Univerzita Palackého v Olomouci
Přírodovědecká fakulta
Katedra geoinformatiky

orange
DATA MINING

ORANGE

Praktický návod do cvičení předmětu Data Mining

Zdena DOBEŠOVÁ

Olomouc 2022

Tato publikace vznikla s podporou Erasmus+ Program, Jean Monnet Module.

Project No. 620791-EPP-1-2020-1-CZ-EPPJMO-MODULE UrbanDM - Data mining and analyzing of urban structures as contribution to European Union studies.

Podpora Evropské komise při tvorbě této publikace nepředstavuje souhlas s obsahem, který odráží pouze názory autorů, a Komise nemůže být zodpovědná za jakékoliv využití informací obsažených v této publikaci.

Za návrh šablony dokumentu děkuji Mgr. Jakubovi Koníčkovi, za laskavé přečtení textu děkuji Bc. Pavlu Novákovi.

Děkuji recenzentům doc. Jiřímu Dvorskému a doc. Pavlu Petrovi za podnětné připomínky k textu.

Odborní recenzenti:

doc. Ing. Pavel Petr, Ph.D.

doc. Mgr. Jiří Dvorský, Ph.D.

1. vydání

DOI: 10.5507/prf.22.2440864

ISBN 978-80-244-6086-4 (online: iPDF)

ÚVOD

Předkládaná učebnice představuje software Orange a jeho použití při řešení praktických příkladů. Text je určen zejména studentům předmětu Data Mining v magisterského programu Geoinformatika a kartografie. Text můžou použít i uživatelé z geovědní praxe, kteří chtějí získat základní praktické dovednosti v oblasti Data Mining a softwaru Orange. Text je doplňkem teoretických přednášek předmětu a je směřován jako praktický návod do cvičení, který má na příkladech ozřejmit teorii. Jen v úvodu některých kapitol je uveden stručný teoretický úvod a odkazy na literaturu, která lze použít jako východisko pro další studium. Text učebnice je využitelný i pro samostudium.

Praktické příklady využívají jednak data, která jsou instalována se softwarem Orange a také data často používána při demonstraci Data Mining metod v řadě učebnic jako je např. dataset kosatců Iris. Nicméně autorka textu se snažila zařadit do procvičování geografická data, jako např. data z Evropské statistické databáze Eurostat či projektu Copernicus Urban Atlas, kdy oba zdroje shromažďují a poskytují data v rámci Evropské unie. Cílem bylo, aby data byla oborově blízká, aktuální a přínosná pro studenty oboru Geoinformatika.

V textu je řada odkazů na další podpůrné materiály k software Orange jako je oficiální manuál nebo videa na YouTube. Protože neustále dochází k aktualizaci softwaru, některé obrázky nebo sdělení v textu nemusí odpovídat právě používané aktuální verzi. Učební text vznikal postupně s použitím verze 3.24 až po verzi 3.30. Mimo průběžně vydávané nové verze softwaru Orange je k dispozici i řada doplňků. S ohledem na osnovu předmětu učebnice postihuje jen některá vybraná témata Data Miningu a některé dostupné doplňky či widgety. Studentům se tak otvírá prostor pro další samostatné zkoumání a práci v softwaru Orange.

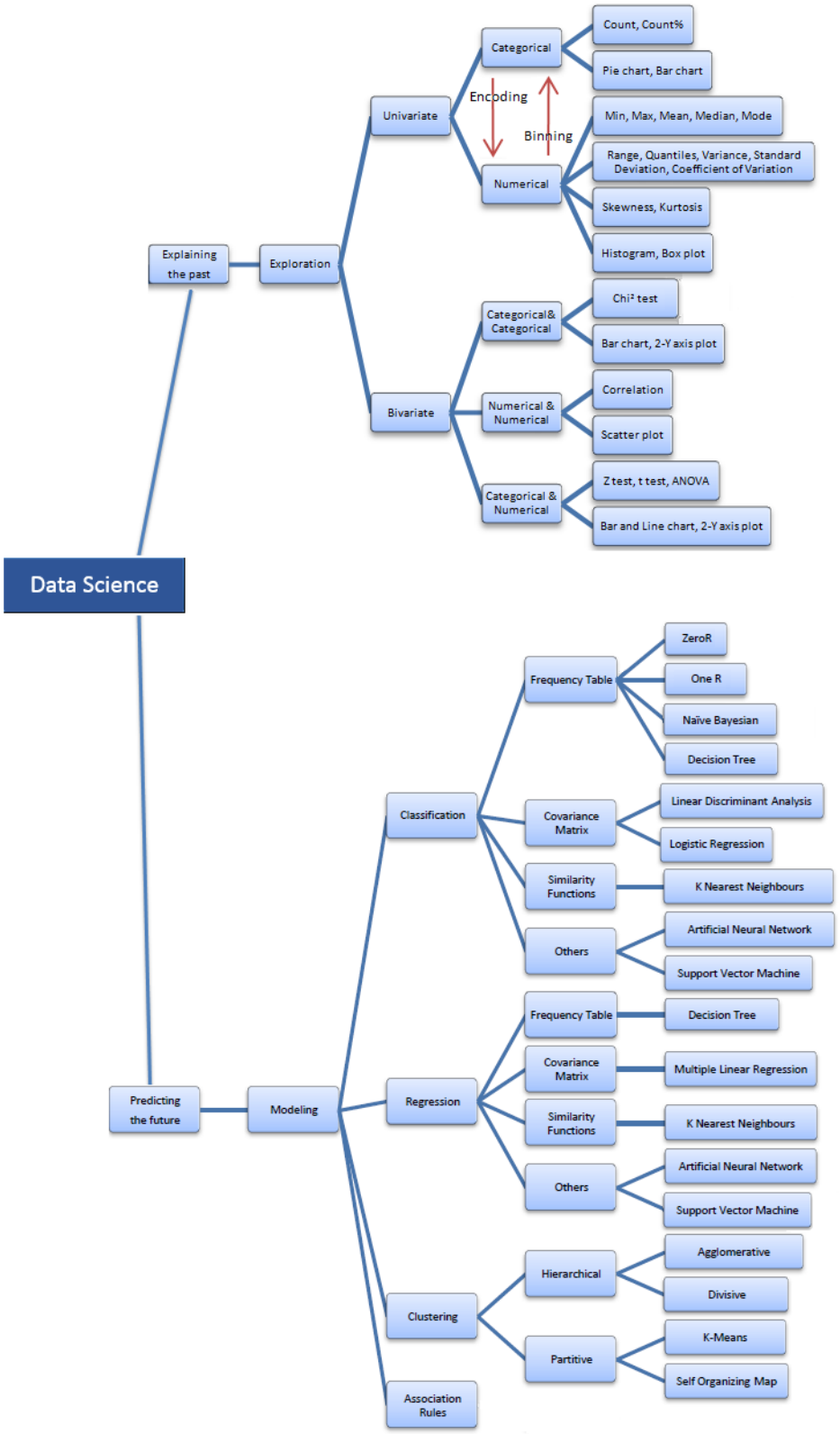
OBSAH

1 Data Science a přehled metod 7
2 Software Orange 8
2.1 Úvodní informace o softwaru 8
2.2 Rozhraní a základy práce 9
3 Popis a hrubá filtrace dat 10
3.1 Základní popis dat 10
3.2 Průzkum dat pomocí krabicového grafu 13
3.3 Detekce odlehlých hodnot 14
3.4 Rozptylový graf a uzel Color 15
3.5 Profilový graf 16
3.6 Diskretizace do intervalů 16
3.7 Oprava chybných dat a sjednocení kategorií 19
3.8 Imputace hodnot 19
3.9 Normalizace a standardizace 21
3.10 Dichtomizace atributů 23
3.11 Smazání nevýznamných atributů 25
4 Korelace a redukce dimenzionality 26
4.1 Korelace 26
4.2 PCA – analýza hlavních komponent 27
4.3 Korespondenční analýza 31
5 Shlukování 35
5.1 Jaccardův koeficient asociace 35
5.2 Výpočet vzdálenosti a hierarchické shlukování 38
5.3 Posouzení konzistence shluku pomocí Silhouette Plot 41
5.3.1 Shlukování podle atributů 43
5.4 Metoda k-Means 44
5.4.1 Interaktivní k-Means 44
5.4.2 Shlukování metodou k-Means 45
5.4.3 Shlukování po redukci dimenzí 48
5.5 Hledání nejbližšího souseda 49
5.6 Metoda DBSCAN 51
5.7 Kohonenova mapa – SOM 54
6 Predikční a klasifikační modely 55
6.1 Lineární regresní model a predikce pomocí modelu 55
6.2 Model k-NN 57

6.3 Rozhodovací stromy ... 59
6.3.1 Strom pro určení druhu kosatce ... 59
6.3.2 Strom hraní tenisu ... 60
6.3.3 Strom pro určení pohlaví z hmotnosti a výšky ... 63
6.3.4 Klasifikace podle rozhodovacího stromu ... 64
6.3.5 Úloha Uchazeči ... 67
6.4 Random Forest – náhodný les ... 69
6.5 Naivní bayesovský klasifikátor ... 71
6.6 Metoda Support Vector Machines ... 73
7 Asociační pravidla ... 76
7.1 Frekventované sady instancí ... 77
7.2 Asociační pravidla pro lokalitu domu ... 79
7.3 Asociační pravidla pro dichotomická data ... 81
8 Neuronové sítě a doplněk Image Analytics ... 84
8.1 Neuronová síť Painters ... 84
8.2 Podobnost map ... 87
8.3 Kategorizace mapy ... 90
8.4 Hledání podobných evropských měst ... 91
9 Prostorová data a doplněk Geo ... 93
9.1 Zobrazení bodových dat v mapě pomocí Geo Map ... 93
9.2 Geokódování pomocí uzlu Geocoding ... 95
9.3 Tvorba kartogramu pomocí uzlu Choropleth map ... 98
9.4 Vymezení převahy jevů v území pomocí rozhodovacího stromu ... 98
10 Časové řady ... 101
10.1 Korelace dvou časových řad ... 101
10.2 Zjištění stacionarity časové řady ... 102
10.3 Rozklad časové řady ... 107
10.4 Autokorelace a predikce hodnot časové řady ... 110
11 Závěr ... 114
12 Použité zdroje ... 115

1 DATA SCIENCE A PŘEHLED METOD

Přehled základních metod Data Science je zobrazen na Obr. 1 (Sayad 2020a). Základní dělení metod čtenáři napomůže v orientaci a porozumění metodám. Některým vybraným metodám je věnován tento studijní text.



Obr. 1 Základní dělení metod (zdroj Sayad S. An Introduction to Data Science, http://www.saedsayad.com/data_mining_map.htm)

2 SOFTWARE ORANGE

Kapitola obsahuje úvodní informace jak získat software Orange a dále stručné seznámení s jeho použitím.

2.1 Úvodní informace o softwaru

Software Orange je vytvářen na University of Ljubljana na Faculty of Computer and Information Science ve Slovinsku, v laboratoři bioinformatiky, původně pod výzkumnou skupinou Biolab (Orange Data Mining 2021a). Orange je svobodný software; lze jej šířit a/nebo upravovat za podmínek GNU General Public License zveřejněné Free Software Foundation.

Orange je určen pro pedagogické účely a výzkum. Orange je **vizuální programovací jazyk pro data mining a interaktivní analýzu** (Demšar *et al.* 2013). Průzkum a zpracování dat se provádí v grafickém prostředí, kde se jednotlivé kroky vytváří jako uzly postupu zpracování, tzv. **workflow**.

Volné stažení instalace softwaru je ze stránek **https://orangedatamining.com**.

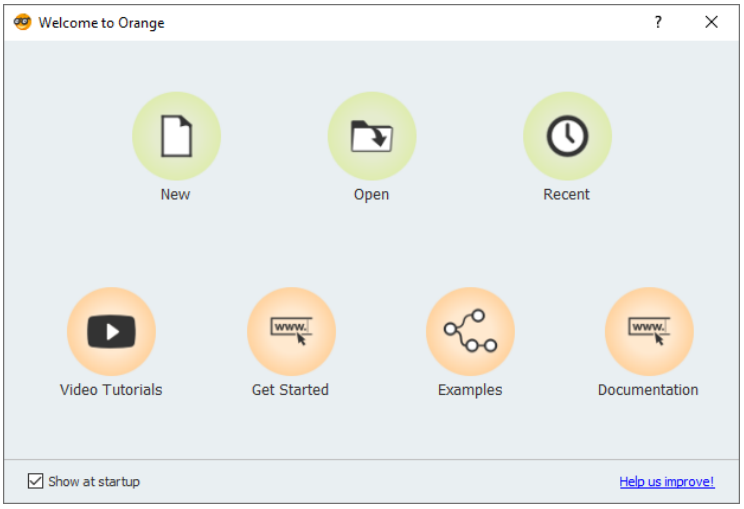
YouTube poskytuje řadu krátkých návodů v délce 2–4 min. Tyto návody jsou přímo zpracovány autorskou výzkumnou skupinou z univerzity ze Slovinska a jsou tak dobrým zdrojem pro rychlé seznámení se softwarem.

Jako první lze doporučit startující sadu návodů:

Getting Started with Orange 01: Welcome to Orange

Getting Started with Orange 02: Data Workflows

Getting Started with Orange 03: Widgets and Channels



Obr. 2 Uvítací dialog, který obsahuje link na první YouTube tutoriál, příklady a link na dokumentaci

Aktuální oficiální manuál je dostupný zde https://buildmedia.readthedocs.org/media/pdf/orange-visual-programming/latest/orange-visual-programming.pdf (Orange Data Mining 2021c). Tento soubor je stále průběžně aktualizovaný a je dostupná vždy poslední platná verze.

Rozcestník nápovědy je dostupný na adrese https://orange3.readthedocs.io/projects/orange-visual-programming/en/master/ (Orange Data Mining 2021b).

Protože videa na YouTube a oficiální manuál dostatečně podrobně popisuje základy práce, zejména popis různých způsobů připojení dat, tento text se soustředí, po krátkém seznámení s rozhraním, přímo na konkrétní operace a příklady.

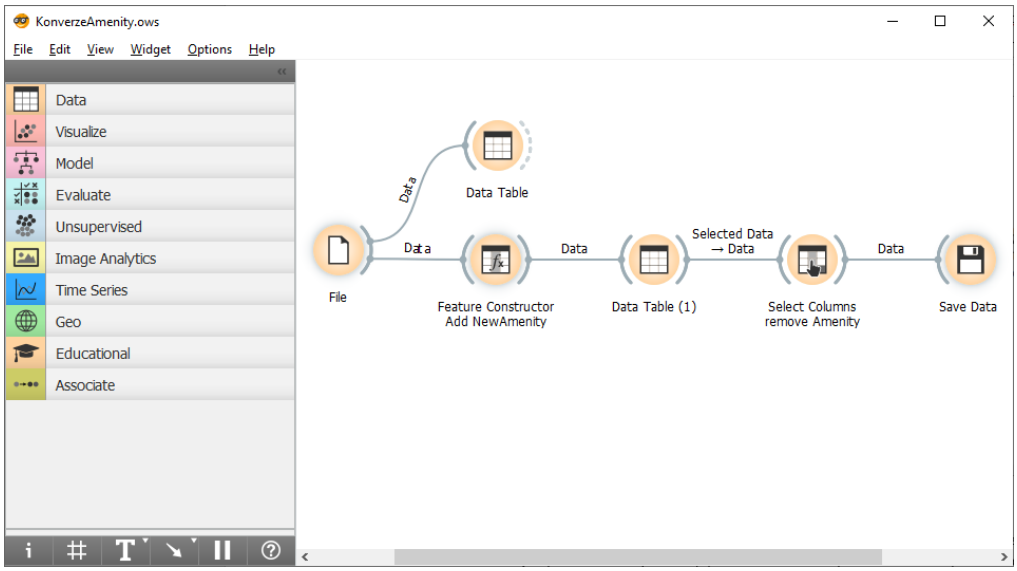
K jednotlivým příkladům učebnice jsou dostupná cvičná data a hotová workflow. V úvodu jednotlivých příkladů je vždy uveden název vstupních dat a příslušného workflow. Někdy více příkladů používá stejná cvičná data a společné workflow.

Data používaná v této učebnici jsou dostupná ke stažení na webu **dobesova.upol.cz/Orange**.

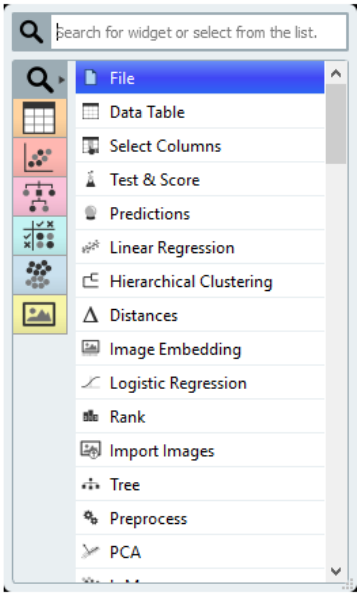
Vstupní data příkladů jsou umístěna v adresáři *Data*, hotové postupy jsou v adresáři *Workflow.*

2.2 Rozhraní a základy práce

Práce v prostředí Orange je intuitivní a rychlá. Jednotlivé operace se vybírají z panelu vlevo a přetahují se na kreslící plochu jako uzly (Obr. 3). Operace jsou v levém panelu sdruženy do skupin jako je *Data*, *Visualize*, *Model* atd. Skupiny jsou barevně odlišeny, kdy uzly ve skupině mají vždy stejnou barvu podle skupiny. Uzly lze také vkládat na plochu přes pravé tlačítko myši, kdy se otevře okno nabídky operací, a přes vstup *Search* lze zadáním několika počátečních písmen rychle vyhledat požadovanou operaci (Obr. 4). Tento český manuál bude operace ve workflow označovat jako **uzly**. Originální manuál označuje uzly jako **widgety**, protože reprezentují samostatné programové kódy s konkrétním určením.



Obr. 3 Základní rozhraní software Orange



Obr. 4 Search okno pro vyhledání uzlu (widgetu)

Postupy řešení úlohy budou v textu označovány jako **workflow**. Postupy se ukládají do souboru s koncovkou **ows,** např. *3_Iris.ows*. Orange používá jako výchozí vstupní datový formát soubory s koncovkou *tab*. Výhodou je možnost načítat vstupní data ze souboru Microsoft Excel, CSV formátu nebo URL adresy. Uzel **Data Sets** umožňuje načíst cvičná data dodávaná se softwarem Orange.

Výhodou programu je, že jednotlivé uzly automaticky přepočítávají výsledky a není tedy nutné spouštět manuálně každý krok nového zpracování při změně dat nebo parametrů. Výchozí volbu uzlů *Send/Apply Automatically* lze odebrat a potom se nové zpracování spouští stiskem tlačítka (např. Obr. 6 dole).

3 POPIS A HRUBÁ FILTRACE DAT

První skupina operací **Data** obsahuje operace na základní načtení, zobrazení a předzpracování dat. Prvním krokem zpracování dat by mělo vždy být základní seznámení se s daty a získání základních popisných charakteristik. Následujícím krokem zpracování je očištění dat od hrubých chyb a transformace dat (Šarmanová 2012) (Berka 2005).

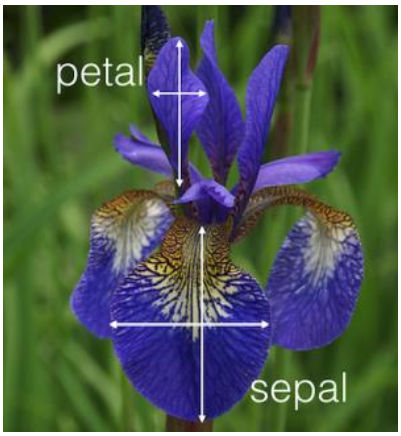
3.1 Základní popis dat

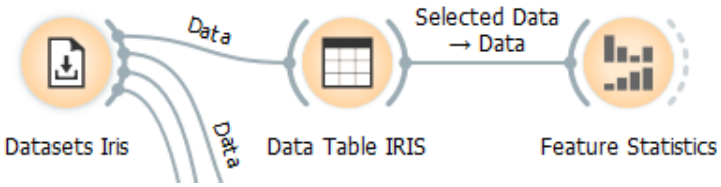
Data lze přidat do workflow pomocí uzlu **File** (libovolný soubor) nebo uzlu **Datasets** (cvičná data Orange). Následně lze obsah dat zobrazit formou tabulky pomocí uzlu **Data Table**. Textový popis uzlu lze uživatelsky měnit.

Příklad 1

Data *3_Iris.xlsx* nebo dataset *Iris.dat,* **workflow** *3_Iris.ows*

Základní popis dat bude v následujících kapitolách vysvětlen na datasetu Iris, který obsahuje údaje o velikosti okvětních a kališních lístků (Obr. 5) pro tři druhy kosatců: *iris-setosa*, *iris-versicolor*, *iris-virginica*. Soubor obsahuje celkem 150 záznamů, od každého druhu 50 jedinců. Tento dataset je i součástí cvičných datasetů instalace software Orange.

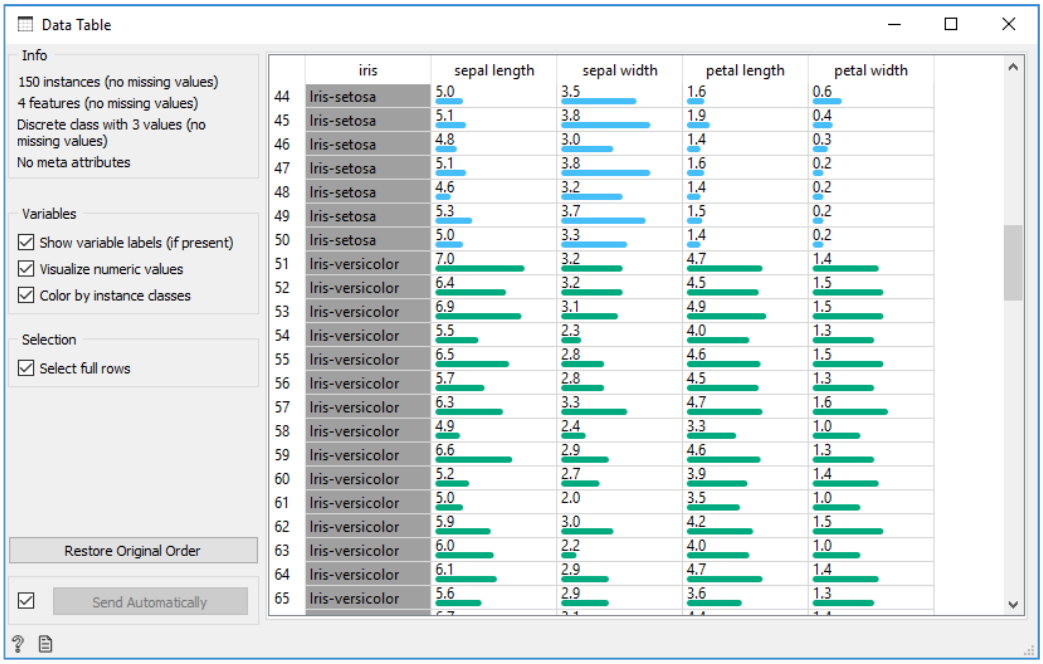




Obr. 5 Květ kosatce s rozměry okvětních a kališních lístků (vlevo) (Kedro 2020) *a workflow pro základní průzkum souboru 3_Iris.xlsx (vpravo)*

Uzel **Data Table** přehledně zobrazí data (Obr. 6). Lze zaškrtnout volbu *Visualize numeric values,* kdy délka barevné podtrhující čáry v tabulce znázorňuje relativní hodnotu čísla vůči ostatním hodnotám téhož pole. Klikem v záhlaví sloupce lze setřídit data vzestupně, či sestupně podle numerických nebo kategoriálních hodnot.

Pro zpracování v dalším uzlu se musí vybrat řádky v tabulce. Klik vlevo nahoře v rohu tabulky vybírá celou tabulku.



	iris	sepal length	sepal width	petal length	petal width
44	Iris-setosa	5.0	3.5	1.6	0.6
45	Iris-setosa	5.1	3.8	1.9	0.4
46	Iris-setosa	4.8	3.0	1.4	0.3
47	Iris-setosa	5.1	3.8	1.6	0.2
48	Iris-setosa	4.6	3.2	1.4	0.2
49	Iris-setosa	5.3	3.7	1.5	0.2
50	Iris-setosa	5.0	3.3	1.4	0.2
51	Iris-versicolor	7.0	3.2	4.7	1.4
52	Iris-versicolor	6.4	3.2	4.5	1.5
53	Iris-versicolor	6.9	3.1	4.9	1.5
54	Iris-versicolor	5.5	2.3	4.0	1.3
55	Iris-versicolor	6.5	2.8	4.6	1.5
56	Iris-versicolor	5.7	2.8	4.5	1.3
57	Iris-versicolor	6.3	3.3	4.7	1.6
58	Iris-versicolor	4.9	2.4	3.3	1.0
59	Iris-versicolor	6.6	2.9	4.6	1.3
60	Iris-versicolor	5.2	2.7	3.9	1.4
61	Iris-versicolor	5.0	2.0	3.5	1.0
62	Iris-versicolor	5.9	3.0	4.2	1.5
63	Iris-versicolor	6.0	2.2	4.0	1.0
64	Iris-versicolor	6.1	2.9	4.7	1.4
65	Iris-versicolor	5.6	2.9	3.6	1.3

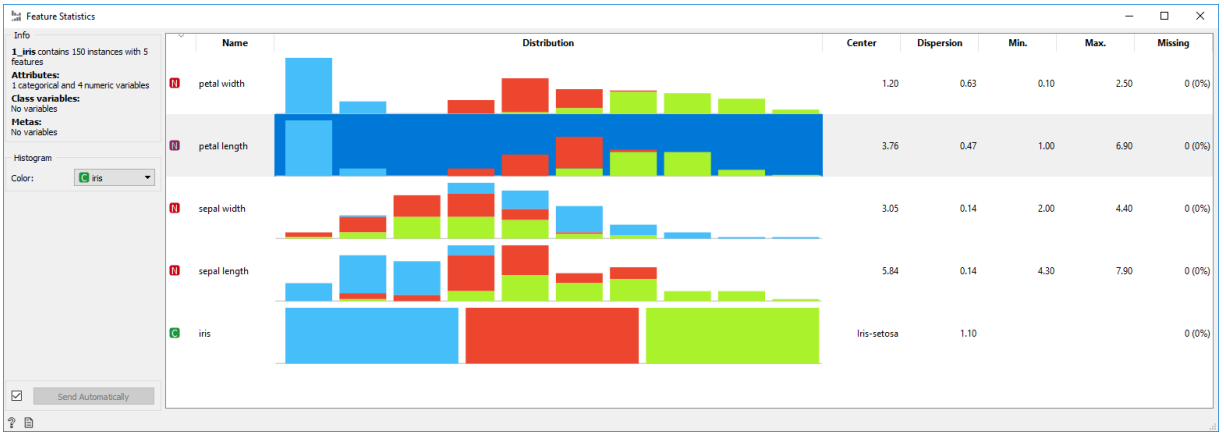
Obr. 6 Zobrazení tabulky dat pomocí uzlu Data Table

Uzel **Feature Statistics** zobrazí přehledně základní statistiky a distribuci hodnot (histogram) jednotlivých atributů (Obr. 7). Pokud data obsahují i kategorii záznamu, zde druh kosatce, tak jsou různé kategorie zobrazeny různou barvou i v grafu distribuce hodnot, sloupec *Distribution*. Atribut s údajem kategorie lze vybrat volbou *Histogram – Color*.

Sloupec *Center* pro kategoriální data zobrazí **modus** (nejčastěji se vyskytující hodnotu, tedy s nejvyšší relativní četností), v případě numerických dat je spočítána **průměrná hodnota**.

Sloupec *Dispersion* pro kategoriální data zobrazí **entropii** hodnot (jak moc jsou různorodé nebo stejnorodé hodnoty). Pro numerická data je spočítán **index disperze D.** Jedná se o normalizovanou míru rozptylu distribuce pravděpodobnosti; je to měřítko, které se používá ke kvantifikaci toho, zda je soubor pozorovaných výskytů seskupen nebo rozptýlen ve srovnání se standardním statistickým modelem. **Index disperze D** se spočítá jako podíl rozptylu (střední kvadratické odchylky) σ^2 a průměru μ dle vzorce (1).

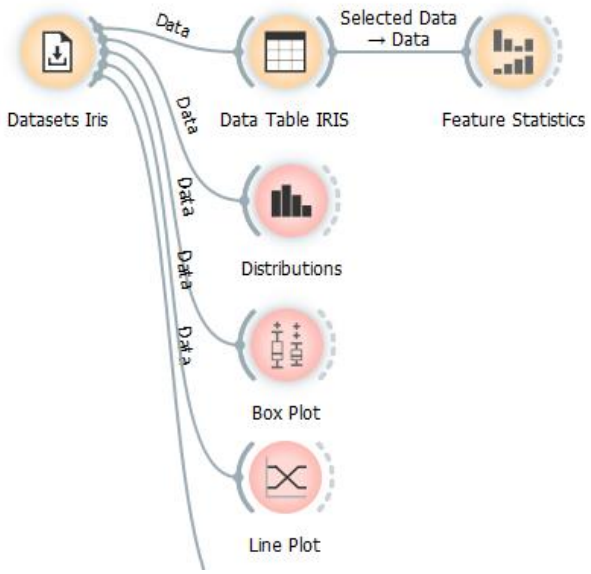
$$D = \frac{\sigma^2}{\mu} \qquad (1)$$



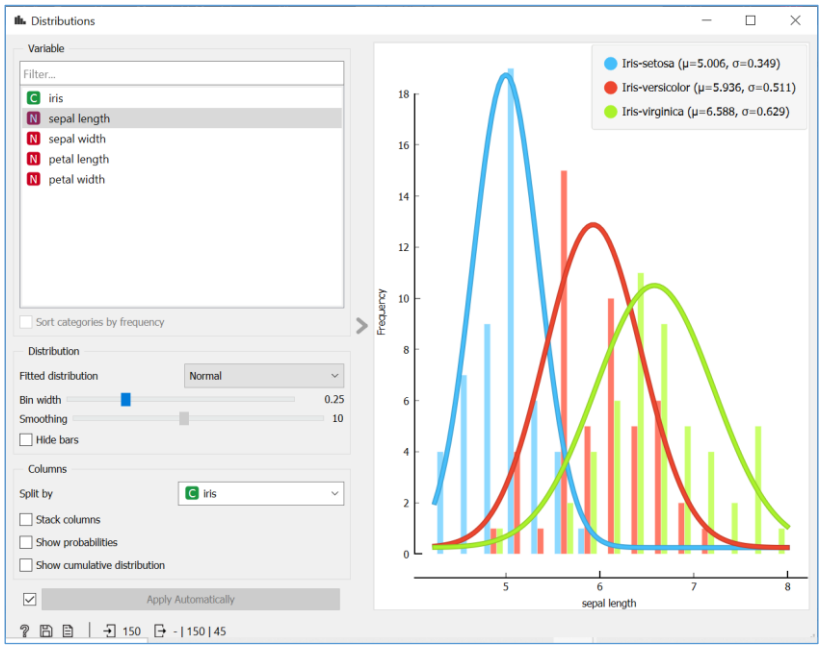
Obr. 7 Výsledky zobrazené pomocí uzlu Feature Statistics pro dataset 3_Iris.xlsx

> *Poznámka. Zkuste samostatně vyšetřit dataset heart-disease, kde je nominální údaj gender. Disperze tohoto údaje (v tomto případě entropie) je 0,62 a Central (zde nejčastější hodnota) je male. Znamená to, že v datasetu převládají muži nad ženami, mužů je tam 62 %.*

Přidejte do workflow další uzel **Distributions** (Obr. 8), který vykreslí distribuci hodnot atributů v grafu, který se nazývá histogram (Obr. 9). Uzel *Distributions* má červenou barvu, na rozdíl od dosud použitých oranžových uzlů ze sekce Data. Tento uzel je ze skupiny uzlů *Visualize* určenou pro zobrazení dat.



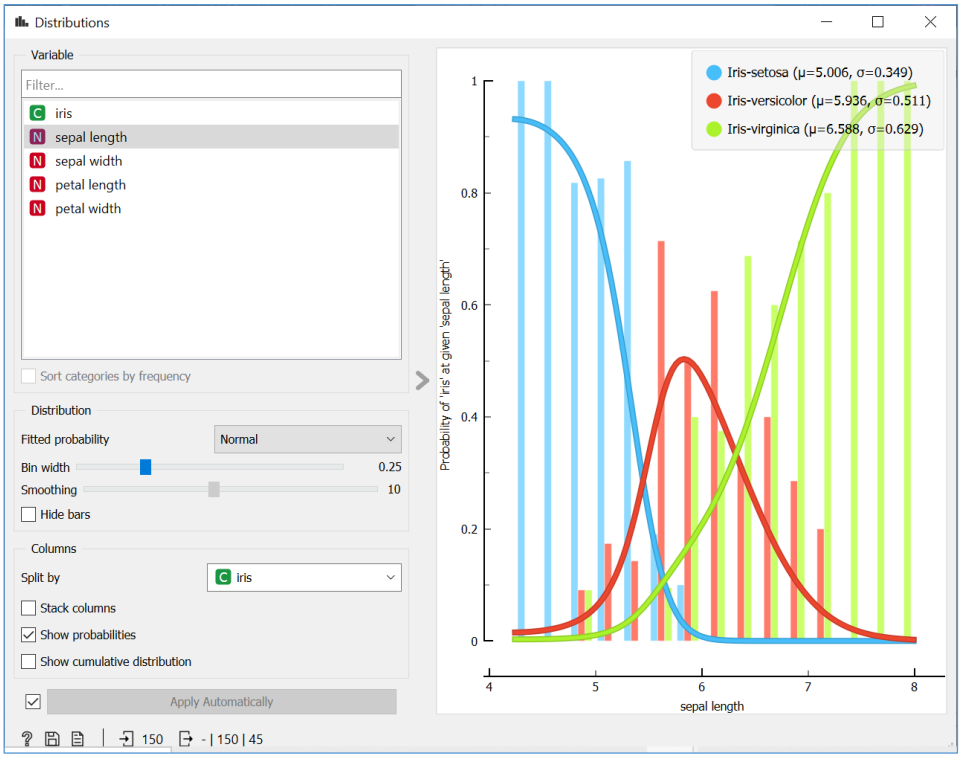
Obr. 8 Uzel Feature Statistics, Distributions, Box Plot a Line Plot ve workflow



Obr. 9 Zobrazení dat pomocí uzlu Distributions

Při odebrání zatržítka *Hide bars* se zobrazí frekvence hodnot ve formě sloupcového grafu v intervalech hodnot (Obr. 10). Posuvníkem *Bin width* se nastaví šířka intervalů – bins. Rozložení hodnot lze porovnávat s různými distribucemi pomocí výběru pod *Fitted Distribution*. Na obrázku je vybráno normální rozdělení.

Navíc lze zobrazit i rozložení pravděpodobnosti při volbě *Show probabilities*. Vyzkoušejte i volbu *Stack columns* a *Show cumulative distribution* pro různé rozměry okvětních a kališních lístků a sledujte rozdíly v jednotlivých druzích kosatců.



Obr. 10 Zobrazení rozložení pravděpodobnosti atributu sepal lenght pro jednotlivé druhy kosatců

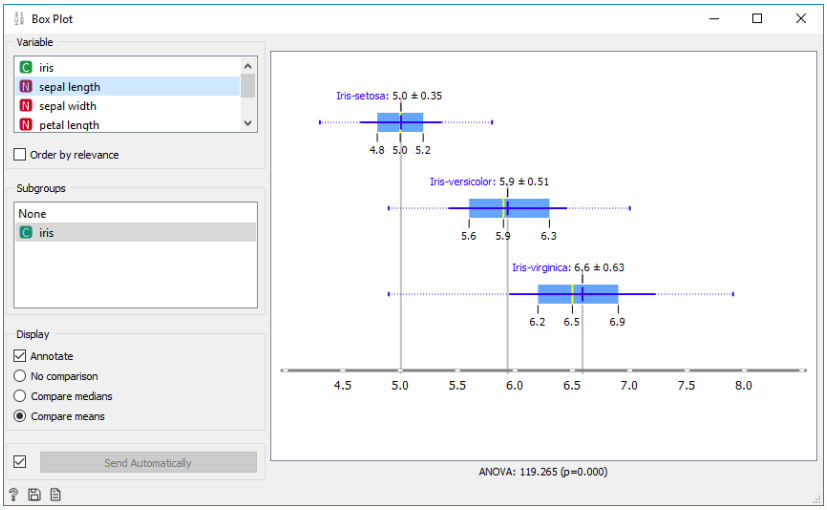
3.2 Průzkum dat pomocí krabicového grafu

Další seznámení s daty Iris lze povést pomocí krabicového grafu. Uzel **Box Plot** zobrazí krabicový graf celého souboru (volba *Subgroups* None) nebo lze zobrazit postupně boxploty všech atributů. Zobrazený boxplot je v Orange netradičně otočený o 90°.

Krabicový graf – boxplot znázorňuje spočítané údaje takto:

- Modré obdélníky (krabice) reprezentují oblast mezi **prvním (25 %) a třetím (75 %) kvartilem**. Tenké vodorovné modré čáry (vousy) reprezentují variabilitu dat pod prvním a nad třetím kvartilem.
- **Mean, průměr je tmavě modrá** svislá čára.
- **Median je žlutá** svislá čárka uvnitř modrého obdélníku.
- **Outlier** (odlehlá hodnota) je spojena tečkovanou vodorovnou čárou, která navazuje na vousy.

Změnou v okně *Subgroups* lze z volby *None* na *iris* zobrazit srovnávací boxplot jednoho atributu pro tři druhy irisů. Volbou *Compare median, Compare means* lze porovnat mediány a průměry, jsou vykresleny svislé šedé čáry.

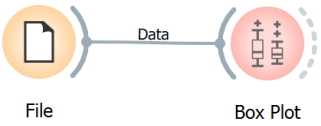


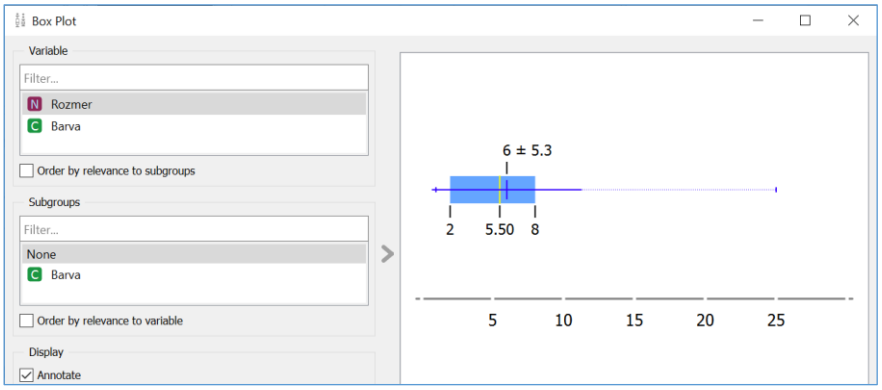
Obr. 11 Zobrazení krabicových grafů pomocí uzlu Box Plot s porovnáním mediánů a průměrů

Příklad 2

Data *3_Boxplot.xlsx,* **workflow** *3_OdlehlaHodnota.ows*

Boxplot lze také dobře využít pro rychlé zjištění odlehlých hodnot (outliers). Vykreslete boxplot i pro následující cvičná data. Zde je evidentní jedna odlehlá hodnota 25. Patrně se jedná o chybu v datech.





Obr. 12 Zjištění odlehlé hodnoty pomocí krabicového grafu

3.3 Detekce odlehlých hodnot

Detekce odlehlých hodnot je proces hledání instancí v souboru dat, které se liší od ostatních. Obecně lze detekci odlehlých hodnot rozdělit na řízenou a neřízenou. Řízená detekce vyžaduje soubor dat s označenými instancemi, který obsahují příznak, zda je záznam „normální“ nebo „abnormální“. Při neřízené detekci příznak chybí. Předpokládá se, že většina instancí v neoznačeném souboru dat je „normální“, a hledá instance, které se od „normálních“ datových bodů liší (Tan 2021).

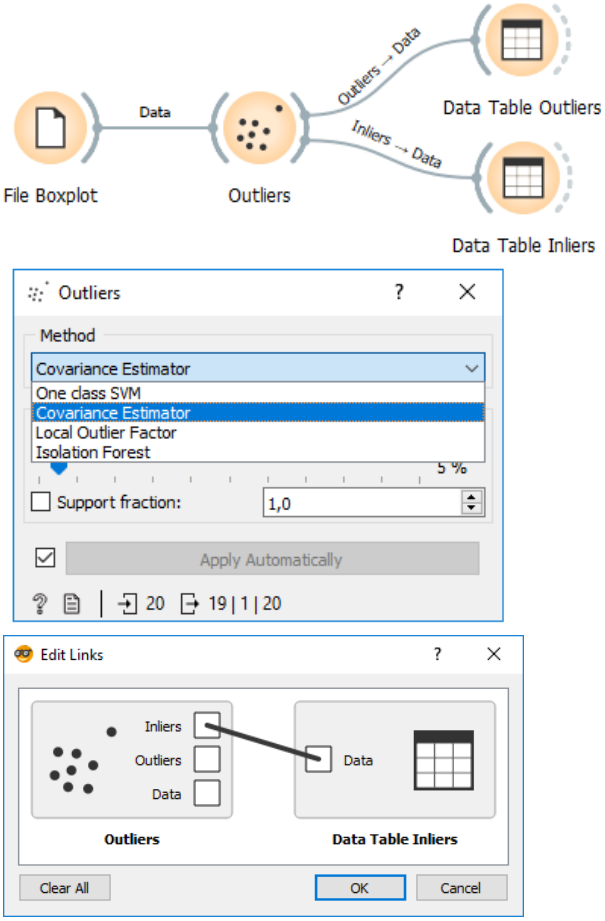
Příklad 3

Data *3_Boxplot.xlsx,* **workflow** *3_Outliers.ows*

Ve vstupních datech se nastaví atribut *Barva* na *skip*. Bude se zjišťovat odlehlá hodnota pro atribut *Rozmer*.

Uzel **Outliers** nabízí čtyři metody detekce: *One Class SVM, Covariance Estimator, Local Outlier Factor a Isolation Forest*. *One-class SVM* s nelineárním kernelem (RBF) funguje dobře i u negausovských rozdělení, zatímco *Covariance Estimator* funguje pouze u dat s gausovským rozdělením. Jedním z efektivních způsobů, jak provádět detekci odlehlých hodnot na středně velkých dimenzionálních souborech dat, je použití algoritmu *Local Outlier Factor*. Tento algoritmus vypočítá skóre odrážející stupeň abnormality pozorování. Měří lokální odchylku hustoty daného datového bodu vzhledem k jeho sousedům. Další metoda vhodná pro multidimenzionální data, je použití náhodných lesů (Isolation Forest) (Orange Data Mining 2021b). Více o principech metod je v kapitole 6.

Výstupem z uzlu **Outliers** je jednak tabulka instancí, které jsou odlehlé hodnoty (outliers) a potom tabulka instancí, které nejsou odlehlé (inliers). Pokud jsou výstupem všechna data, je v nich přidán nový atribut *Outliers* s hodnotami *Yes/No*. Pro jednoduchá data *3_Boxplot.xls* experimentujte s nastavením parametrů metod tak, aby byla identifikována instance s odlehlou hodnotou *Rozmer 25*.



Obr. 13 Workflow s uzlem Outliers a nastavení spojných čar v dialogu Edit Links pro specifikaci výstupu

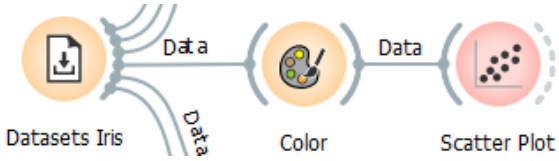
Samostatně zjistěte odlehlé hodnoty v souboru 10_Rail_quartal_EUROSTAT.xlsx, který obsahuje počty přepravených osob na železnici v Evropě z databáze EUROSTAT. Jako odlehlé hodnoty budou identifikované počty pasažérů v roce 2020 a 2021 v době pandemie nemoci covid-19.

3.4 Rozptylový graf a uzel Color

Uzel **Scatter Plot** (rozptylový graf) zobrazuje objekty jako body v pravoúhlých souřadnicích grafu. Každý objekt svou pozicí bodu vyjadřuje hodnotu dvou jeho vybraných popisných atributů vynesených na příslušných osách. Scatter plot se používá k vyšetřování vztahu dvou proměnných.

