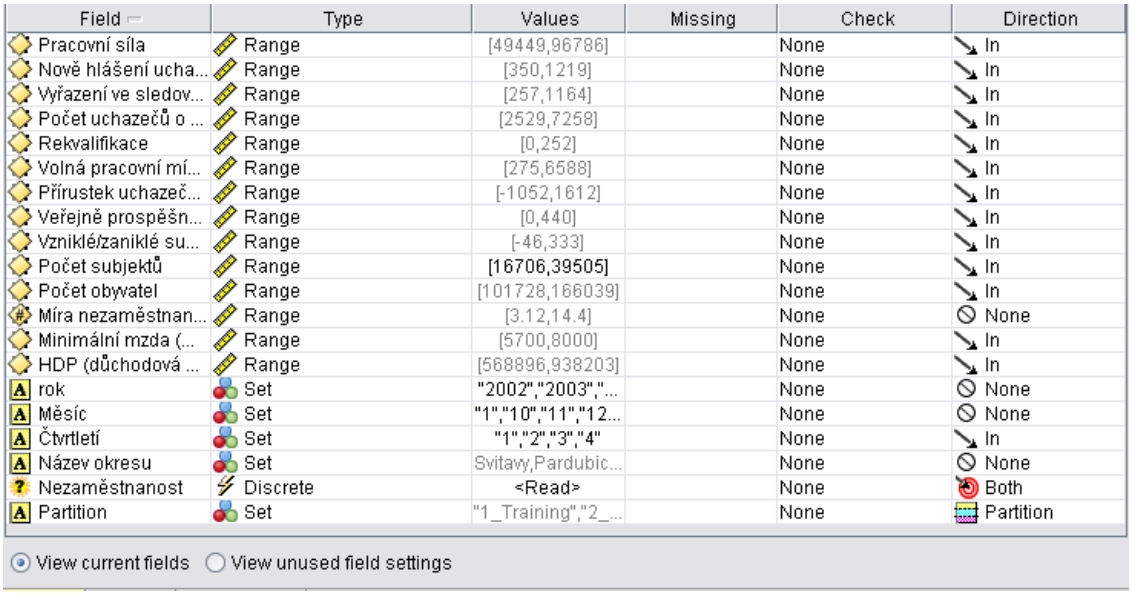
5.1 Rozhodovací stromy

Pro modelování byly z RS, které Clementine nabízí vybrány dva: C5.0 a CR&T.

C5.0

Tento model se využívá v problémových případech, např. pokud mají vstupní data velká čísla, nebo některá data chybí. Pracuje se všemi druhy proměnných. Na vstupu vyžaduje minimálně jednu proměnnou a na výstupu je možné použít pouze jednu proměnnou, která je kategorická.

Před spuštěním učení se RS C5.0 je nutné pro něj nadefinovat vstupy a výstupy, to je provedeno za pomoci uzlu TYPE: Na Obrázku 13 jsou vidět vstupy a výstupy. Protože tento typ RS pracuje na výstupu jen s kategorizovanými daty, bylo proto nutné v přípravě dat převést míru nezaměstnanost na kategorizovanou proměnnou.



Field	Type	Values	Missing	Check	Direction
Pracovní síla	Range	[49449,96786]		None	In
Nově hlášení ucha...	Range	[350,1219]		None	In
Vyřazení ve sledov...	Range	[257,1164]		None	In
Počet uchazečů o ...	Range	[2529,7258]		None	In
Rekvalifikace	Range	[0,252]		None	In
Volná pracovní mí...	Range	[275,6588]		None	In
Přírustek uchazeč...	Range	[-1052,1612]		None	In
Veřejně prospěšn...	Range	[0,440]		None	In
Vzniklé/zaniklé su...	Range	[-46,333]		None	In
Počet subjektů	Range	[16706,39505]		None	In
Počet obyvatel	Range	[101728,166039]		None	In
Míra nezaměstnan...	Range	[3.12,14.4]		None	None
Minimální mzda (...	Range	[5700,8000]		None	In
HDP (důchodová ...	Range	[568896,938203]		None	In
rok	Set	"2002","2003",...		None	None
Měsíc	Set	"1","10","11","12...		None	None
Čtvrtletí	Set	"1","2","3","4"		None	In
Název okresu	Set	Svitavy,Pardubic...		None	None
Nezaměstnanost	Discrete	<Read>		None	Both
Partition	Set	"1_Training","2_...		None	Partition

View current fields View unused field settings

Obrázek 13: Vstupy a výstupy C5.0 [Zdroj: vlastní]

Jako výstup metody je nastaven rozhodovací strom, dále je nastaven způsob tvorby jednoduchý a je vyžadována přesnost RS. Při nastavování všech metod rozhodovacích stromů musí být zaškrtnuté políčko „použití rozdělení dat“ („Use partitioned data“), neboli rozdělení dat na trénovací a testovací. Výstup, tedy RS, je přiložený k diplomové práci v Příloze č. 4.

Konečná analýza predikce pomocí RS C5.0 dopadla velmi dobře. Trénovací data jsou odhadována jen z 3,24% špatně a testovací ze 13,33% špatně. Celková analýza provedená uzlem ANALYSIS je vidět na Obrázku 14. Obrázek RS je v Příloze č. 4.

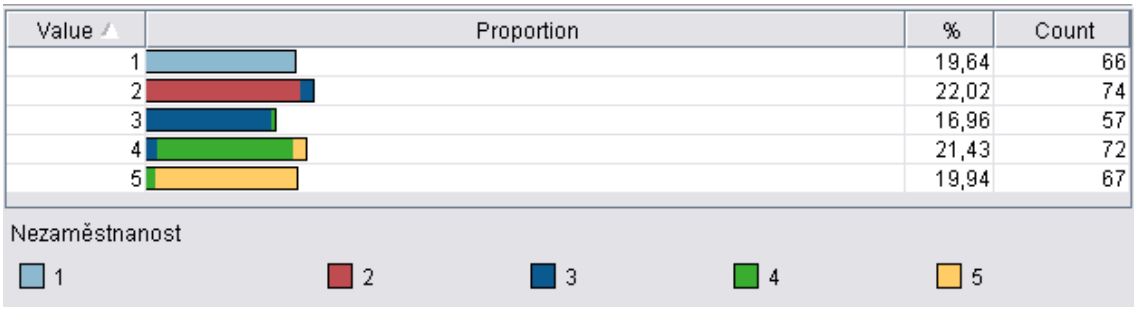
Results for output field Nezaměstnanost

Comparing $C-Nezaměstnanost with Nezaměstnanost

'Partition'	1_Training		2_Testing	
Correct	209	96,76%	104	86,67%
Wrong	7	3,24%	16	13,33%
Total	216		120	

Obrázek 14: Analýza výsledků metody C5.0 [Zdroj: vlastní]

Kromě Obrázku 14 jsou výsledky této metody zobrazeny také pomocí grafu DISTRIBUTION, který je zobrazen na Obrázku 15. Metoda C5.0 se zmýlila 16x, jak je vidět na při rozdělní do jednotlivých skupin není číslo až tak markantní, například 1 – „nízká nezaměstnanost“byla zařazena bez jediné chyby a v ostatních skupinách jsou taky jen drobné chyby.



Obrázek 15: Graf DISTRIBUTION metody C5.0 [Zdroj: vlastní]

C&RT

I tento model je odolný při výskytu problému, jako je chybějící proměnná nebo velká čísla v datovém souboru. Pracuje s kategorizovanými a spojitými proměnnými. Jako vstupní proměnná může být jedna a více proměnných a na výstupu může být jen jedna cílová proměnná, ta může být jak spojitá tak kategorizovaná.

Pro další metodu byla stejně nastavena vstupní data, jako u metody C5.0 (nastavení je na obrázku 13), ale protože meto C&RT umí pracovat na výstupu i se spojitými daty, byla tato metoda otestována i pro predikci původní míry nezaměstnanosti.

Výstupem metody byl zvolen obecný model a hloubka stromu je nastavena na 5 úrovní pod kořenem stromu. Analýza výsledků této metody pro kategorizovaná data je na Obrázku 16. Trénovací data tato metoda predikuje s chybou asi 7,87% a testovací data predikuje chybně z 15%.