Příklad 4

Data *3_Iris.xlsx* nebo dataset *Iris.dat,* **workflow** *3_Iris.ows*

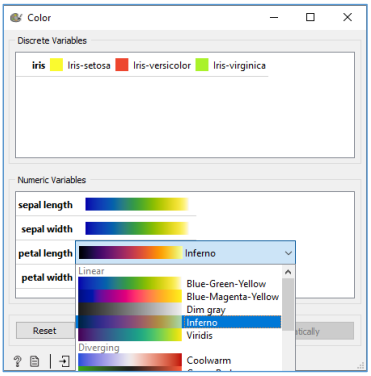


Obr. 14 Workflow s uzlem Color a Scatter Plot

V dialogu uzlu Scatter Plot můžeme pro *Axis x* a *Axis y* vybrat libovolné dvě kombinace atributů popisující rozměry kališních (sepal) a okvětních (petal) lístků. Pokud data obsahují kategorii, zde druh kosatce, tak lze body obarvit podle tohoto atributu nastavením v kolonce **Color**. Stiskem tlačítka *Find Informative Projections* se určí pořadí nejlepších skóre kombinací dvou vstupních atributů, které nejpřesněji určují druh kosatce. Pro klasifikaci kosatců je to *petal length* a *petal width*. Z grafu na Obr. 15 je dobře čitelné, že druhy kosatců se výrazně odlišují podle *petal lenght* a *petal width*, částečně podle *septal width*. Navíc je zřejmé, že druh Iris-setosa má výrazně menší okvětní lístky než další dva druhy kosatců. Některé květy Iris-versicolor a Iris-virginica mají obdobné rozměry.

Obr. 15 Rozptylový graf pro dataset Iris

Uzel **Color** lze předřadit před uzel Scatter Plot (Obr. 14), pokud chceme nastavit vlastní barvy pro znázornění bodů v grafu místo základní modré, červené a zelené barvy. V uzlu **Color** lze individuálně nastavit barvu jednotlivým druhům kosatců. Pro numerické veličiny lze nastavit předvolenou barevnou stupnici (Obr. 16). Uzel Color je použitelný v kombinaci s některými dalšími uzly v Orange.



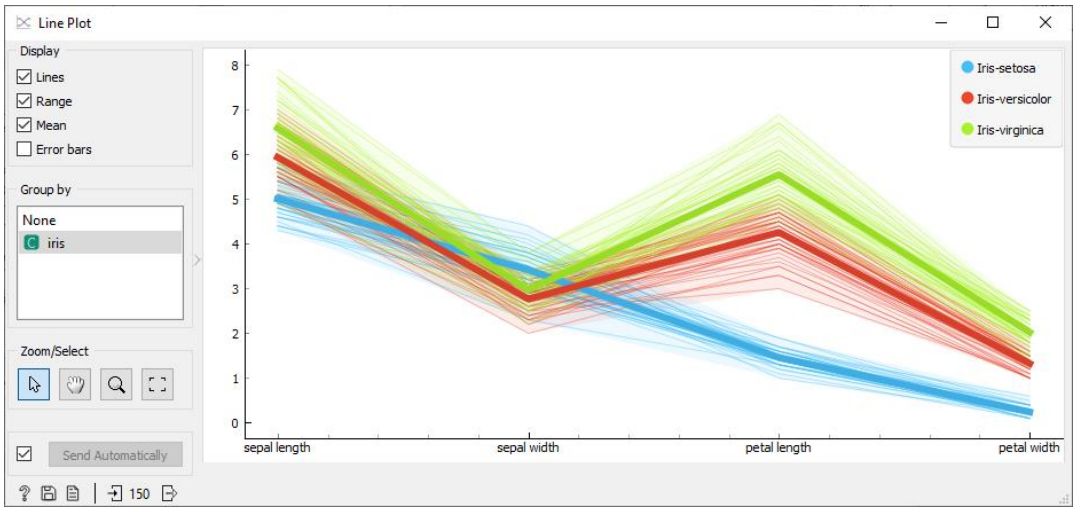
Obr. 16 Nastavení dialogu Color pro změnu přiřazení barev a barevných stupnic

3.5 Profilový graf

K seznámení s daty kromě výše zmíněných uzlů *Feature Statistics, Distribution*, *Box Plot* a *Scatter Plot* také dobře poslouží i liniový graf **Line Plot**, někdy označovaný také jako *profilový graf*, či *graf v paralelních osách* (*Parallel Coordinate Plot*). Tento specifický graf má na vodorovné ose X jednotlivé atributy, na svislé ose Y číselné hodnoty jednotlivých atributů. Na ose Y je dobře viditelný rozptyl hodnot. Graf umožňuje lépe porovnat hodnoty proměnných napříč datovým souborem při srovnatelném měřítku.

Příklad 5

Data *3_Iris.xlsx* nebo dataset *Iris.dat,* **workflow** *3_Iris.ows*



Obr. 17 Paralelní graf čtyř atributů datasetu Iris

Profilový graf ukáže rozdíly v klasifikačních třídách. Z grafu dobře vyčteme, že druhy kosatců se výrazně odlišují podle *petal lenght* a *petal width*, částečně podle *septal width*. Naopak kosatce nelze rozlišit úplně dobře podle *sepal lenght*. Uzel *Line Plot* umožňuje zobrazit kromě linií jednotlivých instancí (volba *Line*) také podbarvit rozsah volbou *Range* a silnou linií zdůraznit i průměrné hodnoty jednotlivých atributů (volba *Mean*). V případě, že data obsahují kategorii (zde druh irisu), lze data podle něj obarvit výběrem názvu kategorie v okně *Group by*.

> *Upozornění 1: Nezáleží na pořadí atributů na ose X a dokonce změna pořadí může vést ke větší názornosti. Pořadí sloupců je ale nutné upravit ve zdrojových datech (nelze v rámci uzlu). Spojné čáry neznamenají spojitost průběhu, ale pouze spojují hodnoty jedné instance v datech (jednoho řádku).*

> *Upozornění 2: V případě velice rozdílného rozsahu hodnot jednotlivých atributů (třeba 1 až 10 a 1 až 1000) nemusí být profilový graf až tak názorný, neboť měřítko a rozsah svislé osy je společný pro všechny atributy. Potom je třeba provést nejprve standardizaci dat. Naopak u datasetu Iris je rozsah hodnot čtyř atributů obdobný, tedy 0,14 až 8 mm a je dobře viditelný rozsah hodnot všech čtyř atributů.*

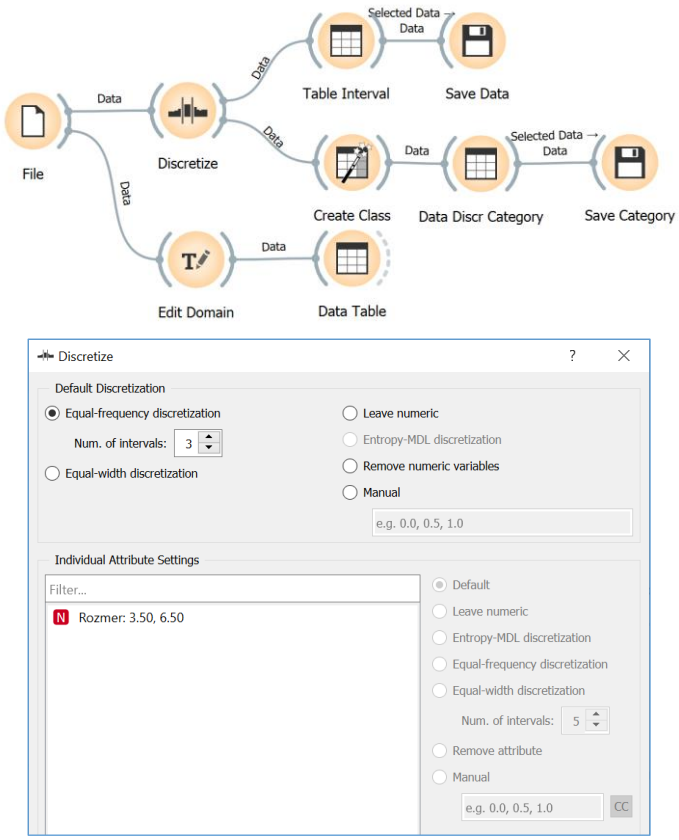
3.6 Diskretizace do intervalů

Mezi jednu z úloh předzpracování dat patří převod hodnot atributu ze spojité domény čísel (reálná čísla, resp. i přirozená nebo celá čísla) do diskrétních kategorií. K tomu slouží uzel **Discretize**. Při převodu spojité veličiny na kategorii může dojít k nechtěnému skrytí závislostí v datech. Je proto nutná dobrá znalost dat.

Příklad 6

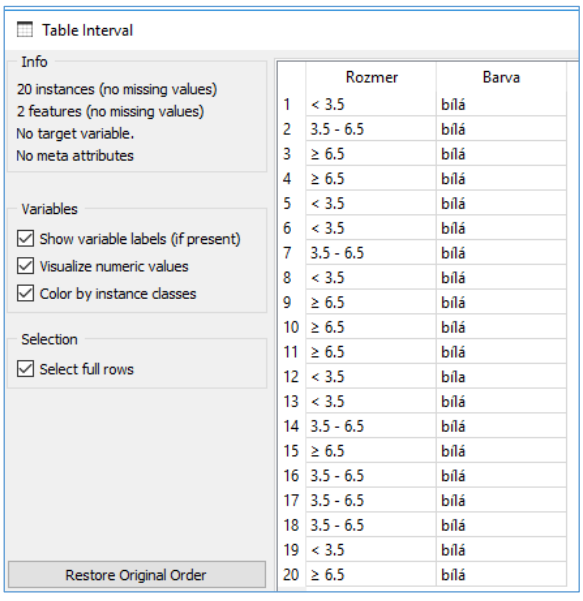
Data *3_Boxplot.xlsx*, **workflow** *3_DiscretCategory.ows*

Číselné hodnoty atributu *Rozmer* převedeme na tři intervaly (tři kategorie). V uzlu **Discretize** zvolíme volbu *Equal-frequency discretization* – podle stejné četnosti do ručně zvoleného počtu intervalů. Zadejte ve volbě *Num. of intervals* hodnotu 3. Hranice intervalů ukazuje automaticky okno *Settings* (pro tři intervaly jsou hranice 3,5 a 6,5). Výsledek diskretizace je viditelný v uzlu datové tabulky **Data Table** (ve workflow pojmenovaný Table Interval).



Obr. 18 Workflow a dialog uzlu Discretize s možnostmi nastavení intervalů

V datové tabulce je viditelné, že původní konkrétní číselné hodnoty atributu *Rozmer* jsou přepsány příslušným intervalem hodnot (Obr. 19).



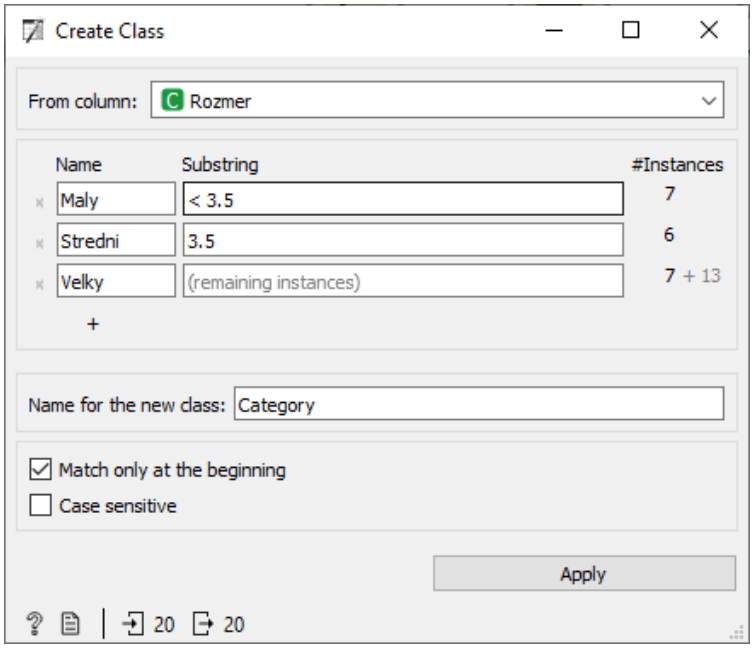
	Rozmer	Barva
1	< 3.5	bílá
2	3.5 - 6.5	bílá
3	≥ 6.5	bílá
4	≥ 6.5	bílá
5	< 3.5	bílá
6	< 3.5	bílá
7	3.5 - 6.5	bílá
8	< 3.5	bílá
9	≥ 6.5	bílá
10	≥ 6.5	bílá
11	≥ 6.5	bílá
12	< 3.5	bílá
13	< 3.5	bílá
14	3.5 - 6.5	bílá
15	≥ 6.5	bílá
16	3.5 - 6.5	bílá
17	3.5 - 6.5	bílá
18	3.5 - 6.5	bílá
19	< 3.5	bílá
20	≥ 6.5	bílá

Obr. 19 Výsledek operace Discretize

Pro další zpracování a analýzu je vhodnější, pokud je výsledek reprezentován kódem kategorie nebo textem. To lze vyřešit přidáním nového atributu s kategorií prostřednictvím uzlu **Create Class.** V dialogu se pomocí zadání podmínek převede ze zdrojového atributu do nové kategorie v nově pojmenovaném atributu. Do kolonky *From column* zadejte zdrojový atribut, zde atribut *Rozmer*, který obsahuje výsledek diskretizace. Nové tři kategorie budou mít označení ve sloupci *Name* hodnoty *Maly, Stredni, Velky* a k nim zadejte podmínky podle obrázku (další kategorie se přidávají přes tlačítko +). Název nového atributu se zadá do kolonky *Name for the new class* zde zadáno *Category* (Obr. 20).

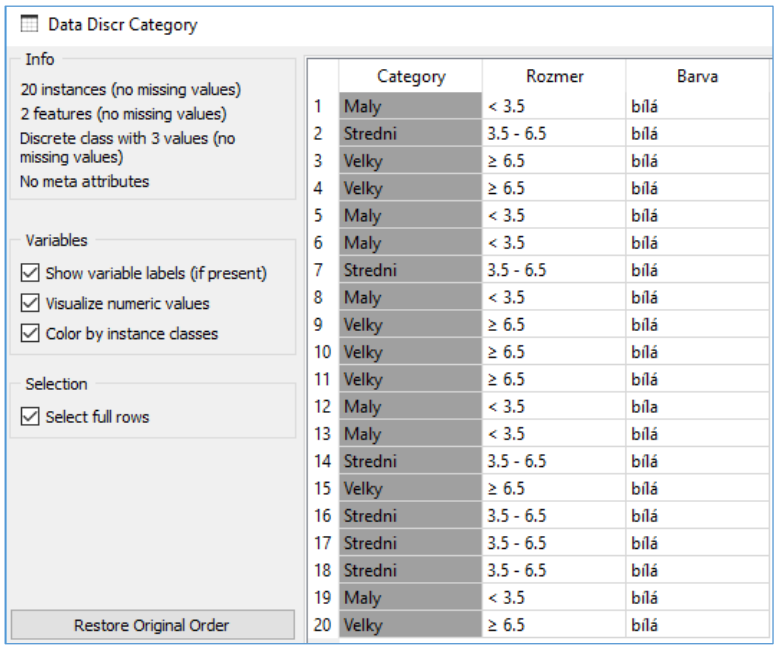
Všimněte si sloupce vpravo *#Instances*, který informuje, kolik záznamů odpovídá právě zadané podmínce. Lze si tak průběžně kontrolovat správnost, zda jsou pokryty všechny záznamy. Celkový počet vstupních a výstupních řádků je uveden v dolní části dialogového okna (Obr. 20).

> *Pozor Orange používá desetinnou tečku místo čárky. V popisném textu bude používána desetinná čárka v souladu s pravidly českého jazyka.*



Obr. 20 Nastavení dialogu Create Class

Výsledek vytvoření nového atributu a vyplněných hodnot kategorií je viditelný v datové tabulce vedle původní hodnoty před diskretizací (Obr. 21).



	Category	Rozmer	Barva
1	Maly	< 3.5	bílá
2	Stredni	3.5 - 6.5	bílá
3	Velky	≥ 6.5	bílá
4	Velky	≥ 6.5	bílá
5	Maly	< 3.5	bílá
6	Maly	< 3.5	bílá
7	Stredni	3.5 - 6.5	bílá
8	Maly	< 3.5	bílá
9	Velky	≥ 6.5	bílá
10	Velky	≥ 6.5	bílá
11	Velky	≥ 6.5	bílá
12	Maly	< 3.5	bíla
13	Maly	< 3.5	bílá
14	Stredni	3.5 - 6.5	bílá
15	Velky	≥ 6.5	bílá
16	Stredni	3.5 - 6.5	bílá
17	Stredni	3.5 - 6.5	bílá
18	Stredni	3.5 - 6.5	bílá
19	Maly	< 3.5	bílá
20	Velky	≥ 6.5	bílá

Obr. 21 Výsledek diskretizace s novým atributem Category a hodnotami třech kategorií

Výsledek úpravy dat lze uložit pomocí uzlu **Save Data** (vyberte formát CSV nebo XLS).

> *Vyzkoušejte i diskretizaci volbu Equal-width discretization, která nastaví stejnou šířku intervalu bez ohledu na četnost hodnot. Pozor, data obsahují odlehlou hodnotu, která významně ovlivní šířku intervalů. Důsledkem je málo instancí v poslední kategorii.*

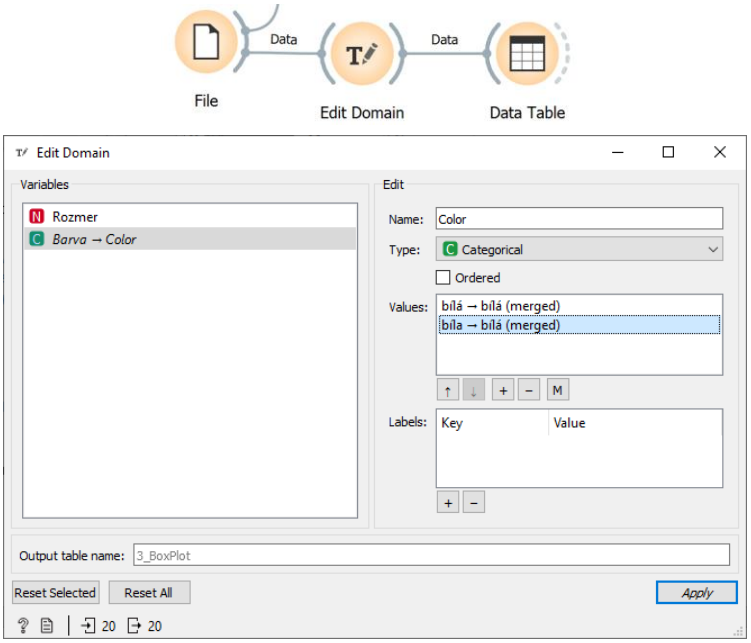
3.7 Oprava chybných dat a sjednocení kategorií

V datech se může vyskytovat překlep v hodnotě textového atributu jako kategorie. Automatickou opravu chybných hodnot lze provést pomocí uzlu **Edit Domain.** Tento uzel slouží i k přejmenování atributu.

Příklad 7

Data *3_Boxplot.xlsx,* **workflow** *3_DiscretCategory.ows*

Cvičná data *3_Boxplot.xlsx* obsahují kategoriální proměnnou *Barva* s hodnotou *bílá*. Data obsahují i chybně zapsanou hodnotu (překlep), a to slovo *bíla*. Uzel **Edit Domain** obsahuje možnost vybrat atribut v levém okně *Variables* (zde Barva), zadat nový název atributu v kolonce *Name* (zde *Color)* a v okně *Values* zadat nové hodnoty, zde tedy nahradit slovo bíla za slovo bílá (Obr. 22). Původní hodnota je nahrazena správnou hodnotou.



Obr. 22 Nastavení editace hodnot domény pomocí uzlu Edit Domain

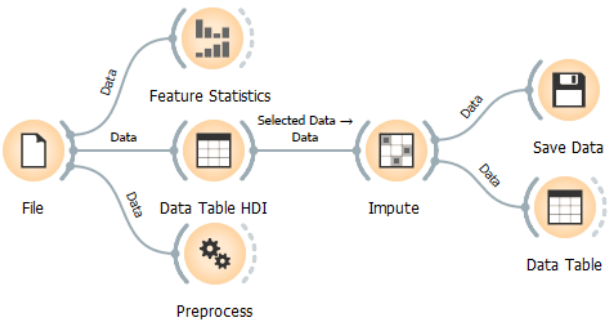
3.8 Imputace hodnot

Doplnění neboli imputace chybějících hodnot lze provést pomocí uzlu **Impute**. Nejprve je ale nutné zjistit, kolik záznamů, a v kterých atributech chybí. Doplnění chybějících dat se musí provádět uvážlivě, protože se tak můžou výrazně pozměnit data, jejich průměr, medián nebo četnosti jednotlivých hodnot.

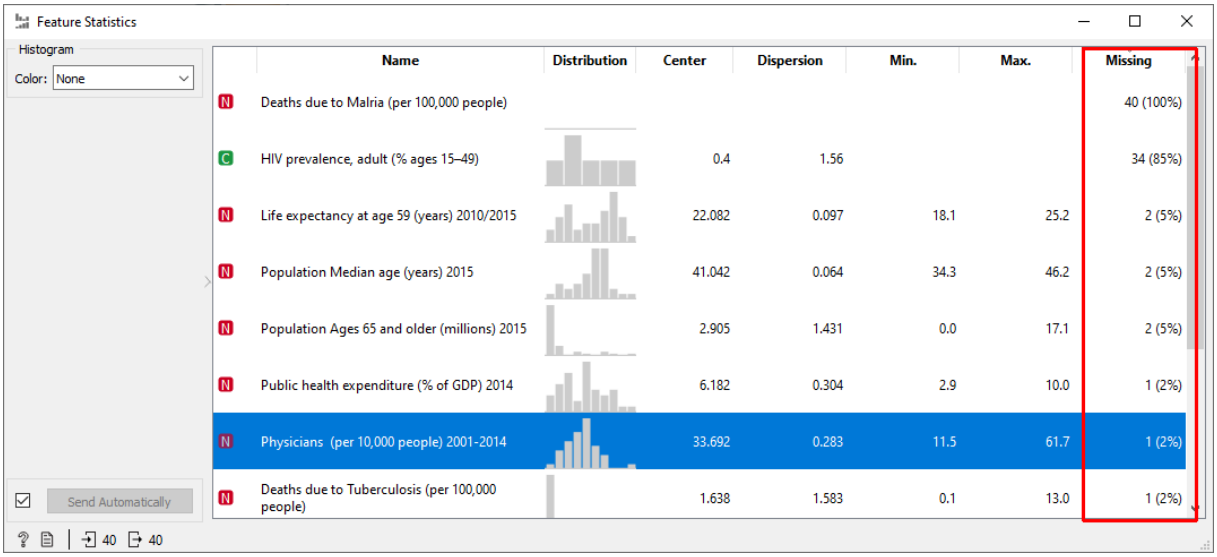
Příklad 8

Data *3_HDI_Europe.xlsx,* **workflow** *3_Impute.ows*

Cvičná data obsahují vybrané statistické ukazatele evropských zemí. Některé údaje chybí. To dobře zjistíme jednak pomocí uzlu Data Table nebo již známým uzlem **Feature Statistics**, kde poslední sloupec **Missing** zobrazuje, kolik údajů pro daný atribut chybí, a kolik to tvoří procent. Setřiďte řádky podle hodnoty sloupce Missing. Vidíme, že chybí jeden záznam o počtu lékařů na 100 tis. obyvatel u Lichtenštejnska, který se pohybuje jinak mezi hodnotami 11,5 až 61,7. Hodnotu atributu *Physicians (per 10,000 people*) doplníme (Obr. 24).



Obr. 23 Workflow s uzlem Impute a Preprocess pro imputaci hodnot

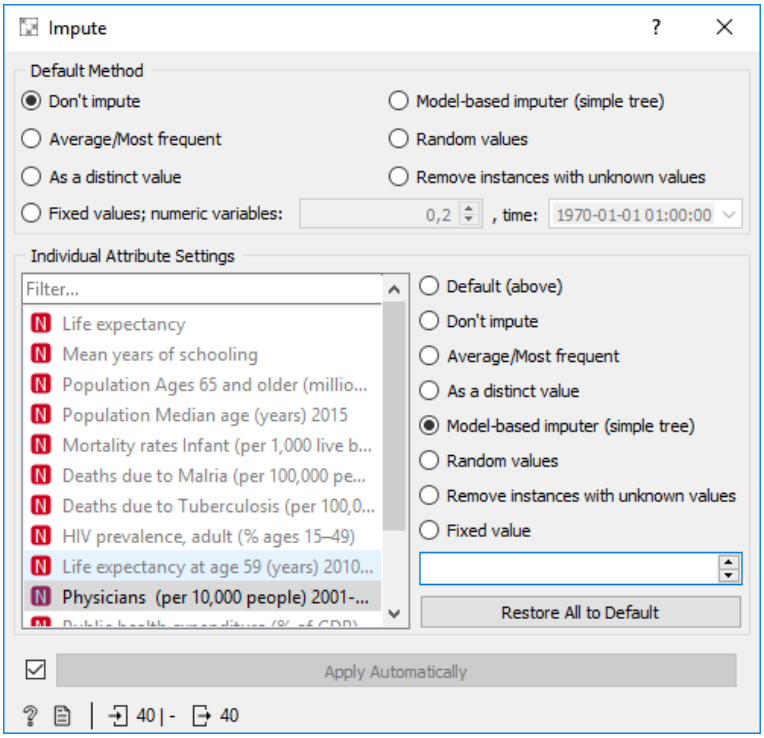


Obr. 24 Zjištění chybějících hodnot pomocí uzlu Feature Statistics

Imputaci provedeme pomocí uzlu **Impute**. Lze zvolit jednu výchozí metodu pro všechny atributy a tu pak nastavit či změnit individuálním nastavením pro každý atribut v spodním okně *Individual Attribute Setting* (Obr. 25). Nejlépe je zvolit jako výchozí hodnotu *Don`t impute* pro všechny atributy a pak detailně nastavit každý atribut. V individuálních volbách se vedle názvu atributu vypisují nastavené individuální volby (někdy to není vidět z důvodu dlouhých názvů atributů).

Nabízí se několik metod imputace:

- Volba *Average/Most frequent* dopočítá průměrnou hodnotu nebo nejčastější kategorii.
- Volba *Random values* doplní náhodné hodnoty.
- Volba *Fixed value* umožňuje zadat konkrétní hodnotu. Pokud zvolíme pesimistický odhad, tak minimální hodnota celého datasetu pro počet lékařů je 11,5. Lze také chybějící hodnotu pouze nahradit označením kódem např. 9999 pomocí právě volby *Fixed value*.



Obr. 25 Nastaveni imputace hodnot pomocí uzlu Impute

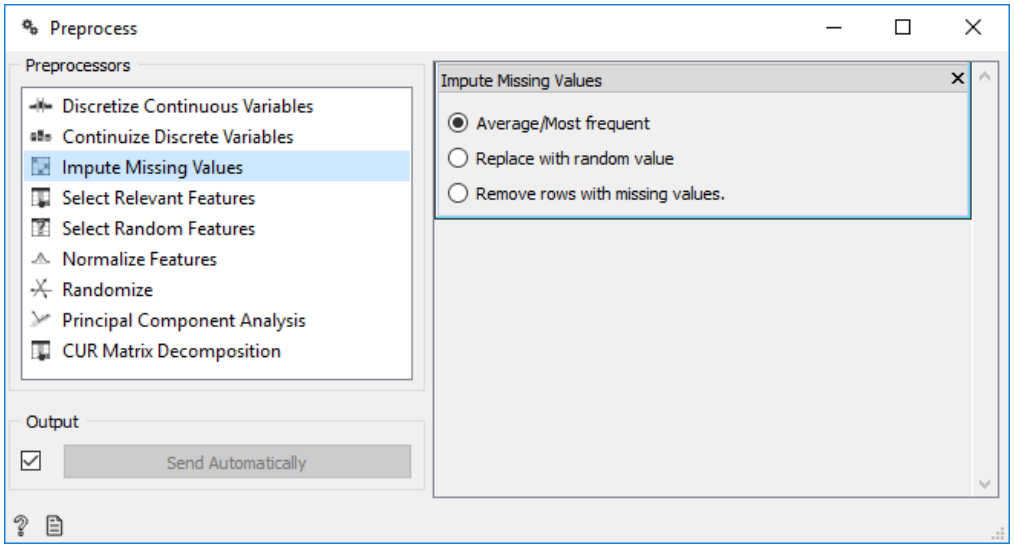
- Volba ***Model-based imputer (simple tree)*** vytvoří model pro předpovídání chybějící hodnoty na základě hodnot ostatních atributů. Pro každý atribut je vytvořen samostatný model. Výchozí model je **1-NN** (1-nearest neighbor klasifikátor), který přebírá hodnotu z nejpodobnějšího záznamu, někdy nazývaná imputace *hot deck* (Joenssen a Bankhofer 2012). Model 1-NN umí předpovědět jak chybějící diskrétní (kategoriální) hodnoty tak spojité hodnoty. Pro údaj o počtu lékařů vypočítá tato metoda hodnotu 33,48.

Experimentujte s různými metodami. Dopočítaná data lze následně uložit uzlem *Save Data* (Obr. 23).

Více operací **předzpracování** se nachází v uzlu **Preprocess**. Je tu také možnost provést základní imputaci hodnot.

Volba **Impute Missing Values** pro imputaci zde nabízí jen tři možnosti: průměr, náhodné číslo a smazání záznamů s chybějícími atributy (tato úprava se v datech často nedoporučuje a neprovádí z důvodu ztráty celých záznamů).

V uzlu Preprocess lze nastavit několik různých úprav dat. Úpravy se přidávají jako samostatná okna do pravého panelu. Potom se úpravy provedou naráz v rámci jednoho uzlu (Obr. 26).



Obr. 26 Možnosti imputace chybějících hodnot pomocí uzlu Preprocess

3.9 Normalizace a standardizace

Normalizaci a standardizaci dat je nutné provést před další analýzou dat, zejména shlukováním.

Standardizací reálného znaku rozumíme odstranění závislosti jednotlivých atributů na různých jednotkách měření (Šarmanová 2012). To znamená, pokud jsou jednotky měření hodnot atributů, které popisují jeden objekt, různé např. metr, kilogram, procenta %, km^2, roky, ceny, bez rozměru atd., je nutná standardizace. Metody založené na výpočtu vzdálenosti (shlukování), použité bez předchozí standardizace, by dávaly výsledky zkreslené vlivem rozdílných jednotek. Dominantní atributy by více ovlivnily shlukování než jiné, které by je ovlivnily málo. Standardizace přepočte hodnoty atributů tak, aby byly souměřitelné (Petr 2014b). Obecný termín *škálování* vystihuje, že operace standardizace se týká jak jednotek veličin, tak i počátku stupnice (Meloun et al. 2012).

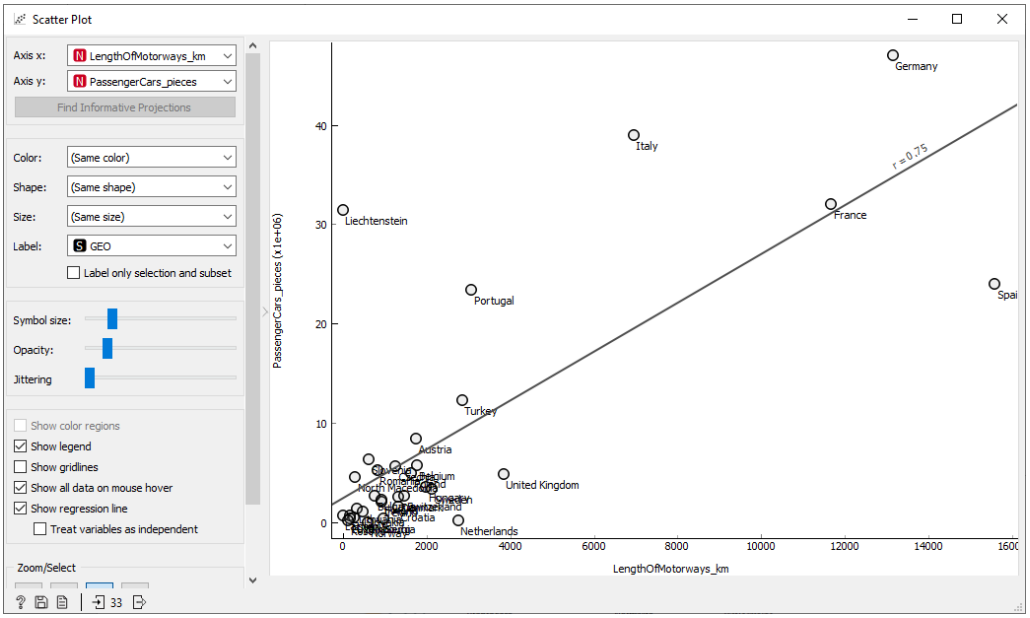
Normalizace odstraňuje u reálnými atributů závislosti dat na velikosti objektů. Např. různý věk dítěte má vliv na jeho výšku, váhu a velikost chodidla. Každý objekt představuje vektor určený množinou svých atributů. Tyto vektory mohou mít různou normu. Různé normy mohou ovlivňovat vyhodnocení podobnosti (shlukování) objektů. Normalizace zajistí, že všechny objekty mají stejnou normu, nejlépe jednotkovou a tudíž stejnou váhu při analýze.

V uzlu **Preprocess** se v Orange nachází jak volby pro normalizaci, tak standardizaci dat.

Příklad 9

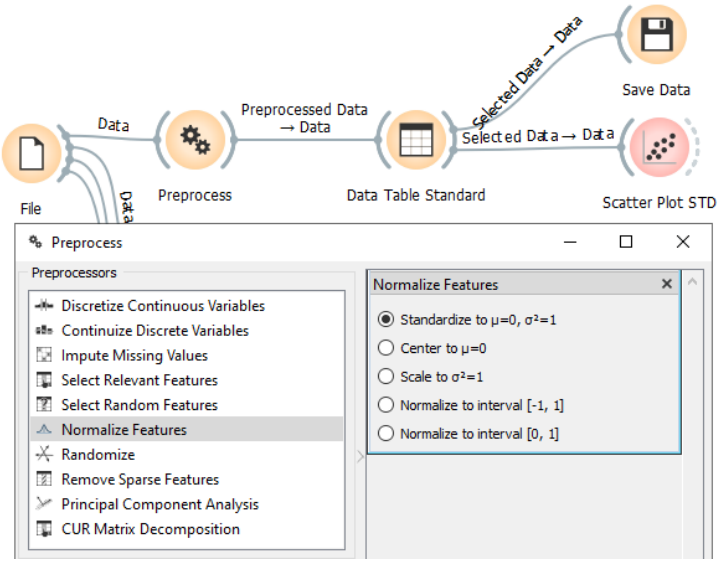
Data *3_EU_Transport2018.xlsx,* **workflow** *3_Normalize.ows*

Vstupní data obsahují údaje o délce dálnic a počtu osobních vozů v jednotlivých státech Evropy v roce 2018 z databáze Eurostat (Eurostat 2021). Měřící jednotky dat jsou tedy různé. Délka dálnic je měřena v kilometrech a počty vozů v kusech. Proto se provádí standardizace. Data nejprve prozkoumejte pomocí uzlů, které již znáte *(Data Table, Boxplot, Scatter Plot*). V Evropě jsou země s vysokou hodnotou délky dálnic (Německo, Španělsko) a naopak jsou země s malým počtem osobních automobilů (Chorvatsko, Kypr). Zajímavé je Lichtenštejnsko, které má malou délku dálnic, a naopak velký počet automobilů. To je viditelné na rozptylovém grafu na Obr. 27.

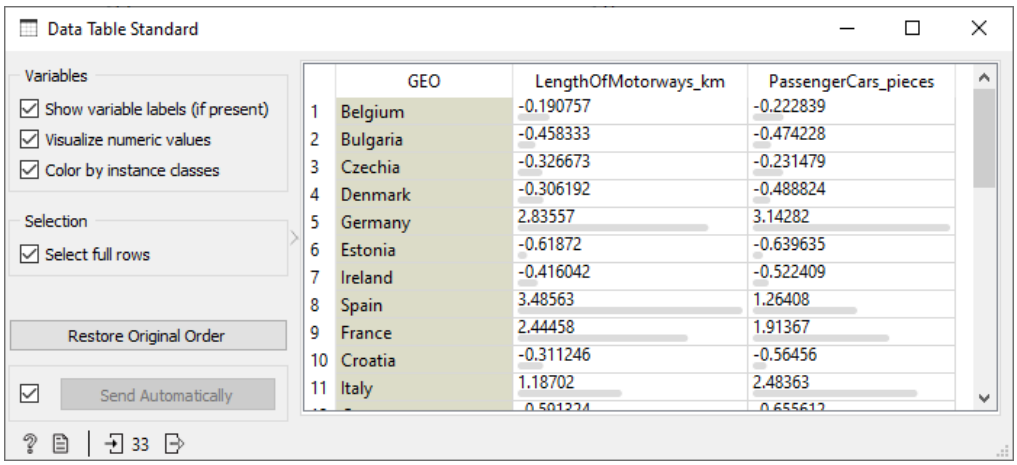


Obr. 27 Scatter Plot vstupních dat o dopravě z databáze Eurostat

V uzlu **Preprocess** zvolte *Normalize Features* a dále vyberte volbu **Standardize** (Obr. 28). Přepočítaná data zobrazte pomocí uzlu Data Table (Obr. 29). Data mají po standardizaci střední hodnotu rovnu nula.



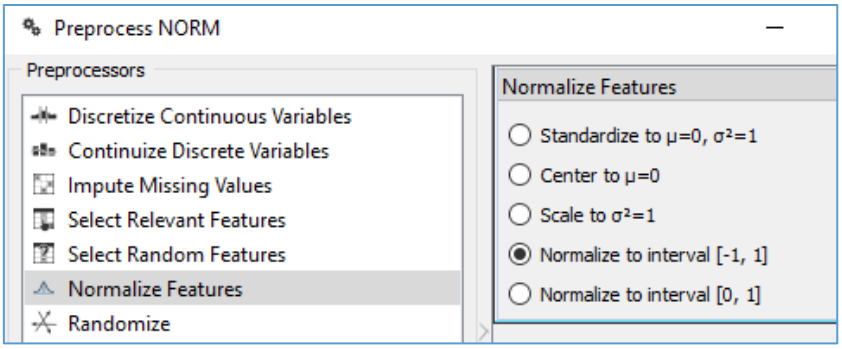
Obr. 28 Standardizace dat v uzlu Preprocess



	GEO	LengthOfMotorways_km	PassengerCars_pieces
1	Belgium	-0.190757	-0.222839
2	Bulgaria	-0.458333	-0.474228
3	Czechia	-0.326673	-0.231479
4	Denmark	-0.306192	-0.488824
5	Germany	2.83557	3.14282
6	Estonia	-0.61872	-0.639635
7	Ireland	-0.416042	-0.522409
8	Spain	3.48563	1.26408
9	France	2.44458	1.91367
10	Croatia	-0.311246	-0.56456
11	Italy	1.18702	2.48363

Obr. 29 Výsledek standardizace

Dále vyzkoušejte v novém uzlu Preprocess volbu *Normalize to interval [-1, 1]*. Data jsou přepočítána do avizovaného intervalu (Obr. 30). Dobrou zprávou je, že uzly pro shlukování *k-Means* a *Hierarchical clustering*, které budou popisovány v následujících kapitolách, přímo nabízejí volbu normalizace dat v rámci uzlu a není třeba data upravovat před shlukováním těmito metodami.



Obr. 30 Normalizace dat do intervalu [-1, 1]

Normalizaci a standardizaci vyzkoušejte ještě jednou na souboru *3_Deti.xlsx*.

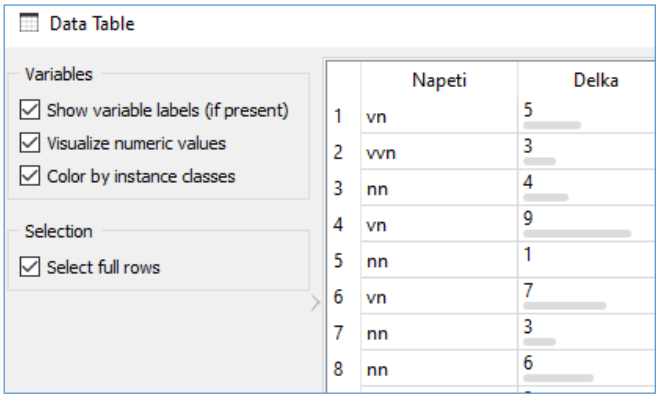
3.10 Dichtomizace atributů

Dichotomizace převádí vícehodnotová kategoriální data na více atributů logického datového typu, tzv. **dummy proměnné**.

Příklad 10

Data *3_ElektrickeVedeni.xlsx*, **workflow** *3_Dichotom.ows*

Zdrojová data obsahují délku a typ jednotlivých vedení elektrického proudu. Jsou rozlišeny tři kategorie *nn-nízké napětí, vn-vysoké napětí a vvn-velmi vysoké napětí* v atributu *Napeti* (Obr. 31). Dichotomizaci provedeme pomocí uzlu **Continuize**.

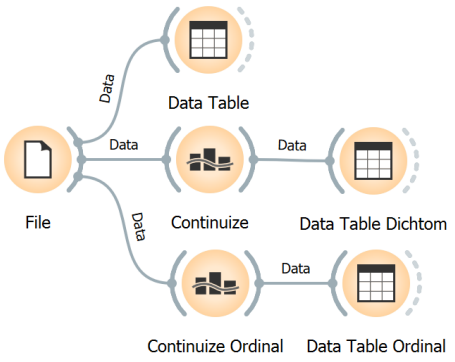


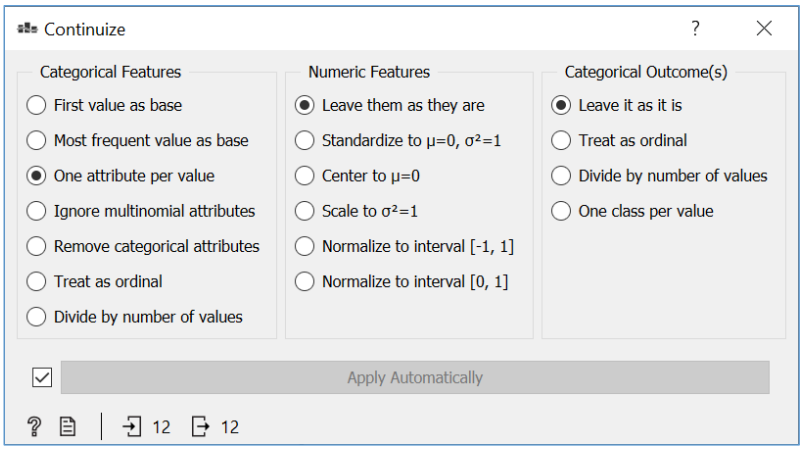
Obr. 31 Část zdrojových dat elektrického vedení s kategorií Napeti

V první levé sekci uzlu **Continuize** (Obr. 32) pro kategoriální data vybereme volbu *One attribute per value*.

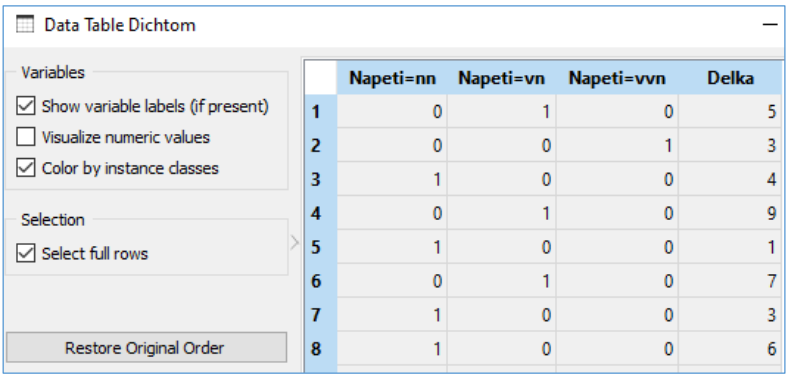
Jeden vstupní atribut *Napeti* je převeden na tři nové atributy s názvy podle hodnot původních *Napeti=nn, Napeti=vn* a *Napeti=vvn*. Uzel automaticky identifikuje počet kategoriálních hodnot a přidá příslušný počet nových atributů, kdy zároveň původní atribut odebere. Výsledek se zobrazí v uzlu *Data Table* a uloží pomocí uzlu *Save Data*.

Prostřední sekce uzlu **Continuize** slouží pro změnu numerických hodnot. V tomto příkladu ji nevyužijeme, bude ponechána výchozí hodnota *Leave them as they are.* Obdobně bude ponechána beze změny sekce *Categorical Outcome(s)* s nastavením *Leave it as it is* (Obr. 32).





Obr. 32 Workflow pro dichotomizaci s uzlem Continuize



	Napeti=nn	Napeti=vn	Napeti=vvn	Delka
1	0	1	0	5
2	0	0	1	3
3	1	0	0	4
4	0	1	0	9
5	1	0	0	1
6	0	1	0	7
7	1	0	0	3
8	1	0	0	6

Obr. 33 Výsledek dichtomizace

Příklad 11

V případě, že existují jen dvě kategorie, tak se nevytváří dva nové atributy, ale automaticky jen jeden. Vyzkoušejte samostatně na datech *4_Houses.xlsx s* volbou *Ignore multinominal attributes* a sledujte výsledek dichotomizace atributu *Water_Distance* (Obr. 34).

	ID_House	Water_Distance
1	1	long
2	2	short
3	3	long
4	4	short
5	5	short
6	6	short
7	7	long
8	8	short

	ID_House	Water_Distance=short
1	1	0
2	2	1
3	3	0
4	4	1
5	5	1
6	6	1
7	7	0
8	8	1

Obr. 34 Vstupní data (vlevo) a výsledek dichotomizace atributu, kde jsou pouze dvě vstupní kategorie (vpravo)

Všimněte si, že uzel Continuize nabízí další užitečné volby (včetně výše probírané standardizace a normalizace). Volba Treat as ordinal nahradí textové kategorie číselným kódem. Vyzkoušejte samostatně.

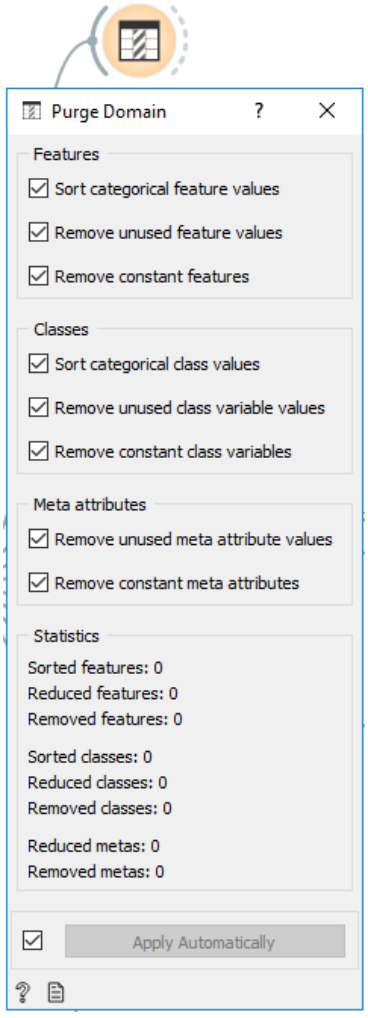
3.11 Smazání nevýznamných atributů

Uzel **Purge Domain** umožní smazat vybrané atributy podle zadaných požadavků. Někdy je užitečné data očistit od nevýznamných atributů nebo záznamů. Zejména u obsáhlých dat je lépe to provést automatizovaně než ručně.

Požadavky na smazání atributů se nastavují v sekci *Features*, požadavky na smazání kategoriálních atributů se nastavují v sekci *Classes.* Třetí sekce nastavuje mazání atributů, které jsou označeny jako *meta*. Statistiku s počtem dotčených atributů okamžitě ukazuje spodní část *Statistics* (Obr. 35).

Samostatně prozkoumejte možnosti. Zajímavé tipy obsahuje Blog Orange (Pretnar 2016b):

https://orangedatamining.com/blog/2016/01/29/tips-and-tricks-for-data-preparation/



Obr. 35 Smazání atributů pomocí Purge Domain

4 KORELACE A REDUKCE DIMENZIONALITY

Analýza hlavních komponent redukuje počet vstupních atributů, pokud spolu silně korelují. Následně je nahradí novými popisnými atributy, které jsou lineární kombinací vstupních atributů. Nové atributy jsou nazývány hlavními komponentami. Nejprve je tedy dobré zjistit, zda existuje korelace atributů.

4.1 Korelace

Příklad 12

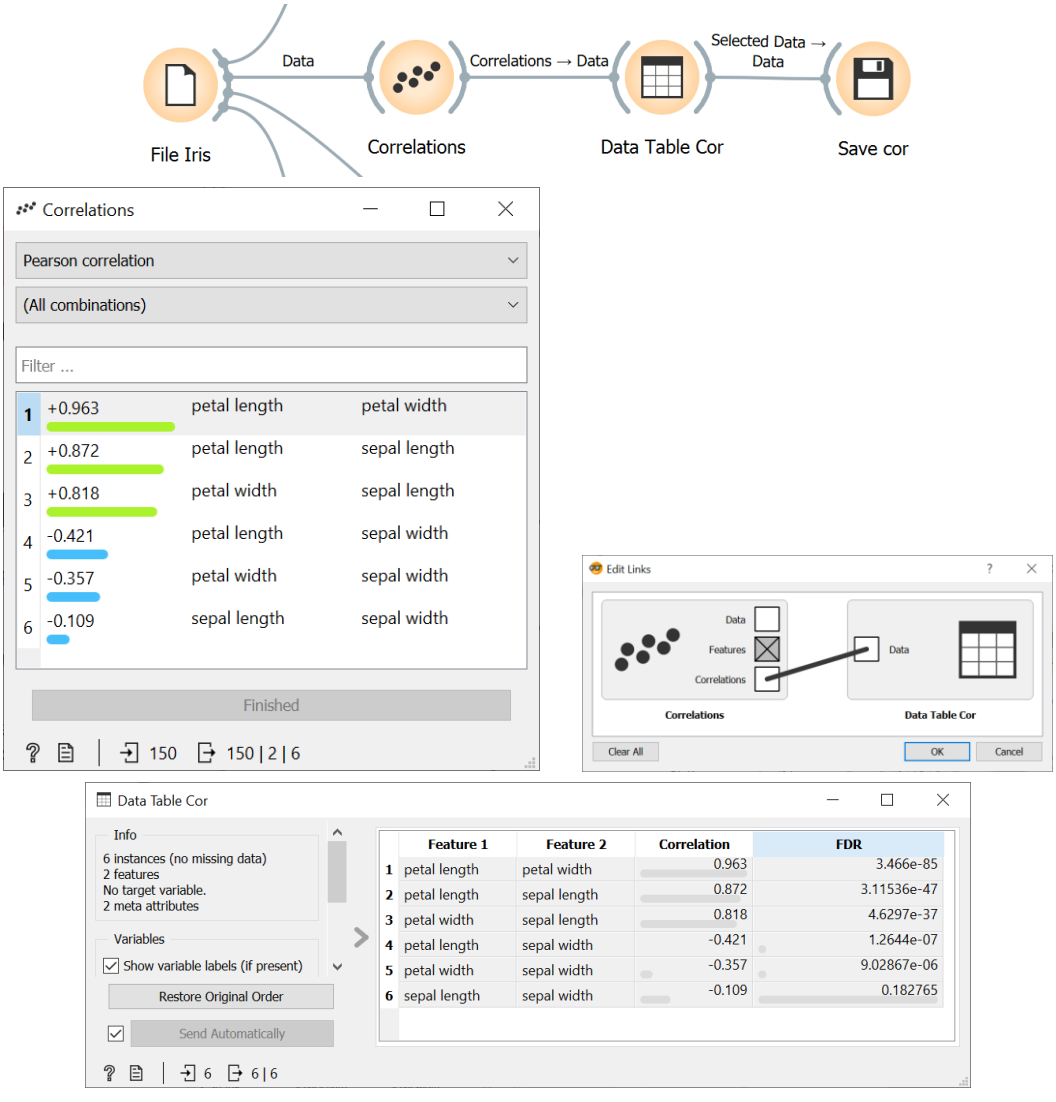
Data *3_Iris.xlsx,* **workflow** *4_PCA_Iris.ows*

Korelace se vyšetřuje pomocí uzlu **Correlations**. V záhlaví dialogu se vybírá typ korelace: *Pearson* (výchozí volba) nebo *Spearman*. Spearmanův korelační koeficient počítá korelaci pro pořadová data.

V druhé volbě je možné pouze vybrat jeden určitý atribut nebo ponechat zobrazit všechny kombinace atributů s vypočtenou korelací. Údaj *Filter* slouží k výběru korelací vyšších než zadaný limit (např. 0,8). Tabulka ukáže po dvojicích Pearsonovy korelační koeficienty všech čtyř vstupních atributů popisující květy kosatců (Obr. 36).

Ve výsledku je šest hodnot korelací (odpovídá šesti kombinacím). Zelenou čárou jsou znázorněny kladné korelace a modrou záporné korelace. Hodnoty korelací jsou seřazeny od největší kladné hodnoty po nejmenší zápornou hodnotu. Zde je evidentní, že *petal length* a *petal width* mají vysokou korelaci 0,963, tzv. vysvětlují stejnou skrytou (latentní) proměnnou a to druh kosatce.

Hodnoty korelací lze zobrazit v tabulce *Data Table*, který je připojen za uzel *Correlations* a uložit pomocí uzlu *Save Data.* Sloupec FDR zobrazuje hodnotu false discovery rate.



Obr. 36 Workflow, výsledné korelace jednotlivých atributů a nastavení signálů v dialogu Edit Links mezi uzly

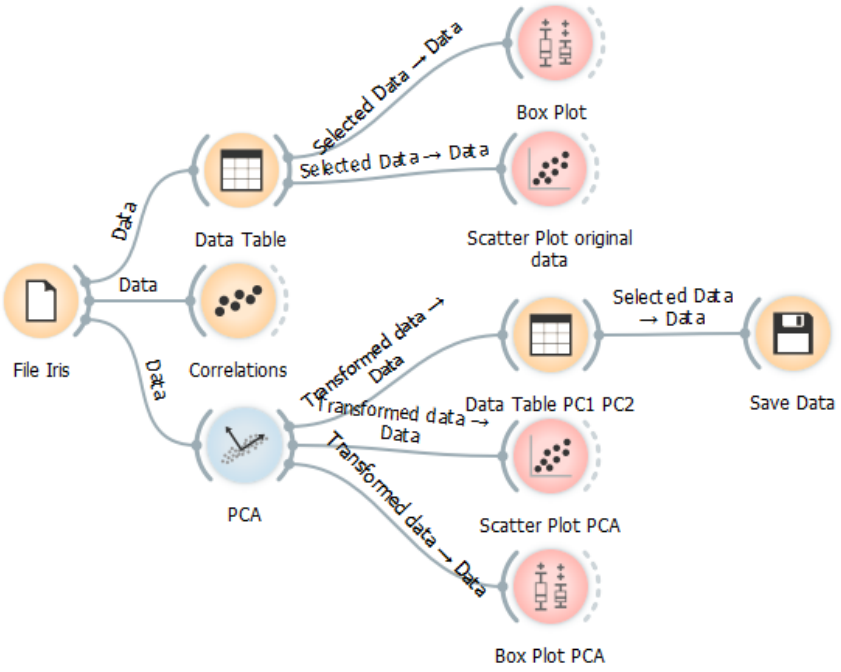
4.2 PCA – analýza hlavních komponent

Analýza hlavních komponent (PCA Principal Component Analysis) patří jak do fáze předpřípravy dat, tak ji lze použít jako jednu z analytických metod (Šarmanová 2012). PCA odstraňuje vzájemnou závislost vstupních atributů a nahrazuje je novými atributy označovanými jako hlavní komponenty PC1, PC2, PC3 atd. Nové hlavní komponenty jsou navzájem nekorelované (nebo minimálně korelované) a odhalují latentní neměřitelné proměnné. Metoda tedy provádí redukci dimenzí, kdy počet původních vstupních atributů (charakteristik) je podstatně snížen na nižší počet nových atributů (komponent), který dostatečně popisuje objekty. Je snaha na závěr najít vhodné a výstižné pojmenování pro tyto nové komponenty podle příspěvku původních atributů. Důležité je vystihnout a interpretovat nové komponenty a určit o čem vlastně vypovídají.

Příklad 13

Data *3_Iris.xlsx,* **workflow** *4_PCA_Iris.ows*

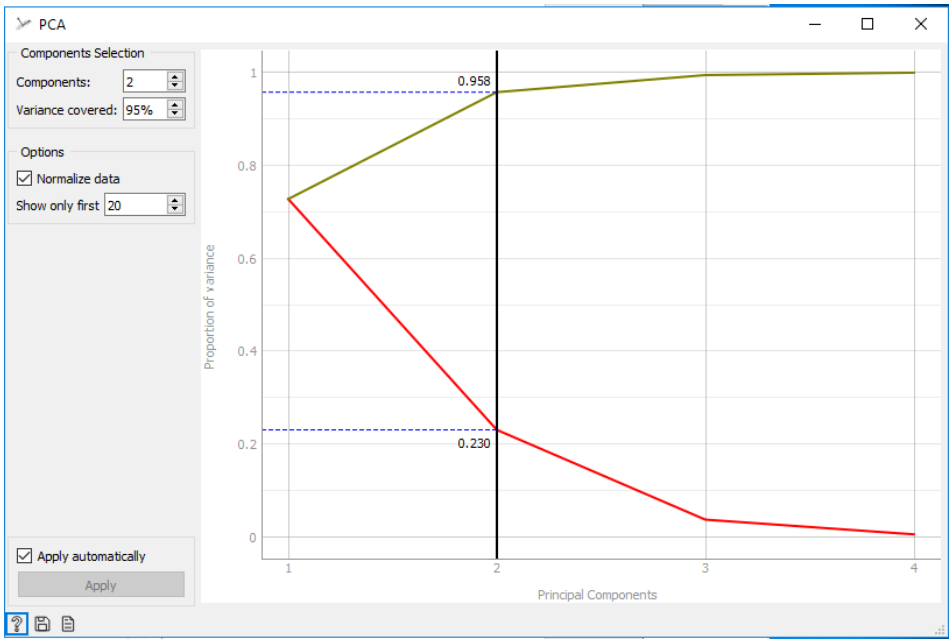
Workflow obsahuje nejprve uzly pro základní inspekci vstupních dat – *Data Table, Box Plot, Scatter Plot a Correlations* (Obr. 37). Modrý uzel **PCA** slouží k transformaci vstupních atributů do nových popisných atributů – hlavních komponent. Uzel se nachází v modré sekci uzlů **Unsupervised** (učení bez učitele). Analýza hlavních komponent může následně pomoci identifikovat v nových komponentách shluky.



Obr. 37 Workflow s transformací pomocí metody PCA

Modrý uzel PCA ukáže Cattelův indexový graf úpatí vlastních čísel (neboli sutinový graf), kde lze interaktivně posouvat svislou černou čárou a nastavit tak počet zvolených hlavních komponent (Obr. 38). Počet lze nastavit i číselně v kolonce *Components*. Určení počtu komponent je provázáno s údajem *Variance covered* (pokrytí rozptylu). Při zvyšování procent pokrytí se automaticky bude zvyšovat počet komponent, naopak při snižování pokrytí se bude počet komponent automaticky snižovat. Vlastní čísla slouží k určení počtu „využitelných" hlavních komponent, jež se zvolí v analýze k dalšímu užívání. Procento a kumulativní procento popisuje proměnlivost v původních atributech, popsanou dotyčnou hlavní komponentou (Meloun et al. 2012).

Vstupní data jsou automaticky před transformací normalizována díky zatržítku *Normalize data* (Obr. 38).



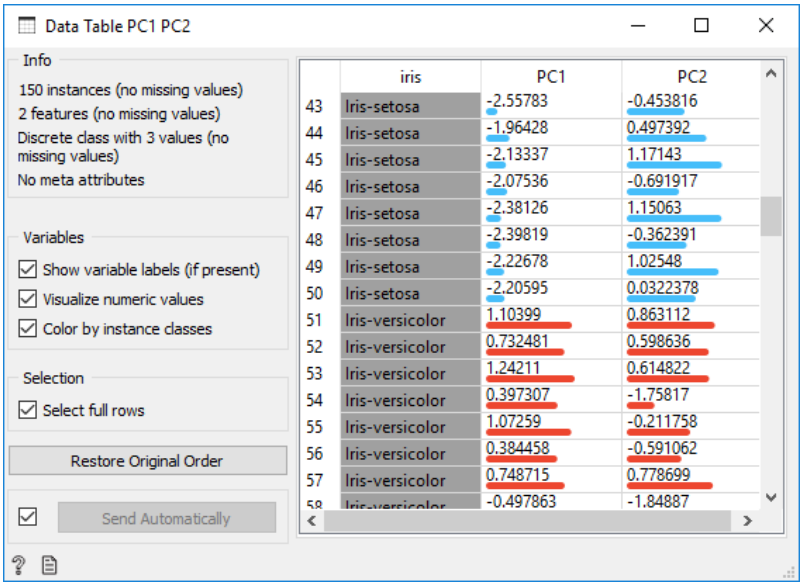
Obr. 38 Sutinový graf vlastních čísel (scree plot)

Červená čára je **sutinový (scree) graf vlastních čísel**. Na ose X je vynesen číselný index vlastního čísla. Osa Y ukazuje, jakou část rozptylu (procent) vysvětluje, které vlastní číslo (pozor není to přímo hodnota vlastního čísla). Pomocí grafu lze určit úpatí, kde se výrazně mění směr prudce klesající křivky. Při zvážení co nejvyššího pokrytí lze rozhodnout o dostačujícím počtu výsledných hlavních komponent.

Zelená čára ukazuje **kumulativní rozptyl**, který pokrývají komponenty v součtu. Maximum je 1, tzn. 100 %.

Pro data kosatců postačují dvě komponenty (vysvětlují 95,8 % rozptylu), tři komponenty vysvětlují 99,5 %.

Transformované nové souřadnice (komponenty) jednotlivých instancí kosatců zobrazuje další uzel *Data Table* (Obr. 39), který je napojen na uzel *PCA*. Tabulka obsahuje původní i nová data. Rozšířená data uložíme uzlem *Save Data*.

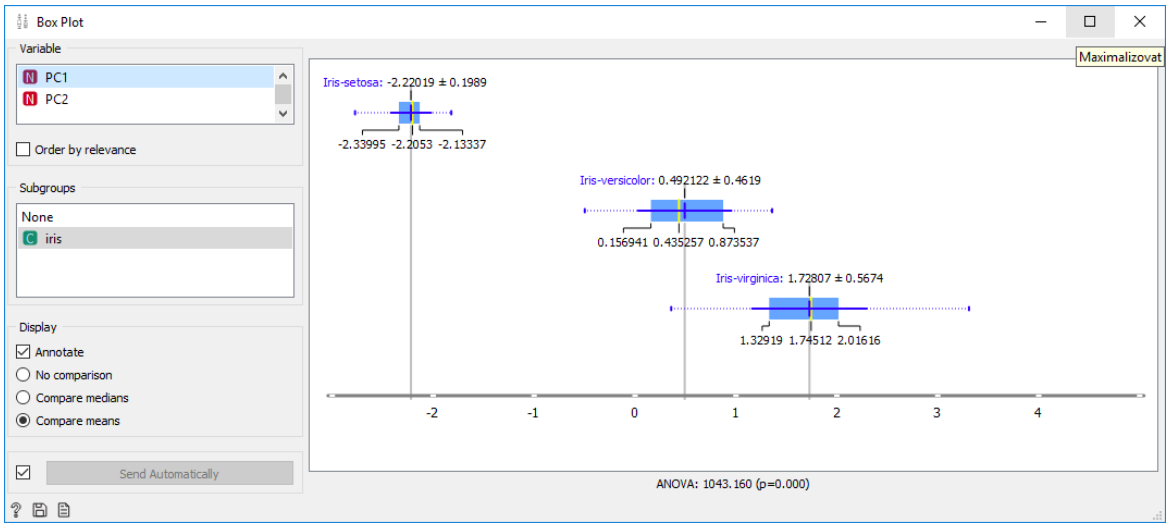


Obr. 39 Přepočítané nové popisné atributy PC1 a PC2 metodou PCA pro květy kosatců

Rozptylový graf **Scatter Plot**, který je připojen za uzel PCA, zobrazí objekty v souřadnicích nových komponent PC1 a PC2. Střed souřadnic 0, 0 je uprostřed grafu. V grafu jsou viditelné tři shluky jedinců (Obr. 40). Zobrazení shluků lze považovat za analytický výstup PCA.

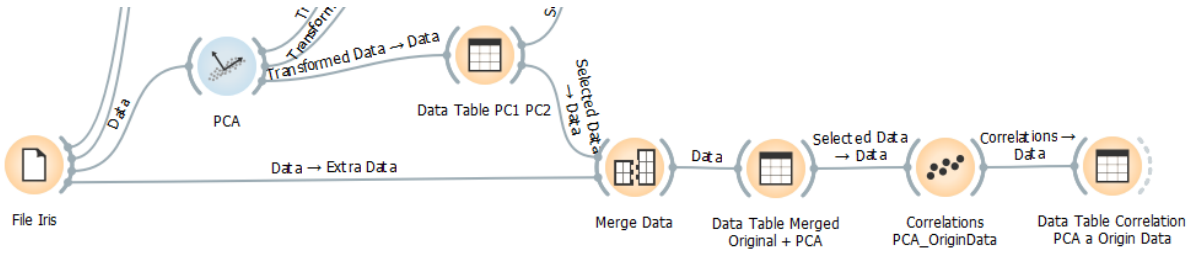
Obr. 40 Rozptylový graf v souřadnicích hlavních komponent PC1 a PC2

Box Plot pro nové souřadnice PC1 a PC2 (Obr. 41) ukazuje, že první komponenta PC1 velice dobře rozlišuje jednotlivé druhy kosatců. V případě komponenty PC2 je překryv krabic větší.



Obr. 41 Box Ploty nové souřadnice PC1 podle druhu kosatce

Pokud by původní data a hodnoty nových komponent nebyly viditelné ve výsledné tabulce u jednotlivých instancí (u starších verzí Orange), lze spojit tabulky pomocí uzlu **Merge**. Výsledek je pak viditelný v následujícím uzlu *Data Table*.



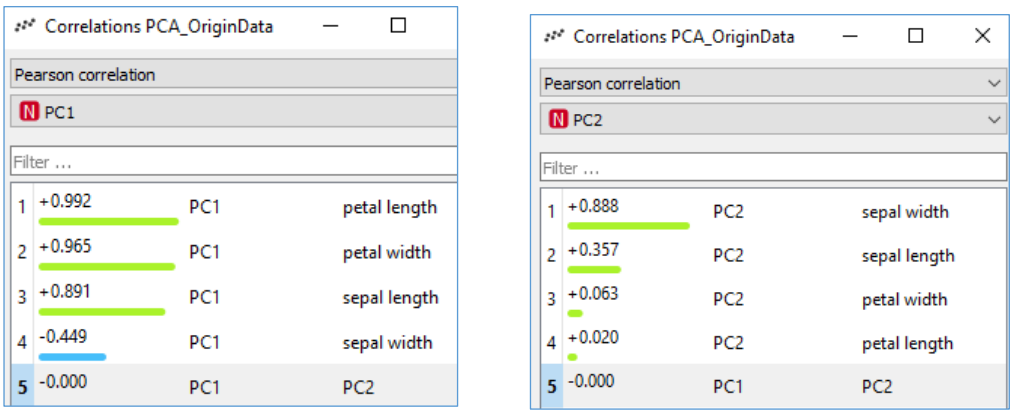
Obr. 42 Spojení originálních a transformovaných dat do jedné tabulky uzlem Merge

	iris	PC1	PC2	sepal length	sepal width	petal length	petal width
43	Iris-setosa	-2.55783	-0.453816	4.4	3.2	1.3	0.2
44	Iris-setosa	-1.96428	0.497392	5.0	3.5	1.6	0.6
45	Iris-setosa	-2.13337	1.17143	5.1	3.8	1.9	0.4
46	Iris-setosa	-2.07536	-0.691917	4.8	3.0	1.4	0.3
47	Iris-setosa	-2.38126	1.15063	5.1	3.8	1.6	0.2
48	Iris-setosa	-2.39819	-0.362391	4.6	3.2	1.4	0.2
49	Iris-setosa	-2.22678	1.02548	5.3	3.7	1.5	0.2
50	Iris-setosa	-2.20595	0.0322378	5.0	3.3	1.4	0.2
51	Iris-versicolor	1.10399	0.863112	7.0	3.2	4.7	1.4
52	Iris-versicolor	0.732481	0.598636	6.4	3.2	4.5	1.5
53	Iris-versicolor	1.24211	0.614822	6.9	3.1	4.9	1.5
54	Iris-versicolor	0.397307	-1.75817	5.5	2.3	4.0	1.3
55	Iris-versicolor	1.07259	-0.211758	6.5	2.8	4.6	1.5
56	Iris-versicolor	0.384458	-0.591062	5.7	2.8	4.5	1.3

Obr. 43 Výsledek spojení původních atributů a nových komponent do jedné tabulky

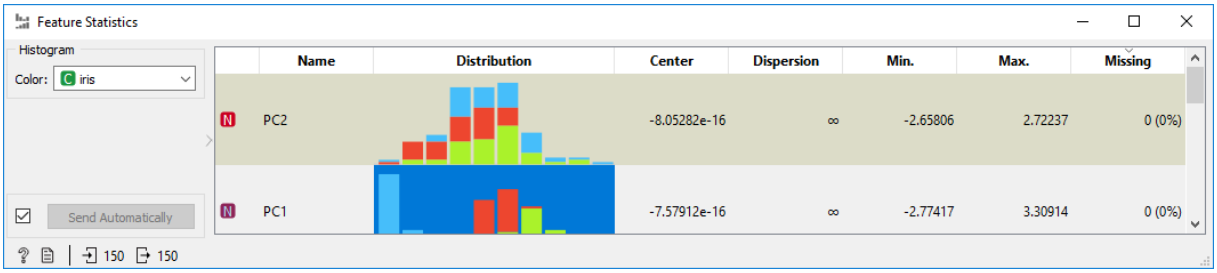
Jak korelují nové komponenty PC1 a PC2 s původními atributy lze spočítat pomocí uzlu **Correlation**, který následuje za tabulkou spojených dat uzlem *Merge*. Korelace ukazují, jak přispívají původní jednotlivé atributy do nových komponent. Vysoká kladná korelace znamená vysoký příspěvek, korelace kolem nuly má malý podíl na hodnotě nové komponenty a záporná korelace přispívá negativně k hodnotě nové komponenty. Do komponenty **PC1** přispívá nejvíce atribut *petal length*, do komponenty **PC2** přispívá nejvíce atribut *sepal width* (Obr. 44).

Výpočet nulové hodnoty korelace mezi PC1 a PC2 jen potvrzuje myšlenku PCA, a to nahrazení původních korelovaných atributů novými atributy PC1 a PC2, které nekorelují. Výpočet korelací při vyhodnocení nahrazuje graf komponentních zátěží, který v Orange není přímo možné vykreslit.



Obr. 44 Korelace původních atributů a nových komponent PC1 (vlevo) a PC2 (vpravo)

Pomocí uzlu *Feature Statistics* zkontrolujte, že průměrná hodnota PC1 a PC2 je téměř nula (sloupec Center, kde je velice malé číslo, Obr. 45). Při porovnání sloupce Min. a Max. je viditelné větší rozsah minima a maxima u první komponenty PC1 než u komponenty PC2, což odpovídá tvrzení, že první komponenta postihuje největší rozptyl.



Obr. 45 Statistický popis nových komponent PC1 a PC2

> *Video návod na YouTube: Getting Started with Orange 09: Principal Component Analysis*

4.3 Korespondenční analýza

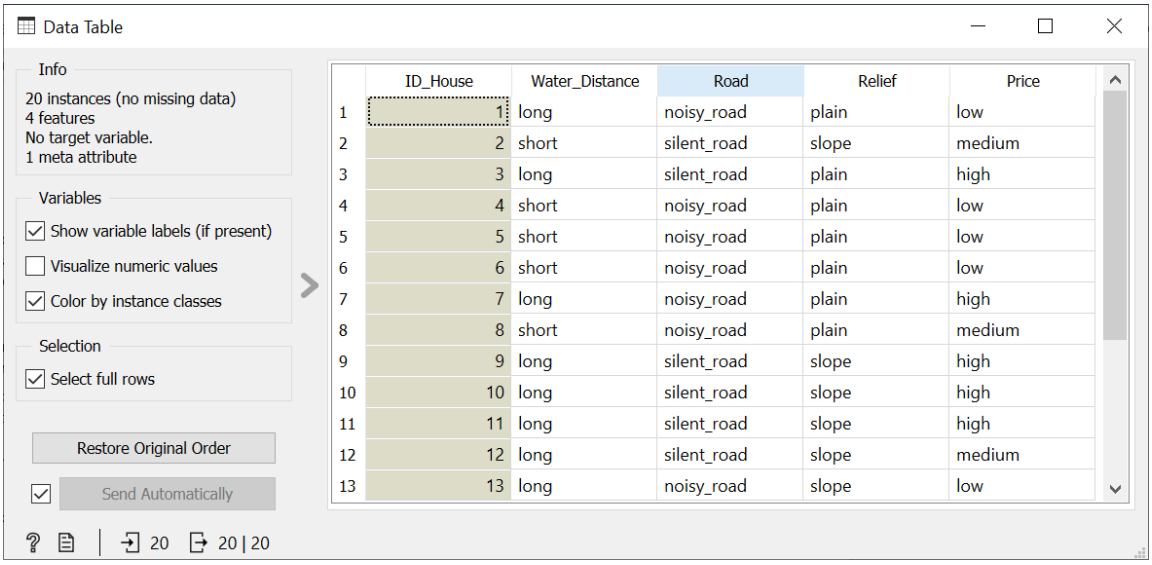
Další metoda k redukci vstupních dimenzí je korespondenční analýza (CA Correspondence Analysis). Je podobná metodě hlavních komponent (PCA), ale provádí na rozdíl od PCA lineární transformaci **kategoriálních** (diskrétních) hodnot. PCA naopak pracuje se spojitými daty. Korespondenční analýza řeší obdobný problém jako metoda hlavních komponent, ve které jsou závislosti původních spojitých proměnných vysvětlovány pomocí menšího počtu latentních komponent. V korespondenční analýze sledujeme vztahy mezi jednotlivými kategoriemi více kategoriálních proměnných. Výsledkem analýzy je tzv. **korespondenční mapa** představující osy nového redukovaného souřadného systému, ve kterém jsou graficky zobrazeny jednotlivé kategorie proměnných (Janoušová et al. 2020a).

Korespondenční analýza redukuje dimenzionalitu hledáním *vlastních vektorů asociační matice*. Podobně jako u dalších ordinačních metod je i v případě korespondenční analýzy snaha získat ordinační osy v klesajícím stupni důležitosti tak, aby se hlavní informace obsažená v tabulce dala shrnout do podprostoru s co možná nejmenším počtem dimenzí.

Příklad 14

Data *4_Houses.xlsx,* **workflow** *4_CorrespondenceAnal_Houses.ows*

Dvacet domů je popsáno polohou vůči vodě *Water_Distance* (short, long), zda stojí u tiché či hlučné silnice *Road* (noisy, silent), zda stojí na svahu či na rovině *Relief* (slope, plain) a poslední údaj je o úrovni ceny domu *Price* (low, medium, high).



	ID_House	Water_Distance	Road	Relief	Price
1	1	long	noisy_road	plain	low
2	2	short	silent_road	slope	medium
3	3	long	silent_road	plain	high
4	4	short	noisy_road	plain	low
5	5	short	noisy_road	plain	low
6	6	short	noisy_road	plain	low
7	7	long	noisy_road	plain	high
8	8	short	noisy_road	plain	medium
9	9	long	silent_road	slope	high
10	10	long	silent_road	slope	high
11	11	long	silent_road	slope	high
12	12	long	silent_road	slope	medium
13	13	long	noisy_road	slope	low

Obr. 46 Vstupní data popisující polohu a cenu domů

Vstupní kategoriální proměnné lze uspořádat do kontingenční tabulky. Ukázka kontingenční tabulky pro data o domech je na Obr. 47. Z tabulky vidíme, že nejvíce 5 domů má nízkou cenu, je blízko vody a je poblíž hlučné

silnice. Dále 3 domy mají vysokou cenu, jsou daleko od vody a nejsou u hlučné silnice. Ostatní počty výskytů domů se pohybují v hodnotách 0, 1 a 2.

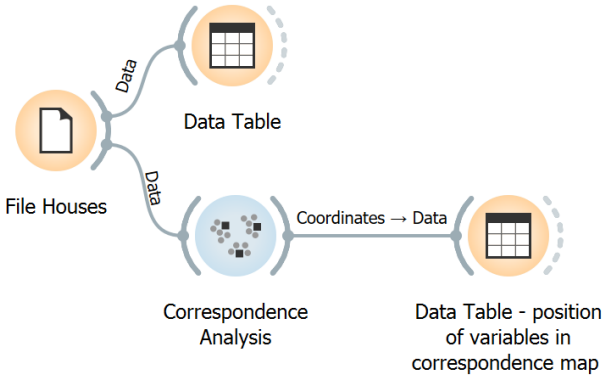
Počet podle ceny

Relief	plain				slope			
Road	silent		noisy		silent		noisy	
Watter_Distance	long	short	long	short	long	short	long	short
Price=high	2		1		3		2	
Price=low			1	5			1	
Price=medium		1		1	1	1	1	

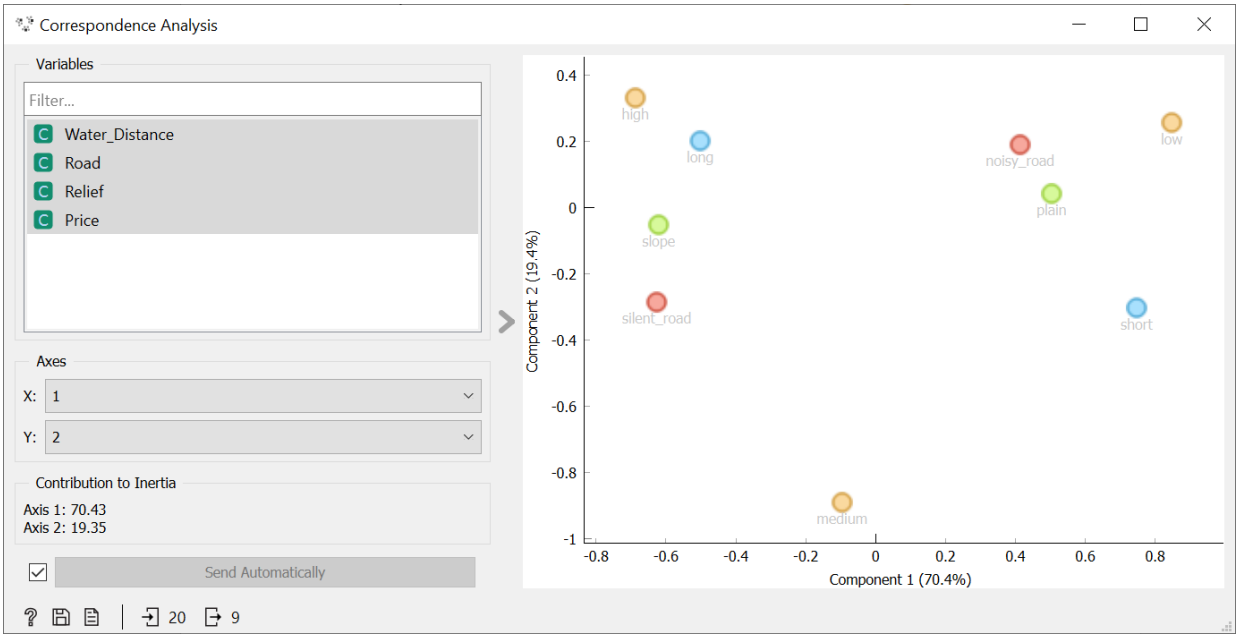
Obr. 47 Kontingenční tabulka zobrazující četnosti domů podle charakteristik

Korespondenční analýza je přínosná zejména při zpracovávání rozsáhlých kontingenčních tabulek (velký počet atributů a velký počet kategorií), kdy je grafické zobrazení formou korespondenční mapy přehlednější než číselné výstupy. Jde hlavně o popisnou a průzkumnou metodu, jejíž součástí není testování statistické významnosti získaných modelů (Janoušová et al. 2020b).

Modrý uzel **Correspondence Analysis** je ze skupiny **Unsupervised** stejně jako uzel PCA. Uzel zobrazí výsledek analýzy ve formě korespondenční mapy, kdy atributy jsou odlišeny barvou. Různé hodnoty téhož atributu mají stejnou barvu (Obr. 49).



Obr. 48 Workflow s uzlem korespondenční analýzy



Obr. 49 Korespondenční analýza s korespondenční mapou pro domy

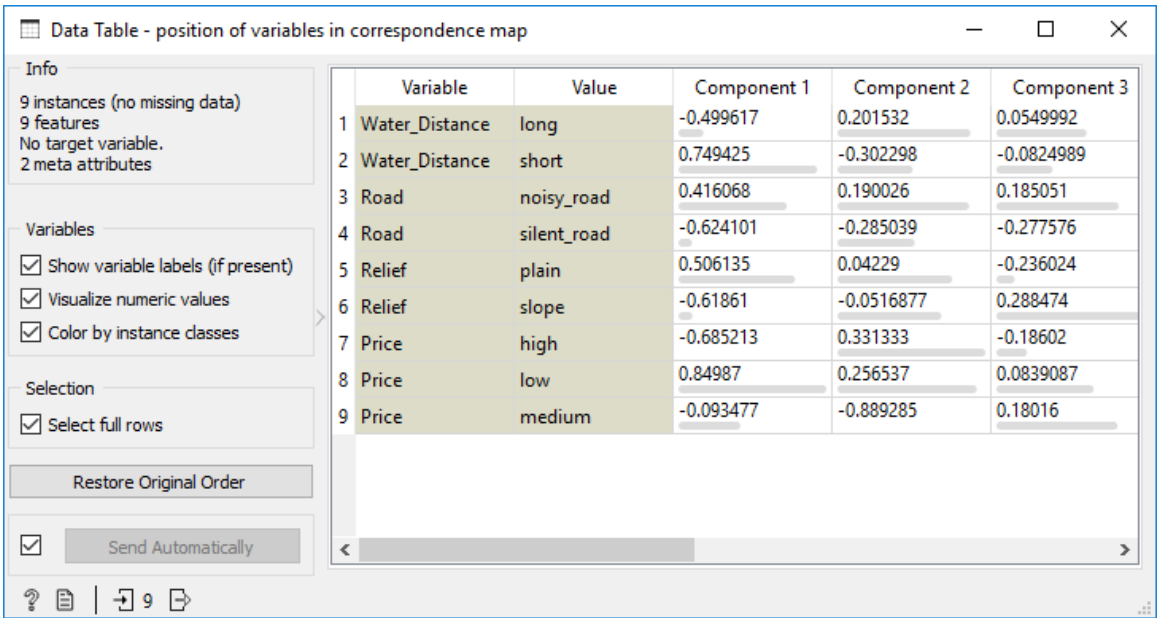
Korespondenční mapa ukazuje, *že* bod *Price = medium* je umístěn dole a nemá významný vztah k dalším veličinám. V korespondenční mapě se vytvořily dvě oblasti souvisejících hodnot. Jedna je vlevo a druhá je vpravo (Obr. 49).

V levé části korespondenční mapy jsou domy s vysokou cenou *Price=high* a je evidentní, že jsou daleko od vody, údaj *Water_distance=long* (dům je patrně mimo záplavovou zónu). Navíc je v levé části i málo tichá silnice *silent_road* a poloha na svahu *slope*. Všechny tyto údaje tedy spolu souvisí.

V pravé části korespondenční mapy jsou blízko sebe domy s nízkou cenou *Price=low* a hlučnou silnicí *noisy_road*. To znamená, že blízkost hlučné silnice snižuje cenu domu. Je patrná i souvislost s polohou na rovině *plai*n. Určitý vliv na nízkou cenu domu má patrně to, že je blízko vody *Water_distance=short*.

V dolní části dialogu je údaj o příspěvku k inerci ***Contribution to Inertia***. Geometricky vyjadřuje inerce stupeň rozptýlení bodů ve vícerozměrném prostoru a můžeme ji chápat jako analogii rozptylu známému ze statistického modelování.

Pozice proměnných v nových souřadnicích *Component1, Component2, Component3, …* z korespondenční mapy zobrazíme pomocí následujícího uzlu *Data Table* (Obr. 50) ve workflow. Výsledek lze uložit obvyklým způsobem pomocí uzlu *Save Data*.



	Variable	Value	Component 1	Component 2	Component 3
1	Water_Distance	long	-0.499617	0.201532	0.0549992
2	Water_Distance	short	0.749425	-0.302298	-0.0824989
3	Road	noisy_road	0.416068	0.190026	0.185051
4	Road	silent_road	-0.624101	-0.285039	-0.277576
5	Relief	plain	0.506135	0.04229	-0.236024
6	Relief	slope	-0.61861	-0.0516877	0.288474
7	Price	high	-0.685213	0.331333	-0.18602
8	Price	low	0.84987	0.256537	0.0839087
9	Price	medium	-0.093477	-0.889285	0.18016

Obr. 50 Pozice proměnných v nových komponentách zjištěných korespondenční analýzou

Upozornění

V korespondenční mapě se můžou interpretovat vzdálenosti mezi řádkovými kategoriemi a vzdálenosti mezi sloupcovými kategoriemi, ne ovšem vzdálenosti mezi řádkovými body a sloupcovými body. Lze ale interpretovat relativní pozici bodu z jedné sady s ohledem ke všem bodům druhé sady. Pro korespondenční mapu obecně platí, že:

- blízkost dvou řádků (sloupců) značí podobný profil v těchto dvou řádcích (pojmem profil označujeme distribuci podmíněné četnosti);
- pokud jsou od sebe řádky či sloupce vzdáleny, jejich profil je značně odlišný;
- blízkost určitého řádku a určitého sloupce znamená, že tento řádek má důležitou váhu v daném sloupci;
- pokud jsou od sebe určitý řádek a sloupec daleko, nejsou v daném sloupci téměř žádná pozorování, která přísluší danému řádku
- body poblíž středu ordinačního diagramu nemají výrazný profil; střed ordinačního diagramu je těžištěm bodů jak řádkových, tak sloupcových kategorií (Janoušová et al. 2020a).

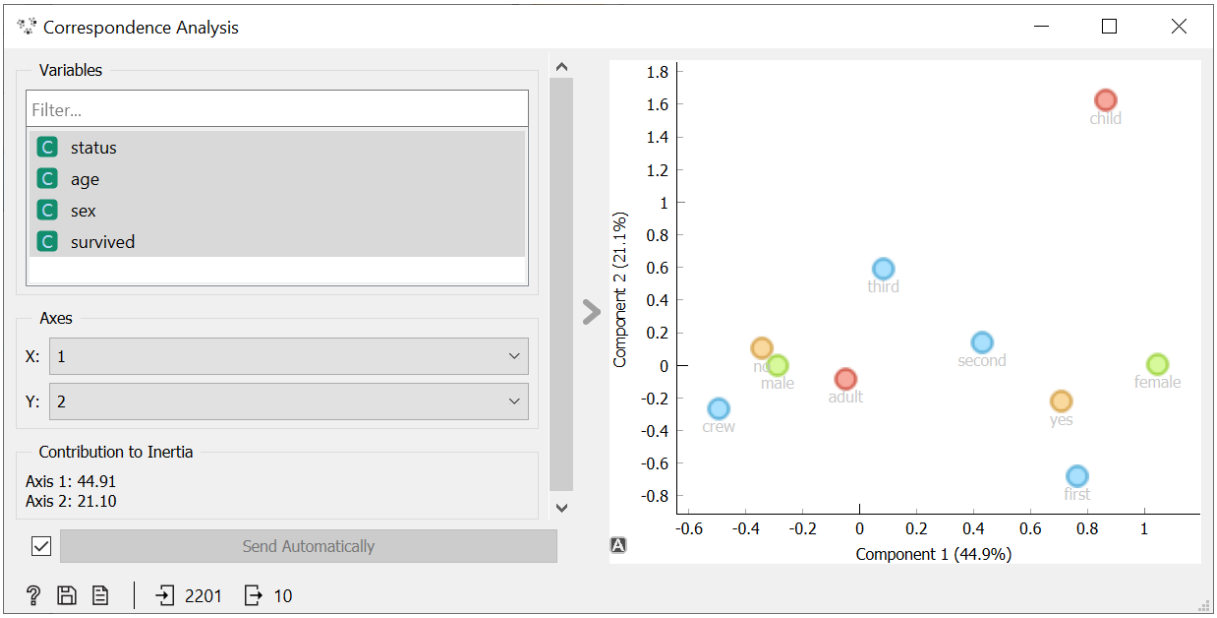
Příklad 15

Data *Titanic.tab,* **workflow** *4_CorrespondenceAnal_Titanic.ows*

Ze cvičných dat Orange lze pro korespondenční analýzu použít dataset **Titanic** (Dawson 1995). Tento dataset obsahuje údaje o cestujících na zaoceánské lodi Titanic. Kromě údaje, zda se zachránili a přežili ztroskotání lodi – atribut *survived* (yes, no), data obsahují údaje o pohlaví *sex* (male, female), věku *age* (adult, child), cestovní třídě *status* (first class, second class, third clas a crew).