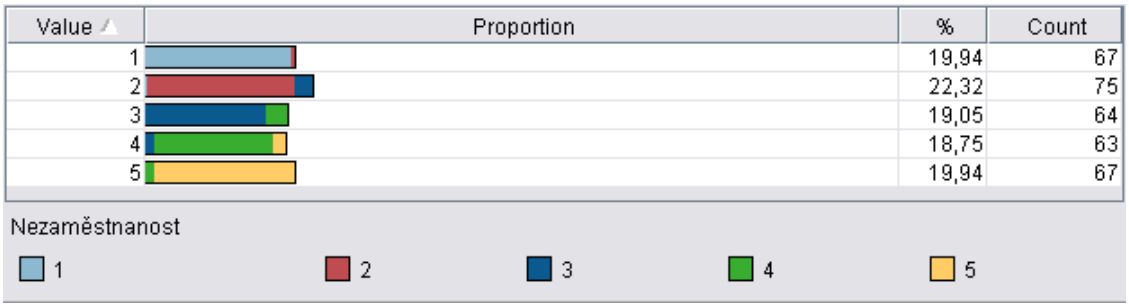
Results for output field Nezaměstnanost

Comparing $R-Nezaměstnanost with Nezaměstnanost

'Partition'	1_Training		2_Testing	
Correct	199	92,13%	102	85%
Wrong	17	7,87%	18	15%
Total	216		120	

Obrázek 16: Analýza výsledků metody C&RT (kategorizovaná data) [Zdroj: vlastní]

Do modelu vstupují stejné proměnné jako do předchozího modelu. Výstupní proměnnou je typ nezaměstnanosti a výsledek je graficky zobrazen na Obrázku 17. Obrázek RS je v Příloze č. 4.



Obrázek 17: Graf DISTRIBUTION metody C&RT [Zdroj: vlastní]

Na dalším Obrázku 18 je zobrazena analýza výsledků stejné metody ale pro spojitá data. Míru nezaměstnanosti odhaduje pro trénovací data s chybou. Nastavení vstupních a výstupních dat je stejné jako pro kategorizovaná data, jen místo vstupní proměnné nezaměstnanost (kategorizovaná) je nastavena vstupní proměnná míra nezaměstnanosti (spojitá).

Results for output field Míra nezaměstnanosti

Comparing $R-Míra nezaměstnanosti with Míra nezaměstnanosti

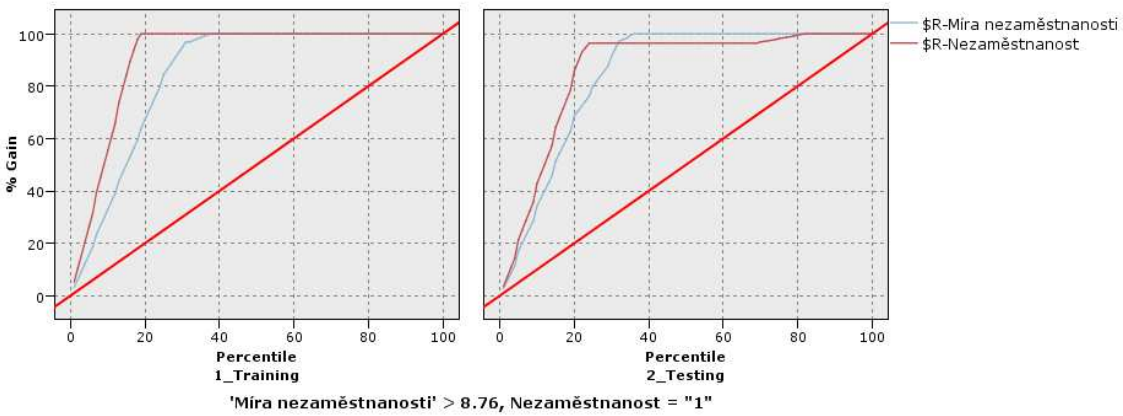
'Partition'	1_Training	2_Testing
Minimum Error	-1,039	-1,416
Maximum Error	1,001	1,851
Mean Error	-0,0	0,0
Mean Absolute Error	0,224	0,366
Standard Deviation	0,318	0,515
Linear Correlation	0,991	0,979
Occurrences	216	120

Obrázek 18: Analýza výsledků metody C&RT (spojitá data) [Zdroj: vlastní]

Na obrázku je vidět, že střední chyba této metody je 0 a že maximální chyba u trénovacích dat je něco málo přes 1 a maximální chyba u testovacích dat je 1,851. Obrázek RS je v Příloze č. 4.

Srovnání odhadu kategorizovaných a spojitých dat metodou C&RT:

Na obrázku 19 je vidět průběh učení metody pro oba typy dat. Na trénovacích datech lépe tato metoda odhaduje data kategorizovaná, kdy bezchybného odhadu dosáhne pod 20 percentilu. Na testovacích datech je lepší metoda C&RT pro data spojitá, kde bezchybného odhadu dosáhne pod 40 percentilu.



Obrázek 19: Graf trénování a testování RS C&RT [Zdroj: vlastní]

RS z obou metod u C&RT i oba typy dat jsou v Příloze č. 4.

5.2 Neuronové sítě

Kromě rozhodovacích stromů byla predikce modelována i pomocí neuronových sítí. NS stejně jako metoda CR&T umí na výstupu pracovat jak s kategorizovanými tak se spojitými daty. I na NS byla predikována jak kategorizovaná tak spojitá data, výsledky jsou okomentovány v závěru této podkapitoly.

NS jsou velmi silným nástrojem, nevyžadují zvláštní matematické a statistické znalosti. Výsledky mají obvykle stejně dobré jako ostatní techniky a občas i mnohem lepší. U NS neexistuje omezení typem dat. Je nutné zadat jeden nebo více vstupů a na výstupu musí být také uvedena minimálně jedna proměnná. Protože uzel NEURAL NET využívá učení s učitelem, podává model informaci o tom, jak jsou výsledky správné.

Pro neuronovou sít bylo otestováno několik trénovacích metod NS. Z nabízených metod programem Clementine jsou vyzkoušeny tyto tři: Multiple, Prune a Dynamic.

Metoda Multiple: vytváří více sítí různých topologií (přesné číslo závisí na trénovacích datech). Po natrénování je vybrán model s nejmenší chybou. [24]

Metoda Prune: využívá metodu postupného odřezávání nejslabších jednotek ve vstupní a skryté vrstvě NS. [24]

Metoda Dynamic: modifikuje standardní strukturu NS přidáním nebo odstraněním uzlů ve skryté vrstvě v průběhu procesu učení. [24]

Výsledky a srovnání všech tří metod jsou vidět na Obrázku 20. Z analýzy je patrné, že nejlépe predikuje metoda Prune, trénovací data predikuje s chybou jen 4,63% a u testovacích vypočítala jen 14,17% hodnot špatně. Ve čtvrté tabulce na obrázku 20 je srovnání všech tří použitých metod. Metody se shodly ve trénovacích datech ve 148 hodnotách z 216 a v testovacích datech v 71 případech ze 120. Metoda Prune dopadla podle analýzy nejlépe ze všech tří testovaných metod.

Results for output field Nezaměstnanost

Individual Models

Comparing $N-Nezaměstnanost with Nezaměstnanost

'Partition'	1_Training		2_Testing	
Correct	191	88,43%	96	80%
Wrong	25	11,57%	24	20%
Total	216		120	

Comparing $N1-Nezaměstnanost with Nezaměstnanost

'Partition'	1_Training		2_Testing	
Correct	206	95,37%	103	85,83%
Wrong	10	4,63%	17	14,17%
Total	216		120	

Comparing $N2-Nezaměstnanost with Nezaměstnanost

'Partition'	1_Training		2_Testing	
Correct	163	75,46%	71	59,17%
Wrong	53	24,54%	49	40,83%
Total	216		120	

Agreement between $N-Nezaměstnanost $N1-Nezaměstnanost $N2-Nezaměstnanost

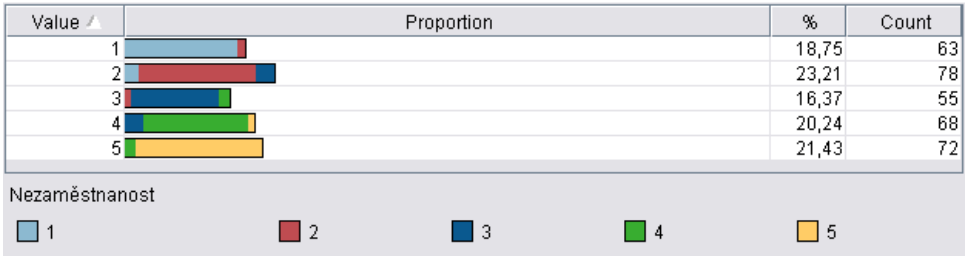
'Partition'	1_Training		2_Testing	
Agree	148	68,52%	71	59,17%
Disagree	68	31,48%	49	40,83%
Total	216		120	

Comparing Agreement with Nezaměstnanost

'Partition'	1_Training		2_Testing	
Correct	147	99,32%	61	85,92%
Wrong	1	0,68%	10	14,08%
Total	148		71	

Obrázek 20: Analýza výsledků predikce NS [Zdroj: vlastní]

Protože model Prune vyšel z analýzy nejlépe na Obrázku 21 je znázorněn graf DISTRIBUTION, který ukazuje, v kterých skupinách se jak zmýlil. Model se zmýlil v 17 případech, ale stejně jako u předchozích dvou jsou to u jednotlivých skupin drobné chyby. Na rozdíl od první metody C5.0 zde není ani jedna skupina nezaměstnanosti, kterou by NS odhadla celou bez chyby.