Z korespondenční mapy na Obr. 51 je patrné v pravé části, že se zachránily (*yes*) převážně ženy (*female*). Dále se zachránili cestující první a druhé třídy (*first, second*). Naopak v levé části mají k sobě blízko kategorie, kdy nepřežili muži (*no, male*) a lidé z posádky (crew). Dospělí *adult* se nachází téměř ve středu 0,0 nových komponent. Tento bod poblíž středu diagramu nemá výrazný profil; střed diagramu je těžištěm bodů jak řádkových, tak sloupcových kategorií.

Příspěvek k inerci první a druhé komponenty zde není tak vysoký jako v předchozím příkladu domů.



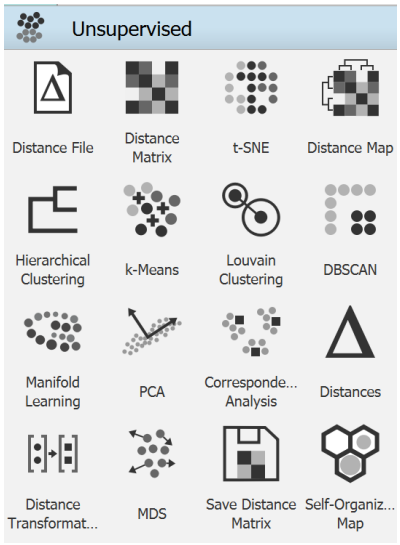
Obr. 51 Korespondenční mapa pro dataset Titanic

> *Při interpretaci korespondenční mapy je třeba dobře sledovat měřítka jednotlivých os X a Y, která můžou mít odlišný rozsah a měřítko, a to může mít vliv na vyhodnocení blízkosti kategorií.*

5 SHLUKOVÁNÍ

Shlukování je typ analýzy, který hledá podobné vícerozměrné objekty, které následně vytváří shluky podobných objektů (Lukasová a Šarmanová 1985). Shlukování se hodí zejména tam, kde objekty projevují přirozenou tendenci se seskupovat (Meloun et al. 2012). Shlukování patří do skupiny učení bez učitele (unsupervised learning). Pro shluky se následně stanovují kritéria a pravidla, podle nichž je možné následně nový objekt přiřadit k existujícímu shluku a zařadit jej tak do odpovídající kategorie (třídy, segmentu) .

Pro metody učení bez učitele je v Orange skupina uzlů označená **Unsupervised**.



Obr. 52 Sekce Unsupervised v Orange

5.1 Jaccardův koeficient asociace

Na základě binární příznaků zjišťuje Jaccardův koeficient asociace objektů (Petr 2014b). Koeficient je zjišťován vždy pro dvojice objektů O_i a O_j , které jsou popsané *m* atributy s binárními hodnotami. Kontingenční tabulku můžeme použít jako nejjednodušší postup pro vyhodnocení vztahu mezi dvěma objekty. Jednotlivé hodnoty kontingenční tabulky odpovídají četnostem kombinací hodnot v datech. Nejprve se vypočítají kumulované počty shodných (*a*), neshodných (*b, c*) a negativně shodných znaků (*d*) pro dva objekty ve formě čtyřpolní tabulky. Počty se následně dosadí do vzorce pro výpočet Jaccardova koeficientu asociace A_J (1). Čím vyšší je hodnota koeficientu, tím jsou si dva objekty podobnější, naopak čím je nižší, tím jsou si nepodobnější. Ve vzorci pro výpočet Jaccardova koeficientu asociace se nebere v úvahu počet negativních shod *d*.

Tabulka1: Čtyřpolní tabulka pro výpočet podobnosti dvou objektů O_i a O_j

O_i / O_j	**1**	**0**	
1	*a*	*b*	r
0	*c*	*d*	s
	k	l	m

$$A_J(O_i, O_j) = \frac{a}{a + b + c} \qquad (2)$$

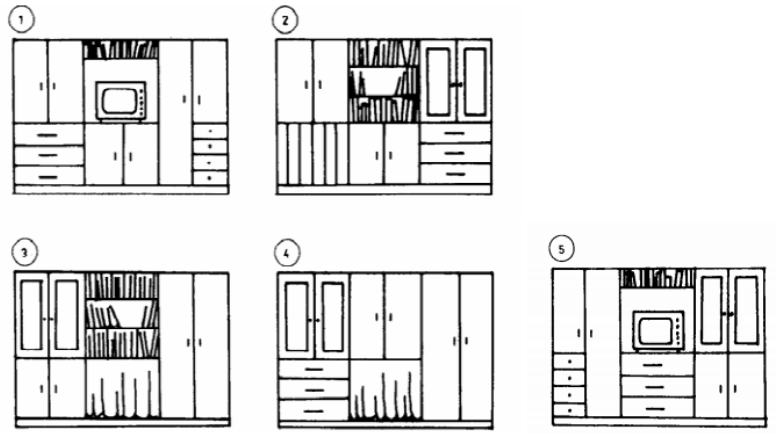
Jaccardův koeficient lze následně použít pro výpočet **vzdálenosti** dvou objektů. Vzdálenost vyjadřuje nepodobnost (*dissimilarity*) dvou objektů. Vzdálenost D_J se vypočítá odečtením Jaccardova koeficientu od hodnoty 1, vyjádřeno vzorcem (2). Potom menší hodnota vzdálenosti znamená větší podobnost objektů, a naopak větší hodnota vzdálenosti znamená nepodobnost objektů.

$$D_J(O_i, O_j) = 1 - A_J(O_i, O_j) \qquad (3)$$

Příklad 16

Data *5_CabinetJaccard.xlsx*, **workflow** *5_JaccardCabinet.ows*

Data obsahují popis pěti různých skříňových sestav do obýváku zobrazených na Obr. 53 (Šarmanová, 2012).



Obr. 53 Skříňové sestavy jako vstupní objekty pro zjištění podobnosti (Šarmanová 2012)

Data obsahují informace, zda skříňová sestava obsahuje nebo neobsahuje **0/1** daný typ skříňky (binární data). Sestava se může skládat z těchto deseti skříněk: dolní, horní, šatní, zásuvky, televize, knihovna, prádelník, příborník, pořadač, závěs. Každá skříňová sestava je popsána 10 atributy. Porovnává se pět různých sestav. Úkol je nalézt nejpodobnější dvojice skříňových sestav.

Data Table

Info

5 instances (no missing data)
10 features
No target variable.
1 meta attribute

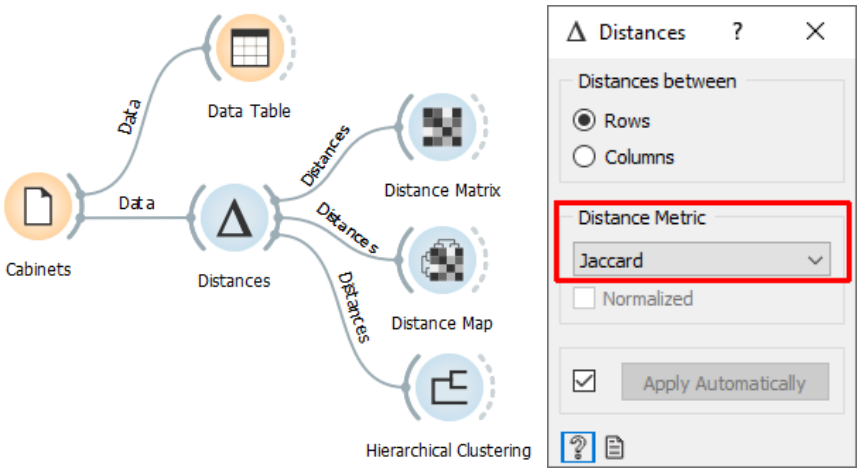
Variables

Show variable labels (if present)
Visualize numeric values

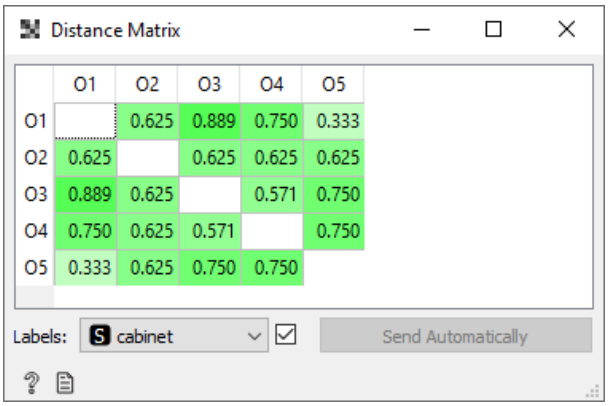
	cabinet	dolní	horní	šatní	zásuvky	televize	knihovna	prádelník	příborník	pořadač	závěs
1	O1	1	1	0	1	1	0	1	0	0	0
2	O2	1	1	0	0	0	1	1	1	1	0
3	O3	1	0	1	0	0	1	0	1	0	1
4	O4	0	1	1	0	0	0	1	1	0	1
5	O5	1	0	0	1	1	0	1	1	0	0

Obr. 54 Vstupní popisná data pěti skříňových sestav

Ve workflow v uzlu **Distances** vybereme volbu *Distance Metric* **Jaccard**. Výsledné Jaccardovy koeficienty nejsou zobrazeny přímo, ale jsou automaticky napočítány a odečteny od hodnoty 1 jako nepodobnosti podle vzorce (3). Vzdálenosti je možné vidět v uzlu **Distance Matrix**, který je připojen za uzel *Distances*. Uzel *Distance Matrix* zobrazuje přímo číselné hodnoty koeficientů nepodobnosti (vzdálenosti) s odstupňovanou sytostí zelené barvy. Alternativně lze zobrazit matici koeficientů v grafické formě pomocí uzlu **Distance Map.** Lze volit různé barevné stupnice. V případě šedé stupnice sytější barva znázorňuje vyšší hodnotu podobnosti.



Obr. 55 Workflow a nastavení metriky v uzlu Distances pro vzdálenost

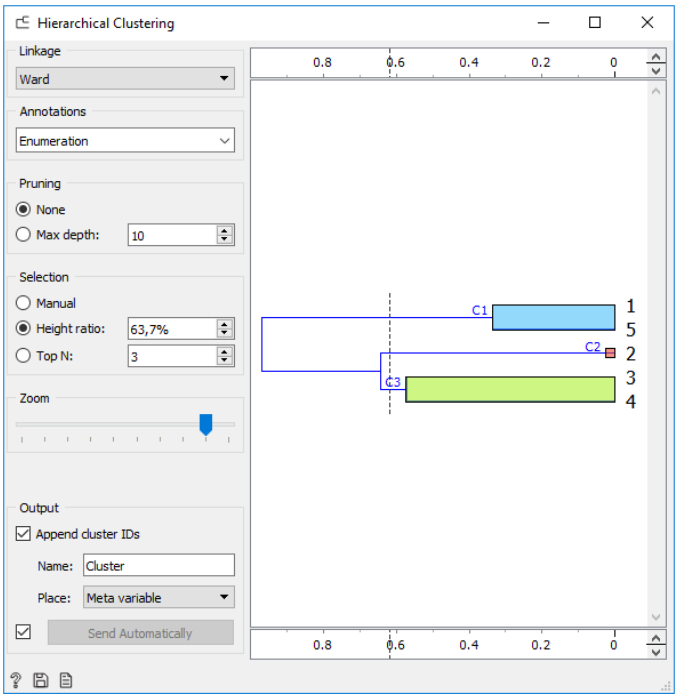


Obr. 56 Matice vzdáleností jako výsledek nepodobnosti podle Jaccardova koeficientu

Nejnižší hodnota vzdálenosti (tj. podobnosti) 0,333 je mezi skříňovou sestavou O1 a O5, a tudíž jsou si nejpodobnější. Další hodně podobnou dvojicí jsou sestavy O3 A O4 s hodnotou 0,571. Nejméně podobná si je sestava O1 a O3, kde je vzdálenost 0,889. Vzdálenosti lze také zobrazit uzlem *Distance Map*.

Obr. 57 Výsledné vzdálenosti ve formě mapy vzdáleností

Lépe čitelné podobnosti jsou při zobrazení pomocí dendrogramu (Obr. 58). Uzel **Hierarchical Clustering** provede hierarchické shlukování. Metoda spojování je v kolonce *Linkage* nastavena na **Ward**.



Obr. 58 Výsledek hierarchického shlukování ve formě dendrogramu

Z výsledku na Obr. 58 je patrný stejný výsledek jako v matici vzdálenosti a to, že nejpodobnější je si skříňová sestava O1 a O5. Potom si je podobná navzájem skříňová sestava O3 a O4 a skříňová sestava číslo O2 je úplně jiná než ostatní, připojuje se v dendrogramu jako poslední ke dvojici O3 a O4.

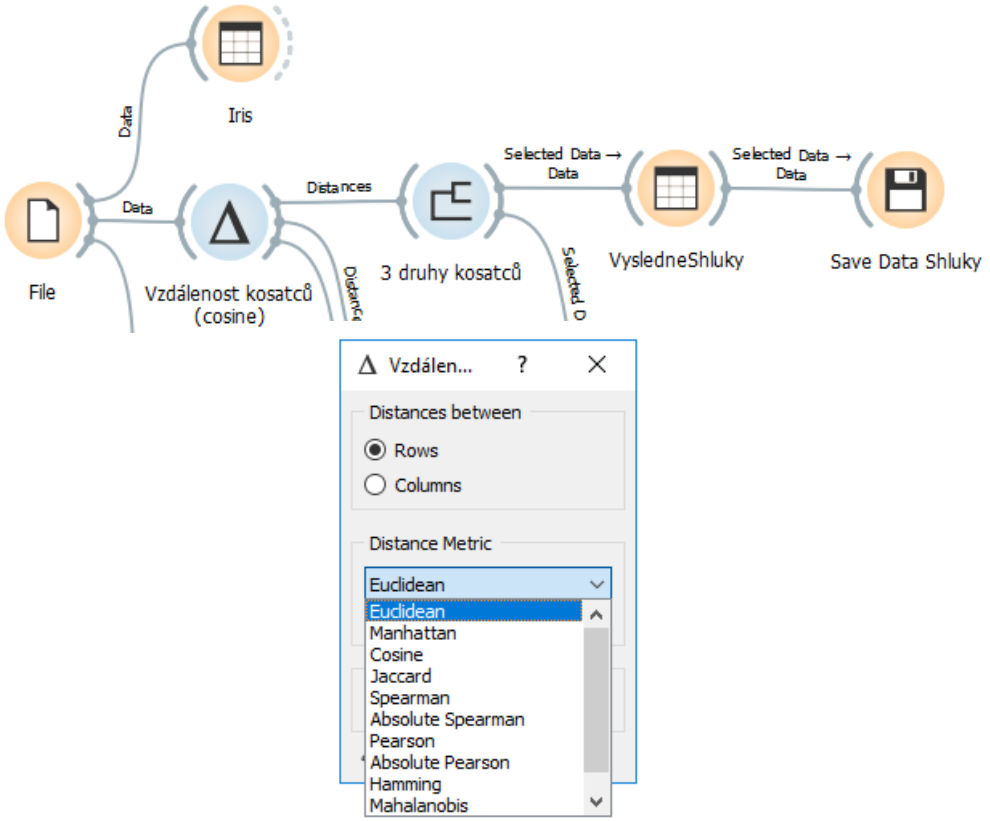
> *Samostatně zjistěte pomocí Jaccardova koeficientu, jak jsou si podobná evropská města podle různých druhů dopravy (metro, tramvaj, lodní doprava, bike sharing, …). Města jsou jednoduše popsána existencí/neexistencí různých druhů dopravy. Použijte soubor se zdrojovými daty 5_DopravaMesta.xlsx.*

5.2 Výpočet vzdálenosti a hierarchické shlukování

Příklad 17

Data *3_Iris.xlsx*, **workflow** *5_Iris_Hierarchical.ows*

Uzel **Distances** umožňuje nastavit různé způsoby výpočtu vzdálenosti. Jsou to vzdálenosti*: Euclidean, Manhattan, Mahalanobis, Cosine, Jaccad, (Absolute) Spearman, (Absolute) Pearson, Hamming a Bhattacharyya*. Pro numerická vstupní data zvolíme Euklidovskou vzdálenost – *Euclidean*. Pro malý počet vstupních atributů (zde čtyři) je Euklidovská vzdálenost ještě vhodná. Pro vyšší počet vstupních atributů se již nepoužívá.



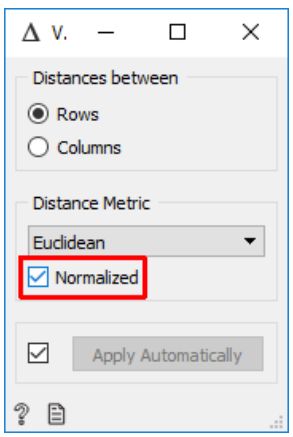
Obr. 59 Workflow a výběr metriky v uzlu Distance

Pro volbu *Euclidean* je automaticky zaškrtnuta volba *Normalized*. Tzn., že data jsou již v rámci tohoto uzlu předzpracována, je provedena jejich normalizace (zde je určitě žádoucí, neboť můžou být různě staré květy kosatců a tzn. různě narostlé).

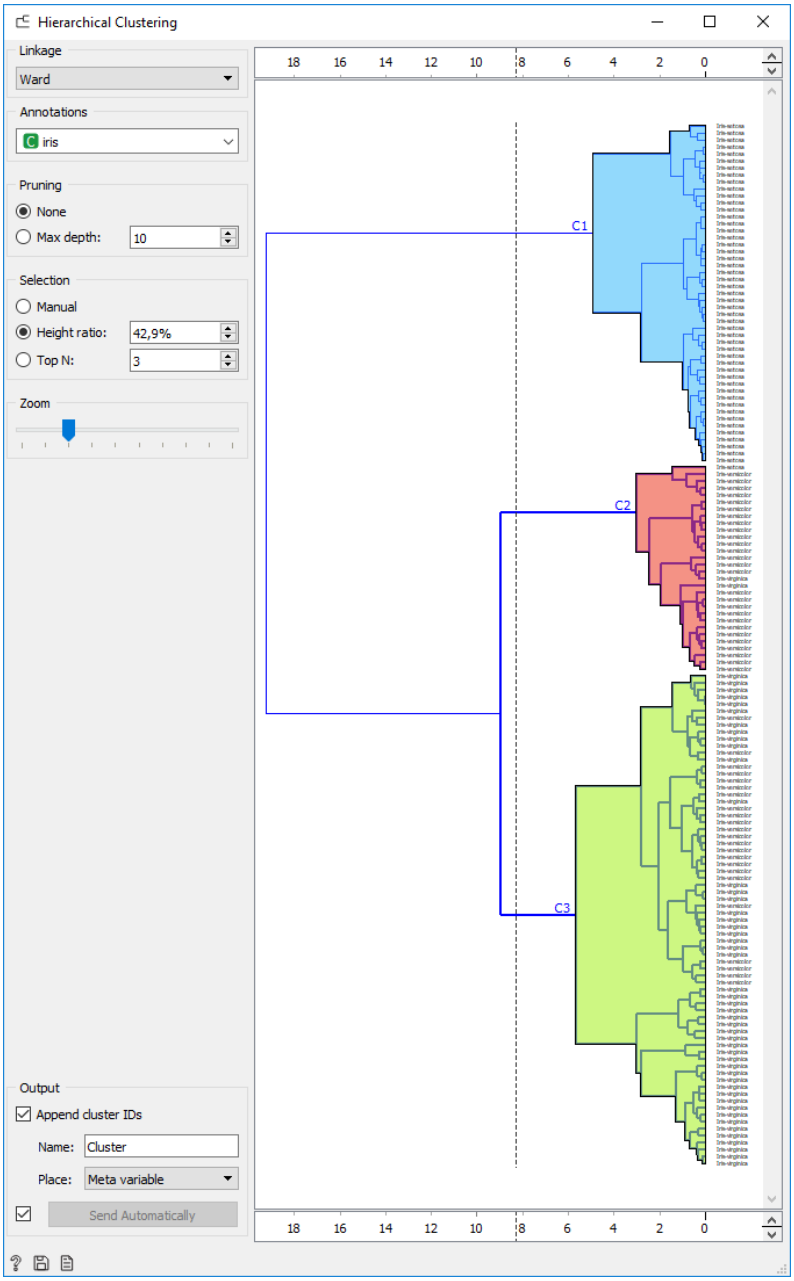
Následně je provedeno hierarchické shlukování uzlem **Hierarchical Clustering** (Obr. 59). Metoda spojování je v kolonce *Linkage* nastavena na **Wardovu** metodu a ruční nastavení 3 shluků lze měnit posunem svislé čárkované čáry v dendrogramu (neboli stromu shluků).

Wardova metoda minimalizuje heterogenitu uvnitř jednotlivých shluků pomocí analýzy rozptylu (Ward 1963).

> *Vyzkoušejte i další způsoby spojení Linkage: Single, Average, Weighted a Complete.*



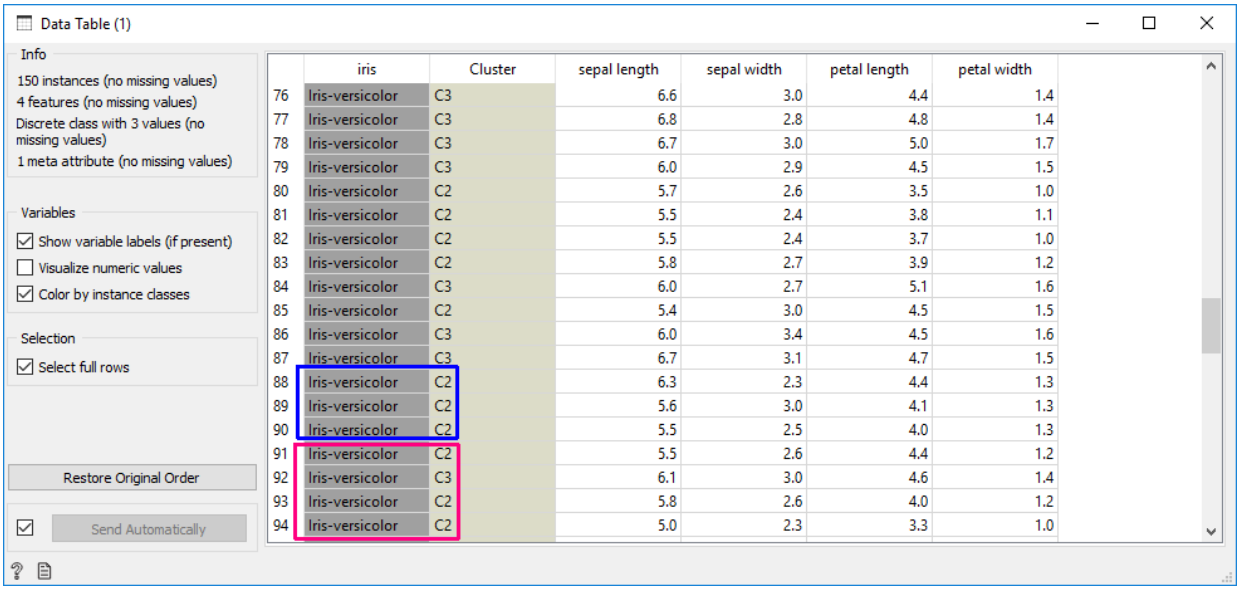
Obr. 60 Volba normalizace atributů před výpočtem vzdáleností



Obr. 61 Hierarchické shlukování Wardovou metodou a výsledný dendrogram

Pro zobrazení přiřazení do výsledných kategorií je přidán uzel *Data Table*, kde je viditelný nový sloupec *Cluster* s kategorií označenou C1, C2 nebo C3. Zjištěné shluky lze uložit do výstupního souboru XLSX pomocí uzlu *Save*.

Je zřejmé, že některé kosatce z druhé a třetího druhu nejsou zařazeny správně – porovnejte atributy *Iris a Cluster* (Obr. 62). Hierarchické shlukování je v tomto případě „učení bez učitele“, shlukování nevyužívá informaci o druhu kosatce. Je využívána jen pro závěrečné porovnání výsledku shlukování. Druh kosatce je v uzlu *File* nastaven na roli *target*.



	iris	Cluster	sepal length	sepal width	petal length	petal width
76	Iris-versicolor	C3	6.6	3.0	4.4	1.4
77	Iris-versicolor	C3	6.8	2.8	4.8	1.4
78	Iris-versicolor	C3	6.7	3.0	5.0	1.7
79	Iris-versicolor	C3	6.0	2.9	4.5	1.5
80	Iris-versicolor	C2	5.7	2.6	3.5	1.0
81	Iris-versicolor	C2	5.5	2.4	3.8	1.1
82	Iris-versicolor	C2	5.5	2.4	3.7	1.0
83	Iris-versicolor	C2	5.8	2.7	3.9	1.2
84	Iris-versicolor	C3	6.0	2.7	5.1	1.6
85	Iris-versicolor	C2	5.4	3.0	4.5	1.5
86	Iris-versicolor	C3	6.0	3.4	4.5	1.6
87	Iris-versicolor	C3	6.7	3.1	4.7	1.5
88	Iris-versicolor	C2	6.3	2.3	4.4	1.3
89	Iris-versicolor	C2	5.6	3.0	4.1	1.3
90	Iris-versicolor	C2	5.5	2.5	4.0	1.3
91	Iris-versicolor	C2	5.5	2.6	4.4	1.2
92	Iris-versicolor	C3	6.1	3.0	4.6	1.4
93	Iris-versicolor	C2	5.8	2.6	4.0	1.2
94	Iris-versicolor	C2	5.0	2.3	3.3	1.0

Obr. 62 Porovnání přiřazení do shluku C2 a C3 ve sloupci Cluser s druhem kosatce (sloupec iris)

Vyzkoušejte v uzlu **Distance** volbu *Cosine* a porovnejte výsledek hierarchického shlukování, resp. kolik kosatců je zařazeno nyní špatně. Postup lze navrhnout jako druhou větev s uzlem **Distances** a **Hierarchical Clustering** a potom je možnost porovnat dendrogramy ve dvou samostatných oknech navzájem, jak se liší přiřazení kosatců do shluků.

Evidentně druhý způsob výpočtu vzdálenosti ***Cosine*** vychází lépe než vzdálenost *Euclidean*. Kosinová vzdálenost je vhodná pro vícedimenzionální data. Kosinová podobnost bere v úvahu úhel φ mezi vektory, které směřují z počátku souřadnicového systému k bodům v prostoru (Hendl 2012) (Wikipedia 2020a). Bod reprezentuje jeden vstupní objekt v n-dimenzionálním prostoru. Vzdálenost mezi objekty je potom definována jako kosinová nepodobnost *1 – cos (φ)*. V případě, že je úhel mezi vektory dvou objektů φ = 0°, tak je cos (0) = 1, objekty jsou si podobné a vzdálenost je nula. V případě, že je úhel φ = 90°, tak je cos (90) = 0 a objekty mají vzdálenost 1, jsou si nepodobné.

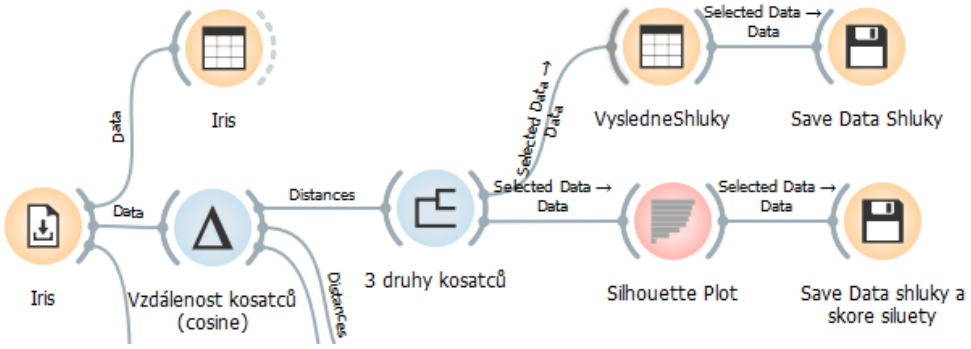
5.3 Posouzení konzistence shluku pomocí Silhouette Plot

Uzel **Silhouette Plot** nabízí graf siluety, který umožňuje posouzení konzistence jednotlivých shluků. Jedná se o grafické vyjádření kvality shluku. Číselné skóre siluety pro každou instanci je měřítkem toho, jak podobná je instance vlastnímu shluku, kam je zařazena, ve srovnání s jinými shluky. Instance jsou sestupně seřazeny v každém shluku podle svého skóre siluety. Skóre siluety blízko 1 označuje, že instance dat je blízko středu shluku. Instance mající skóre siluety blízko 0 je na hranici mezi dvěma shluky. Pokud má mnoho instancí nízkou nebo zápornou hodnotu, může být zvoleno příliš mnoho nebo příliš málo shluků a je třeba zvolit jiný počet shluků.

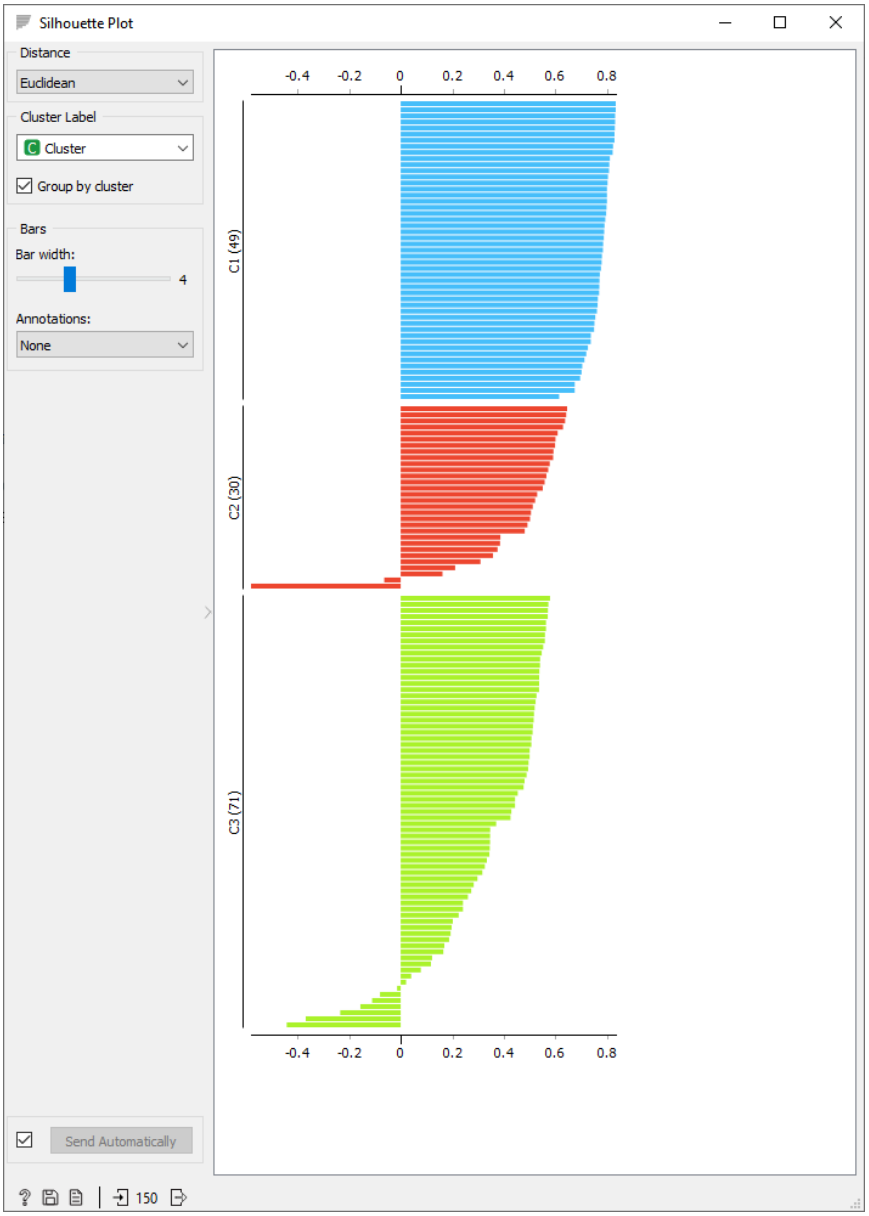
Příklad 18

Data *3_Iris.xlsx*, **workflow** *5_Iris_Hierarchical.ows*

Uzel Silhouette Plot je zařazen za uzel hierarchického shlukování (Obr. 63). Podle výběru počtu shluků je vykreslen příslušný graf siluety (Obr. 64). Tloušťka sloupce lze v grafu měnit parametrem *Bar width*. Pro kosatce je evidentní, že první shluk C1 je konzistentní a všechny kosatce jsou dosti podobné středu shluku, skóre převážně 0,8. V případě dalších dvou shluků je patrné, že někteří jedinci nejsou shluku moc podobní, resp. jsou do shluku zařazení chybně.



Obr. 63 Workflow s uzlem Silhouette Plot pro posouzení výsledku shlukování



Obr. 64 Silhouette Plot pro posouzení instancí a jejich vzdálenosti ke středu shluku

Pokud chceme posuzovat výsledek shlukování, je nutné přepnout v kolonce *Cluster Label* z volby *Iris* na volbu *Cluster*. Výsledné číselné skóre siluety včetně označení shluku C*x* lze uložit pomocí uzlu *Save Data* připojeného za uzel *Silhouette Plot*.

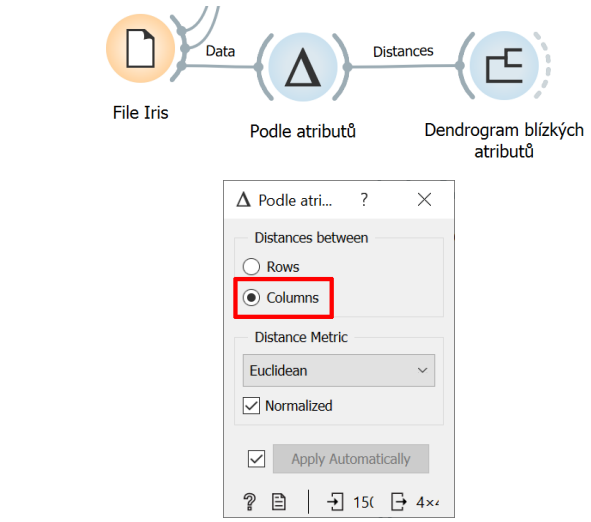
Silhouette plot lze vykreslit přímo i pro zdrojová data, když je známá kategorie jako zde, a to druh kosatce Iris (nemusí být tedy připojen jen za uzel shlukování). U jednotlivých druhů kosatců je viditelné, že někteří jedinci nemají typické rozměry pro svůj druh a mají skóre siluety nulové, či záporné. I tento uzel umožnuje nastavit různou metriku pro výpočet vzdálenosti*: Euclidean, Cosine, Manhattan.* Podle metriky se bude měnit i silueta shluků.

> *Další zdroje:*
>
> *https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/visualize/silhouetteplot.html*
>
> *https://orangedatamining.com/blog/2016/03/23/all-i-see-is-silhouette/* (Pretnar 2016a)

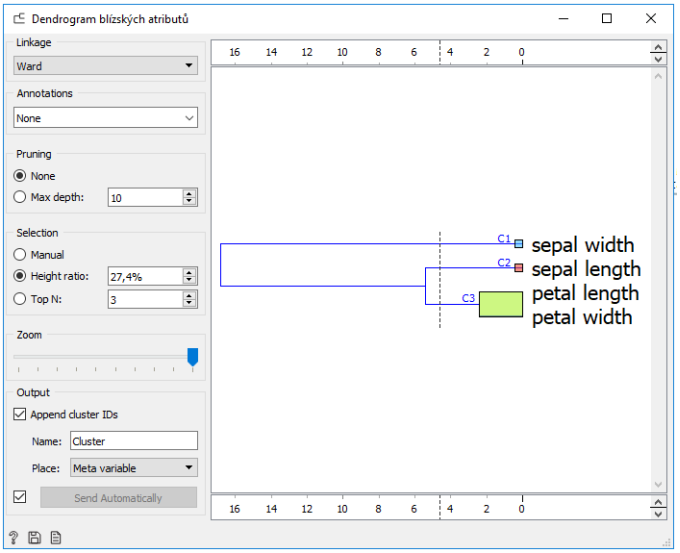
5.3.1 Shlukování podle atributů

Příklad 18 - pokračování

V uzlu *Distance* zvolíme *Columns*. Výsledný dendrogram ukáže, které atributy jsou si na 150 objektech podobnější (Obr. 66). Je zřejmé, že rozměry *Petal* jsou si podobné (blízké) a tudíž mají velký vliv na rozlišení kosatců, naopak rozměr *Sepal width* je výrazně odlišnější od ostatních rozměrů. Potvrzuje se výsledek vysoké korelace rozměrů petal z kapitoly 4.1. Oba rozměry petal také nejvíce přispívají do první komponenty PC1 (kap. 4.2). Sepal width naopak přispívá ke komponentě PC2.



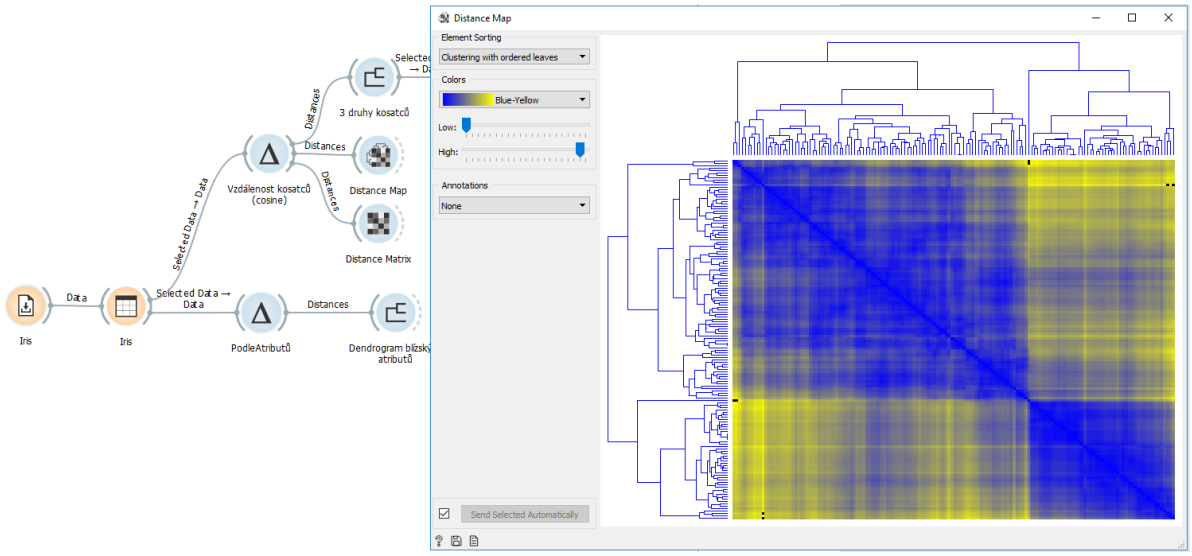
Obr. 65 Nastavení výpočtu vzdálenosti podle sloupců



Obr. 66 Výsledný dendrogram podle hodnot atributů květních lístků

Do workflow přidejte uzly **Distance Map** a **Distance Matrix**, které byly již použity a popsány v prvním příkladu kapitoly 5. 1 pro hledání podobnosti skříňových sestav. Uzly spočítají a znázorní vzdálenosti jednotlivých kosatců navzájem. Distance Map vykreslí i odpovídající dendrogram. Výslednou matici lze uložit pomocí uzlu *Save Distance Matrix*.

Přepočítání matice vzdáleností při změně metriky pro 150 instancí může chvíli trvat, což je znázorněno červenou tečkou u uzlu.



Obr. 67 Matice vzdáleností zobrazená pomocí uzlu Distance Map

5.4 Metoda k-Means

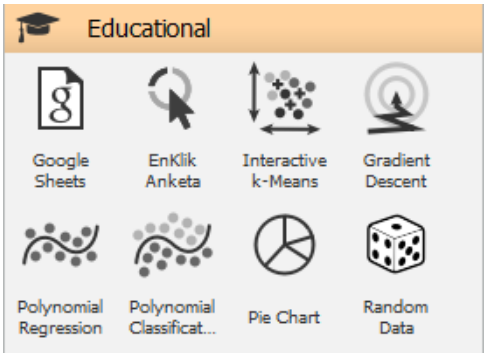
Metoda k-Means (k-středová) je nejužívanější shlukovací metodou, která používá pevně zadaný počet shluků. Algoritmus je jednoduchý, přirozený a relativně rychle konverguje. Patří mezi optimalizační metody, kdy se počáteční rozklad zlepšuje tak, aby kriteriální funkce součtu vzdálenosti objektů příslušných ke středu shluku nabyla minima (Šarmanová 2012).

5.4.1 Interaktivní k-Means

Příklad 19

Data *5_Kriminalita.xlsx (*list *TriKriminality),* **workflow** *5_InteraktivKmeans.ows*

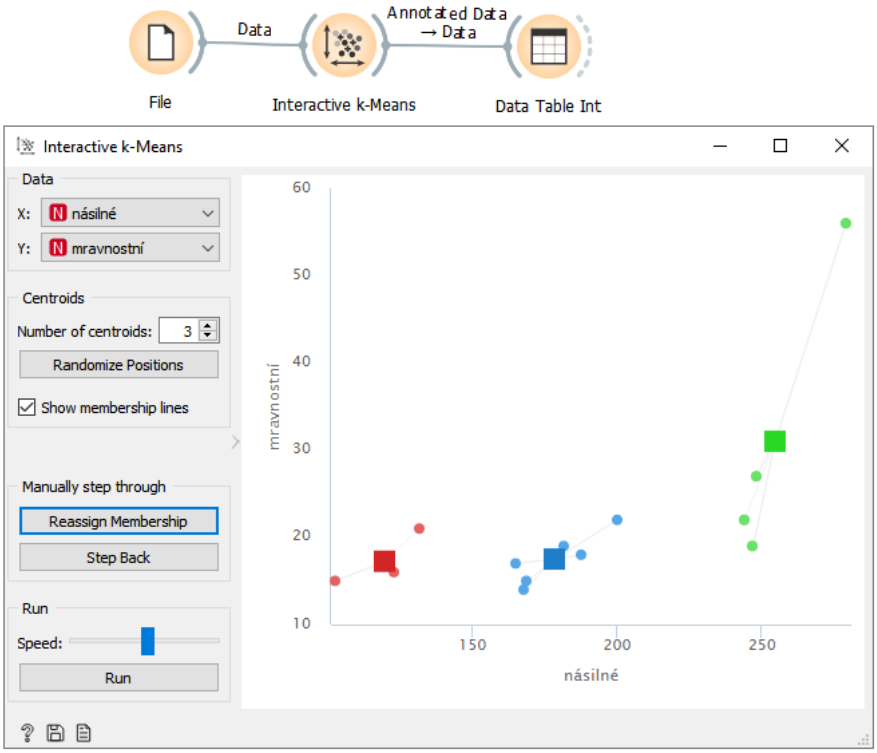
Uzel **Interactive k-Means** je uzel určený k názorné výuce metody k-Means. Tento uzel je dostupný po instalaci doplňku **Educational**. Instalace se provede volbou **Options** a **Add-ons**. Pro instalaci doplňku je třeba spustit Orange s administrátorskými právy a následně po instalaci doplňku Orange restartovat. Do levého okna rozhraní je přidána oranžová skupina Educational, kde je osm uzlů pro výuku (Obr. 68).



Obr. 68 Skupina uzlů ve skupině Educational

Uzel **Interactive k-Means** umožňuje v souřadnicích dvou vstupních atributů experimentovat s počtem shluků (Obr. 69). Lze buď náhodně generovat výchozí pozici středů, nebo myší měnit pozici výchozích středů shluků znázorněných čtverečky. Po ruční změně pozice středu lze spustit znovu přiřazení objektů do shluku stiskem *Recompute Centroids* a následně se tlačítko změní na volbu *Reassign Membership*. Tyto kroky lze několikrát zopakovat, než jsou středy a shluky stabilní. Vztažné čáry ukazují příslušnost bodů-objektů ke středu shluku. Nastavte i hodnotu třeba pěti shluků a experimentujte.

Po ustálení shluku přepněte na zobrazení dvojice jiných atributů na ose X a Y. Je viditelné, že shluky se jeví jinak.