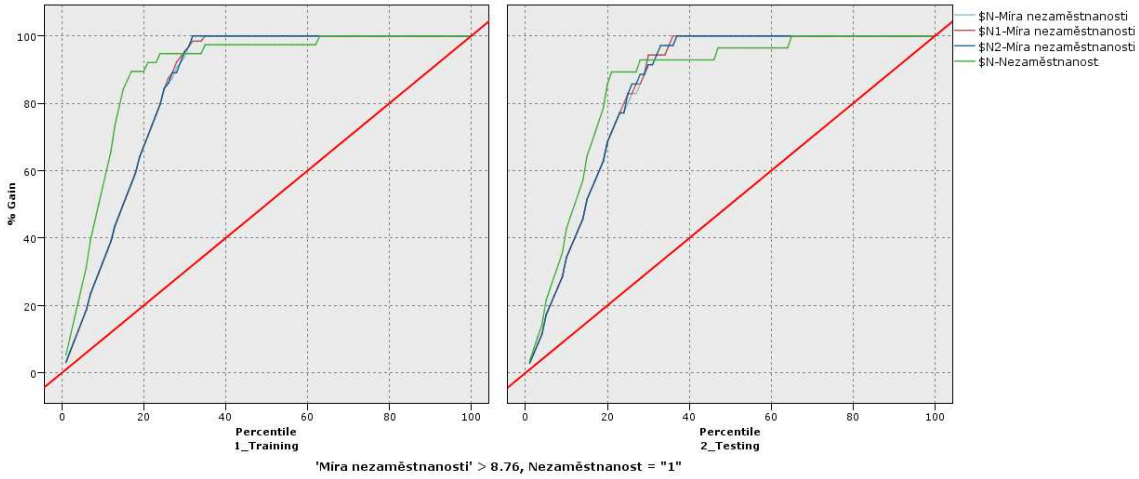


Obrázek 21: Graf DISTRIBUTION metody Prune NS [Zdroj: vlastní]

Protože NS umí na výstupu pracovat i se spojitými daty, je pomocí NS provedena i predikce původní míry nezaměstnanosti (spojitá data).

Srovnání predikce spojitých a kategorizovaných dat NS:

Na Obrázku 22 je vidět průběh učení NS pro oba typy dat. Pro trénovací i testovací data dopadla predikce NS lépe na spojitých datech, pro všechny tři metody, Multiple, Prune a Dynamic. U testovacích dat bezchybného odhadu dosáhnou všechny trénovací metody NS okolo 30 percentilu. Nejlépe u spojitých dat, i když jen o velmi málo, dopadla stejně jako v případě kategorizovaných dat metoda Prune.



Obrázek 22:Graf trénování a testování NS [Zdroj: vlastní]

5.3 Zjištění vlivu změny metodiky výpočtu míry nezaměstnanosti

Na začátku diplomové práce je uvedeno, že v červenci roku 2004 změnilo MPSV ČR metodiku výpočtu míry nezaměstnanosti, je zde i analýza zda a jak se změní míra správnosti odhadu nezaměstnanosti. Jestli změna metodiky ovlivnila učení nebo ne bylo zjišťováno na výsledcích nejlepší metody, která pro kategorická data byla C5.0. Dále byl vliv zkoumán pro nejlepší trénovací metodu NS, kterou je metoda Prune.

U obou metod byl vliv změny zjišťován dvěma různými způsoby:

1. způsob: pomocí uzlu SELECT byly vybrány jen ty objekty které byly danou metodou určeny správně. Dále pak za pomoci grafu DISTRIBUTION bylo zjištěno kolik hodnot v jednotlivých letech bylo určeno správně.

Na Obrázku 23 je vidět graf pro kategorizovaná data pro metodu C5.0. Z grafu je zřejmé, že daná změna neměla na určování nezaměstnanosti vliv. Ve všech sledovaných letech je podobný počet hodnot správně.

Value	Proportion	%	Count
2002		13,74	43
2003		14,06	44
2004		14,06	44
2005		14,06	44
2006		13,74	43
2007		15,34	48
2008		15,02	47

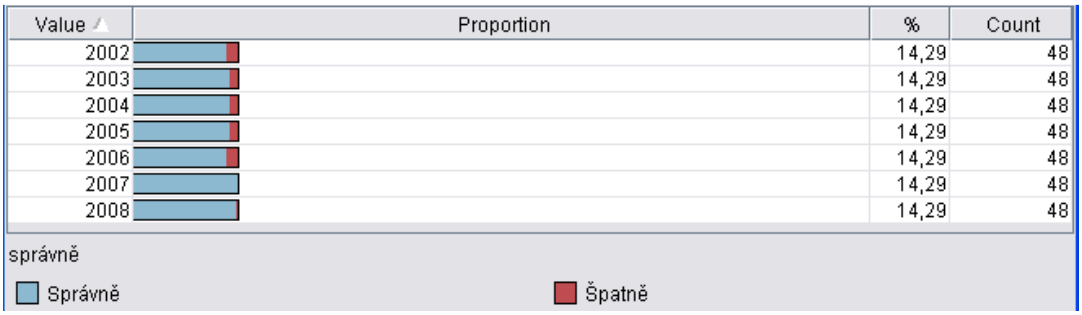
Obrázek 23: Správné hodnoty v jednotlivých letech – C5.0 [Zdroj: vlastní]

Také z následujícího Obrázku 24 je patrné, že změna metodiky neměla na určování nezaměstnanosti vliv. Obrázek ukazuje počty hodnot v jednotlivých letech určených správně NS s trénovací metodou Prune.

Value	Proportion	%	Count
2002		13,24	38
2003		15,33	44
2004		13,94	40
2005		13,94	40
2006		14,29	41
2007		15,33	44
2008		13,94	40

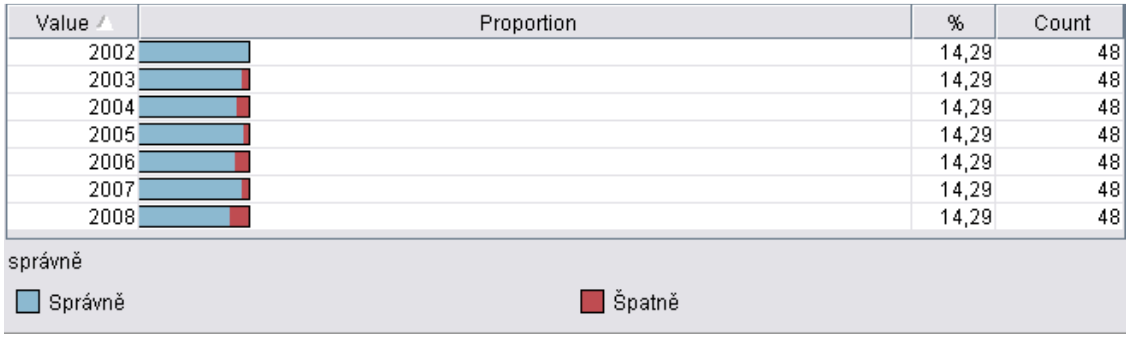
Obrázek 24: Správné hodnoty v jednotlivých letech – NS: Prune [Zdroj: vlastní]

2. způsob: pomocí uzlu DERIVE byl nadefinován další sloupec který má proměnné True (správně) a False (špatně), z toho sloupce je vygenerován graf DISTRIBUTION. Z grafu je snadné určit, zda změna výpočtu míry nezaměstnanosti má nebo nemá na odhadnutí nezaměstnanosti vliv. Na Obrázku 25 je vidět, že daná metoda odhadovala kategorizovaná data s obdobným procentem chybně před změnou v roce 2004 i po změně.