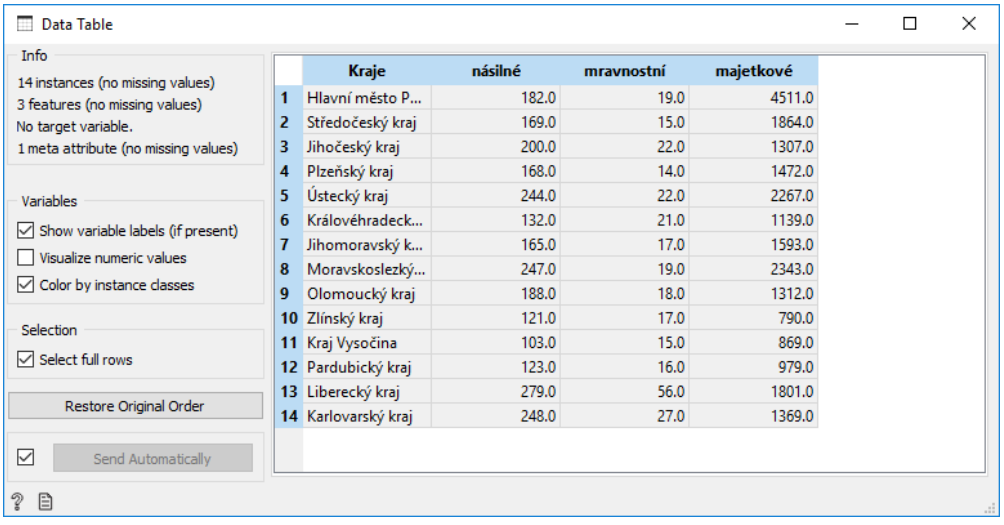
Obr. 69 Interaktivní experimentování s metodou k-Means

5.4.2 Shlukování metodou k-Means

Příklad 20

Data *5_Kriminalita.xlsx,* **workflow** *5_Kmeans_Hierarch.ows*

Data obsahují údaje o počtu trestných činů v jednotlivých krajích ČR. Vyberte list *TriKriminality*, kde jsou pouze údaje za násilné, mravnostní a majetkové trestné činy v absolutních počtech (Obr. 70). Variantně lze data přepočítat na poměrové ukazatele vydělením počtem obyvatel v jednotlivých krajích.

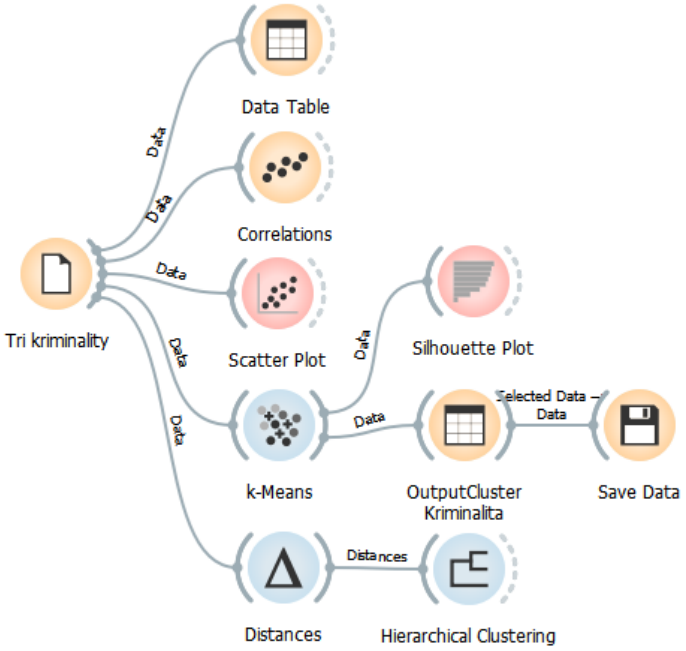


	Kraje	násilné	mravnostní	majetkové
1	Hlavní město P...	182.0	19.0	4511.0
2	Středočeský kraj	169.0	15.0	1864.0
3	Jihočeský kraj	200.0	22.0	1307.0
4	Plzeňský kraj	168.0	14.0	1472.0
5	Ústecký kraj	244.0	22.0	2267.0
6	Královéhradeck...	132.0	21.0	1139.0
7	Jihomoravský k...	165.0	17.0	1593.0
8	Moravskoslezský...	247.0	19.0	2343.0
9	Olomoucký kraj	188.0	18.0	1312.0
10	Zlínský kraj	121.0	17.0	790.0
11	Kraj Vysočina	103.0	15.0	869.0
12	Pardubický kraj	123.0	16.0	979.0
13	Liberecký kraj	279.0	56.0	1801.0
14	Karlovarský kraj	248.0	27.0	1369.0

Obr. 70 Zdrojová data kriminality

Nejprve data prozkoumáme pomocí uzlu **Scatter Plot** a **Correlation** (Obr. 71). Korelace násilných a mravnostních trestných činů je poměrně vysoká a to 0,662. Naopak ostatní korelace jsou nižší. Majetkové trestné činy jsou odlišného charakteru než mravností a násilné trestné činy a směřují proti vlastnictví (např. krádež, zpronevěra) a majetku jako celku (např. podvod). Mezi majetkové trestné činy patří i trestné činy ohrožující nehmotné statky, jako jsou osobní informace, autorská práva apod. (např. neoprávněný přístup k počítačovému systému a nosiči informací) (Policie ČR 2020). Majetkové trestné činy jsou převážně páchány jinou skupinou lidí než násilné a mravnostní činy. To vysvětluje vysokou korelaci mravnostních a násilných činů.

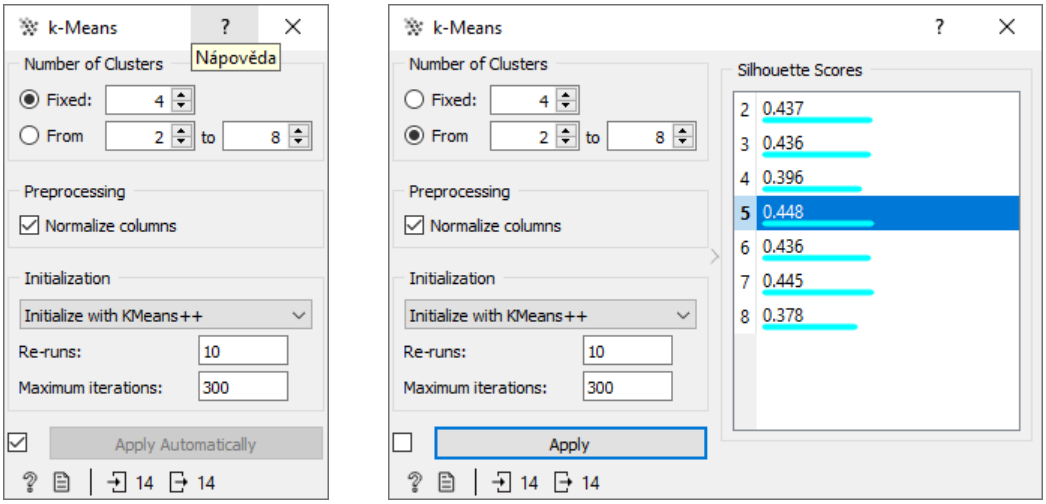
Z dat budeme zkoumat, které kraje ČR jsou si podobné počtem různých druhů kriminálních činů.



Obr. 71 Workflow s uzlem k-Means

V modrém uzlu **k-Means** se nastavuje buď pevně zvolený počet shluků (*Fixed*) nebo lze nastavit interval pro automatické vyhledání vhodného počtu shluků algoritmem (Obr. 72). Při automatickém vyhledání (*From … to*) je vhodný počet shluků automaticky zvýrazněn v okně vpravo s maximálním údajem skóre siluety *Silhouette Scores*. Pro více než 5 000 objektů není Silhouette skóre počítáno z důvodu náročnosti výpočtu.

Kromě zadání počtu shluků je nutné u této metody určit **počáteční centroidy shluků**. K tomu slouží volba ***Initialization***, kde lze zvolit *Initialize with KMeans++* nebo *Random initialization*. Inicializace *KMeans++* vybere první střed náhodně a následný střed je zvolen ze zbývajících bodů s pravděpodobností úměrnou druhé mocnině vzdálenosti od nejbližšího středu. Volba *Random initialization* nejprve přiřadí středy náhodně a poté jsou aktualizovány dalšími iteracemi.

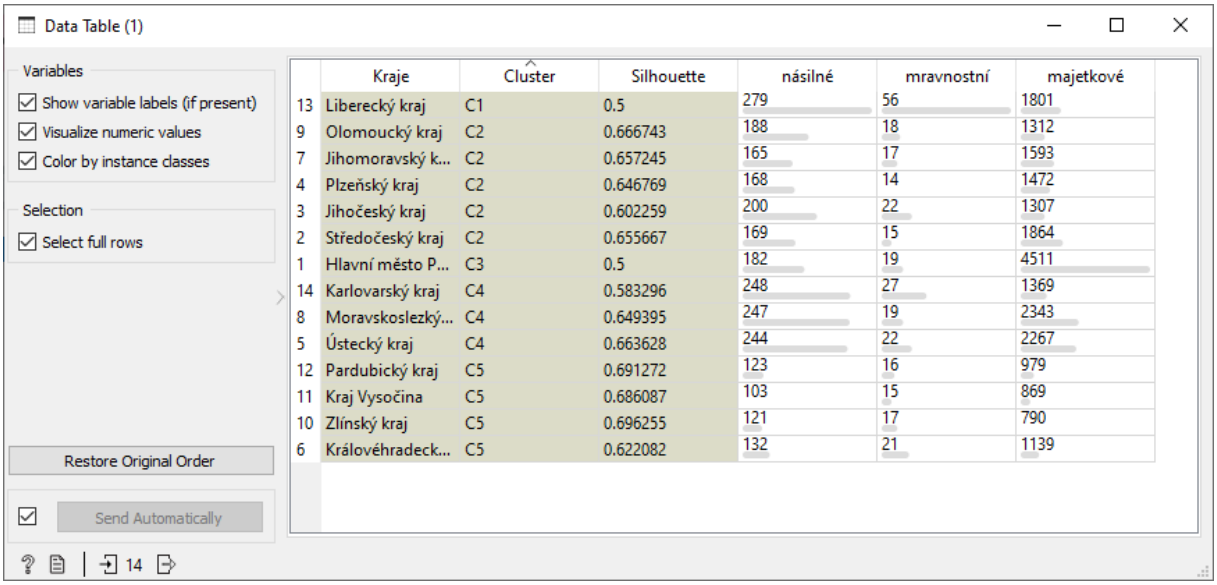


Obr. 72 Nastavení uzlu k-Means na pevný počet shluků (vlevo), vyhodnocení vhodného počtu shluků pomocí Silhoutte Scores (vpravo)

Lze také experimentovat s hodnotou ***Re-runs,*** která udává kolikrát je algoritmus spuštěn z náhodných počátečních středů; následně se použije výsledek s nejnižším součtem čtverců v rámci shluku. Hodnota ***Maximum iterations*** nastavuje ručně maximální počet iterací rámci každého spuštění algoritmu.

Po tyto data je nutné zatrhnout volbu *Normalize columns* v sekci Preprocessing.

Výsledné shluky krajů určené metodou k-Means lze vidět ve sloupci *Cluster* v uzlu *Data Table* připojené za uzel *k-Means* (Obr. 73). Seřaďte shluky podle označení shluku C*x* ve sloupci *Cluster*. Praha je evidentně vždy sama v samostatném shluku C3 z důvodu výrazně vyšších hodnot majetkových trestných činů. Také Liberecký kraj vytváří jeden shluk C1, patrně z důvodu vysokého počtu mravnostních trestných činů. Zajímavá je potom podobnost jednotlivých krajů. Shlukování je možné spouštět opakovaně s různými volbami inicializace a pozorovat vliv na výsledek shlukování.

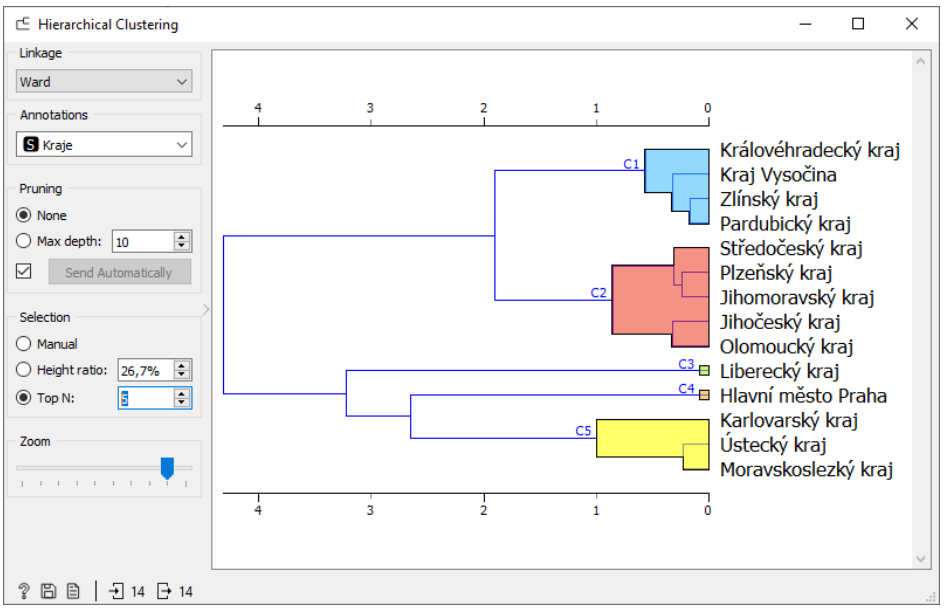


	Kraje	Cluster	Silhouette	násilné	mravnostní	majetkové
13	Liberecký kraj	C1	0.5	279	56	1801
9	Olomoucký kraj	C2	0.666743	188	18	1312
7	Jihomoravský k...	C2	0.657245	165	17	1593
4	Plzeňský kraj	C2	0.646769	168	14	1472
3	Jihočeský kraj	C2	0.602259	200	22	1307
2	Středočeský kraj	C2	0.655667	169	15	1864
1	Hlavní město P...	C3	0.5	182	19	4511
14	Karlovarský kraj	C4	0.583296	248	27	1369
8	Moravskoslezský...	C4	0.649395	247	19	2343
5	Ústecký kraj	C4	0.663628	244	22	2267
12	Pardubický kraj	C5	0.691272	123	16	979
11	Kraj Vysočina	C5	0.686087	103	15	869
10	Zlínský kraj	C5	0.696255	121	17	790
6	Královéhradeck...	C5	0.622082	132	21	1139

Obr. 73 Výsledek shlukování krajů podle trestných činů metodou k-Means

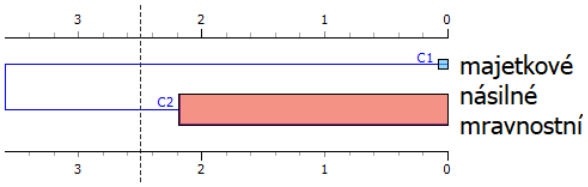
> *Graf siluety lze vykreslit pomocí uzlu Silhouette Plot i pro výsledek shlukování metodou k-Means (nejen pro hierarchické shlukování). Viz uzel ve workflow na Obr. 71.*

Následně můžeme zkusit ještě hierarchické shlukování těchto dat a porovnat výsledky metody k-Means a hierarchického shlukování (Obr. 74). Pro volbu *Distances* před hierarchickým shlukováním volte *Euclidean*.



Obr. 74 Hierarchické shlukování do pěti shluků

Vyzkoušejte shlukování podle atributů jako v předchozím příkladu pro kosatce (volba *Columns* v uzlu *Distance*). Toto shlukování nám ukáže blízkost atributů, tedy druhů trestných činů. Výsledek odpovídá zmíněnému faktu v úvodu příkladu o různých skupinách osob, které páchají trestné činy. Dále to lze interpretovat tak, že majetkové činy jsou páchány převážně v jiných krajích než násilné a mravnostní činy a naopak.



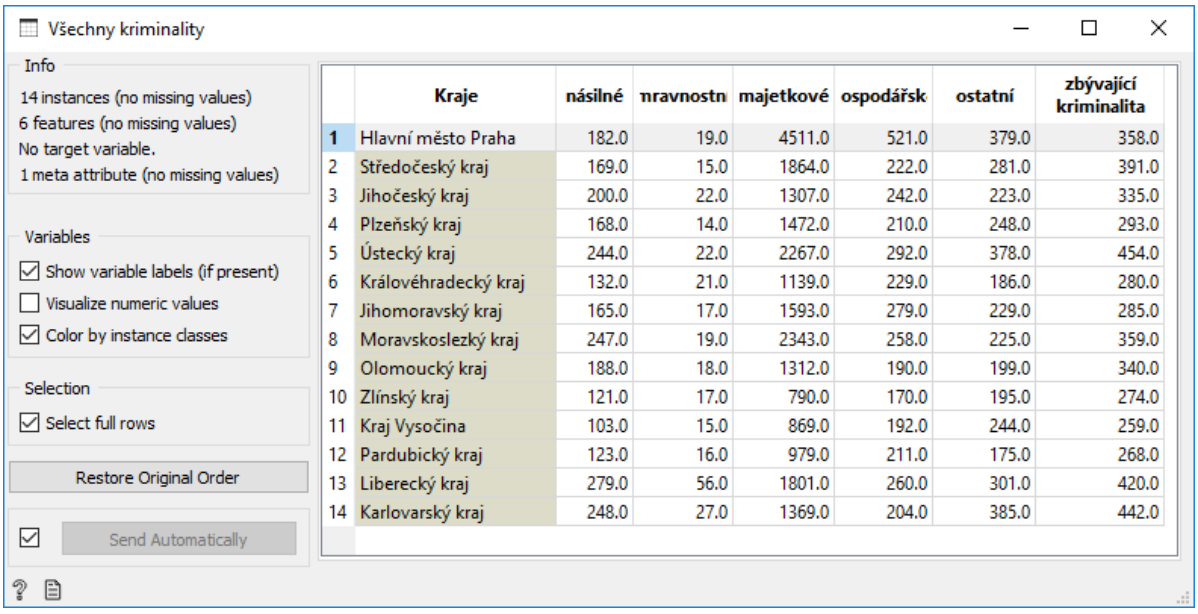
Obr. 75 Shlukování podle druhů trestných činů

5.4.3 Shlukování po redukci dimenzí

Příklad 21

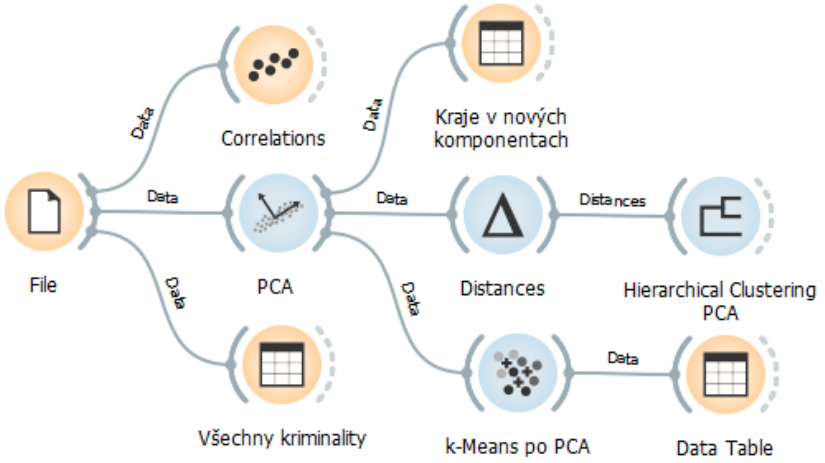
Data *5_Kriminalita.xlsx,* **workflow** *5_Kmeans_Kriminalita.ows*

Zdrojová data *5_Kriminalita.xlsx* obsahují list *VsechnyDruhyKriminality,* kde je uvedeno více druhů kriminálních činů než na listu *TriKriminality*. Zpracování je realizováno jako druhé workflow ve stejném souboru. Opět zjistíme korelaci jednotlivých kriminalit. Zde je nejvyšší korelace mezi hospodářskými a majetkovými trestnými činy – 0,947. Nad takto silně korelovanými atributy nelze přímo provést shlukování. Proto provedeme redukci dimenzí pomocí uzlu PCA na tři hlavní komponenty (Obr. 77). Následně provedeme hierarchické shlukování (Obr. 78) nebo použijeme metodu k-Means, kde nejlepší siluetu mají 3 shluky.

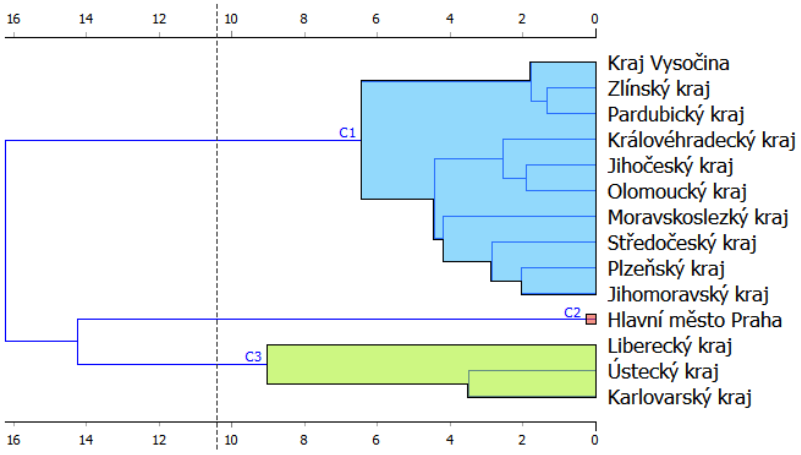


	Kraje	násilné	nravnostn	majetkové	ospodářsk	ostatní	zbývající kriminalita
1	Hlavní město Praha	182.0	19.0	4511.0	521.0	379.0	358.0
2	Středočeský kraj	169.0	15.0	1864.0	222.0	281.0	391.0
3	Jihočeský kraj	200.0	22.0	1307.0	242.0	223.0	335.0
4	Plzeňský kraj	168.0	14.0	1472.0	210.0	248.0	293.0
5	Ústecký kraj	244.0	22.0	2267.0	292.0	378.0	454.0
6	Královéhradecký kraj	132.0	21.0	1139.0	229.0	186.0	280.0
7	Jihomoravský kraj	165.0	17.0	1593.0	279.0	229.0	285.0
8	Moravskoslezký kraj	247.0	19.0	2343.0	258.0	225.0	359.0
9	Olomoucký kraj	188.0	18.0	1312.0	190.0	199.0	340.0
10	Zlínský kraj	121.0	17.0	790.0	170.0	195.0	274.0
11	Kraj Vysočina	103.0	15.0	869.0	192.0	244.0	259.0
12	Pardubický kraj	123.0	16.0	979.0	211.0	175.0	268.0
13	Liberecký kraj	279.0	56.0	1801.0	260.0	301.0	420.0
14	Karlovarský kraj	248.0	27.0	1369.0	204.0	385.0	442.0

Obr. 76 Zdrojová dat všech druhů trestných činů



Obr. 77 Hierarchické shlukování krajů podle druhů kriminálních činů po transformaci pomocí PCA



Obr. 78 Dendrogram krajů po redukci dimenzí pomocí PCA pro více druhů trestných činů

5.5 Hledání nejbližšího souseda

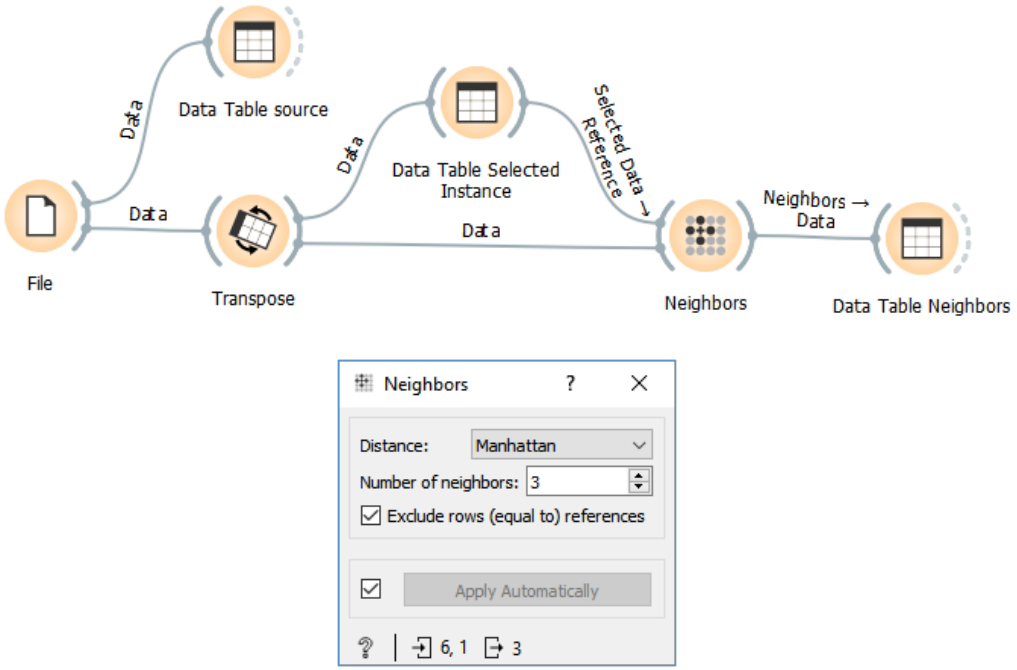
Příklad 22

Data *5_LanduseCount.xlsx,* **workflow** *5_NeighborLandUse.ows*

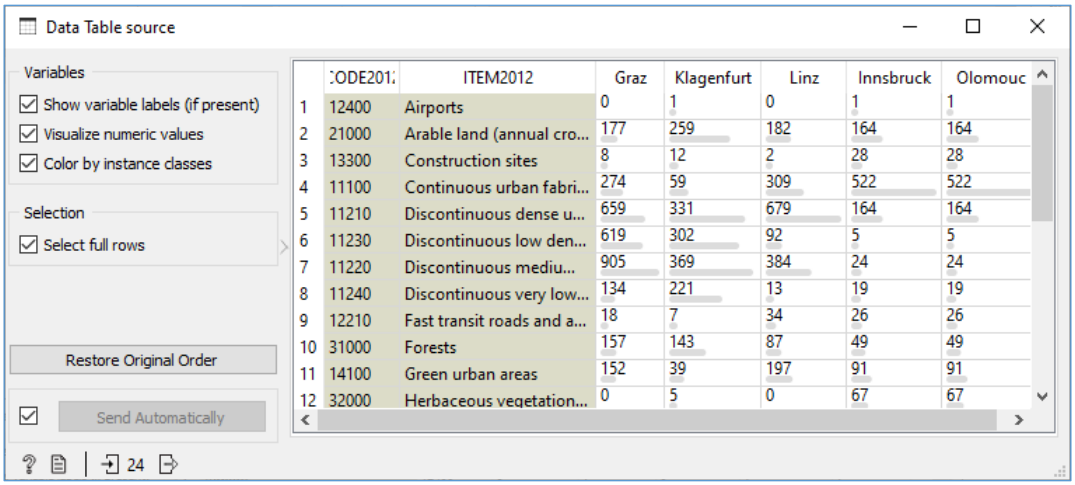
Uzel **Neighbors** hledá z množiny objektů nejbližší objekt k zadanému vzorovému objektu. Uzel má dva vstupy, první vstup obsahuje vzorový objekt, která je určena výběrem jednoho řádku v tabulce. K tomu objektu se hledají nejbližší sousedi. Druhý vstup obsahuje tabulku s více objekty, mezi kterými se hledají nejbližší sousedi.

V uzlu **Neighbors** se nastavuje požadovaných počet hledaných sousedů, kolonka *Number of neighbours* (Obr. 79). Může se hledat pouze jeden soused nebo i více sousedů. Důležité je použít vhodnou metriku *Distance*. Zde je zvolena metrika **Manhattan**, která je vhodná pro vícedimenzionální data. Další možnosti výpočtu vzdálenosti jsou: Euclidean, Mahanalobis, Cosine, Jaccard, Spearman, Absolute Spearman a Absolute Pearson.

Zdrojová data obsahují několik evropských měst. Ke každému městu máme údaj o počtu polygonů podle jednotlivých kategorií využití území – landuse (Airports, Arable land, Urban fabric, ...). Zdrojová data byla získána z Copernicus Urban Atlas (Copernicus Programme 2020). Z prostorových vrstev byly následně pro jednotlivé kategorie napočítány počty polygonů podle typu landuse. Některé hodnoty jsou nulové. Například město Linz má nulovou hodnotu v kategorii Airports, neboť letiště se nachází až za hranicí města. Protože hledáme podobná města je nejprve nutné data transponovat tak, aby jeden řádek představoval jedno město. Sloupce jsou potom jednotlivé kategorie landuse.

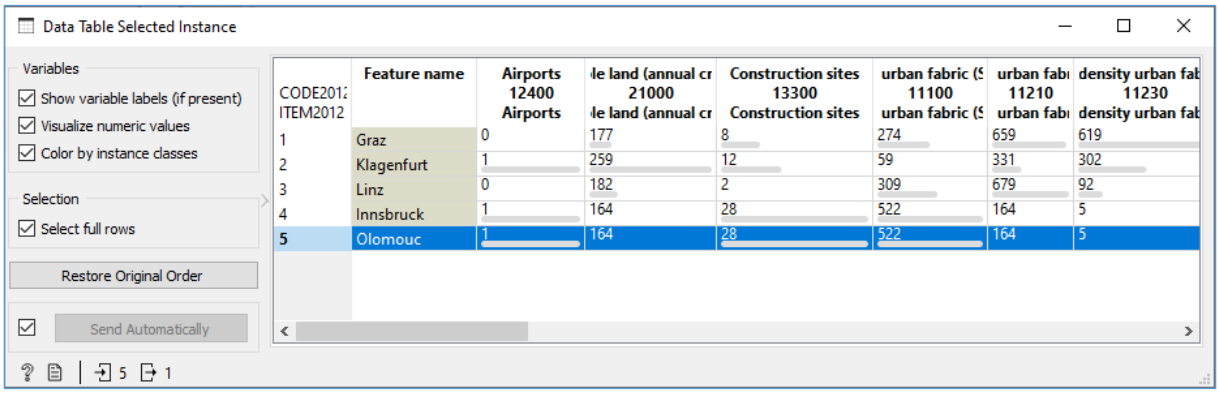


Obr. 79 Workflow s uzlem Neighbors pro zjištění nejbližších instancí



	CODE201	ITEM2012	Graz	Klagenfurt	Linz	Innsbruck	Olomouc
1	12400	Airports	0	1	0	1	1
2	21000	Arable land (annual cro...	177	259	182	164	164
3	13300	Construction sites	8	12	2	28	28
4	11100	Continuous urban fabri...	274	59	309	522	522
5	11210	Discontinuous dense u...	659	331	679	164	164
6	11230	Discontinuous low den...	619	302	92	5	5
7	11220	Discontinuous mediu...	905	369	384	24	24
8	11240	Discontinuous very low...	134	221	13	19	19
9	12210	Fast transit roads and a...	18	7	34	26	26
10	31000	Forests	157	143	87	49	49
11	14100	Green urban areas	152	39	197	91	91
12	32000	Herbaceous vegetation...	0	5	0	67	67

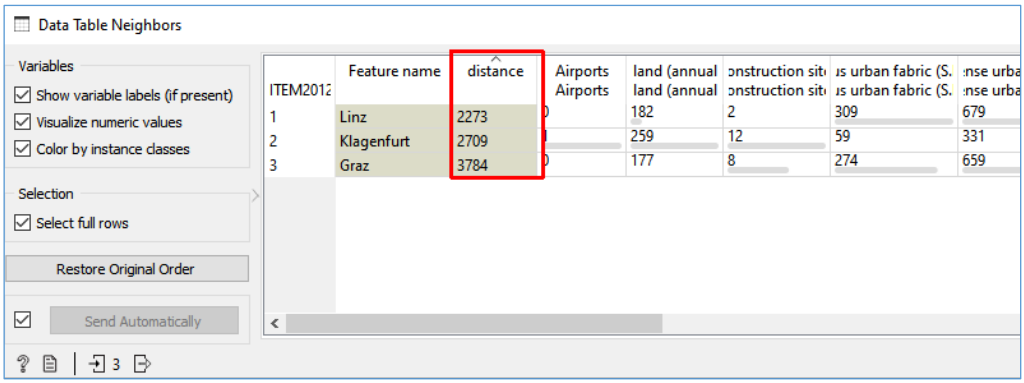
Obr. 80 Zdrojová data před transponováním



CODE2012 ITEM2012	Feature name	Airports 12400 Airports	le land (annual cr 21000 le land (annual cr	Construction sites 13300 Construction sites	urban fabric (S 11100 urban fabric (S	urban fabı 11210 urban fabı	density urban fab 11230 density urban fab
1	Graz	0	177	8	274	659	619
2	Klagenfurt	1	259	12	59	331	302
3	Linz	0	182	2	309	679	92
4	Innsbruck	1	164	28	522	164	5
5	Olomouc	1	164	28	522	164	5

Obr. 81 Výběr vzorové instance, pro kterou je hledán nejbližší soused

První vstup do uzlu Neighbors je tabulka se vzorovým objektem, který je zadaný výběrem řádku (Obr. 79). Druhý vstup je stejná tabulka, kde jsou všechna města. Výsledek je dobře viditelný v uzlu *Data Table* za uzlem *Neighbors*. V tomto uzlu je seznam sousedních (podobných) objektů, kde je i údaj o spočítané vzdálenosti ve sloupci ***distance***. Podobná města je možné seřadit podle vzdálenosti, aby bylo zřejmé pořadí nejpodobnějších měst.



Obr. 82 Výsledek hledání sousedů s údajem spočítané vzdálenosti

Pokud v dialogu Neigbors odebereme zatržítko *Exclude rows equal to references,* bude ve výsledné tabulce uveden i vzorový objekt a vzdálenost k tomuto objektu bude 0 (Obr. 82). Výhoda zobrazení jak vzorového objektu, tak nejbližších objektů je možnost porovnat hodnoty dílčích atributů, což může být nápomocné při interpretaci výsledků.

Samostatně zpracujte hledání podobného státu EU pro data *3_EU_Transport2018.xlsx*. Zde je nutná nejprve standardizace a normalizace vstupních dat pro údaj délky dálnic a počtu osobních automobilů. Bylo by vhodné vypustit ze zpracování Lotyšsko (Latvia). Má nulovou délku dálnic podle statistik Eurostat, což není ve skutečnosti pravda. Použijte workflow *5_NeighborTransport.ows.*

5.6 Metoda DBSCAN

Metoda DBSCAN (Density-Based Spatial Clustering with Noise) je založena na neparametrickém algoritmu shlukování založeném na hustotě. Seskupuje objekty (reprezentovány body), které jsou těsně spojeny dohromady vzhledem k celé sadě objektů v určitém prostoru. Body, které leží samy v regionech s nízkou hustotou (jehož nejbližší sousedé jsou příliš daleko), jsou označeny jako odlehlé body – šum.

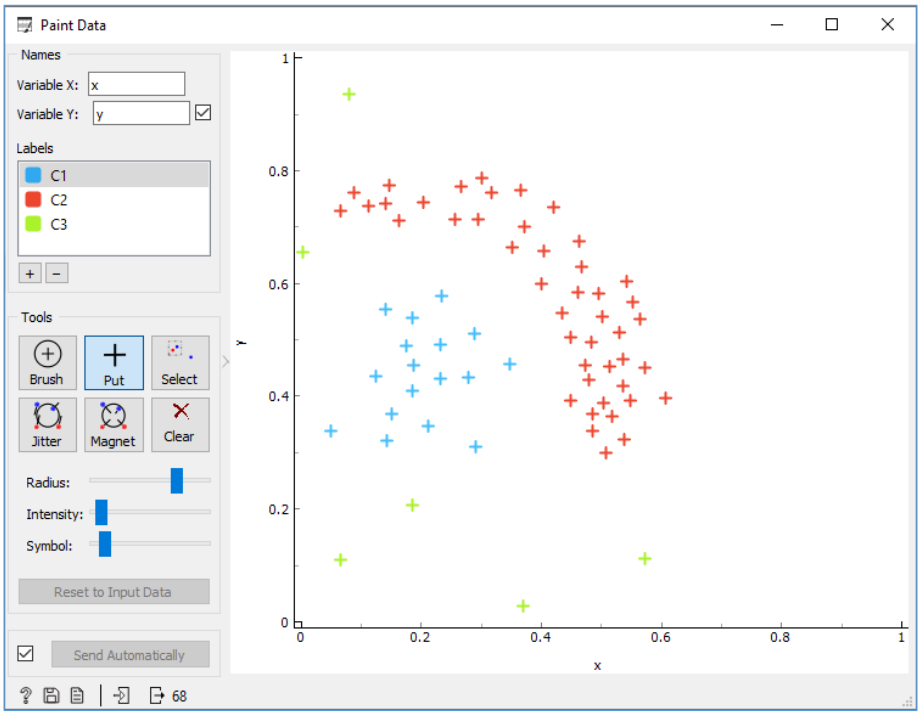
Shlukování na vstupu uvažuje parametr ε, který určí poloměr sousedství vzhledem k nějakému bodu. Druhý parametr je počet okolních bodů P. Bod je označen jako **základní bod**, pokud v jejich **okolí ε** je dosažitelné minimálně **P bodů**. Tyto body jsou označeny jako **jádrové body**. Další body se označují jako dosažitelné, pokud existuje cesta mezi body, kdy je splněn poloměr sousedství mezi jednotlivými body na cestě. O bodech se říká, že jsou přímo dosažitelné ze základních bodů. Body, které nesplňují podmínku počtu bodů v okolí, jsou okrajové body shluku (Wikipedia 2020b).To znamená, že počáteční bod a všechny body na cestě se stanou jádrovými body, patřící do stejného shluku. Všechny body nedosažitelné z jiného bodu jsou odlehlé hodnoty nebo šum.

Tento algoritmus dobře odlišuje shluky, které nejsou lineárně separovatelné. Problémem algoritmu je, pokud různé shluky mají rozdílnou hustotu. Algoritmus očekává stejnou hustotu objektů ve všech shlucích. Existují vylepšení základního algoritmu.

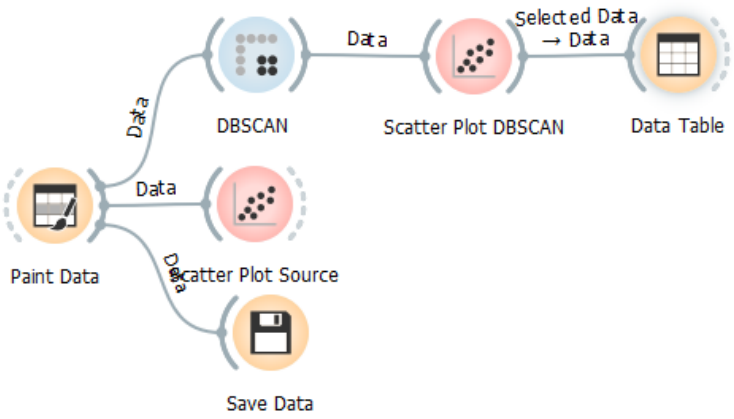
Příklad 23

Data *5_PaintedData.xlsx,* **workflow** *5_DBSCAN.ows*

Pro účely vyzkoušení této metody jsou vytvořena fiktivní cvičná data pomocí uzlu ***Paint Data***. Vytvoříme obloukový shluk **C1** červených bodů, kulový shluk **C2** modrých bodů a třetí kategorii **C3** odlehlých zelených bodů v souřadnicích X a Y v intervalu <0, 1> jako na Obr. 83. Budeme zjišťovat, zda se metodě DBSCAN podaří správně separovat shluky ve cvičných datech. Shluky bodů nejsou lineárně separovatelné přímkou.



Obr. 83 Cvičná data vytvořená pomocí uzlu Paint Data

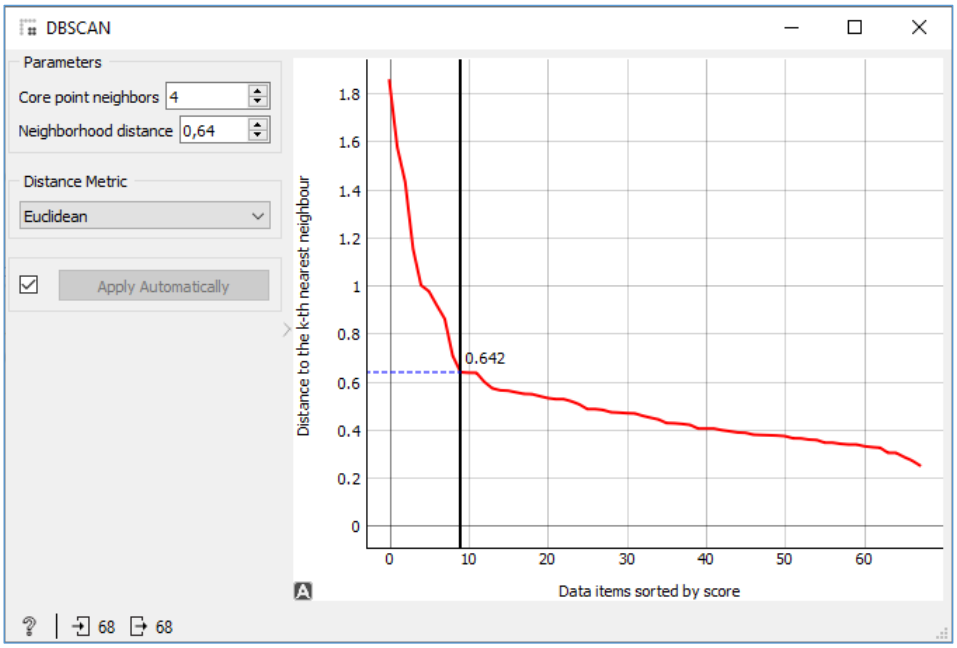


Obr. 84 Workflow s uzlem DBSCAN

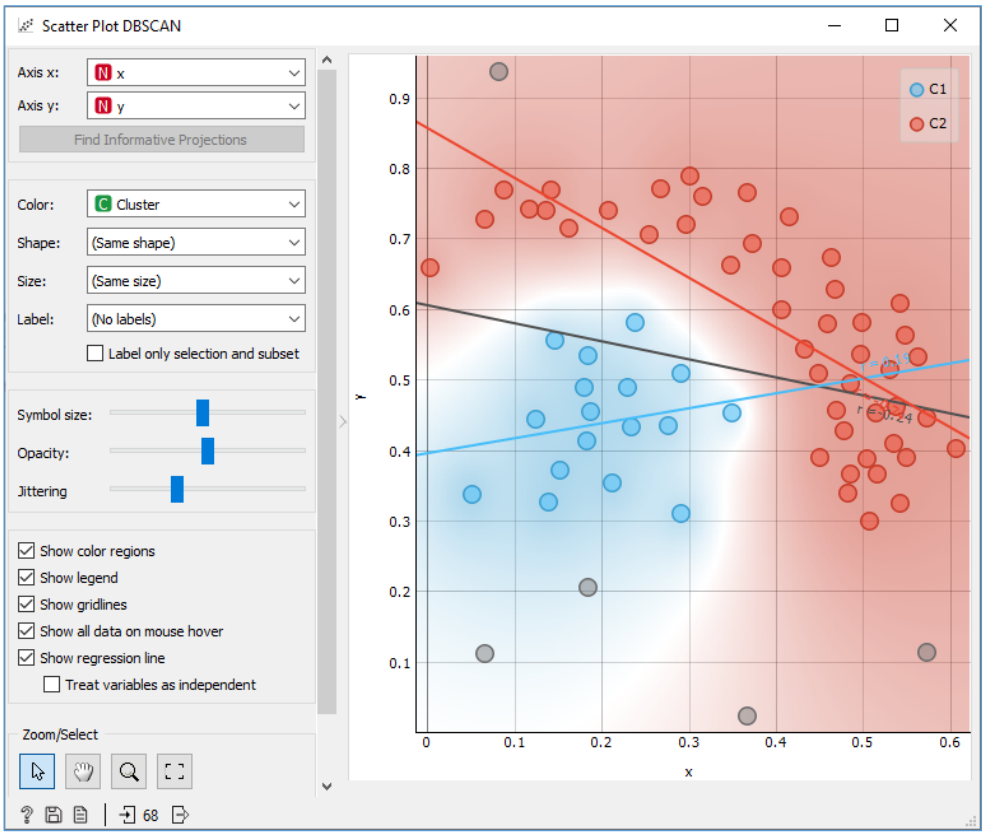
Uzel **DBCAN** ukazuje seřazený graf vzdáleností pro vybraný počet sousedů **P** (Obr. 85). Posunem černé čáry se nastavuje poloměr vzdálenosti **ε** k nejbližším jádrovým bodům. Můžeme experimentovat jak s počtem bodů, tak vzdáleností. Doporučuje se zvolit první výrazný zlom na sestupné křivce.

Výsledek shlukování lze průběžně ověřovat pomocí uzlu **Scatter Plot**, kdy se nastaví obarvení bodů (volba Color) podle nově spočítaného atributu **Cluster** (Obr. 86). Barevně jsou zobrazeny i regiony a regresní přímky každého shluku. Šedé body jsou odlehlé body, které se neshlukují, tj. šumové body. Výsledek lze porovnat s rozptylovým grafem vstupních hodnot. Metoda DBSCAN celkem správně vymezila příslušnost bodů ke shlukům.