Obrázek 25: Odhadnuté hodnoty v jednotlivých letech – C5.0 [Zdroj: vlastní]

Stejně tak na Obrázku 26 je vidět, že vliv změny výpočtu míry nezaměstnanosti není žádný.



Value	Proportion	%	Count
2002		14,29	48
2003		14,29	48
2004		14,29	48
2005		14,29	48
2006		14,29	48
2007		14,29	48
2008		14,29	48

Obrázek 26: Odhadnuté hodnoty v jednotlivých letech – NS: Prune [Zdroj: vlastní]

Vliv byl zjišťován pouze u kategorizovaných dat, protože u spojitých dat by takovéto zjišťování bylo složité. U spojitých dat je velmi málo pravděpodobné, že by kterákoliv metoda dokázala odhadnout míru nezaměstnanosti na setiny správně. Vliv by bylo možné zkoumat pomocí odchylky mezi původní mírou nezaměstnanosti a odhadovanou mírou nezaměstnanosti.

6 ANALÝZA VÝSLEDKŮ

Protože pro odhad nezaměstnanosti bylo využito v modelování pomocí vybraných metod učení s učitelem, je hodnocení modelu založeno na testování shody nalezených hodnot s informací učitele.

Pro zjištění shody hodnot s informací učitele byl využit stejně jako u dílčích výsledků uzel ANALYSIS, která zobrazuje u jednotlivých metod počet správně (Correct) a počet špatně (Wrong) zvolených hodnot z celkového počtu hodnot. Dál bylo pro zjištění výsledků využito grafické zobrazení uzlu EVALUATION, které ukazuje jak rychle a jak dobře se je daná metoda schopná naučit podle poskytnutých dat predikovat zvolenou hodnotu.

Na Obrázku 27 jsou analýzy výsledků všech použitých RS a NS. Pro dané modelování je nejlepší metoda RS C5.0, která u testovacích dat odhaduje hodnotu nezaměstnanosti s chybou 13,33%.

Results for output field Nezaměstnanost

Individual Models

Comparing $C-Nezaměstnanost with Nezaměstnanost

'Partition'	1_Training		2_Testing	
Correct	209	96,76%	104	86,67%
Wrong	7	3,24%	16	13,33%
Total	216		120	

Comparing $R-Nezaměstnanost with Nezaměstnanost

'Partition'	1_Training		2_Testing	
Correct	199	92,13%	102	85%
Wrong	17	7,87%	18	15%
Total	216		120	

Comparing $N-Nezaměstnanost with Nezaměstnanost

'Partition'	1_Training		2_Testing	
Correct	206	95,37%	103	85,83%
Wrong	10	4,63%	17	14,17%
Total	216		120	

Agreement between $C-Nezaměstnanost $R-Nezaměstnanost $N-Nezaměstnanost

'Partition'	1_Training		2_Testing	
Agree	187	86,57%	90	75%
Disagree	29	13,43%	30	25%
Total	216		120	

Comparing Agreement with Nezaměstnanost

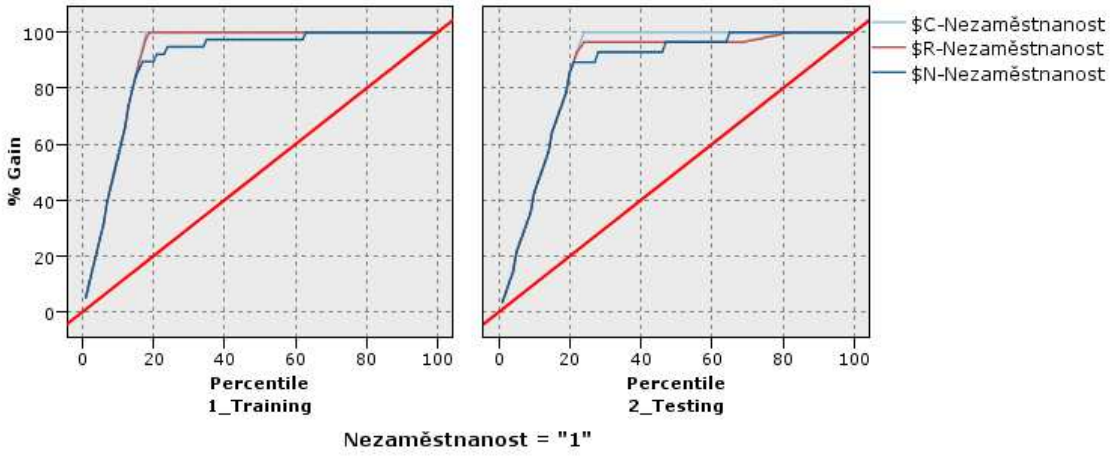
'Partition'	1_Training		2_Testing	
Correct	187	100%	85	94,44%
Wrong	0	0%	5	5,56%
Total	187		90	

Obrázek 27: Analýza predikce jednotlivých metod [Zdroj: vlastní]

Poslední dvě tabulky na obrázku ukazují, v kolika případech se metody shodly. Je vidět, že každá metoda pracuje na jiném principu, protože u testovacích dat je tato shoda jen 75% a z toho je správných 94,44%, což znamená, že společně správně odhadnutých hodnot bylo

v testovacích datech jen 85 ze 120 (cca 71%). Ostatní hodnoty byly odhadnuty špatně nebo se metody neshodují.

Následující Obrázek 28 ukazuje grafické zobrazení učení jednotlivých metod. I na tomto zobrazení je patrné, že nejlepší metodou pro modelování je metoda C5.0 (světle modrá). Bezchybné predikce u testovacích dat tato metoda dosáhne hned po 20 percentilu. Ostatním dvěma metodám trvá naučení se na těchto vstupních datech déle. Metoda C&RT (vínová) bezchybně predikuje u testovacích dat až na 80 percentilu a NS (tmavě modrá) po 60 percentilu.



Obrázek 28: Grafické zobrazení výsledků učení [Zdroj: vlastní]

Předcházející zhodnocení bylo pro kategorizovaná data upravené nezaměstnanosti. Protože bylo provedeno i modelování pro data spojitá, jsou zde popsány výsledky i tohoto modelování, byť nebyla pro tato data využita metoda RS C5.0, která neumí pracovat na výstupu se spojitými daty.

Na Obrázku 29 je analýza výsledků predikce původní míry nezaměstnanosti (spojitá data). Z obou metod, které byly na predikci použity je vidět, že NS s trénovací metodou Prune predikuje míru nezaměstnanosti s menšími maximálními a minimálními chybami, než RS C&RT, a absolutní střední chyba je pro testovací data jen 0,18 zatímco u RS C&RT je 0,366.

Results for output field Míra nezaměstnanosti

Individual Models

Comparing $N-Míra nezaměstnanosti with Míra nezaměstnanosti

'Partition'	1_Training	2_Testing
Minimum Error	-0,646	-0,73
Maximum Error	0,722	0,568
Mean Error	-0,012	-0,025
Mean Absolute Error	0,169	0,18
Standard Deviation	0,22	0,231
Linear Correlation	0,996	0,996
Occurrences	216	120

Comparing $R-Míra nezaměstnanosti with Míra nezaměstnanosti

'Partition'	1_Training	2_Testing
Minimum Error	-1,039	-1,416
Maximum Error	1,001	1,851
Mean Error	-0,0	0,0
Mean Absolute Error	0,224	0,366
Standard Deviation	0,318	0,515
Linear Correlation	0,991	0,979
Occurrences	216	120

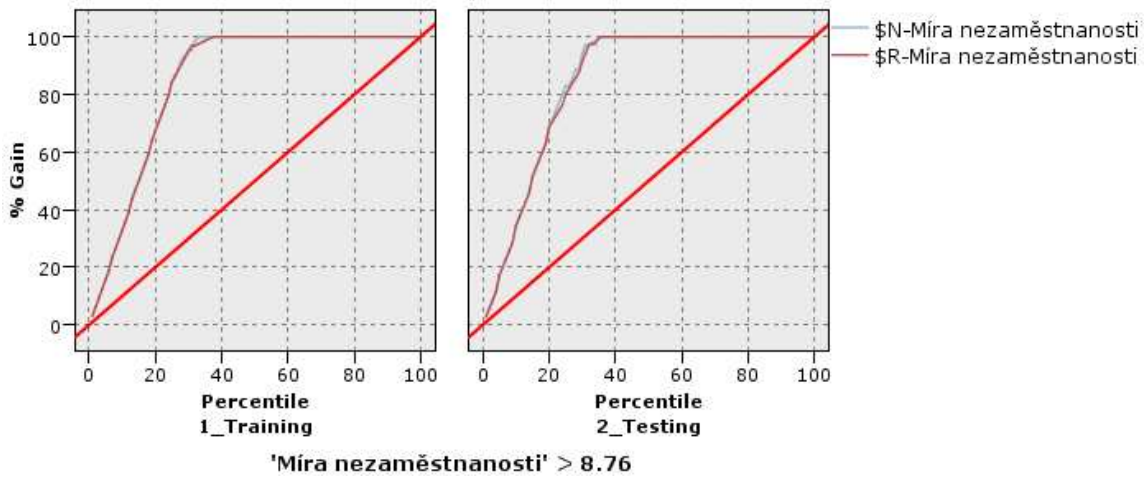
Agreement between $N-Míra nezaměstnanosti $R-Míra nezaměstnanosti

Comparing Agreement with Míra nezaměstnanosti

'Partition'	1_Training	2_Testing
Minimum Error	-0,654	-1,025
Maximum Error	0,665	0,954
Mean Error	-0,006	-0,013
Mean Absolute Error	0,157	0,227
Standard Deviation	0,209	0,317
Linear Correlation	0,996	0,992
Occurrences	216	120

Obrázek 29: Analýza výsledků (spojitá data) [Zdroj: vlastní]

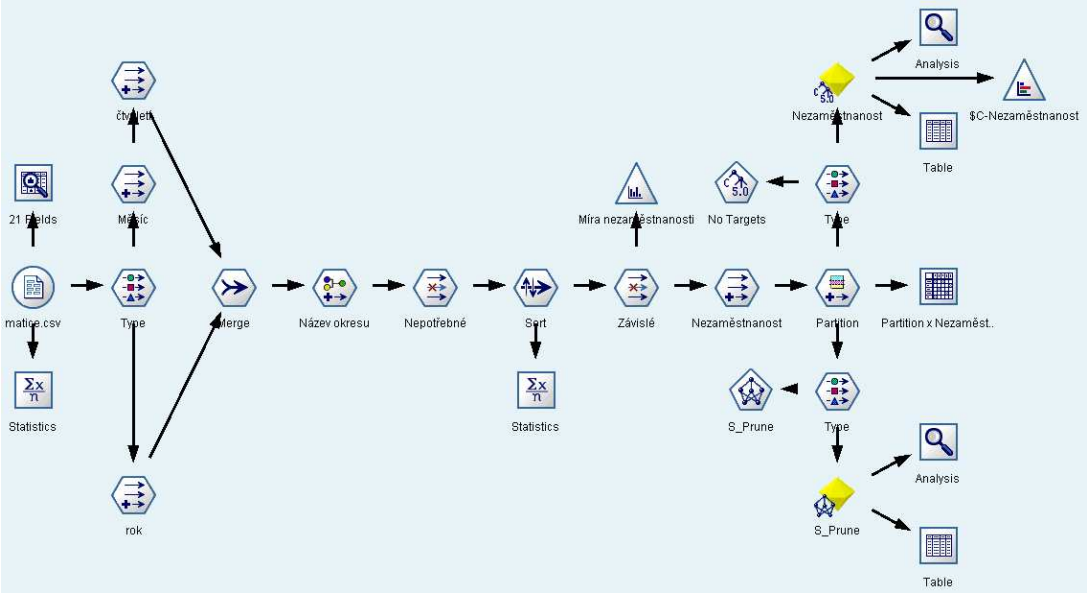
Obrázek 30 ukazuje grafické zobrazení učení obou metod. I zde je patrné, že lepši metodou pro predikci spojitých dat je NS, rozdíl není tak markantní jako v předchozím případě, ale bezchybné predikce dosáhne NS o něco málo dříve než RS C&RT.



Obrázek 30: Grafické zobrazení výsledků učení (spojitá data) [Zdroj: vlastní]

Z analýzy modelování jasně vyšlo, že nejlepší metodou pro modelování nezaměstnanosti, rozdělené do kategorií, je nejlepší metoda rozhodovacích stromů C5.0, která ale neumí pracovat na výstupu se spojitými daty, pro původní míru nezaměstnanosti je tedy z použitých metod nejlepší neuronová síť s trénovací metodou Prune.

Po celkové analýze výsledků by pak výsledný model pro zjišťování nezaměstnanosti respektive míry nezaměstnanosti mohl v programu Clementine vypadat jako Stream na Obrázku 31.



Obrázek 31: Výsledný návrh modelu [Zdroj: vlastní]

V modelu je zahrnuta celá příprava dat a nejlepší metody jak pro kategorizované výstupní proměnné tak pro výstupní proměnné spojité.

7 ZÁVĚR

V práci jsem se zabývala modelováním míry nezaměstnanosti a ji přímo či nepřímo ovlivňujících ekonomických ukazatelů na regionální úrovni pomocí rozhodovacích stromů a neuronových sítí. Z rozhodovacích stromů byly zvoleny metody C5.0 a C&RT, z neuronových sítí pak trénovací metody Multiple, Prune a Dynamic. Celý problém byl řešen v softwaru vytvořeném firmou SPSS Clementine 11.1.

Firma SPSS se v poslední době snaží prosadit svůj software i ve státní správě. Zástupkyně ředitele firmy SPSS, Petra Čiháková, představila na konferenci ISSS v Hradci Králové hlavní cíl společnosti. Cílem softwaru firmy SPSS je zjednodušit složité, zdlouhavé a náročné procesy, které musí pracovníci státní správy řešit dennodenně. Návštěvníkům ukázali především aplikace v oblasti snižování rizik a podvodného chování, **predikce vývoje nezaměstnanosti**, predikce zločinu a další.

Cílem práce bylo nalezení nejvhodnějšího modelu odhadování nezaměstnanost respektive míry nezaměstnanosti. Z výsledné analýzy vyplývá, že nejlepší ze zkoumaných metod je metoda rozhodovacích strom C5.0. Tato metoda však pracuje na výstupu jen s kategorizovanými daty, nikoli se spojitými. Z metod využitých pro spojitá data, C&RT a NS, dopadla nejlépe NS s trénovací metodou Prune.

Na základě výsledků metody C5.0 pro kategorizovaná data a NS: Prune pro data spojitá lze konstatovat, že oba navržené modely jsou schopny velmi dobře odhadovat na základě vstupních ekonomických ukazatelů nezaměstnanost, respektive míru nezaměstnanosti pro jednotlivé okresy Pardubického kraje.

Cíl práce byl splněn. Výsledný navržený model naleznete na straně 43 v části 6. Analýza výsledků. Tento model je tvořen z nejlepších testovaných metod.

8 POUŽITÁ LITERATURA

[1] CzechTrade. *Hlavní faktory regionálního rozvoje ČR - Kulturní potenciál, Cestovní ruch a Veřejná správa* [online]. 2003 [cit. 2009-02-10]. Dostupný z WWW: <http://www.businessinfo.cz/cz/clanek/rozvojregionu/regionalni-usporadani-a-regiony/1001179/9043/>.

[2] SASKOVÁ, R. *Předzpracování dat z vybraných regionů zaměřených na volný čas (rozvoj kulturních a sportovních příležitostí; spokojený občan)*: Univerzita Pardubice, 2008. 61 s. Diplomová práce. Dostupný z WWW: <http://hdl.handle.net/10195/29588>.

[3] BLAŽEK, J, UHLÍŘ, D. *Teorie regionální rozvoje: nástin, kritika, klasifikace.* Praha: Karolinum, 2002. 211 s. ISBN 80-246-0384-5.

[4] [KADEŘÁBKOVÁ, A. *Základy makroekonomické analýzy*. Praha: Linde, 2003. 175 s. ISBN 80-86131-36-X.]

[5] LIŠKA, V. *Mikroekonomie*. Praha: Professional Publising, 2004. 628 s. ISBN 80-86419-54-1.

[6] BERÁNKOVÁ, K. *MPSV harmonizuje vykazování míry nezaměstnanosti s EU* [online]. 2004 [cit. 2009-03-20]. Dostupný z WWW: <www.mpsv.cz/files/clanky/272/090804a.pdf>.