Data v uzlu *Data Table* obsahují i nový sloupec **Cluster** a navíc sloupec **DBSCAN Core** s hodnotou, zda je bod ve výsledku jádrový nebo ne (hodnota 0/1).



Obr. 85 Nastavení uzlu DBSCAN



Obr. 86 Výsledek shlukování metodou DBSCAN

5.7 Kohonenova mapa – SOM

Úlohu shlukování lze řešit také pomocí neuronové sítě. Teuvo Kohonen v roce 1982 popsal somoorganizující síť (Self Organizing Map – SOM), která se skládá pouze ze dvou vrstev neuronů. První vrstva jsou vstupní neurony a druhá je výstupní vrstva neuronů. Každý neuron vstupní vrstvy je spojen se všemi neurony výstupní vrstvy. Díky složitosti spojení dokáže neuronová síť najít i složitější než lineární vztahy. Bohužel, z naučené neuronové sítě není jednoduché získat interpretaci či získat jednoduchý předpis závislosti mezi vstupními – nezávislými a výstupními závislými proměnnými. Toto nevadí, pokud se síť využije pro predikci nových případů. Jedná se o učení bez učitele.

Příklad 24

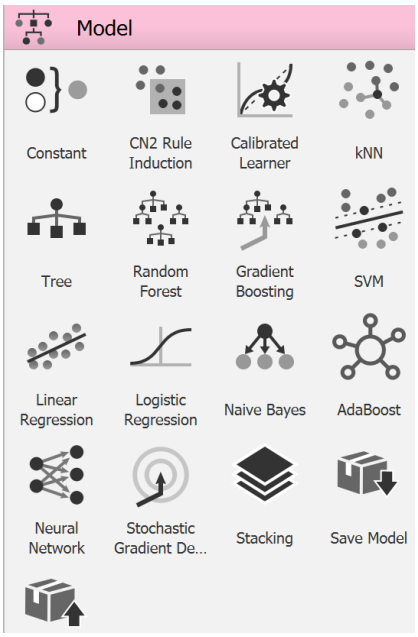
Data *6_MuzZena2020.xlsx (list Data18_2020),* **workflow** *5_SOM_MuzZena.ows*

Workflow použije uzel **Self-Organizing Map**. Ve formě hexagonální nebo čtvercové mřížky je znázorněna výstupní vrstva neuronů. Rozměr mřížky lze uživatelsky nastavit. Stisknutím tlačítka **Start** se spustí trénování sítě. Výsledek je znázorněn ve formě kruhových diagramů. Průměr kruhu znázorňuje počet případů v konkrétním shluku. V případě těchto cvičných dat známe výslednou kategorii a tak barva znázorňuje kategorii muž nebo žena, resp. výsečemi složení jednotlivých shluků. Klikem lze vybrat konkrétní shluk (buňka se vybarví modře) a v následující tabulce *Data Table* se zobrazí odpovídající vstupní instance (Obr. 87).

Obr. 87 Dialog Self-Organizing Map s výslednou mřížkou – neurony a tabulka s instancemi příslušející vybranému shluku

6 PREDIKČNÍ A KLASIFIKAČNÍ MODELY

Ze vstupních dat lze vytvářet různé **modely**. Příkladem modelů je regresní model, k-NN (k-Nearest Neighbor) model, model rozhodovacího stromu, model logistické regrese, naivní Bayesův klasifikátor, model SVM (Support Vector Machine), umělá neuronová síť (ANN), hierarchické i nehierarchické shlukovací modely a řada dalších. Účelem hotových modelů, které jsou sestavené na známých či historických datech, je predikce budoucích hodnot nebo kategorií pro nová data. V rozhraní Orange je pro sestavení predikčních modelů v levém panelu samostatná sekce růžových uzlů **Model**, kde se nachází jednotlivé uzly pro tvorbu predikčních modelů (Obr. 88). V sekci **Unsupervised** jsou uzly pro sestavení modelů patřící do skupiny učení bez učitele (např. shlukování). Různé modely pracují různým způsobem a je nutné vědět, jak jsou konstruovány, co nám říkají o datech a jak je správně interpretovat.



Obr. 88 Sekce Model v rozhraní software Orange

6.1 Lineární regresní model a predikce pomocí modelu

Lineární regrese hledá pro známé dvojice pozorování $[x_i, y_i]$ parametry funkční závislosti $y = f(x)$.

Hledají se tedy parametry rovnice

$$y = q_1 x + q_0 + e \qquad (4)$$

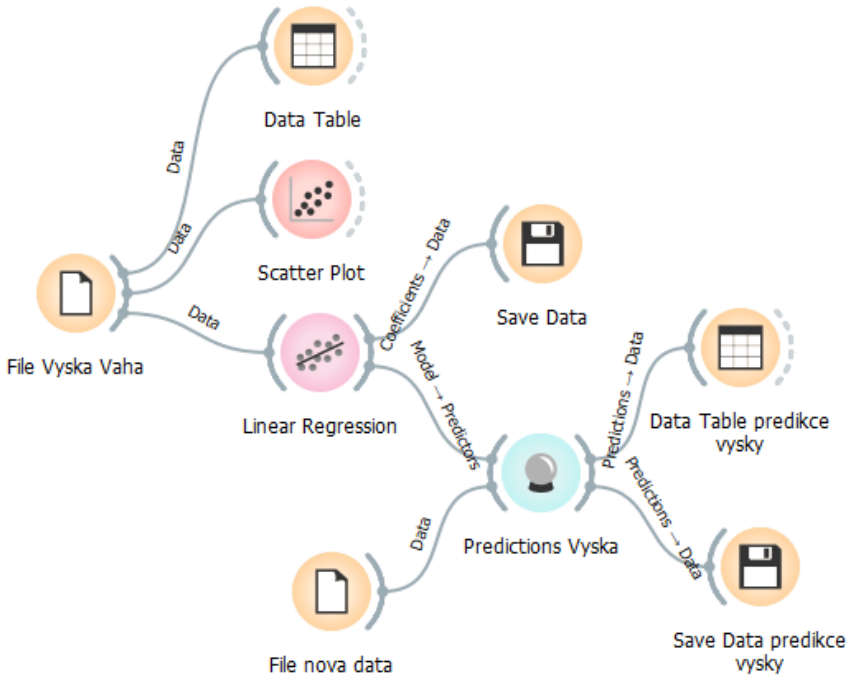
V tomto příkladu bude ukázán postup, jak sestavit lineární regresní model a následně jej použít pro predikci pomocí uzlu **Prediction**. Tento ukázkový postup a uzel **Prediction** lze použít i pro jiné modely dostupné v Orange, než je jen regresní model. Predikce je tedy ukázána na jednom z nejjednodušších modelů, a to lineárním regresním modelu.

Příklad 25

Data *6_MuzZena2020.xlsx (list Data18_2020),* **workflow** *6_PredictionRegrese.ows*

Na vstupu jsou použitá data s údajem o výšce a váze osob. Ve vstupním uzlu *File* je sloupec Výška ponechán jako *Feature*, sloupec Hmotnost je nastaven jako *Target* a sloupec Pohlaví je nastaven na roli *Skip*. Data lze vykreslit pomocí uzlu **Scatter Plot**, kdy na ose X je Výška, na ose Y je Hmotnost. Je patrné, že čím je osoba vyšší, tím má vyšší hmotnost (Obr. 90 nahoře). Je důležité zjistit odlehlé hodnoty a případně je z dat odstranit, neboť lineární regrese je na ně citlivá.

Pomocí růžového uzlu **Linear Regression** je sestaven model, který ze vstupních dat určuje lineární regresní přímku závislosti hmotnosti na výšce osoby (Obr. 89). Koeficienty lineární regrese lze uložit pomocí následného uzlu **Save Data**, kdy je do souboru uložena hodnota **koeficientu Intercept** -57,12575384 a hodnota **směrnice** regresní přímky 0,7202919 (zde pod názvem sloupce Výška).



Obr. 89 Workflow predikce hmotnosti pomocí lineární regrese

Výstup hotového modelu lineární regrese je připojen k modrému uzlu **Prediction**. Do tohoto uzlu jako druhý vstup použijte uzel **File** s novými daty (XLSX list *Neznáme pohlaví*). Pomocí modelu lineární regrese sestaveného na existujících datech se určuje hmotnost pro dvě osoby z jejich známé výšky. V datech je i uvedena jejich skutečná hmotnost. Lze tak porovnat v dialogovém okně uzlu Prediction, jak se predikce hmotnosti liší od skutečné hodnoty (Obr. 90 dole). Predikované hodnoty jsou poněkud nižší než skutečné hodnoty hmotnosti.

Výsledek predikce lze uložit pomocí uzlu **Save Data** a tak se zaznamená pro dvě nové osoby predikované závislé hodnoty hmotnosti pomocí lineárního regresního modelu pro další případné zpracování.

Obr. 90 Scatter plot vstupních dat a výsledek predikce hmotnosti pomocí modelu lineární regrese pro dvě nové osoby

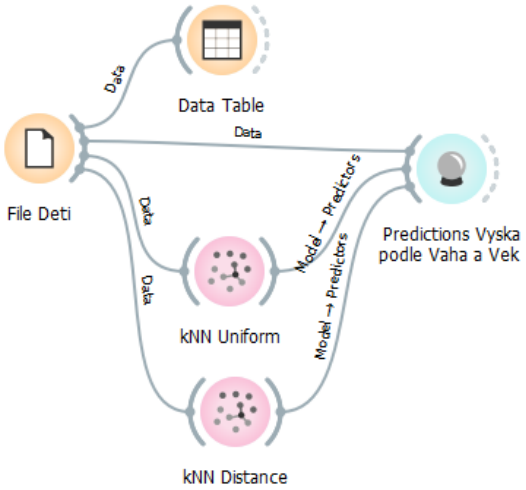
6.2 Model k-NN

Model k-Nearest Neighbors (k-nejbližších sousedů) hledá pro novou instanci výslednou kategorii nebo hodnotu podle k-nejbližších sousedů z trénovacích dat. Model může být použit jak pro klasifikaci (určení třídy) nebo regresi (určení číselné hodnoty). V případě klasifikace je výsledná kategorie určena podle převažující kategorie sousedů. V případě regrese je výsledná hodnota spočítána a predikována jako průměrná hodnota k-sousedních prvků. Jde o úlohu učení s učitelem.

Příklad 26

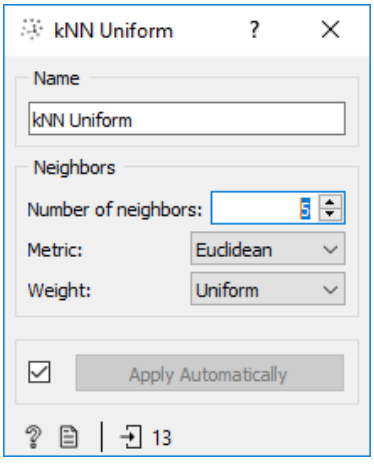
Data 3_*Deti.xlsx,* **workflow** *6_kNN.ows*

V této úloze se pokusíme predikovat výšku dítěte z jeho známé hmotnosti a věku. Nastavte v uzlu **File** atribut Vyska na roli *target*, identifikátor Dite a Nemoc na roli *skip*, zůstanou pouze dva zdrojové atributy Vek a Hmotnost v roli *feature*. Je zřejmé, že s věkem roste hmotnost a výška dítěte. Uzel kNN nepotřebuje v Orange žádné předzpracování. Kromě odstranění instancí s neznámou cílovou hodnotou, imputací hodnot průměrnou hodnotou, tak uzel zejména normalizuje a standardizuje data na standardní odchylku o velikosti 1.



Obr. 91 Workflow s uzlem kNN

Výsledek modelu k-NN závisí nejen na nastaveném počtu sousedů, ale i na nastavené metrice vzdálenosti a na váze okolních instancí. Z důvodu experimentování jsou ve workflow dva uzly **kNN** s různým nastavením váhy (Obr. 91). Volba *Metric* v uzlu kNN umožňuje nastavit metriky *Euclidean, Manhattan, Chebysev, Mahalanobis*. Vzhledem k tomu, že predikujeme jen ze dvou atributů *věk* a *hmotnost*, tak v obou uzlech je vybrána euklidovská metrika. Pro více dimenzí je vhodné vybrat jinou metriku. V obou uzlech je nastaven shodný *počet sousedů* na hodnotu **5**. V uzlu **kNN Uniform** je nastaven stejná váha *Weight* okolních instancí, tedy volba *Uniform* (Obr. 92). V uzlu **kNN Distance** je volba *Weight* nastavena na *Distance*, znamená to, že bližší sousedi mají vyšší váhu.

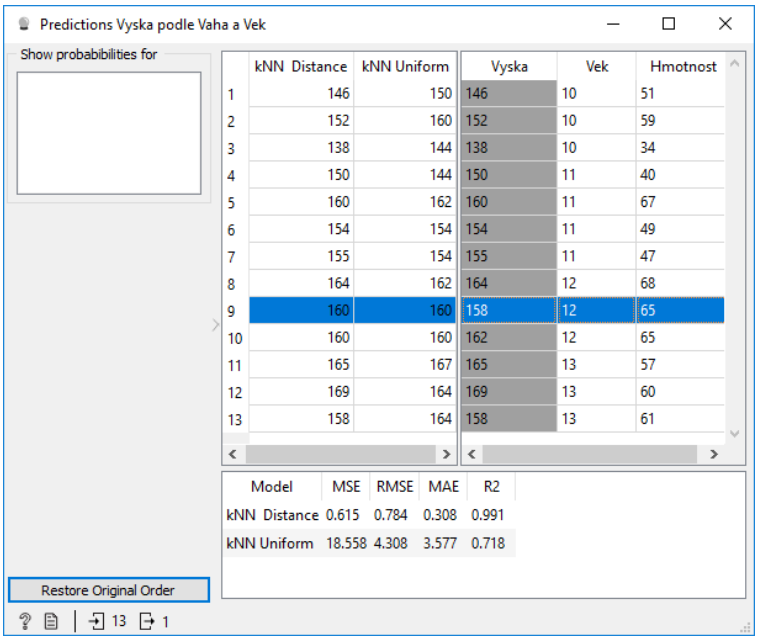


Obr. 92 Nastavení uzlu kNN Uniform

Do uzlu Prediction jsou kromě výstupu dvou modelů k-NN připojena i zdrojová trénovací data (Obr. 91). Pro predikci instance se tedy bere v úvahu 5 okolních sousedů včetně instance sebe samé. Můžeme tedy porovnat, jak se modely shodují se skutečnými hodnotami.

Pro model **kNN Distance**, kde je nastavena váha *Distance,* vidíme, že predikce výšky se ve většině případů shoduje s původní hodnotou (Obr. 93). Nejmenší vzdálenost (a tedy největší váhu) má prvek sám na sebe. Výsledek je způsoben malým počtem záznamů trénovacích dat, respektive unikátností záznamů. V datech jsou pouze dva shodné záznamy dětí číslo 9 a 10, kdy obě děti mají stejný věk 12 roků a stejnou hmotnost 65 kg, ale jinou výšku (158 a 162 cm). Pro tyto záznamy se tedy predikuje průměrná hodnota výšky 160 cm, jiná než jejich původní. Zde vidíme, že zafungoval výpočet průměrné hodnoty sousedů. Dítě číslo 9 je vyznačeno modrou barvou na Obr. 93.

Pro druhý model **kNN Uniform** je viditelné, že všech pět sousedů má stejný vliv na výpočet průměrné výšky. Predikované výšky se tedy u jednotlivých záznamů liší o několik málo centimetrů. Zdá se tedy, že model má i menší přesnost a horší charakteristiky spočítané v dolní části okna (MSE, RMSE, MAE). Výsledek je ale ovlivněn jednoduchostí vstupních dat zmíněných výše.



	kNN Distance	kNN Uniform	Vyska	Vek	Hmotnost
1	146	150	146	10	51
2	152	160	152	10	59
3	138	144	138	10	34
4	150	144	150	11	40
5	160	162	160	11	67
6	154	154	154	11	49
7	155	154	155	11	47
8	164	162	164	12	68
9	160	160	158	12	65
10	160	160	162	12	65
11	165	167	165	13	57
12	169	164	169	13	60
13	158	164	158	13	61

Model	MSE	RMSE	MAE	R2
kNN Distance	0.615	0.784	0.308	0.991
kNN Uniform	18.558	4.308	3.577	0.718

Obr. 93 Výsledek predikce dvou metod k-NN v uzlu Prediction s porovnáním skutečné výšky dětí

> *Speciální případ predikčního modelu k-NN a to zjednodušeného 1-NN je možné použít pro imputaci chybějících hodnot. Zmínku lze nalézt dříve v textu, a to v kapitole 3.7.*

6.3 Rozhodovací stromy

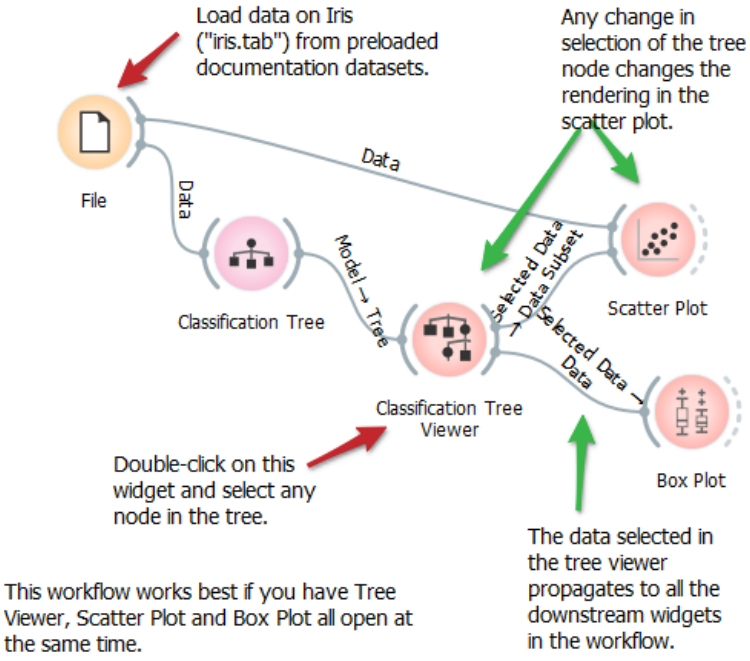
Rozlišujeme dva druhy stromů: regresní a klasifikační stromy. Určující je typ predikované veličiny. Pokud se predikuje veličina (target value) ze souvislého oboru hodnot, typicky z domény reálných čísel, pak se jedná o **regresní stromy**. Pokud se predikuje kategorie neboli diskrétní konečná množina hodnot, např. typ, druh zločinu, označuje strom jako **klasifikační strom**. Používá se společné označení **CART** (Clasification And Regression Tree) (Quinlan 1986), (Quinlan 1993). Rozhodovací stromy se používají pro klasifikaci a predikci, jedná se **o učení s učitelem** (supervised learning) (Petr 2014a). Orange pro konstrukci rozhodovacího stromu používá algoritmus, který počítá informační zisk (gain) pro kategorické veličiny a MSE (mean squared error*)* pro výsledné číselné hodnoty.

6.3.1 Strom pro určení druhu kosatce

Příklad 27

Data *Iris.tab*, **workflow** *6_Tree-scatterplot.ows (ukázkové workflow dodávané s Orange)*

Na těchto datech lze trénovat určení druhu kosatce, tzn. klasifikace (učení s učitelem). Bude vytvářen klasifikační strom, který určuje druh kosatce.

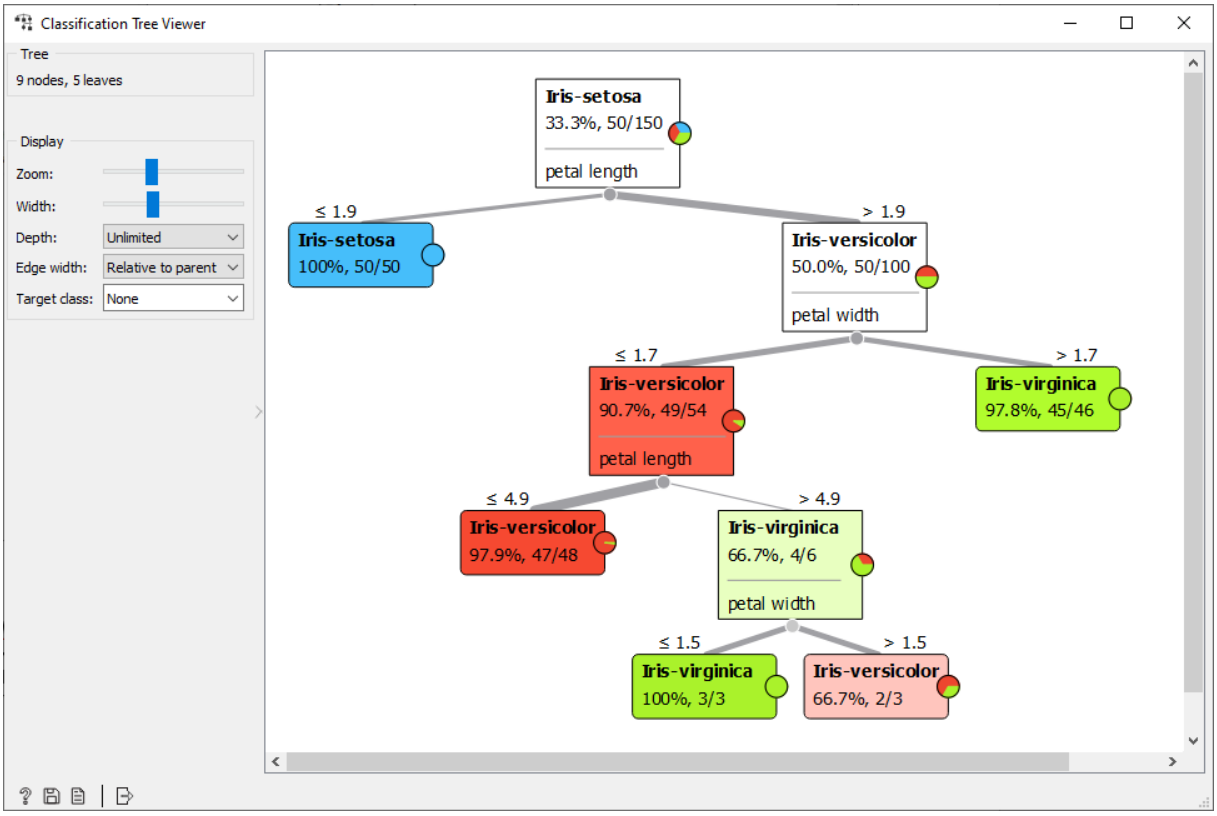


Obr. 94 Ukázkové workflow dodávané s Orange včetně komentářů

Uzel **Classification Tree** slouží k nastavení požadovaných parametrů výsledného stromu. Výsledný rozhodovací strom se zobrazí pomocí uzlu **Classification Tree Viewer**. Uzly stromu jsou znázorněny čtyřúhelníky, kde na první řádce je tučně uveden název predikované klasifikační třídy. Malé kruhové diagramy v pravé části každého uzlu naznačují, jak je každý uzel heterogenní. První nejvyšší uzel obsahuje všechny data, tudíž je uzel bílý a kruhový diagram obsahuje tři stejně velké barevné výseče. Modrý uzel vlevo *Iris-setosa* je homogenní a obsahuje 100 % kosatců stejného druhu 50/50. Pro určení tohoto druhu postačuje velikost *petal length* ≤ 1.9. Z výsledného stromu je zřejmé, že atribut *petal length* má největší vliv na určení druhu. Je dokonce jediným určujícím atributem pro *Iris-setosa*.

Také sytostí barevné výplně uzlu stromu se naznačuje úroveň homogenity uzlu (čím sytější, tím jsou data stejnorodější). Cílový uzel, který je 100% homogenní má vždy zaoblené rohy čtyřúhelníku.

Lze vybrat klikem jeden uzel nebo list ve stromě, ten se označí silným černým okrajem. Následně jsou v uzlu **Scatter plot** znázorněny objekty, které jsou zařazeny do uzlu rozhodovacího stromu. Takto lze postupně prohlížet jednotlivé listy nebo uzly stromu a jim odpovídající objekty.



Obr. 95 Rozhodovací strom určení druhu kosatce

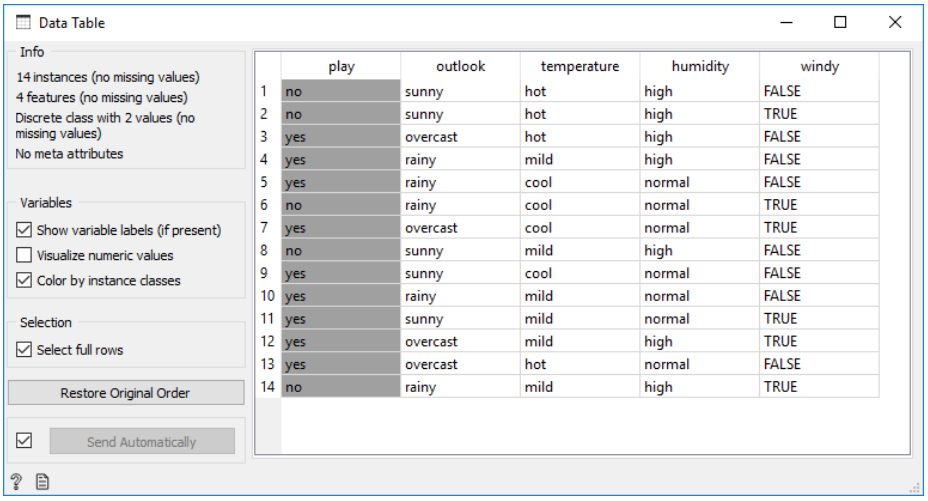
Všimněte si, že se v klasifikačním stromě nikde neuplatní pro určení druhu kosatce atributy popisující rozměry *sepal length* a *sepal width*. Dvojice čísel ukazuje počet jedinců, které odpovídají příslušnému druhu a za lomítkem je celkový počet jedinců v uzlu. Např. zelený poslední list úplně vpravo obsahuje údaj 45/46. Tzn., že podle pravidel vedoucích k tomuto listu je 45 jedinců určeno správně jako *Iris-virginica* a jeden jedinec je určen špatně a není *Iris-virginica*, celkem je tedy 46 jedinců, úspěšnost kritérií je 97,8 %.

6.3.2 Strom hraní tenisu

Příklad 28

Data *6_Tenis-nominal.xls,* **workflow** *6_Tree_Tenis.ows*

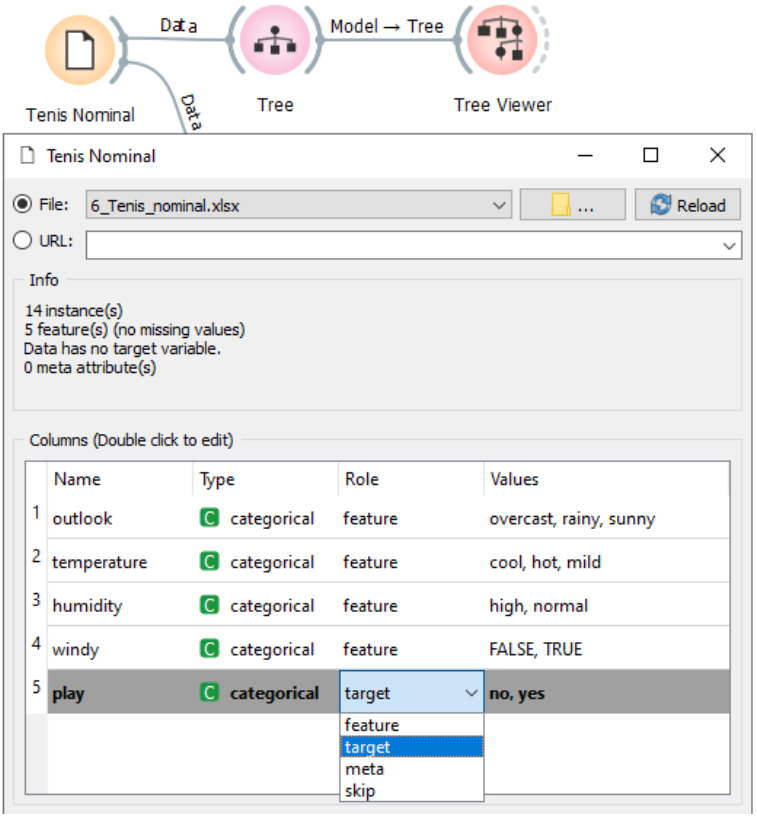
Tato data obsahují jen nominální hodnoty 14 různých situací, kdy se hrál / nehrál tenis (atribut *play*) v závislosti na různých povětrnostních podmínkách – *outlook, temperature, humidity, windy*. Obsah dat je zobrazen pomocí uzlu *Data Table*.



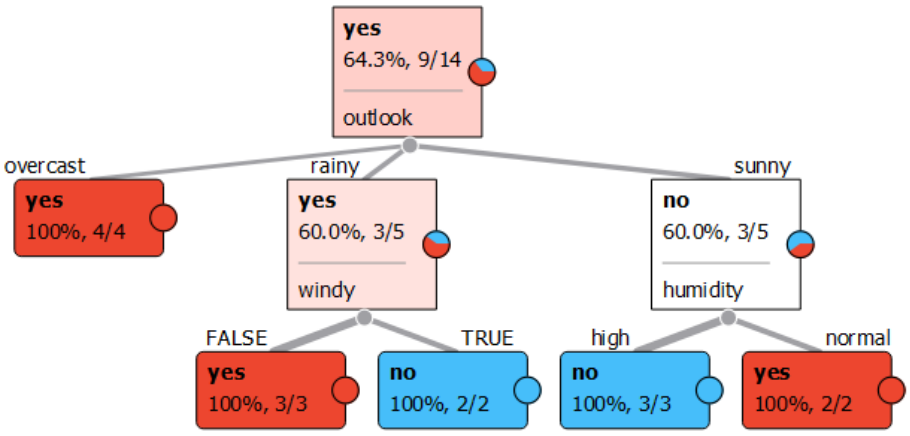
	play	outlook	temperature	humidity	windy
1	no	sunny	hot	high	FALSE
2	no	sunny	hot	high	TRUE
3	yes	overcast	hot	high	FALSE
4	yes	rainy	mild	high	FALSE
5	yes	rainy	cool	normal	FALSE
6	no	rainy	cool	normal	TRUE
7	yes	overcast	cool	normal	TRUE
8	no	sunny	mild	high	FALSE
9	yes	sunny	cool	normal	FALSE
10	yes	rainy	mild	normal	FALSE
11	yes	sunny	mild	normal	TRUE
12	yes	overcast	mild	high	TRUE
13	yes	overcast	hot	normal	FALSE
14	no	rainy	mild	high	TRUE

Obr. 96 Zdrojová data pro rozhodovací strom

Nejprve je nutné nastavit, že atribut *play* je výsledná kategorie. V uzlu *File* změníte hodnotu u atributu ***play*** v sloupci **Role** na hodnotu ***target***. Nastavení cílového klasifikačního atributu je nezbytné před konstrukcí klasifikačního stromu. Následně se přidá do workflow uzel *Tree* a *Tree Viewer*.



Obr. 97 Workflow a nastavení vstupních dat

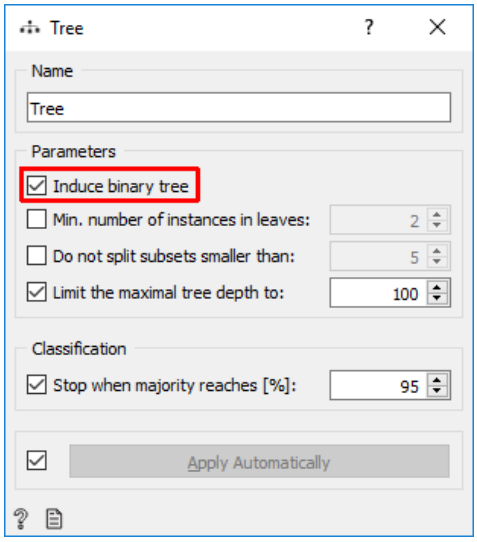


Obr. 98 Výsledný strom

Všechny listy jsou homogenní a čtverce, které je reprezentují, mají oblé roky.

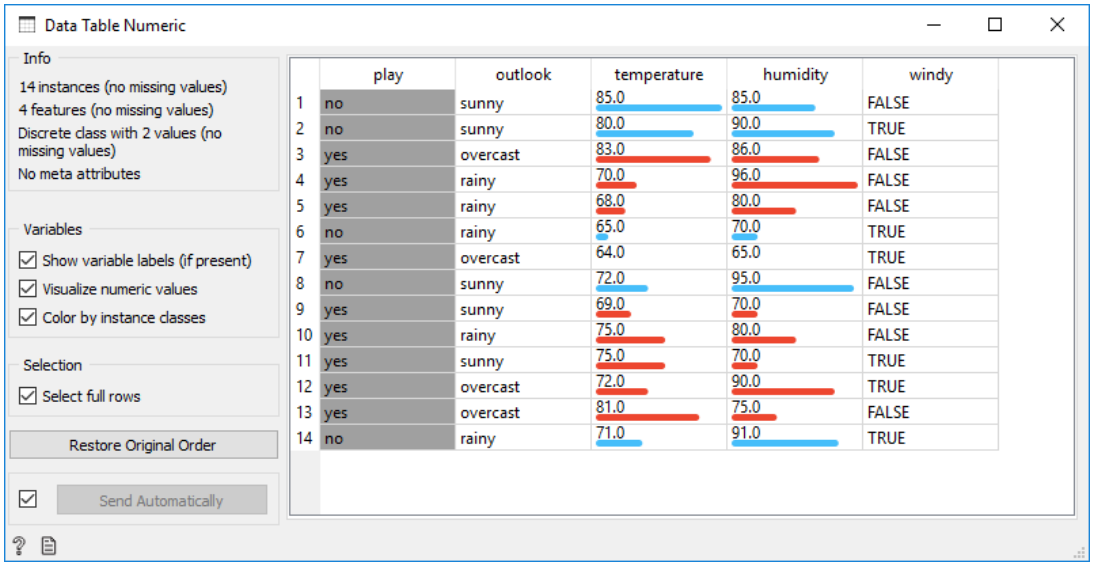
Vyzkoušejte v dialogu uzlu *Tree* vynucení volby jen binárního stromu – *Induce binary tree.* Strom se změní.

*Rozhodovací strom patří do skupiny učení s učitelem. Vstupní data je vhodné v tomto případě rozdělit na trénovací a testovací množinu, pokud je dostatečný počet instancí. Následné testování natrénovaného modelu pomocí testovacích dat zjišťuje, jak velká je chyba modelu. Uzel Tree používá celou vstupní množinu. Pokud chceme rozdělit data na trénovací a testovací množinu, tak předřadíme uzlu Tree uzel **Data Sample**, který umožnuje různé rozdělení vstupních dat. Ukázka použití a nastavení dialogu uzlu Data Sample je v kapitole 6.6, Obr. 122.*



Obr. 99 Nastavení parametrů pro vytvoření rozhodovacího stromu

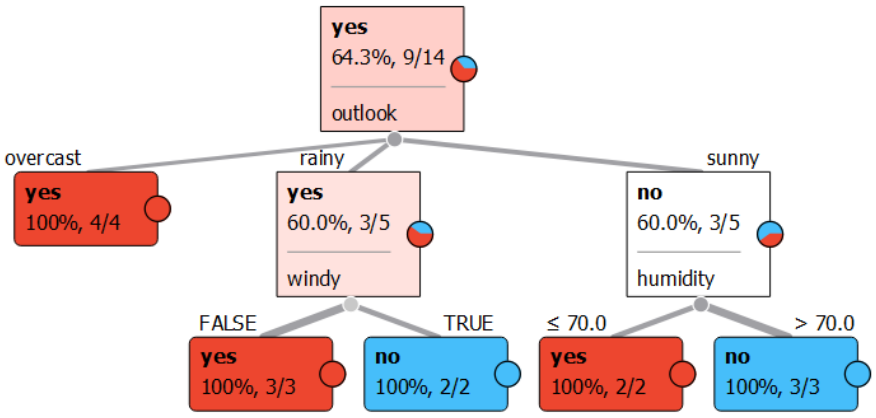
Navrhněte druhé stejné workflow, kde zdrojem dat bude soubor *6_Tenis_numeric.xls.* Zde jsou číselné údaje o teplotě (temperature) a vlhkosti (humidity). Barvy čar odpovídají klasifikační třídě *play*. Modrá barva je NO, červená YES.



	play	outlook	temperature	humidity	windy
1	no	sunny	85.0	85.0	FALSE
2	no	sunny	80.0	90.0	TRUE
3	yes	overcast	83.0	86.0	FALSE
4	yes	rainy	70.0	96.0	FALSE
5	yes	rainy	68.0	80.0	FALSE
6	no	rainy	65.0	70.0	TRUE
7	yes	overcast	64.0	65.0	TRUE
8	no	sunny	72.0	95.0	FALSE
9	yes	sunny	69.0	70.0	FALSE
10	yes	rainy	75.0	80.0	FALSE
11	yes	sunny	75.0	70.0	TRUE
12	yes	overcast	72.0	90.0	TRUE
13	yes	overcast	81.0	75.0	FALSE
14	no	rainy	71.0	91.0	TRUE

Obr. 100 Vstupní data se dvěma spojitými veličinami teplota a vlhkost

Opět se neuplatní teplota, ale objeví se rozhodovací pravidlo, které testuje vlhkost na větší nebo menší než 70 %.



Obr. 101 Výsledný rozhodovací strom pro spojité veličiny

6.3.3 Strom pro určení pohlaví z hmotnosti a výšky

Zdrojem dat je dotazník na adrese, který také můžete vyplnit: https://tinyurl.com/Muz-Zena.

V dotazníku osoba vždy zadá svoji výšku, hmotnost a pohlaví. Pohlaví je klasifikační třída, kterou rozhodovací strom určuje na základě výšky a hmotnosti. Data jsou průběžně doplňována anonymně od skutečných osob od roku 2017 zejména studenty předmětu Data Mining. Skupina osob je věkově homogenní (rozmezí 22–27 roků). Sběr dat a využití ve výuce popisuje článek *Teaching decision tree using a practical example* (Dobesova 2020b).

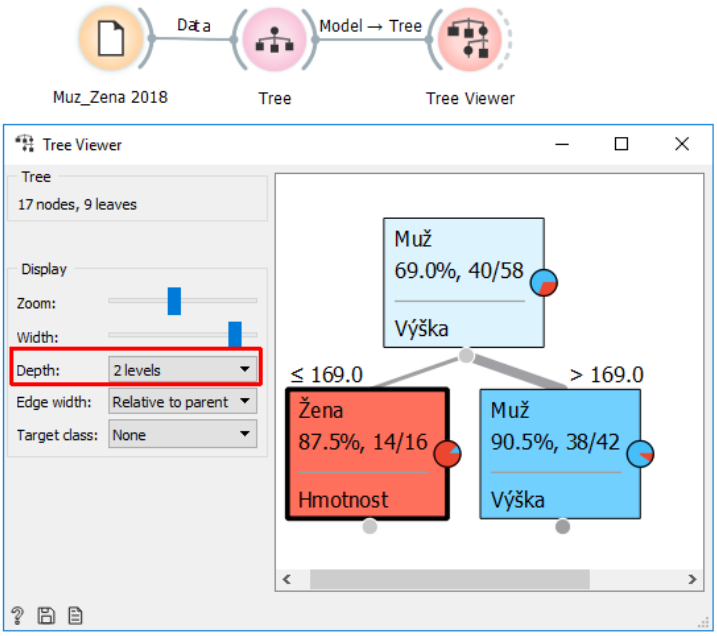
Příklad 29

Data *6_MuzZena2018.xlsx* (list 2018) nebo si aktuální data si stáhněte z odkazu: https://tinyurl.com/MZ-odpovedi. Pozor v datech je jedna chyba (záměna hmotnosti a výšky), nalezněte ji a opravte ji.

Workflow *6_Tree_MuzZena.ows*

Nejprve se v uzlu *File* nastaví atribut Pohlaví jako cílová klasifikační třída – *target*. Sestavte workflow jako v předchozích příkladech z uzlu ***Tree*** a ***Tree Viewer***. Trénovací data v souboru *6_MuzZena2018.xlsx* obsahují 18 žen a 40 mužů, celkem tedy 58 osob.

V tomto příkladu lze velice pěkně experimentovat s omezením na počet úrovní ve stromu v údaji o hloubce stromu ***Depth***. Následující obrázek ukazuje nejjednodušší strom s **dvěma úrovněmi**, kde je rozhodovacím kritériem pouze výška 169 cm (hmotnost se nebere v úvahu). Tento strom obsahuje šest **špatně zařazených osob** (2 muže klasifikuje jako ženu, 4 ženy klasifikuje jako muže). Jednotlivé větve lze rozvinout kliknutím na šedý bod uprostřed spodní hrany obdélníku. Lze tak interaktivně rozvíjet detailní části stromu.



Obr. 102 Rozhodovací strom do druhé úrovně větvení

Pravidla (rules) pro tento nejjednodušší strom budou následující:

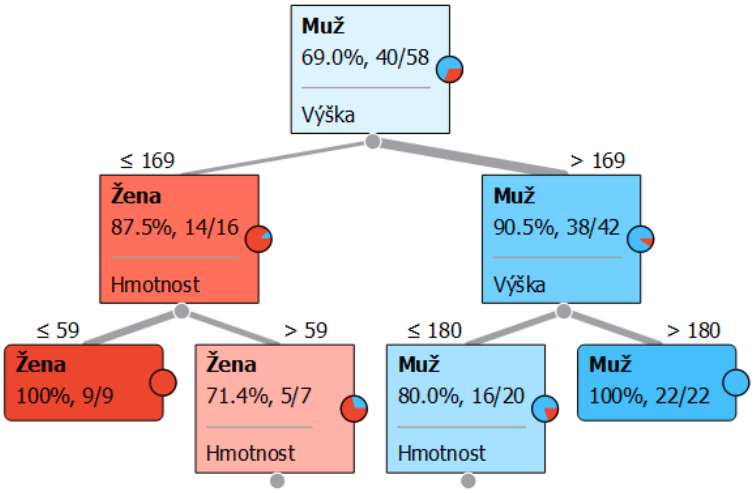
R1: Výška ≤ 169 cm ⇒ Žena

R1: Výška > 169 cm ⇒ Muž

Pro zdrojová data *6_MuzZena2018.xlsx* spočítejte cvičně **Entropii** celého vstupního datasetu.

V souboru je 18 žen a 40 mužů, celkem soubor tvoří 58 záznamů.

Pro **tři úrovně** stromu (Obr. 103) jsou dva listy s 100 % úspěšností klasifikace. Dva listy obsahují úspěšnost menší než 100 %, zde se nachází šest špatně klasifikovaných osob. Navíc je zřejmé, že u žen se nově objevila podmínka testující hmotnost 59 kg. V modré větvi se podruhé testuje kritérium výšky a to hodnota 180 cm. Při výšce 180 cm je stoprocentní jistota v našich datech, že osoba je muž. U osob menších než 180 cm se mezi muži vykytují 4 špatně zařazené ženy. Světle modrý a červený obdélník lze pomocí šedého bodu rozvinout o další úrovně stromu.



Obr. 103 Rozhodovací strom do třetí úrovně větvení

Zapište jednotlivá pravidla R pro vygenerovaný strom:

R1: Výška ≤ 169 cm and Hmotnost ≤ 59 ⇒ Žena (100%)

R2: .....

R3: ....

R4: .....

Zkuste nastavit počet úrovní na čtyři nebo pět a přepište opět strom do formy pravidel. Při zvyšování počtu úrovní diskutujte, kdy se již jedná o **overfiting** (přeučený strom), tzn. strom reaguje na velice odlišné lidi (outliers – vysoká žena nebo malý muž). Při nižším počtu úrovní se jedná o prořezaný strom (**prunning**).

Zapište pravidla pro více úrovní:

R .....

R .....

R .....