[7] *Hrubý domácí produkt (HDP)* [online]. 2009 [cit. 2009-03-20]. Dostupný z WWW: <http://www.czso.cz/csu/redakce.nsf/i/hruby_domaci_produkt_(hdp)>.

[8] *Informace o minimální mzdě* [online]. 2009 [cit. 2009-03-20]. Dostupný z WWW: <http://www.mpsv.cz/cs/4973>.

[9] ČAPEK, J. *Modelování ekonomickcých a sociálních procesů : pro kombinovanou formu studia.* Pardubice : Univerzita Pardubice, 2006. 103 s. ISBN 80-7194-838-1.

[10] PETR, *P. Data Mining : Díl I.* Pardubice : Univerzita Pardubice, 2006. 144 s. ISBN 80-7194-886-1.

[11] POŠÍK, P. *Data Mining* [online]. 2005 [cit. 2009-02-10]. Dostupný z WWW: <http://cyber.felk.cvut.cz/gerstner/teaching/zbd/DataMining1-hout.pdf>.

[12] *Clementine – data mining, prediktivní analýzy, prediktivní modelování.* [online] 2003-2008 [cit. 2009-03-13]. Dostupný z <http://www.spss.cz/sw_clemetnine.htm>

[13] *Společnost SPSS* [online]. 2003-2008 [cit. 2009-03-13]. Dostupný z WWW: <http://www.spss.cz/spss.htm>.

[14] RYCHLÝ, M. *Klasifikace a predikce* [online]. 2006 [cit. 2009-03-13]. Dostupné z: < http://www.fit.vutbr.cz/~rychly/docs/classification-and-prediction/classification-and-prediction.pdf >

[15] BERKA, P. *Dobývání znalostí z databází*. Praha: Academia, 2003. ISBN 80-200-1062-9.

[16] SPSS. *SPSS Classification Trees* [online]. 2008 [cit. 2009-03-13]. Dostupné z: <http://www.spss.cz/sw_mcla.htm>.

[17] *Rozhodovací stromy* [online]. 2002 [cit. 2009-03-20]. Dostupný z WWW: <http://datamining.xf.cz/view.php?cisloclanku=2002102802>.

[18] HAYKIN, S. *Neural Networks: A Compregensive Foundation*. New Jersey: Prentice-Hall, Inc., 1999. 842 s. ISBN 0-13273350-1.

[19] BERRY, M.J.A., LINOFF, G.: Mastering *Data Mining-The Art and Sciebce of Customer Relationship Management.* John Wiley & Sons: New York, 2000, ISBN 0-471-33123-6.

[20] BÍLÁ, J.: Umělá *inteligence a neuronové sítě v aplikacích.* Vydavatelství ČVUT: Praha, 1998, ISBN 80-01-01769-9.

[21] *Statistiky nezaměstnanosti* [online]. 2002-2008 [cit. 2009-01-31]. Dostupný z WWW: <http://portal.mpsv.cz/sz/stat/nz/mes>.

[22] *Statistický bulletin* [online]. 2001-2009 [cit. 2009-01-30]. Dostupný z WWW: <http://www.pardubice.czso.cz/>.

[23] *Veřejná databáze ČSÚ* [online]. 2007-2009 [cit. 2009-02-01]. Dostupný z WWW: <http://vdb.czso.cz/vdbvo/maklist.jsp?kapitola_id=22&vo=tabulka>.

[24] FILIPOVÁ , J, MICHÁLEK, K, PETR, P. *Identifikace automatických přístupů internetových obchodů s využitím metod web usage miningu* [online]. 2007 [cit. 2009-03-22]. Dostupný z WWW: <http://hdl.handle.net/10195/32234>. ISSN 1211-555X.

SEZNAM ZKRATEK

ČR	Česká republika
ČSÚ	Český statistický úřad
EU	Evropská unie
HDP	Hrubý domácí produkt
MPSV ČR	Ministerstvo práce a sociálních věcí České Republiky
NS	Neuronová síť
RS	Rozhodovací stromy

PŘÍLOHA Č. 1: TABULKA VSTUPNÍCH DAT

Označení	Okres	Pracovní síla	Nově hlášení uchazeči o zaměstnání	Vyřazení ve sledovaném měsíci	Vyřezení ve sledovaném měsíci a umístěni	Počet uchazečů o zaměstnání	Rekvalifikace	Volná pracovní místa	Přírůstek uchazečů o zaměstnání	Veřejně prospěšné práce (vytvořená místa)
7/2002	1	50013	781	447	360	4586	115	506	334	40
8/2002	1	50013	571	464	377	4693	85	581	107	34
9/2002	1	50013	806	841	660	4658	140	572	-35	59
10/2002	1	49449	528	771	612	4415	170	500	-243	63
11/2002	1	49449	518	491	410	4442	167	421	27	34
12/2002	1	49449	799	289	223	4952	105	312	510	0
1/2003	1	50026	856	607	438	5201	148	305	249	1
2/2003	1	50026	494	509	414	5186	174	312	-15	9
3/2003	1	50026	450	700	592	4936	158	275	-250	21
4/2003	1	50808	528	794	665	4670	155	298	-266	51
5/2003	1	50637	437	613	503	4494	147	310	-176	60
6/2003	1	50637	592	494	413	4592	137	297	98	63
7/2003	1	51920	775	414	318	4953	115	370	361	72

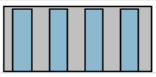
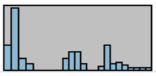
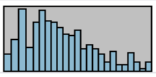
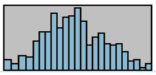
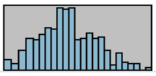
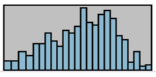
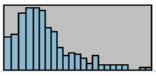
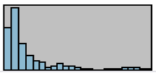
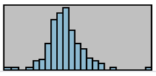
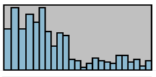
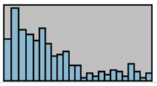
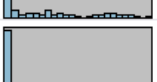
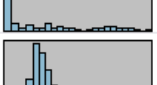
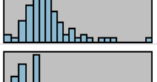
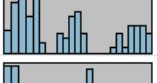
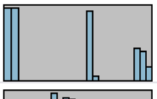
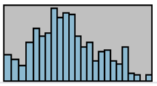
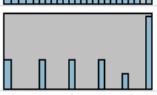
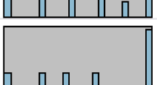
Označení	Veřejně prospěšné práce (umístění uchazeči)	Absolventské praxe (vytvořená místa)	Absolventské praxe (umístění uchazeči)	Vzniklé/ zaniklé subjekty	Počet subjektů	Počet obyvatel	Míra nezaměstnanosti	Minimální mzda (měsíční)	Minimální mzda (hodinová)	HDP (důchodová metoda)
7/2002	40	21	21	83	19234	104884	9.17	5700	33.9	606158
8/2002	34	21	21	83	19234	104884	9.4	5700	33.9	606158
9/2002	59	9	9	83	19234	104884	9.31	5700	33.9	606158
10/2002	63	8	8	42	19360	104901	8.93	5700	33.9	610721
11/2002	34	9	9	42	19360	104901	8.98	5700	33.9	610721
12/2002	0	9	9	42	19360	104901	10.01	5700	33.9	610721
1/2003	1	3	3	79	19598	104821	10.4	6200	36.9	602365
2/2003	9	3	3	79	19598	104821	10.37	6200	36.9	602365
3/2003	21	1	1	79	19598	104821	9.87	6200	36.9	602365
4/2003	51	1	1	127	19979	105046	9.19	6200	36.9	650992
5/2003	60	0	0	127	19979	105046	8.87	6200	36.9	650992
6/2003	63	0	0	127	19979	105046	9.07	6200	36.9	650992
7/2003	72	0	0	134	20381	104879	9.54	6200	36.9	647824

PŘÍLOHA Č. 2: STATISTICKÁ ANALÝZA ATRIBUTŮ

	Pracovní síla	Nově hlášení uchazeči o zaměstnání	Vyřazení ve sledovaném měsíci	Vyřazení ve sledovaném měsíci a umístění	Počet uchazečů o zaměstnání	Rekvalifikace	Volná pracovní místa	Přírůstek uchazečů o zaměstnání	Veřejně prospěšné práce (vytvořená místa)
Minimum	49449	350	257	141	2529	0	275	-1052	0
Maximum	96786	1219	1164	959	7258	252	6588	1612	491
Rozpětí	47337	869	907	818	4729	252	6313	2664	491
Součet	21860297	223732	226645	160533	1671782	21578	384833	-4107	48636
Průměr	65060.41	665.87	674.54	477.78	4975.54	64.22	1145.34	-12.22	144.75
Modus	51308	450	613	330	4693	52	542	-21	77
Medián	61949	624.5	662.5	465	5056.5	54	707	-30	111.5
Počet	336	336	336	336	336	336	336	336	336

Veřejně prospěšné práce (umístění uchazeči)	Absolvenstské praxe (vytvořená místa)	Absolventské praxe (umístění uchazeči)	Vzniklé/zaniklé subjekty	Počet subjektů	Počet obyvatel	Míra nezaměstnanosti	Minimální mzda (měsíční)	Minimální mzda (hodinová)	HDP (důchodová metoda)
0	0	0	-46	16706	101728	3.12	5700	33.9	568896
440	268	268	333	39505	166039	14.4	8000	48.1	938203
440	268	268	379	22799	64311	11.28	2300	14.2	369307
40787	12816	12788	17457	8631327	42649185	2591.24	2378280	14184	250468704
121.39	38.14	38.06	51.96	25688.47	126932.1	7.71	7078.21	42.21	745442.57
27	0	0	24	37205	104565	**X**	8000	48.1	568896
94	0	0	40	23048	121565	7.46	7185	42.5	751238.5
336	336	336	336	336	336	336	336	336	336

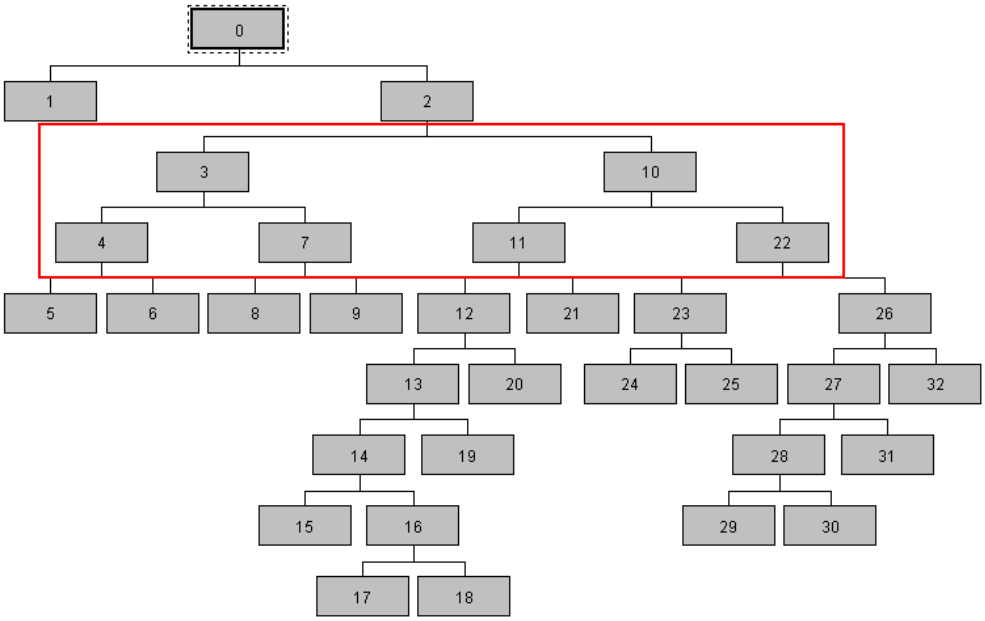
PŘÍLOHA Č. 3: KVALITATIVNÍ ANALÝZA VSTUPNÍCH DAT

Field	Graph	Type	Valid
Označení		Set	336
Okres		Set	336
Pracovní síla		Range	336
Nově hlášení uchazeči o zaměstnání		Range	336
Vyřazení ve sledovaném měsíci		Range	336
Vyřezení ve sledovaném měsíci a umístěni		Range	336
Počet uchazečů o zaměstnání		Range	336
Rekvalifikace		Range	336
Volná pracovní místa		Range	336
Přírustek uchazečů o zaměstnání		Range	336
Veřejně prospěšné práce (vytvořená místa)		Range	336
Veřejně prospěšné práce (umístění uchazeči)		Range	336
Absolvenstské praxe (vytvořená místa)		Range	336
Absolventské praxe (umístění uchazeči)		Range	336
Vzniklé/zaniklé subjekty		Range	336
Počet subjektů		Range	336
Počet obyvatel		Range	336
Míra nezaměstnanosti		Range	336
Minimální mzda (měsíční)		Range	336
Minimální mzda (hodinová)		Range	336
HDP (důchodová metoda)		Range	336

Obrázek 32: Kvalitativní analýza vstupních dat [Zdroj: vlastní]

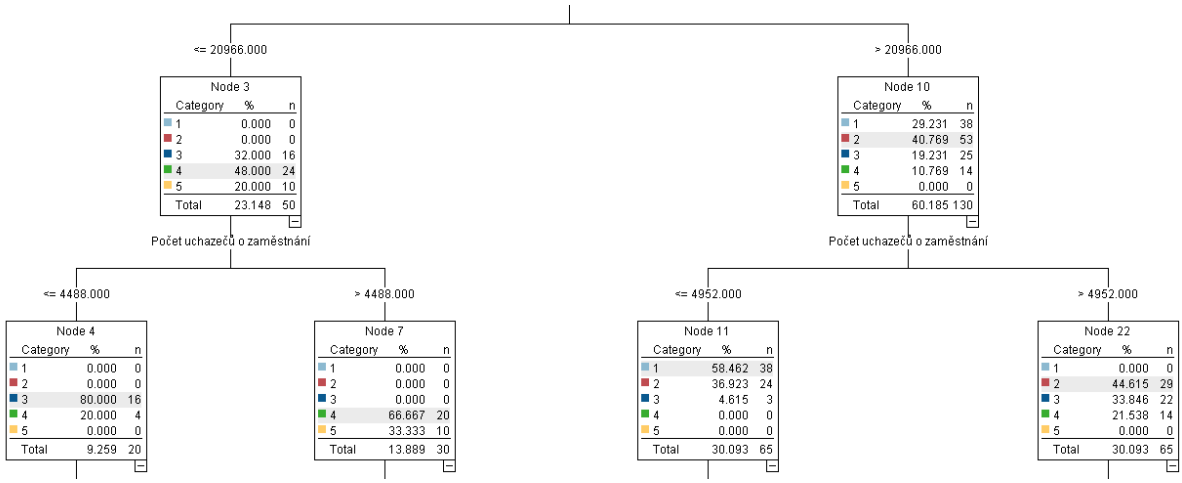
PŘÍLOHA Č. 4: ROZHODOVACÍ STROMY

RS C5.0 – kategorizovaná data



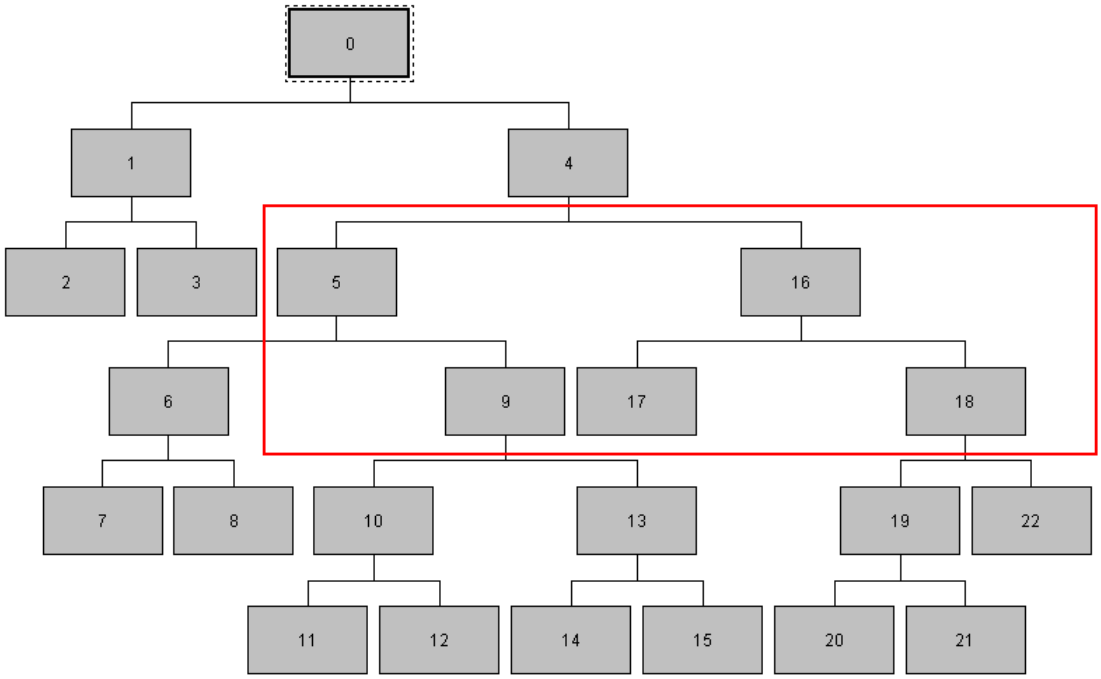
Obrázek 33: RS C5.0 - kategorizovaná data [Zdroj: vlastní]

Na Obrázku 33 je celý RS vytvořený metodou C5.0 pro kategorizovaná data na výstupu, následující Obrázek 34 ukazuje jen červeným rámečkem vyznačenou část RS.