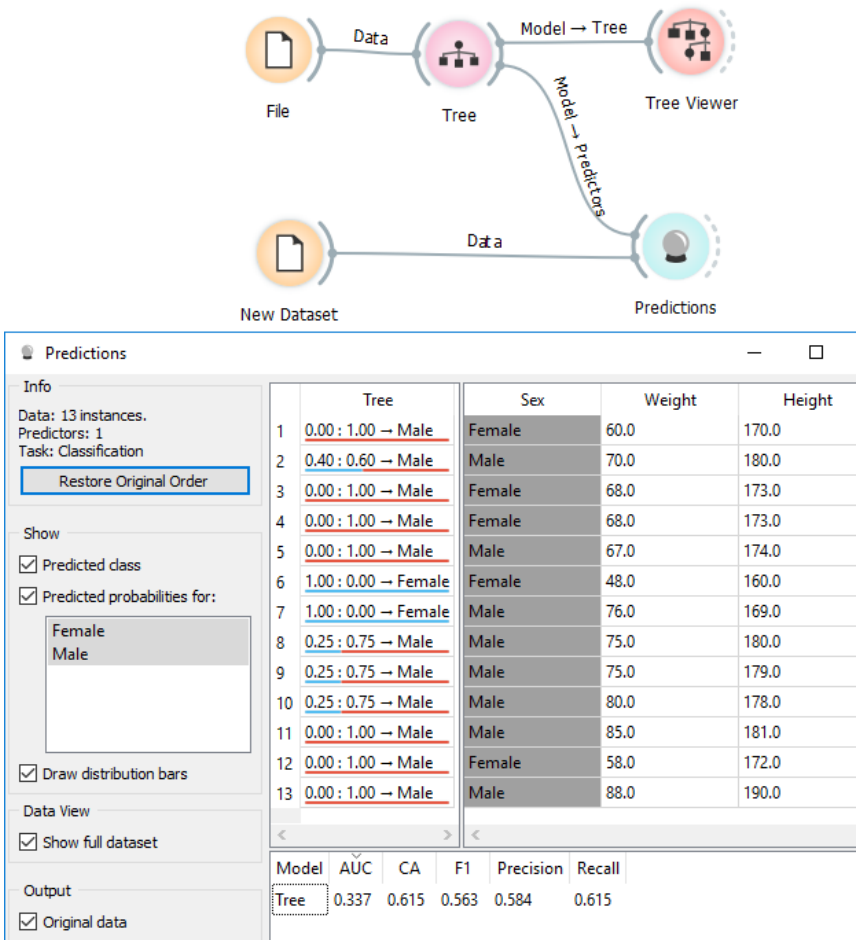
R .....

6.3.4 Klasifikace podle rozhodovacího stromu

Data na listu 2019 v souboru *6_MuzZena2018.xlsx* obsahují údaje o 13 osobách z roku 2019. Tato data se mohou použít pro klasifikaci pohlaví a výsledné porovnání, zda klasifikovaná hodnota odpovídá skutečnosti.

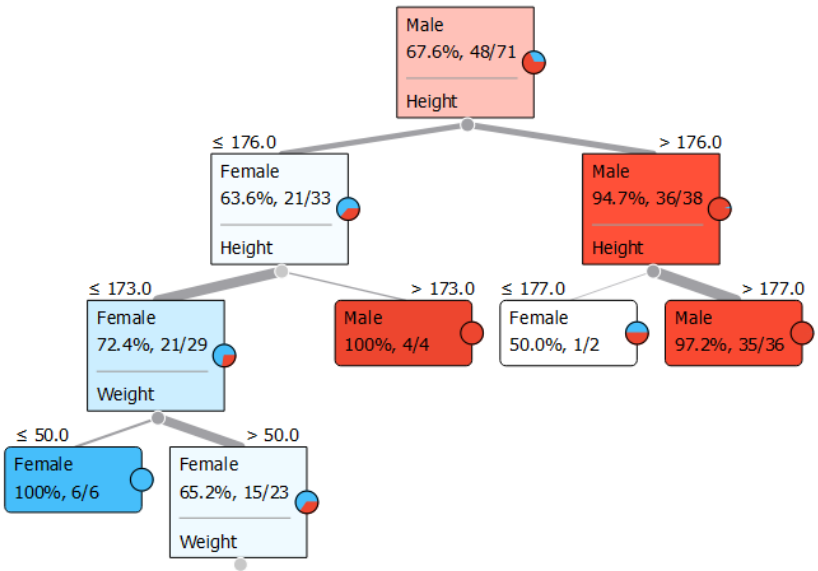
Do postupu se přidá nový uzel *File*, kde se připojí list 2019, a přidá se uzel ***Prediction***. Výsledné okno v tomto uzlu ukazuje porovnání predikovaného pohlaví a skutečného pohlaví (Obr. 104). Ze 13 osob je 5 osob predikováno chybně. Ve čtyřech případech se jedná o chybné určení žen jako muže, a to z toho důvodu, že jejich výška byla vyšší než první pravidlo 169 cm. Levý sloupec vypisuje poměr pravděpodobností klasifikace do dvou tříd. Poměr je znázorněni i poměrem délky barevných čar.

Např. druhá řádka na Obr. 104 ukazuje poměr *0.40 : 0.60 –> Male.* To znamená, že osoba spadá do kategorie žena s pravděpodobností 0,4 a do kategorie muž s pravděpodobností 0,6. Výsledná kategorie je tedy Male.



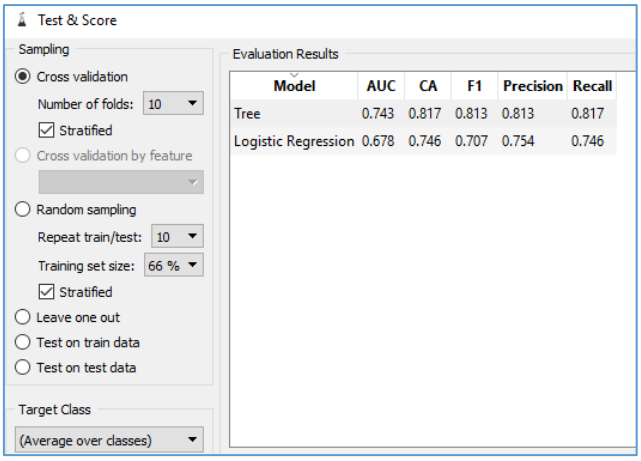
Obr. 104 Výsledek predikce pohlaví

Na závěr vyzkoušejte sestrojit strom ze všech záznamů o osobách (Obr. 105). Data jsou obsažena na listu *All* ve stejném Excel sešitu, kde je celkem 71 osob. Popište, jak se změnil strom.



Obr. 105 Strom pro celý dataset dvou roků

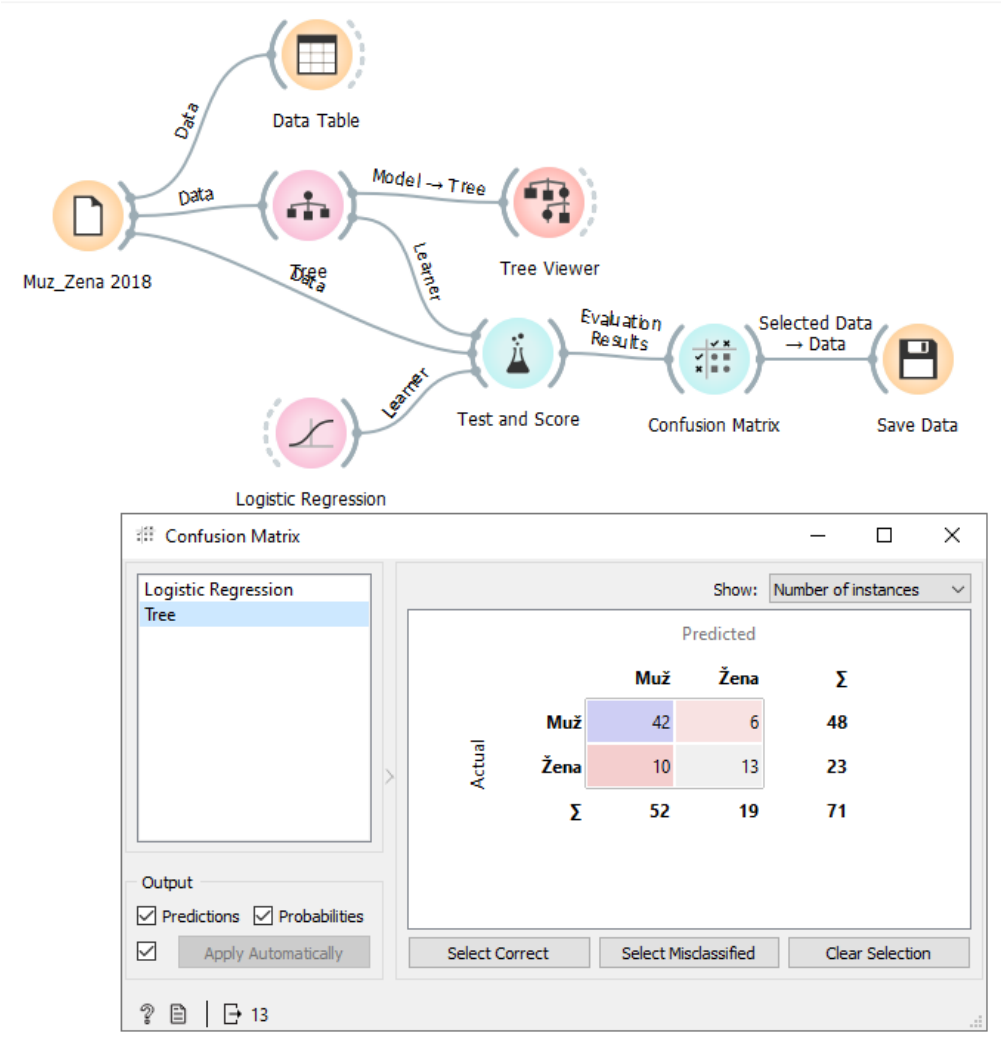
Zajímavé je porovnat, jak správně klasifikuje rozhodovací strom oproti logistické regresi. Do workflow přidáme uzel **Test & Score** a dále uzel logistické regrese. Připojíme do uzlu *Test & Score* uzel zdrojových dat, *Logistic Regression* a uzel *Tree* (Obr. 107).



Obr. 106 Uzel Test & Score

V dialogu uzlu *Test & Score* je zřejmé, že přesnost rozhodovacího stromu je 81 % a je vyšší než logistická regrese pro soubor dat o 71 osobách (Obr. 106).

Výsledek porovnání je navíc dobře viditelný i v uzlu ***Confusion Matrix*** (matice záměn), kde je přehledně uvedena matice s počtem správně a nesprávně zařazených instancí. Vpravo lze přepínat mezi oběma predikčními modely *Logistic Regression* a *Tree* (Obr. 107). Je evidentní, že strom je úspěšnější v určování mužů než žen.



Obr. 107 Workflow s porovnáním rozhodovacího stromu proti logistické regresi a matice záměn

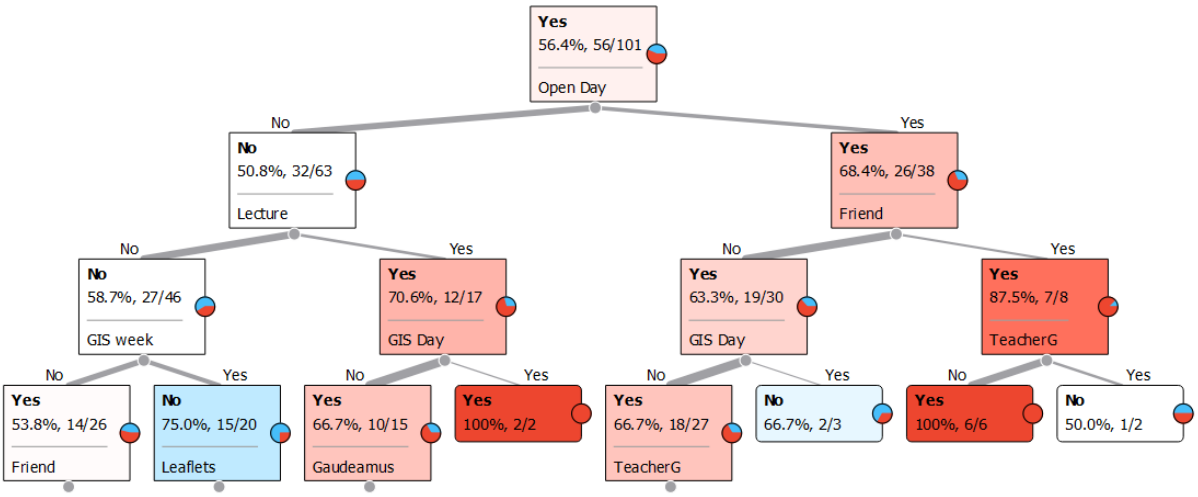
6.3.5 Úloha Uchazeči

Příklad 30

Data *6_Uchazeci.xls (list 2016+2017),* **workflow** *6_Uchazeci_Dotaznik.ows*

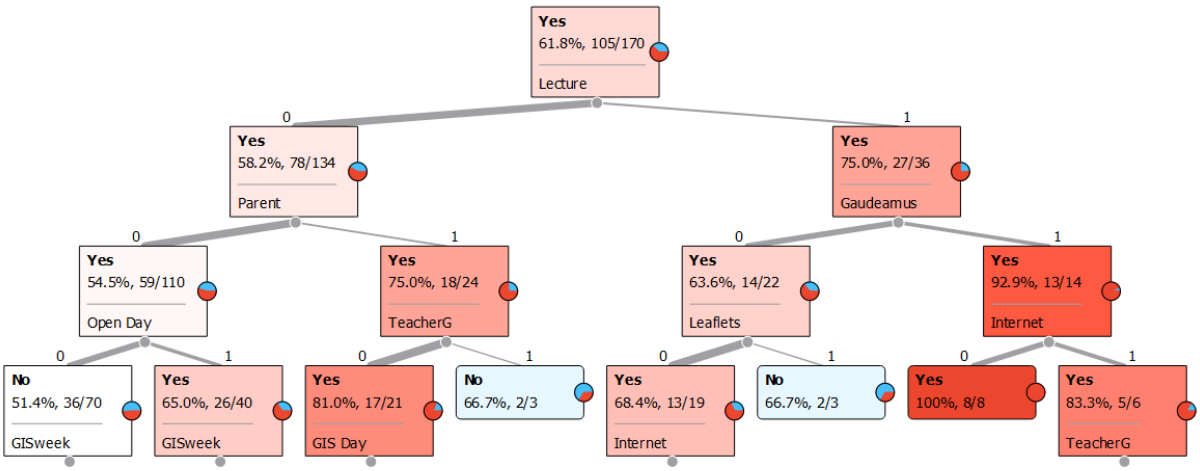
Z dat dotazníkového šetření se zkoumal vliv informací o studovaném oboru na rozhodnutí uchazečů nastoupit či nenastoupit ke studiu bakalářského oboru Geoinformatika. Údaj, o nastoupení ke studiu (Yes/No), je cílovou klasifikační třídou. První data z dotazníku jsou za rok 2016 a 2017. Celkem jsou k dispozici odpovědi od 101 uchazečů, z nichž 56 nastoupilo ke studiu. Popis sběru dat a vyhodnocení je uvedeno v článku *Using decision trees to predict the likelihood of high school students enrolling for university studies* (Dobesova a Pinos 2019).

Rozhodovací strom pro první dva roky je následující:



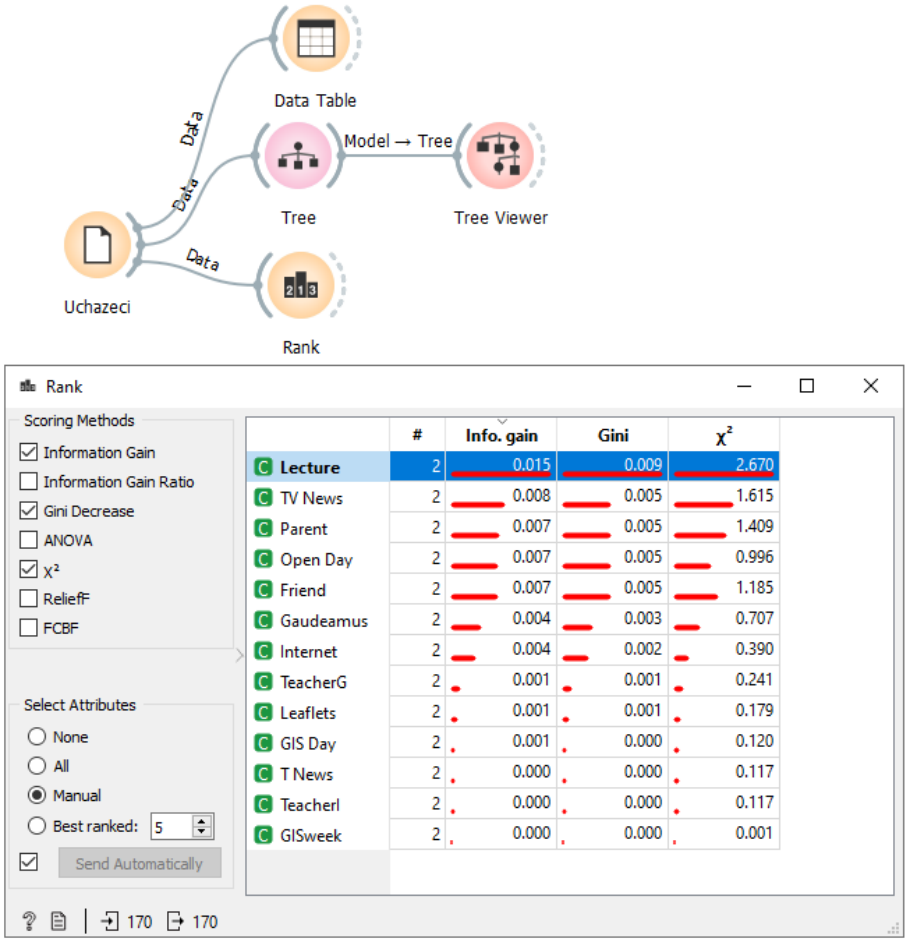
Obr. 108 Rozhodovací strom z dat dvou roků 2016 a 2017

Druhý list v xlsx souboru jsou data z dotazníků již za 4 roky, od roku 2016 až roku 2019. Celkový počet odpovědí je 170 a z toho 105 studentů nastoupilo ke studiu.



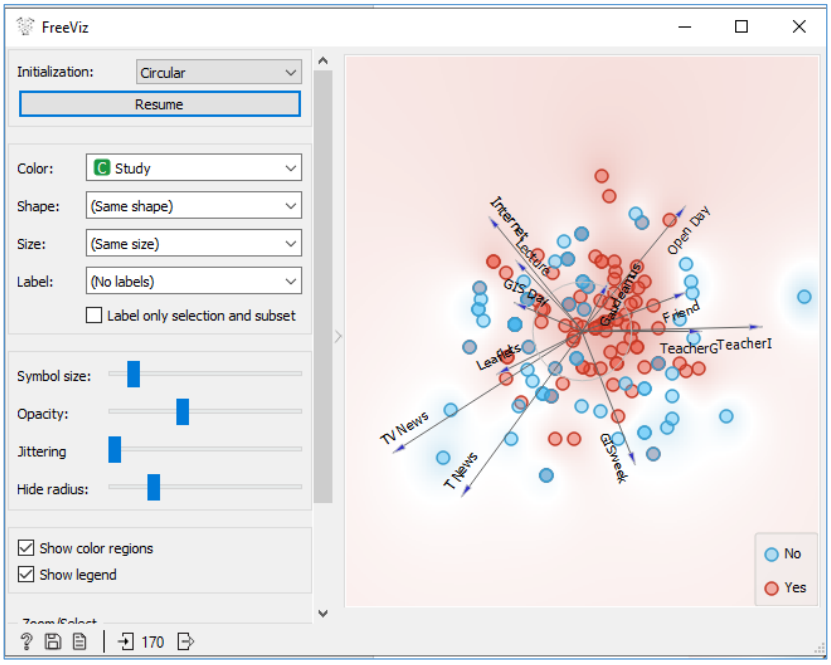
Obr. 109 Rozhodovací strom pro data ze čtyř roků

Z rozhodovacího stromu lze usoudit, které akce mají největší význam, a které přináší jistotu v kladném rozhodnutí nastoupit ke studiu. Stejně tak lze usuzovat, které informační akce naopak nevedou k ovlivnění nástupu ke studiu (modré obdélníky) – student nenastoupil ke studiu. Pro vyhodnocení vlivu atributů lze použít uzel **Rank**. Tento uzel spočítá a seřadí atributy podle významnosti, kdy je spočítán informační zisk. Nejvyšší je informační zisk je pro atribut *Lecture*, z toho důvodu je prvním uzlem.



Obr. 110 Vyhodnocení významnosti atributů uzlem Rank

Pro tato data lze vyzkoušet i uzel **FreeViz**. Tento uzel v mnoha souřadnicích zobrazí vliv atributů na výslednou kategorii. Výpočet nějakou dobru trvá. Lze navzájem porovnat výsledek uzlu *Rank* a *FeeViz*. Ze všech výsledků vyplývá, že vliv mají organizované akce typu Den otevřených dveří (Open Day), Gaudeamus, či doporučení rodičů nebo kamarádů.



Obr. 111 Vizualizace dat a cílové třídy pomocí FreeViz

6.4 Random Forest – náhodný les

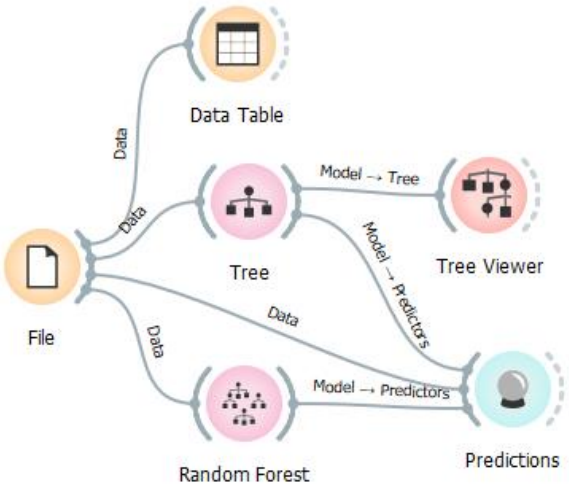
Náhodný les (random forest) používá k predikci či klasifikaci celou sadů rozhodovací stromů. Každý strom je vytvořen z náhodně generovaného vzorku dat z původního datasetu tzv. bootstrapingem. To je statistická metoda založená na **výběrech s opakováním z jednoho datového souboru**. Tímto způsobem se vytvoří **velké množství simulovaných** výběrů (tzv. bootstrapových výběrů) (Červová 2020). Potom se pro každý dílčí dataset vytvoří jeden rozhodovací strom. Při predikci pro novou instanci se potom vybírá z výsledků všech rozhodovacích stromů ten výsledek, který má nejčastější hodnotu (modus) (Mbaabu 2020) (Breiman 2001). Použití náhodného lesa minimalizuje efekt přetrénování, který může nastat při použití jen jednoho rozhodovacího stromu.

Příklad 31

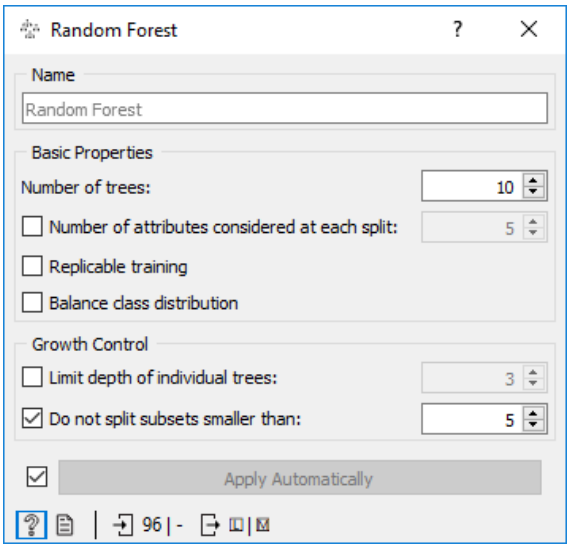
Data *6_MuzZena_2020.xlsx,* **workflow** *6_RandomForest_MuzZena2020.ows*

Vstupní data byla již použita v úloze určení pohlaví osoby z její váhy a výšky. Workflow se skládá z uzlu *File*, kde se připojí sesbíraná data o osobách, atribut pohlaví je nastaven jako cílová klasifikační třída – target. Dále je ve workflow již známý uzel **Tree** a **Tree Viewer**. Do workflow se vloží nový uzel **Random Forest** ze skupiny uzlů *Model*. V uzlu Random Forest lze nastavit počet stromů, kdy výchozí hodnota je 10 a další omezení, např. hloubky stromu atd.

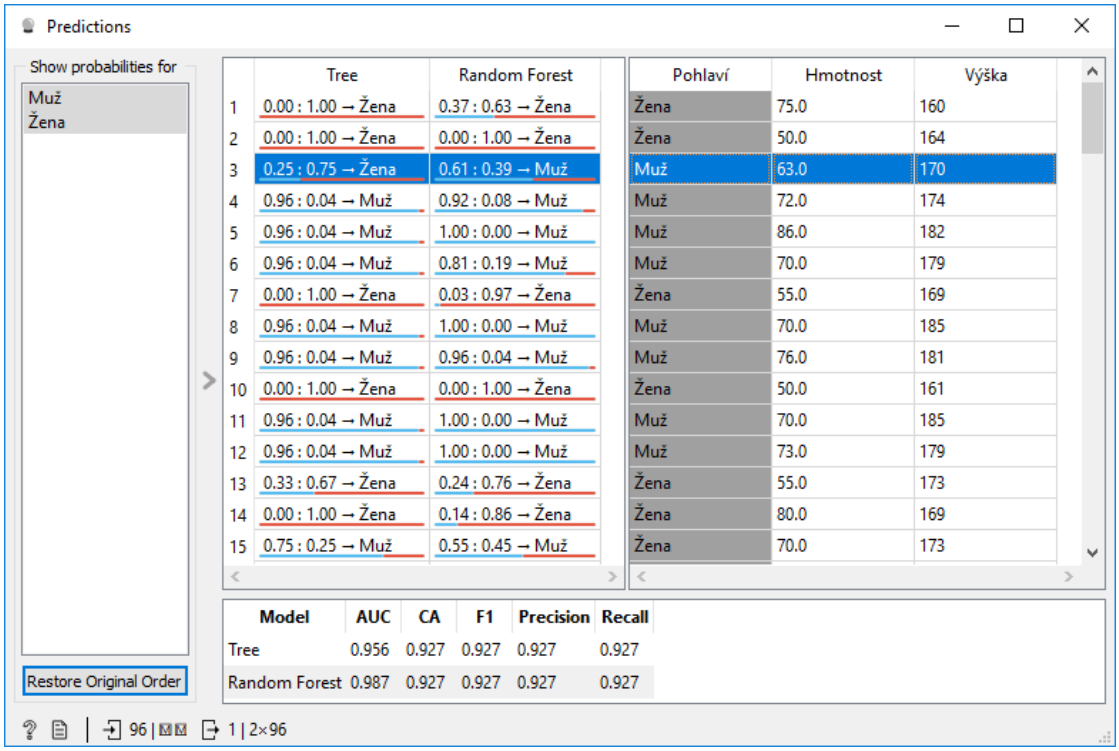
Porovnání skutečné hodnoty pohlaví a predikované hodnoty je pomocí modrého uzlu **Prediction**. Do uzlu Prediction jsou pro porovnání připojena vstupní data a dále výsledek jednoho rozhodovacího stromu a výsledek náhodného lesa.



Obr. 112 Workflow s uzlem Random Forest pro predikci pohlaví



Obr. 113 Dialog uzlu Random Forest



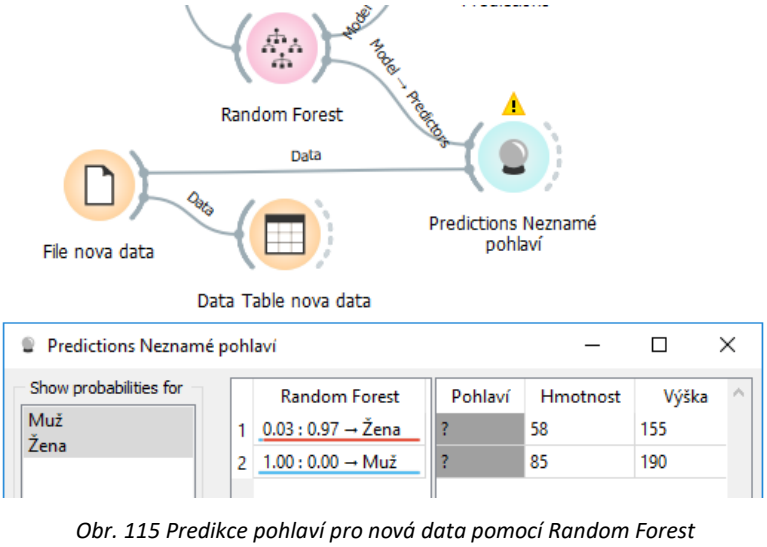
Obr. 114 Výsledky predikce dvou modelů v uzlu Prediction

Výsledný uzel **Prediction** ukazuje v levém okně jak výsledky predikce dvou modelů – sloupec *Tree* a *Random Forest* a zároveň pro jednotlivé záznamy ukazuje i původní data (Obr. 114). U predikcí jsou uvedeny i číselně pravděpodobnosti výsledku. Ve většině případů se predikce navzájem shodují a shodují se i s původní hodnotou. V některých případech je ale *Random forest* v predikci lepší než *Tree*. Viz řádek číslo 3, kdy pro skutečné pohlaví Muž rozhodovací strom *Tree* predikuje špatně pohlaví jako Žena, ale Random Forest predikuje pohlaví správně a to Muž.

V dolní části levého okna je výsledné hodnocení kvality obou modelů. Pro model *Tree* má charakteristika **AUC** (Area Under Curve) hodnotu 0,956 ale pro model *Random Forest* je vyšší hodnota **AUC** a to **0,987**. Znamená to tedy, že model *Random Forest* je lepší než jeden model *Tree* s jedním stromem.

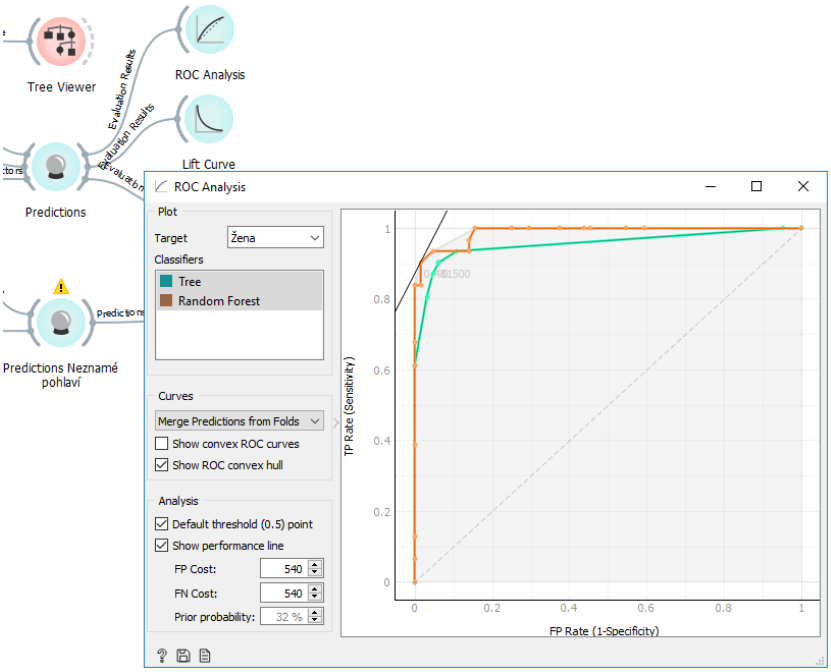
Zkuste experiment, zda lze dosáhnout lepších výsledků pomocí zvyšováním počtu stromů na více než je výchozích **10 stromů** v dialogu uzlu Random Forest.

Random Forest lze použít pro predikci pro nová data osob, kde neznáme pohlaví. Upravte workflow podle následujícího Obr. 115, kdy se do workflow přidá další uzel **Prediction**. Do tohoto uzlu se připojí výstup uzlu **Random Forest** a uzel **File** s XLSX listem *Nezname pohlavi* s novými daty.



Obr. 115 Predikce pohlaví pro nová data pomocí Random Forest

ROC křivku pro porovnání modelů lze vykreslit uzlem **ROC analysis**, který připojíme za první uzel *Prediction.* Z vykreslených křivek je patrné, že strměji stoupá křivka pro model *Random Forest* než pro model *Tree* pro kategorii *Žena*. Model *Random Forest* tedy vychází v predikci lépe. Zkuste přepnout v kolonce *Target* i na kategorii *Muž* a vyhodnotit ROC křivky.



Obr. 116 ROC křivky pro dva modely Tree a Random Tree pro výslednou kategorii Žena

6.5 Naivní bayesovský klasifikátor

Bayesovská klasifikace vychází z Bayesovy věty o podmíněných pravděpodobnostech. Jde o učení s učitelem, kdy z trénovacích dat lze spočítat všechny pravděpodobnosti (*četnosti*) výsledných klasifikačních tříd *P(H)*, které jsou v trénovacích datech. Tato apriorní pravděpodobnost *P*(*H*) odpovídá znalostem o zastoupení jednotlivých tříd bez ohledu na další data popisující instance. Bayesův vztah pro výpočet podmíněné pravděpodobnosti, že platí hypotéza H, pokud pozorujeme evidenci E, je

$$P(H \mid E) = \frac{P(E \mid H) * P(H)}{P(E)} \qquad (5)$$

Podmíněná (aposteriorní) pravděpodobnost *P*(*H*|*E*) vyjadřuje, jak se změní pravděpodobnost hypotézy, pokud nastala evidence E. Při hledání nejpravděpodobnější hypotézy *H* z *t* možných hypotéz nás zajímá ta, která má nejvyšší pravděpodobnost a nezajímá nás konkrétní hodnota pravděpodobnosti, hledá se tedy pouze ta maximální. Ve vzorci (5) se tedy vypouští jmenovatel a zjednodušuje se na vzorec

$$P(E \mid H) = \max_t((P(E \mid H_t) * P(H_t)) \qquad (6)$$

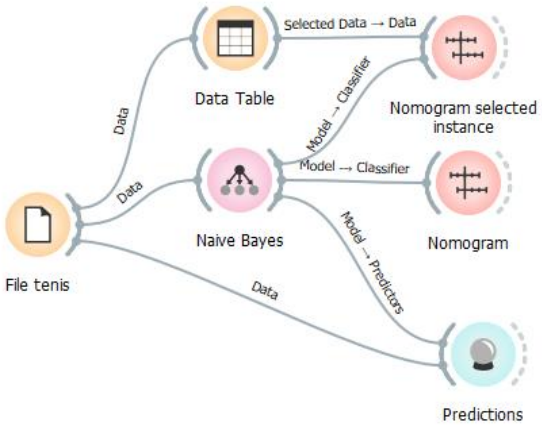
Hledáme tak hypotézu, která má největší věrohodnost (maximum likehood) ze všech hypotéz H_t.(Berka 2005).

Naivní bayesovský klasifikátor je označován jako naivní z toho důvodu, že předpokládá podmíněnou nezávislost vstupních atributů, což bývá v reálných případech málokdy splněno (Pretnar 2019). Naopak často jsou vstupní data popsána řadou korelovaných atributů vykazující redundanci (když prší, bývá i vysoká vlhkost). Přesto vykazuje tento klasifikátor překvapivě dobré vlastnosti v úspěšnosti klasifikace.

Příklad 32

Data *6_Tenis_nominal.xlsx,* **workflow** *6_NaiveBayesTenis.ows*

Vstupní data jsou známá z úlohy rozhodovacích stromů. Máme dvě klasifikační třídy – *play* (yes, no), atribut play je nastaven jako **target**. Jejich pravděpodobnost je *P(no)* = 5/14 = 0,357 a *P(yes)* = 9/14 = 0,642. Máme tedy dvě hypotézy.



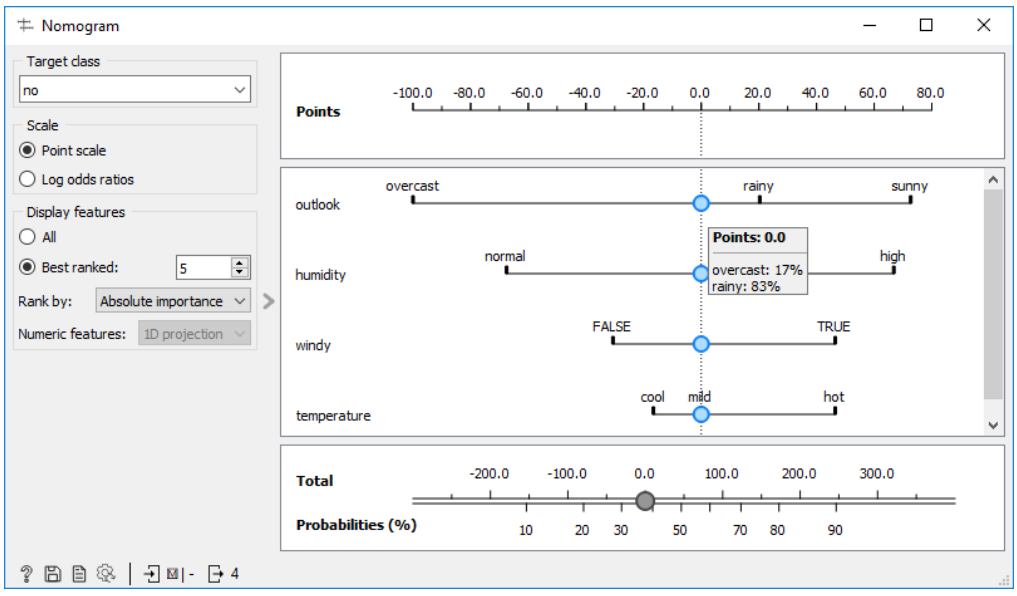
Obr. 117 Workflow s modelem naivní bayesovský klasifikátor

Pro zobrazení se používá uzel **Nomogram,** který je připojen za uzel **Naive Bayes**. Nomogram zobrazuje vliv atributů na pravděpodobnost třídy z trénovacích dat. V levém panelu lze přepínat *Target class* mezi *yes* a *no.* Měřítko *Scale* přepněte na volbu *Point Scale*. V pravém okně jsou zobrazeny atributy v pořadí od nejvlivnějších na výslednou třídu.

Po najetí kurzorem na modrý bod se zobrazí počet bodů a konkrétní pravděpodobnosti v procentech. Vyhodnocení kategorie *no* je na Obr. 118. Je patrné, že rozhodující vliv na nehraní tenisu má atribut *outlook*, který v 83 % má hodnotu *rainy* a v 17 % *overcast*, jen jedenkrát se hrál tenis za slunečného počasí. Modrý bod je umístěn proporčně na vodorovné ose kategorií a lze vyhodnocovat jeho relativní pozici ke kategoriím. Např. modrý bod *windy* je blíže k hodnotě *False* než k hodnotě *True*, tzn., že častěji se hraje tenis za bezvětří. Při velkém počtu atributů lze omezit počet jejich zobrazení v levém okně (zde 5) a různě je řadit. Modrými body lze interaktivně posouvat a zjišťovat tak pravděpodobnost nové situace. Pravděpodobnost lze odečíst na dolní stupnici nad šedým bodem a to jak v procentech, tak hodnotu celkového součtu bodů.

Ve workflow je použit ještě jeden **Nomogram**, který je napojen na uzel Data Table zdrojových dat. Při výběru konkrétního řádku v tabulce je zobrazena tato instance vzhledem k pravděpodobnosti celé trénovací množiny.

Uzel **Prediction** porovná výsledek klasifikačního modelu Naive Bayes a skutečných tříd trénovacích dat. V dolní části jsou uvedeny parametry klasifikačního modelu (AUC je 0,922, Precission 0,801, ...).



Obr. 118 Nomogram pro bayesovský klasifikátor a výslednou třídu „no“

Vyzkoušejte Naivní baeysovský klasifikátor i na číselných vstupních datech, soubor 6_Tenis_numeric.xlsx. Nomogram umisťuje modré body na osy s číselným měřítkem. Dále rozšiřte samostatně workflow o model rozhodovacího stromu a random forest a připojte je do uzlu Prediction. Můžete navzájem porovnat kvalitu modelů pomocí AUC, přesnosti atd. Který je nejlepší?

6.6 Metoda Support Vector Machines

Metoda Support Vector Machines (**SVM**), česky *metoda podpůrných vektorů* hledá **nadrovinu**, která rozdělí vstupní objekty tak, aby došlo k maximální separaci objektů (Cortes a Vapnik 1995). Jedná se o metodu učení s učitelem, tzn., že je ve vstupních datech známá výsledná klasifikační třída nebo výsledná číselná hodnota. Metodu lze použít jak pro klasifikační úlohu, tak pro regresní úlohu predikce pro nové neznámé objekty. Pro regresi je metoda označována jako Support Vector Regression (SVR).

Objekty jsou popsány *n* vstupními atributy. Potom dělící nadrovinou je myšlena hranice, která rozdělí *n-1* dimenzionální prostor na dva opačné poloprostory, které optimálně dělí původní prostor. Jako *podpůrné vektory* jsou označovány souřadnice vstupních objektů, které mají minimální vzdálenost k nadrovině. Zároveň tyto podpůrné vektory maximalizují vzdálenost (označenou jako *margin of tolerance* – epsilon – práh necitlivosti) mezi instancemi rozdílných tříd. Hlavní idea metody je tedy individuální vyhledání nadroviny, minimalizace chyby a maximalizace prahu necitlivosti (vzdálenosti mezi třídami a dělící nadrovinou), s vědomí přípustnosti určité chybovosti, kdy se určité objekty zařadí chybně (Sayad 2020b).

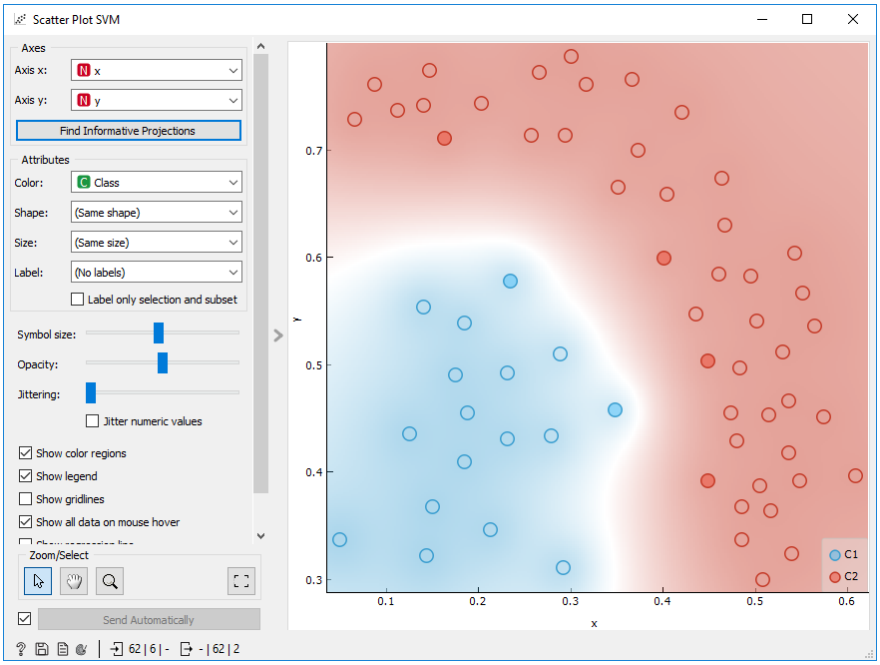
V nejjednodušším případě jsou vstupní objekty ve 2D lineárně separovatelné přímkou, či lineární nadrovinou. Ve složitějších případech dochází k zakřivení průběhu nadroviny. Pro obtížně separovatelná data se využije nelineární SVM, které využívají jádrové transformace – **kernel**. Tato operace transformuje vstupní data do prostoru o vyšší dimenzi, kde už je možné třídy lineárně separovat a dále aplikovat optimalizační algoritmus pro nalezení rozdělující nadroviny. K transformaci mohou být využity různé funkce, například polynomická, sigmoidní nebo radiální RBF. Složitější jádrové transformace mohou napomoci k lepšímu odlišení tříd, a tedy přesnější klasifikaci. Orange používá pro uzel SVM implementaci knihovny LIBSVM (Chang a Lin 2011).