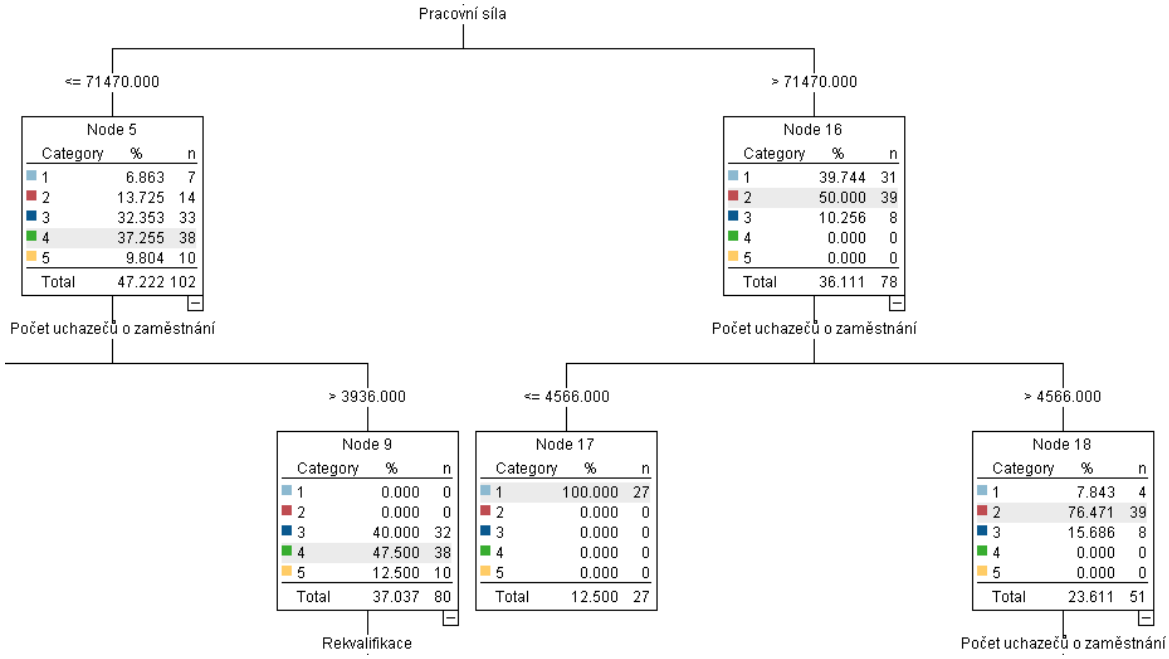
Obrázek 34: Detail RS C5.0 - kategorizovaná data [Zdroj: vlastní]

RS C&RT – kategorizovaná data



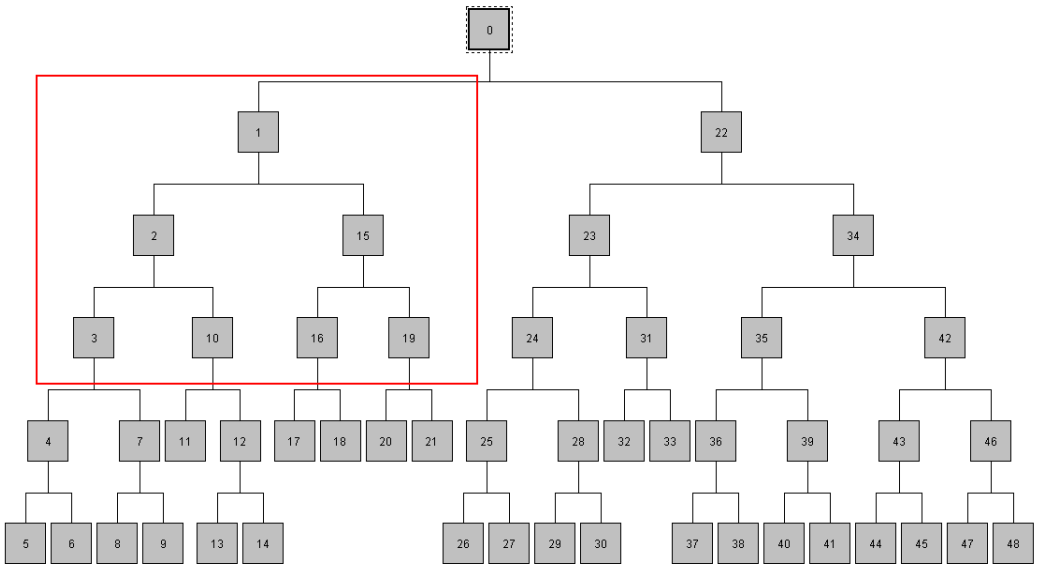
Obrázek 35: RS C&RT - kategorizovaná data [Zdroj: vlastní]

Na Obrázku 35 je celý RS vytvořený metodou C&RT pro kategorizovaná data na výstupu, následující Obrázek 36 ukazuje jen červeným rámečkem vyznačenou část RS.



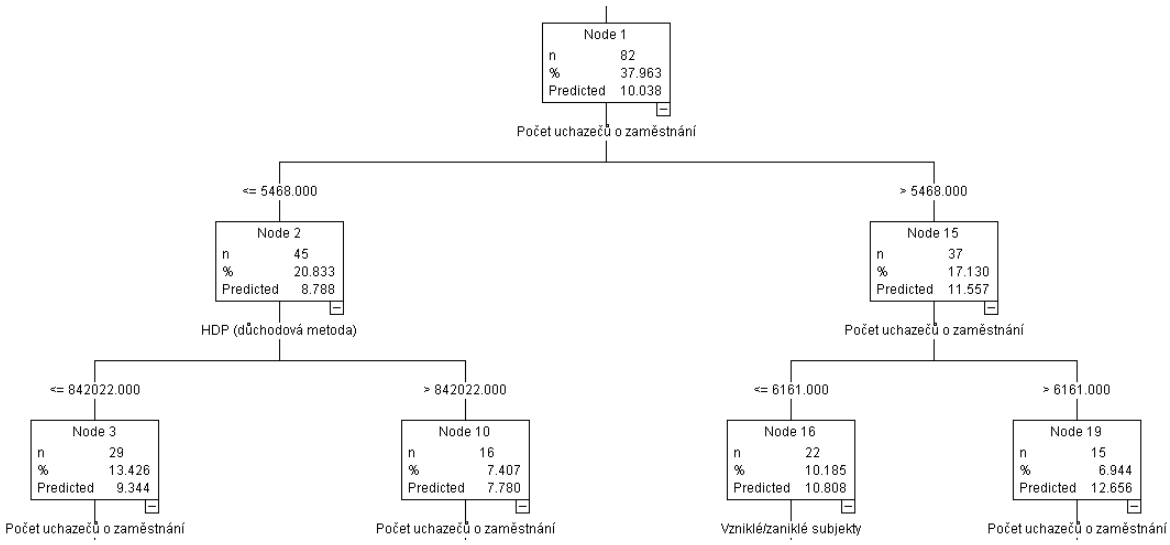
Obrázek 36: Detail RS C&RT - kategorizovaná data [Zdroj: vlastní]

RS C&RT – spojitá data



Obrázek 37: RS C&RT - spojitá data [Zdroj: vlastní]

Na Obrázku 37 je celý RS vytvořený metodou C&RT pro spojitá data na výstupu, následující Obrázek 38 ukazuje jen červeným rámečkem vyznačenou část RS.



Obrázek 38: Detail RS C&RT - spojitá data [Zdroj: vlastní]