Příklad 33

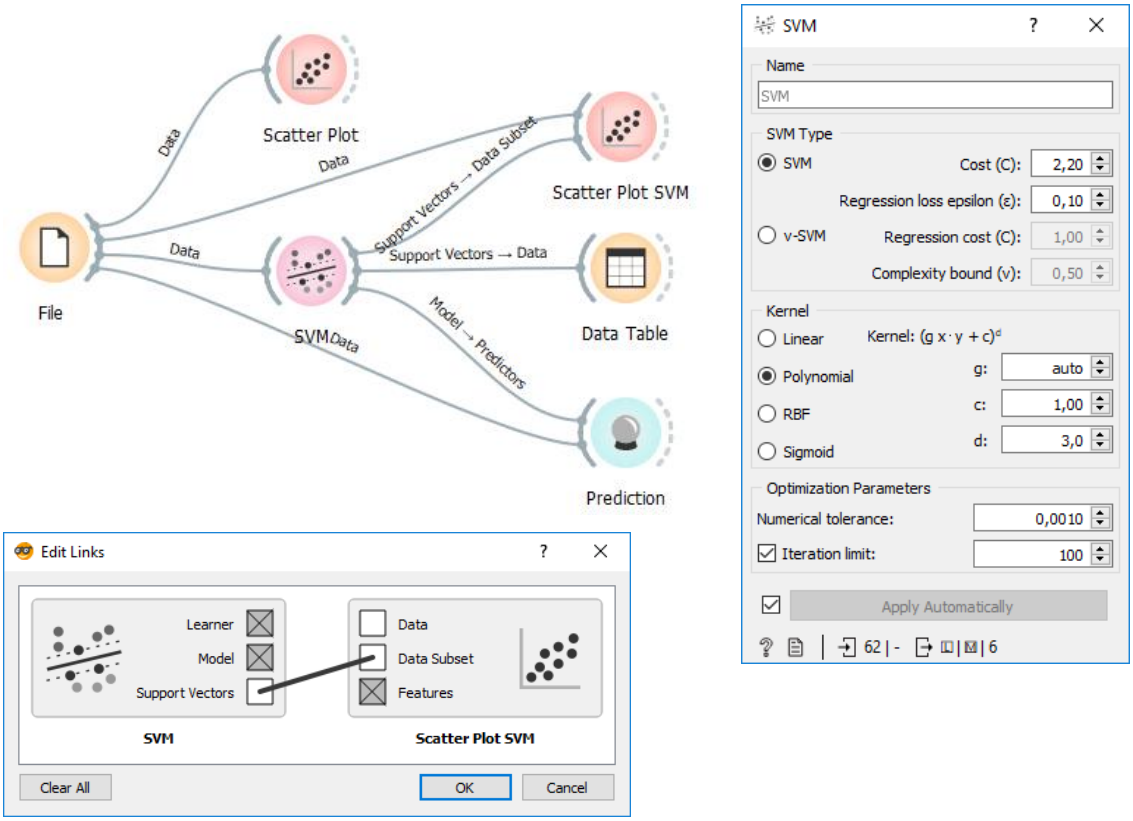
Data *6_PaintedDataSVM.xlsx*, **workflow** *6_SVM.ows*

Vstupní data obsahují jednoduchá cvičná data 62 objektů, ze dvou klasifikačních tříd C1 a C2, atribut Class je nastaven jako *Target*. První třída tvoří téměř kruhový shluk obklopený druhým obloukovým shlukem. Vstupní objekty jsou vykresleny uzlem *Scatter Plot*. Je evidentní, že tyto dvě třídy nejsou jednoduše lineárně separovatelné. Metoda SVM je realizována uzlem **SVM** ze skupiny *Model*. Tento uzel představuje hotový natrénovaný model. Výstupem uzlu je tabulka původních objektů, které jsou určeny jako podpůrné vektory, které jsou viditelné pomocí uzlu *Data Table*.

V dialogu uzlu **SVM** se nastavuje typ modelu (SVM nebo v-SVM – různý způsob minimalizace chybové funkce). Při nastavení vyšší hodnoty penalizace *C – Cost* se ve výsledku snižuje počet podpůrných vektorů pro použitá cvičná data. Pro hodnotu C = 2,2 a polynomiální kernel je to šest vektorů. Nastavení epsilon ε má vliv jen u regrese. Dále se nastavuje *Kernel* pro transformaci do vyššího prostoru. Vyzkoušejte experimentálně jak volbu *Linear*, tak volbu *Polynomial*. Při různém Kernelu se také mění počet výstupních podpůrných vektorů.



Obr. 119 Rozptylový graf původních objektů a podpůrných vektorů (plné body)

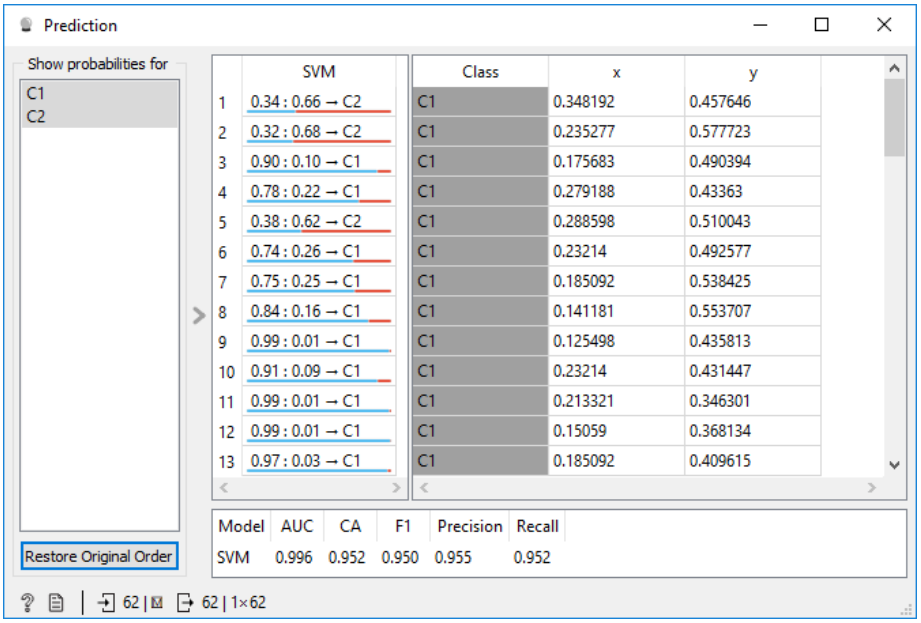


Obr. 120 Workflow s uzlem SVM (nahoře), vyhodnocením predikce uzlem Prediction, nastavení dialogu SVM (vpravo) a nastavení propojení do uzlu Scatter Plot SVM (vlevo)

Zobrazení původních objektů a podpůrných vektorů je ve workflow společným uzlem *Scatter Plot SVM*, který je na Obr. 119. Šest modrých a červených plných bodů jsou podpůrné vektory. Je evidentní, že jsou blízko dělící křivky. Prázdné body jsou ostatní vstupní objekty. Rovnice nadroviny není přímo dostupná, ale natrénovaný model je uložen v uzlu SVM.

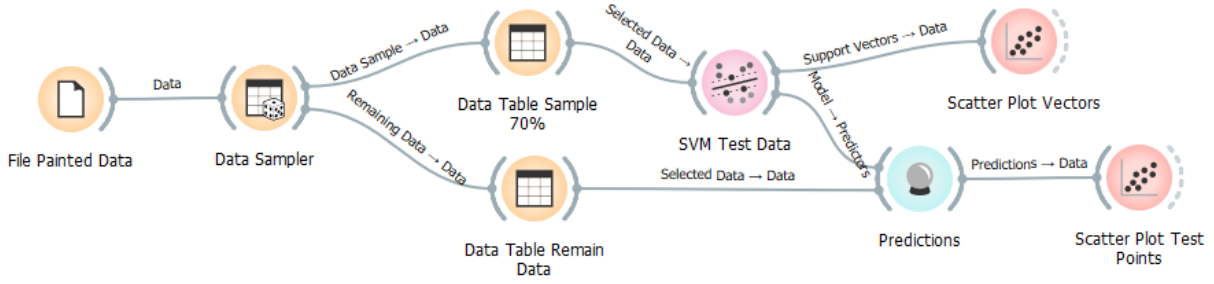
Posouzení kvality modelu můžeme provést pomocí uzlu **Prediction**. Pro lineární kernel je přesnost nižší (AUC=0,996, Precision=0,995, …) než pro polynomický kernel (AUC=1, Precision=1) se stejným počtem podpůrných vektorů. Na Obr. 121 je patrný špatně zařazený první objekt do třídy C2 při lineárním kernelu.

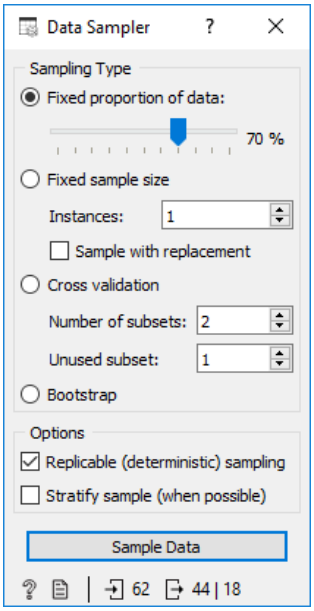
Pro nové případy lze použít uzel *Prediction* obdobně jako v předchozích příkladech (zde již neřešeno).

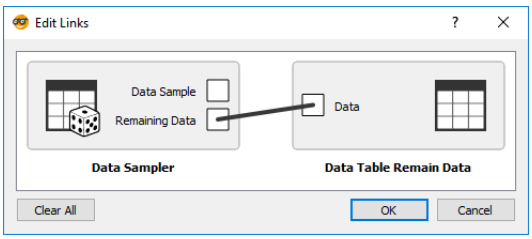


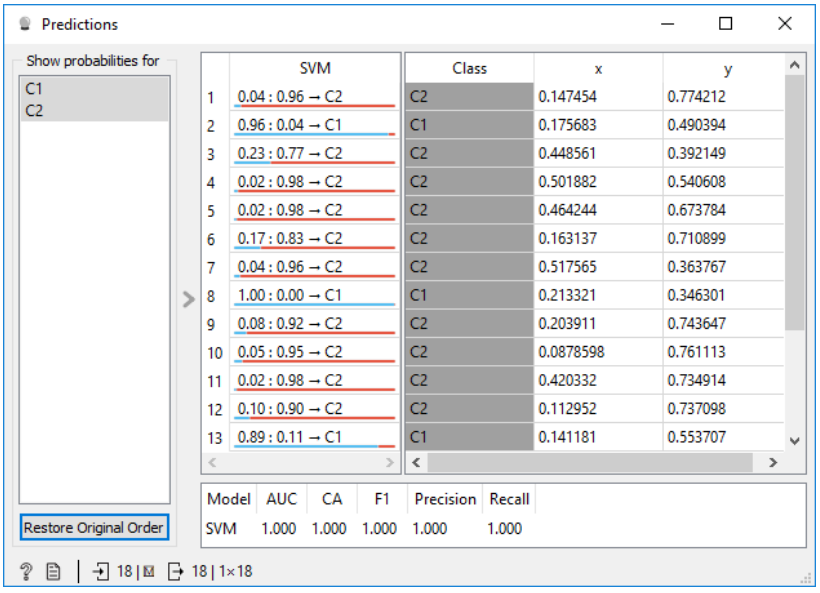
Obr. 121 Výstupní predikce podle modelu SVM s lineárním kernelem a vyhodnocení přesnosti

Pro vytváření modelu a jeho kontrolu lze rozdělit vstupní množinu na **trénovací data** a **testovací data**. K tomu slouží uzel **Data Sampler**. Vstupní data se obecně dělí v poměru 3:2 nebo 3:1 na data trénovací pro vytvoření modelu a zbylá data na testování a ověření správnosti modelu. V tomto případě jsou data rozdělena v poměru 70 % a 30 %, tj. 44 a 18 záznamů. Rozdělená data si lze přímo zobrazit v uzlech *Data Table*. Pozor je nutné správně nastavit obě propojení uzlů v dialogu *Edit Links* na *Sample Data* a *Remain Data.* Trénovací data potom vstupují do uzlu *SVM*. Nastavení modelu SVM je stejné jako v předchozím případě: Cost (C) = 2,2 a polynomický kernel. Body podpůrných vektorů se liší od předchozího případu, kdy byl pro model použit celý dataset 62 vstupních objektů. Zde je stanoveno 7 podpůrných bodů. V uzlu *Prediction* lze porovnat výsledek predikce třídy C1 a C2 podle modelu se skutečnou hodnotou pro 18 testovací objektů. Predikce je v tomto případě bezchybná. Výsledek predikce lze zobrazit i pomocí uzlu *Scatter Plot.*





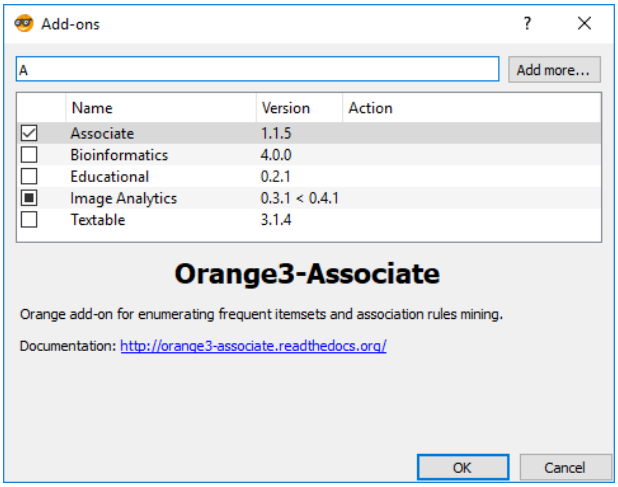




Obr. 122 Použití uzlu Data Sampler pro rozdělení vstupních dat a testování predikce modelu SVM na testovacích datech

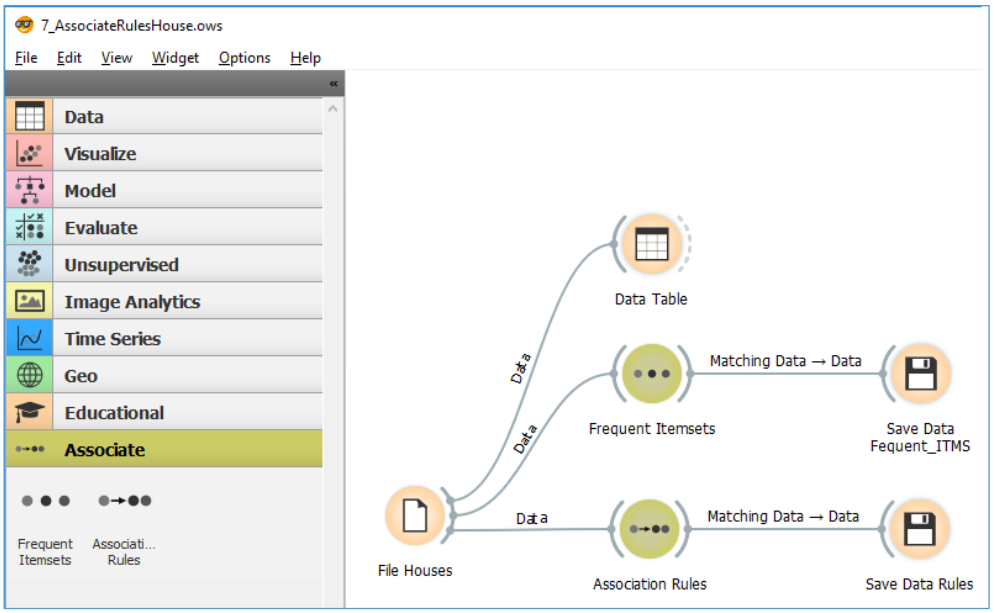
7 ASOCIAČNÍ PRAVIDLA

Pro tvorbu asociačních pravidel je nutné doinstalovat samostatný doplněk **Associate**. Doplněk si doinstalujte z menu volbou **Options** a **Add-ons…** Následně se vybere z nabídky doplněk *Associate*. Pro instalaci jsou vyžadována administrátorská práva a následný restart aplikace Orange.



Obr. 123 Instalace doplňku

Po instalaci se objeví vlevo dole nová zelená skupina **Associate**. Tato skupina nabízí dva zelené uzly *Frequent Itemsets* a *Association Rules*.



Obr. 124 Workflow pro frekventované sady a asociační pravidla

> *Popis doplňku je na https://orange3-associate.readthedocs.io/en/latest/widgets/associationrules.html (Biolab 2016). Popis není součást oficiálního manuálu.*

Asociace jsou závislé vztahy mezi objekty, atributy, proměnnými nebo výskyty, které jsou často skryty ve velkých objemných datových sadách. Asociace jsou velice běžné a tak patří do metod data miningu (Chattamvelli 2011). Orange poskytuje dva algoritmy pro indukci asociačních pravidel, standardní **Apriori algoritmus** popsaný autory Agrawal a Srikant (Agrawal a Srikant 1994) pro analýzu řídké matice dat (analýza nákupního košíku – Market Basket Analysis – MBA). Taková data mají pro jednotlivé atributy pouze hodnotu ano/ne. Např. existence druhů dopravy ve městě je popsána samostatnými atributy: tramvaj (ano/ne), trolejbus (ano/ne), metro (ano/ne). V tomto případě se jedná o **dichotomická data**.

Druhý **algoritmus FP-growth** je variantou Apriori algoritmu pro kategorické hodnoty atributů v datové sadě. Například jeden atribut *doprava* nabývá některou z hodnot „tramvaj / trolejbus / metro / lanovka / bus“. Tento typ kategorických dat lze převést v případě potřeby na dichotomická data.

Oba algoritmy podporují dolování frekventovaných datových sad (itemsets).

Vstupní datové sady se označují jako **transakční data**. Jedna řádka tabulky, záznam (instance) je označován jako transakce (Petr 2014a).

Orange podporuje i vstupní formát *baskets* pro řídká data, který je zejména vhodný pro texty (více o formátech na https://docs.biolab.si//2/reference/rst/Orange.data.formats.html).

7.1 Frekventované sady instancí

Uzel ***Frequent Itemsets*** hledá frekventované transakce (instance), to znamená ty, které jsou velice časté a typické pro danou sadu vstupních dat.

Příklad 34

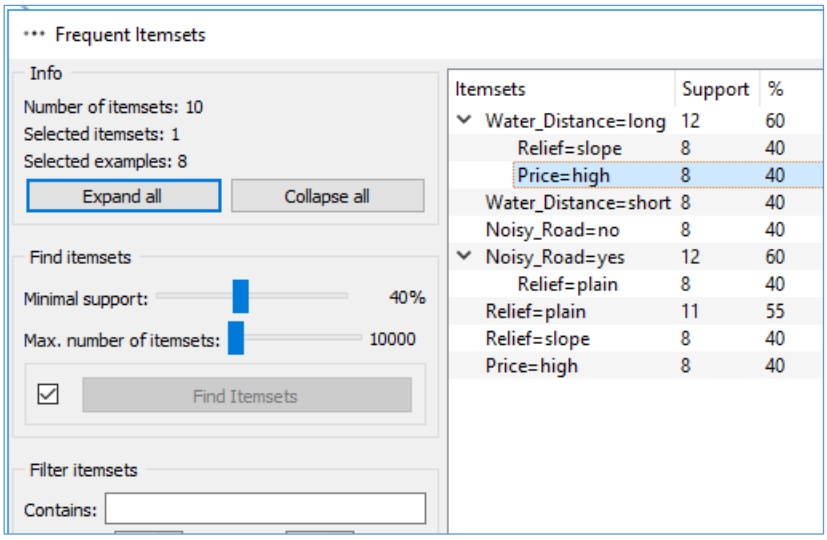
Data *4_Houses.xlsx*, **workflow** *7_AssociateRulesHouse.ows*

Data obsahují 20 domů, které jsou popsány čtyřmi atributy: vzdálenost od vody, vzdálenost od hlučné silnice, tvar reliéfu a cena. Všechny atributy obsahují dvě nebo tři kategorie vyjádřené textem. *ID_House* je číslu domu, kterému je nutné nastavit roli *meta* údaj ve stupních datech v uzlu *File*, aby se nebral v úvahu pro hledání pravidel.

	ID_House	Water_Distance	Noisy_Road	Relief	Price
1	1.0	long	yes	plain	low
2	2.0	short	no	splope	medium
3	3.0	long	no	plain	high
4	4.0	short	yes	plain	low
5	5.0	short	yes	plain	low
6	6.0	short	yes	plain	low
7	7.0	long	yes	plain	high
8	8.0	short	yes	plain	medium
9	9.0	long	no	slope	high
10	10.0	long	no	slope	high
11	11.0	long	no	slope	high
12	12.0	long	no	slope	medium
13	13.0	long	yes	slope	low
14	14.0	long	yes	slope	high
15	15.0	long	yes	slope	high
16	16.0	long	yes	slope	medium
17	17.0	long	no	plain	high
18	18.0	short	yes	plain	low
19	19.0	short	yes	plain	low
20	20.0	short	no	plain	medium

Obr. 125 Zdrojová data s charakteristikou domů

Použité workflow je na Obr. 124, náhled dat je na Obr. 125. V uzlu *Frequent Itemsets* se nastaví minimální podpora pro nalezení frekventovaných záznamů. Čím vyšší je podpora, tím častější případy hledáme (při vyšších hodnotách podpory bude logicky vyhovovat méně záznamů). V příkladu je nastavena minimální podpora 40 %. To znamená, že chceme nalézt všechny frekventované sady instancí, které se vyskytují právě nebo častěji než ve 40 % záznamech.



Obr. 126 Vyhodnocení frekventovaných sad instancí

Na Obr. 126 je v levém okně *Info* uvedeno, kolik frekventovaných sad bylo celkem nalezeno (odpovídá počtu řádků v pravém přehledu). V případě 40 % je to 10 frekventovaných sad.
Výsledek v pravém okně je možné číst takto: sloupec *Support* udává počet instancí (případů), které splňují hodnotu atributu (podmínka v prvním sloupci) a sloupec *%* vyjadřuje přepočítanou podporu na procenta z celého datového souboru. Některé podmínky jsou jednoduché, jiné složené.

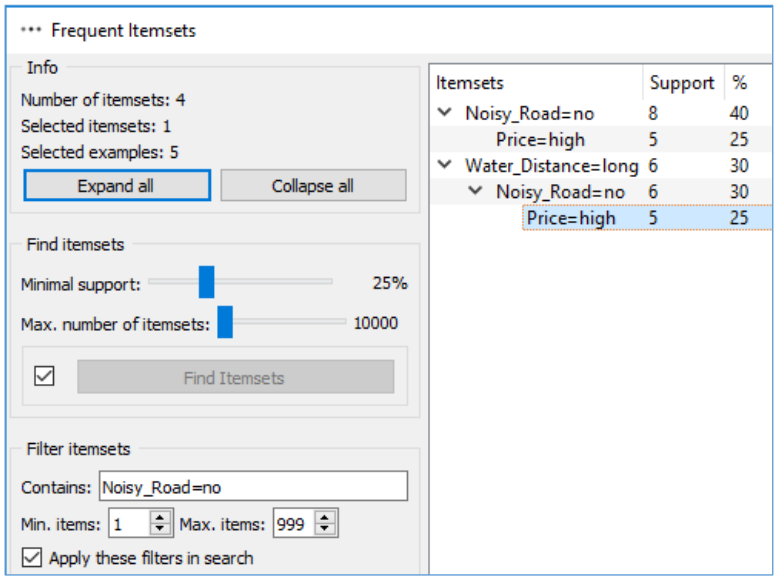
První frekventovaná sada je: Ve 12 případech (které tvoří 60 % datové sady) se domy nachází daleko od vody. Toto pravidlo se současně účastní složeného pravidla, kde lze říct, že v 8 případech (40 %) jsou domy daleko od vody a zároveň se nachází na svahu. Nebo 8 domů je daleko od vody a má zároveň vysokou cenu. Podmínka v prvním sloupci, která je odsazená, se vyhodnocuje jako „platí zároveň“, podmínky, které se nachází na stejné úrovni odsazení, se vyhodnocují jako „nebo“.
Následují jednoduchá pravidla a jedno složené pravidlo s podporami mezi 40 až 60 %.
Experimentujte s nastavením *Minimal support* k vyšším a nižším hodnotám. Při hodnotě větší než 60 % není nalezena žádná frekventovaná sada.

Výsledné instance domů, které splňují podmínku, lze vybrat (modře zvýrazněný řádek v okně *Frequent Itemsets* na Obr. 126) a pomocí uzlu *Save Data* uložit do tabulky v Excelu. Výhodou je, že se uloží nejen hodnoty atributů, ale i *ID_House* a je tak zřejmé, který dům splňuje vybrané frekventované pravidlo.

	A	B	C	D	E
1	**Water_Distance**	**Noisy_Road**	**Relief**	**Price**	**ID_House**
2	long	no	plain	high	3
3	long	yes	plain	high	7
4	long	no	slope	high	9
5	long	no	slope	high	10
6	long	no	slope	high	11
7	long	yes	slope	high	14
8	long	yes	slope	high	15
9	long	no	plain	high	17

Obr. 127 Záznamy, které odpovídají vybrané frekventované sadě.

Navíc lze v dialogu *Freqent Itemsets* filtrovat v levé části dole pomocí podmínky *Contains*. Filtrem lze omezit výsledek na výskyt hodnoty některého atributu, např. Noisy_Road = no. Následující výsledek ukazuje všechna pravidla od minimální podpory 25 %.



Obr. 128 Filtrování pomocí podmínky Contains

Do filtru lze zadat i více atributů oddělených čárkou, např. Noisy_Road=no, Price=high. Filtr může mít i podobu jen názvů atributů bez podmínky, např. Noisy_Road, Price.

7.2 Asociační pravidla pro lokalitu domu

Widget *Associate Rules* implementuje dolovací algoritmus FP-growth (frequent pattern) s „bucketing optimization FP" (Han et al. 2004) (Agarwal et al. [b.r.]). Algoritmus indukuje pravidla pro celou sadu záznamů a přeskakuje pravidla, kde následník neodpovídá jedné z hodnot klasifikační třídy.

Opět použijeme stejná cvičná data jako v předchozí kapitole *7_Houses.xlsx,* postup je také ve stejném workflow.

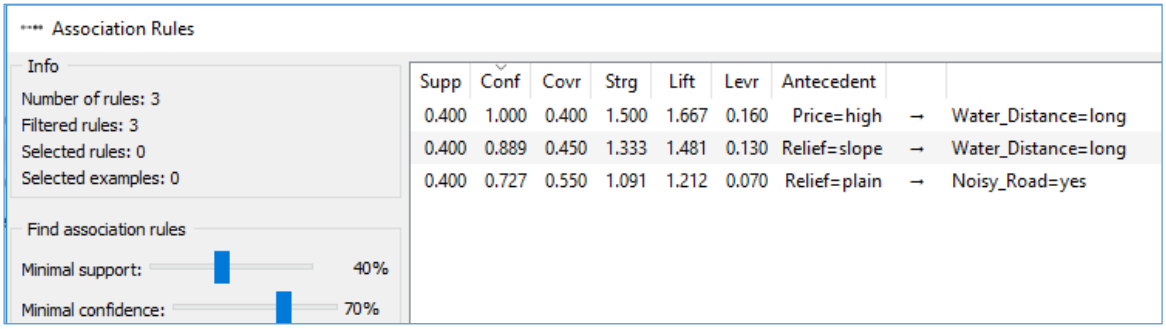
V uzlu **Association rules** nastavte v dialogu parametry pravidel:

Minimal support na nízkou hodnotu, např. 40 %, ***Minimal confidence*** na 70 %. Confidence (spolehlivost) vždy nastavujeme na více než 50 %. Můžete ponechat počet pravidel ***Max. number of rules*** a následně s ním experimentovat.

Výsledná tabulka zobrazuje k jednotlivým pravidlům údaje ve sloupcích **Support** (podpora), **Confidence** (spolehlivost), **Coverage** (pokrytí), **Strength** (sílu), **Lift** a **Leverage** (vliv). Lift je pravděpodobnost P výskytu zároveň předpokladu (Ant) a závěru (Consequent), která je podělena pravděpodobností výskytu předpokladu násobená pravděpodobností výskytu závěru (rovnice 7). Pokud je *Lift = 1*, tak se předpoklad a závěr pravidla pravděpodobně spolu nevyskytují. Při *Lift > 1* je zřejmé, že předpoklad s následkem se mohou vykytovat zároveň (Wikipedia 2020c).

$$Lift = \frac{P(Ant \cap Cons)}{P(Ant) * P(Cons)} \qquad (7)$$

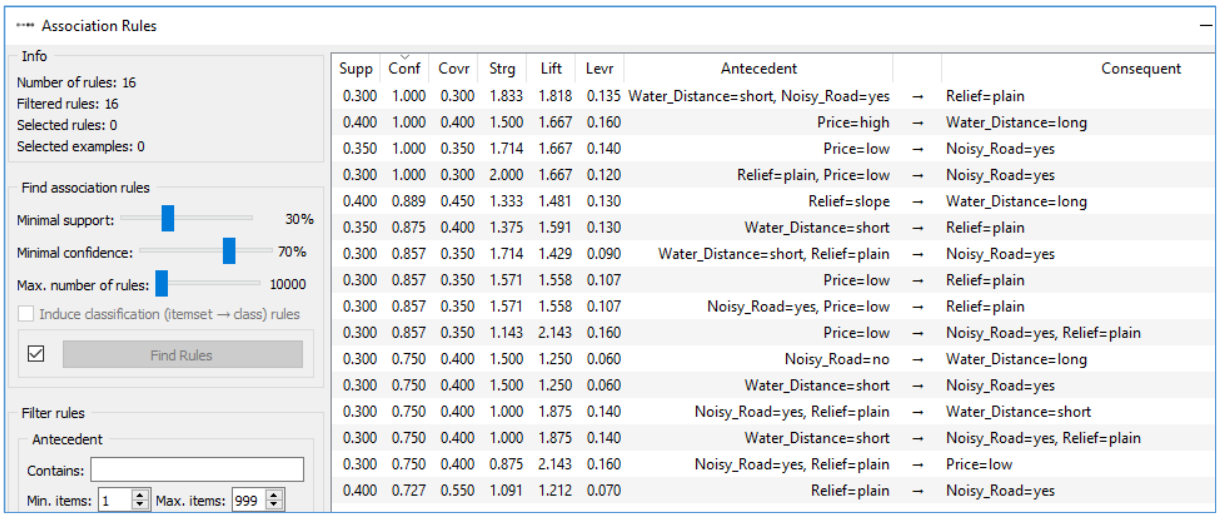
Podle jednotlivých sloupců lze řadit pravidla vzestupně nebo sestupně. Hledáme vždy pravidla, která mají největší podporou (jsou nejčastější) nebo naopak pravidla s nejvyšší spolehlivostí. Pokud je spolehlivost 100 %, tak to znamená, že platí jak předpoklad, tak závěr.



Obr. 129 Výsledná asociační pravidla

První asociační pravidlo na Obr. 129 má největší podporu (0.4) a platí ve všech případech (spolehlivost 1). Pravidlo vyjadřuje, že dům má vysokou cenu a je daleko od vody (nehrozí záplava objektu). Pravidlo je oboustranně platné, neboť má spolehlivost 100 %. Druhé pravidlo s podporou 0,4 a spolehlivostí 0,889 vyjadřuje, že dům, který je na svahu, je ve většině případů i daleko od vody.

Na Obr. 130 se při snížení podpory na 30 % (při spolehlivosti 70 %) objevuje méně časté pravidlo, ale opět oboustranně platné (100% spolehlivosti), a to, že budovy jsou blízko vody, blízko rušné silnice a nachází se na rovině. Z dalších pravidel lze vybírat podle účelu úlohy a interpretovat pravidla. Lze snižovat *Support* a zvyšovat *Confidence*. Při nízké podpoře a vysoké spolehlivosti nalézáme spolehlivé výjimky.



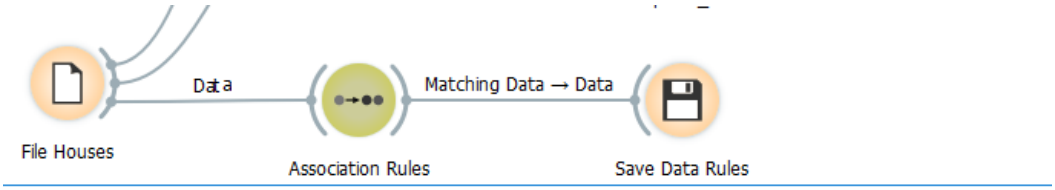
Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.300	1.000	0.300	1.833	1.818	0.135	Water_Distance=short, Noisy_Road=yes	→	Relief=plain
0.400	1.000	0.400	1.500	1.667	0.160	Price=high	→	Water_Distance=long
0.350	1.000	0.350	1.714	1.667	0.140	Price=low	→	Noisy_Road=yes
0.300	1.000	0.300	2.000	1.667	0.120	Relief=plain, Price=low	→	Noisy_Road=yes
0.400	0.889	0.450	1.333	1.481	0.130	Relief=slope	→	Water_Distance=long
0.350	0.875	0.400	1.375	1.591	0.130	Water_Distance=short	→	Relief=plain
0.300	0.857	0.350	1.714	1.429	0.090	Water_Distance=short, Relief=plain	→	Noisy_Road=yes
0.300	0.857	0.350	1.571	1.558	0.107	Price=low	→	Relief=plain
0.300	0.857	0.350	1.571	1.558	0.107	Noisy_Road=yes, Price=low	→	Relief=plain
0.300	0.857	0.350	1.143	2.143	0.160	Price=low	→	Noisy_Road=yes, Relief=plain
0.300	0.750	0.400	1.500	1.250	0.060	Noisy_Road=no	→	Water_Distance=long
0.300	0.750	0.400	1.500	1.250	0.060	Water_Distance=short	→	Noisy_Road=yes
0.300	0.750	0.400	1.000	1.875	0.140	Noisy_Road=yes, Relief=plain	→	Water_Distance=short
0.300	0.750	0.400	1.000	1.875	0.140	Water_Distance=short	→	Noisy_Road=yes, Relief=plain
0.300	0.750	0.400	0.875	2.143	0.160	Noisy_Road=yes, Relief=plain	→	Price=low
0.400	0.727	0.550	1.091	1.212	0.070	Relief=plain	→	Noisy_Road=yes

Obr. 130 Pravidla seřazená podle nejvyšší spolehlivosti Confidence

Navíc lze generování pravidel filtrovat pomocí filtrů na obsah v části ***Antecedent*** anebo ***Consequent.***

V sekci ***Antecedent*** nastavte požadavek, že musí obsahovat (Contains) *Noisy_Road=yes.* V případě více filtrů je lze oddělit čárkou. Stejně tak v sekci ***Consequent*** lze nastavit požadavek, že musí obsahovat (Contains) *Price=low*. V obou částech lze nastavit rozmezí délky předchůdce a následníka (výchozí hodnoty jsou 1 až 999). Pozor často se generují redundantní pravidla, takže je třeba pečlivě vybírat a vyhodnocovat je.

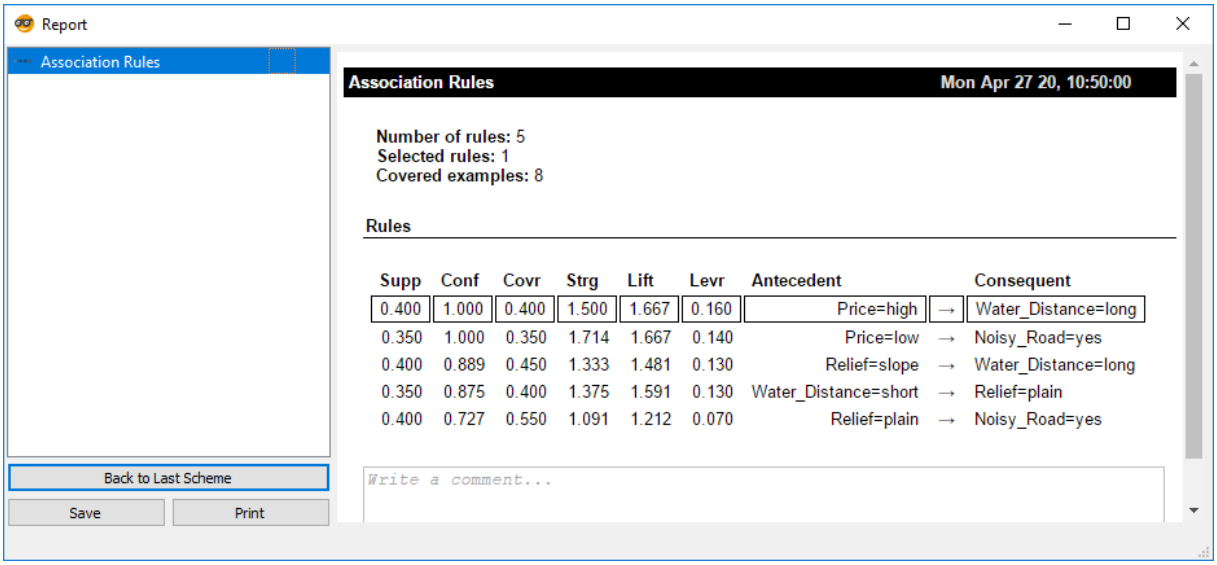
Záznamy, které odpovídají vybranému pravidlu v tabulce (označeny modře), lze uložit do XLSX souboru pomocí uzlu **Save Data**. Pozor neukládá se vlastní vybrané pravidlo, ale záznamy, které jej splňují. Nejlépe je vybrat vždy jedno konkrétní pravidlo a k němu uložit odpovídající záznamy, které jej splňují. Lze takto uložit postupně vždy pro jedno pravidlo soubor odpovídajících vstupních instancí.



Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.300	1.000	0.300	1.833	1.818	0.135	Water_Distance=short, Noisy_Roa...	→	Relief=plain
0.400	1.000	0.400	1.500	1.667	0.160	Price=high	→	Water_Distance=long

Obr. 131 Výběr asociačního pravidla pro uložení odpovídajících instancí

Výsledná pravidla lze také **exportovat do souboru** pomocí ikony **Report** (vlevo dole vedle ikony nápovědy). Lze volit výstupní formát PDF, HTML nebo report. Okno Report je ukázáno na obrázku Obr. 132. Je zde vybráno první pravidlo a výsledkové okno oznamuje, že toto pravidlo pokrývá 8 vstupních instancí.



Obr. 132 Report pravidel do souboru

V uzlu **Asscoiation Rules** lze zaškrtnout i volba *Induce classification rules (itemset → class).* Uzel potom generuje pravidla, která mají na pravé straně pravidla (consequent) kategorickou hodnotu. Pro využití této volby je třeba v uzlu *File* nastavit jedno klasifikační pole jako *Target*.

7.3 Asociační pravidla pro dichotomická data

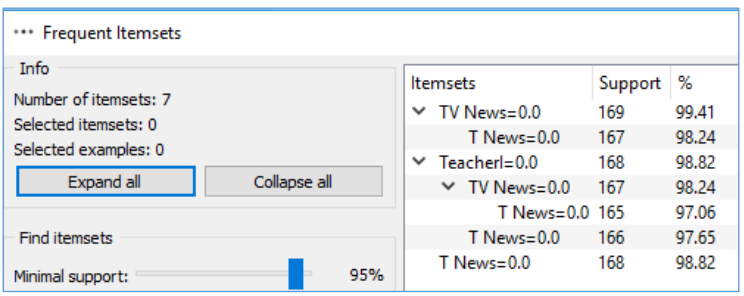
Indukce frekventovaných sad i asociačních pravidel lze provést i pro dichotomická data, která obsahují pro jednotlivé atributy pouze údaj Ano/Ne (nebo 1/0).

Příklad 35

Data *6_Uchazeci.xlsx*, **workflow** *7_AssociateRulesStudents.ows*

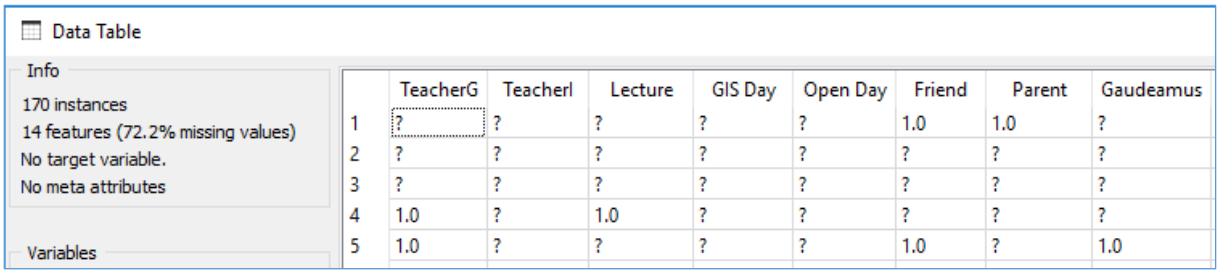
Využijeme data z dotazníků od uchazečů o studium z lekce o rozhodovacích stromech. Jedná se o řídkou matici, neboť převažují negativní odpovědi. Použijte data *6_Uchazeci.xlsx*. Nejprve použijeme list, kde jsou údaje za čtyři roky a potom list s názvem *Sparse*. List *Sparse* obsahuje pouze hodnoty 1. Buňky s nulovými hodnotami jsou prázdné.

Pokud použijeme první list s nulami, tak budou jako nejčastější generovány „negativní“ frekventované sady a pravidla. Negativní pravidlo má tvar $\neg X \Rightarrow \neg Y$ (resp. $X \Rightarrow \neg Y, \neg X \Rightarrow Y$) a vyjadřuje, že pokud se nevyskytuje X, nevyskytuje se ani Y (Chattamvelli 2011). Užitečnost negativních pravidel se musí zvážit podle řešené úlohy. Z hlediska interpretace účinnosti propagačních akcí nám tato pravidla neodhalí, jaká akce má vliv na zájem o studium. Pouze je vidět s podporu přes 95 %, že se student nedozví informaci ani z učitelských novin, ani od učitele předmětu Informatika ani z televize, či rádia.



Obr. 133 Frekventované sady informací pro uchazeče

Z pohledu intepretace je lepší použít list *Sparse* – řídkou matici, kde hodnota nula je nahrazena otazníkem.

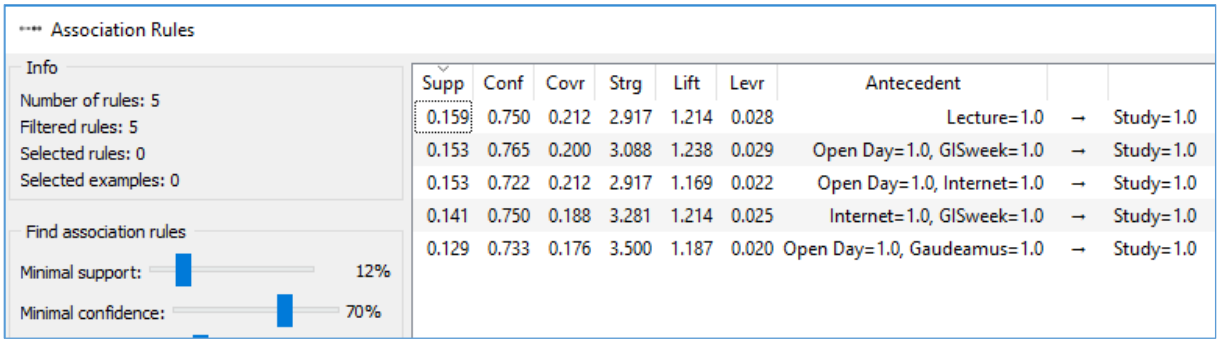


	TeacherG	TeacherI	Lecture	GIS Day	Open Day	Friend	Parent	Gaudeamus
1	?	?	?	?	?	1.0	1.0	?
2	?	?	?	?	?	?	?	?
3	?	?	?	?	?	?	?	?
4	1.0	?	1.0	?	?	?	?	?
5	1.0	?	?	?	?	1.0	?	1.0

Obr. 134 Řídká matice z dotazníku o informacích pro uchazeče

Výsledná pravidla s podmínkou, že pravá strana musí obsahovat pole *Study*, lze již lépe interpretovat.

Z pravidel je zřejmé, že na nástup ke studiu má vliv návštěva uchazeče na Dnu otevřených dveří (Open Day), veletrhu Gaudeamus, realizace přednášky na škole (Lecture) a informace z Internetu (Dobesova 2019a).



Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		
0.159	0.750	0.212	2.917	1.214	0.028	Lecture=1.0	→	Study=1.0
0.153	0.765	0.200	3.088	1.238	0.029	Open Day=1.0, GISweek=1.0	→	Study=1.0
0.153	0.722	0.212	2.917	1.169	0.022	Open Day=1.0, Internet=1.0	→	Study=1.0
0.141	0.750	0.188	3.281	1.214	0.025	Internet=1.0, GISweek=1.0	→	Study=1.0
0.129	0.733	0.176	3.500	1.187	0.020	Open Day=1.0, Gaudeamus=1.0	→	Study=1.0

Obr. 135 Výsledná pravidla pro Study=1

Příklad 36

Data *7_Houses_Dichotom.xlsx*, **workflow** *7_AssociateRulesHouseDich.ows*

K dispozici je soubor domů *7_Houses_Dichotom.xlsx,* kde jsou původní data převedena na dichotomická. U atributů jsou hodnoty 1 a 0 u ceny jsou pouze ponechány hodnoty 1, aby ve výsledcích nebyla „negativní pravidla“. Zkuste samostatně indukovat frekventované sady i asociační pravidla. Výsledky by měly být stejné jako pro původní soubor. Pozor na intepretaci, kdy např. Relief=slope=0 znamená, že se dům nachází na rovině.

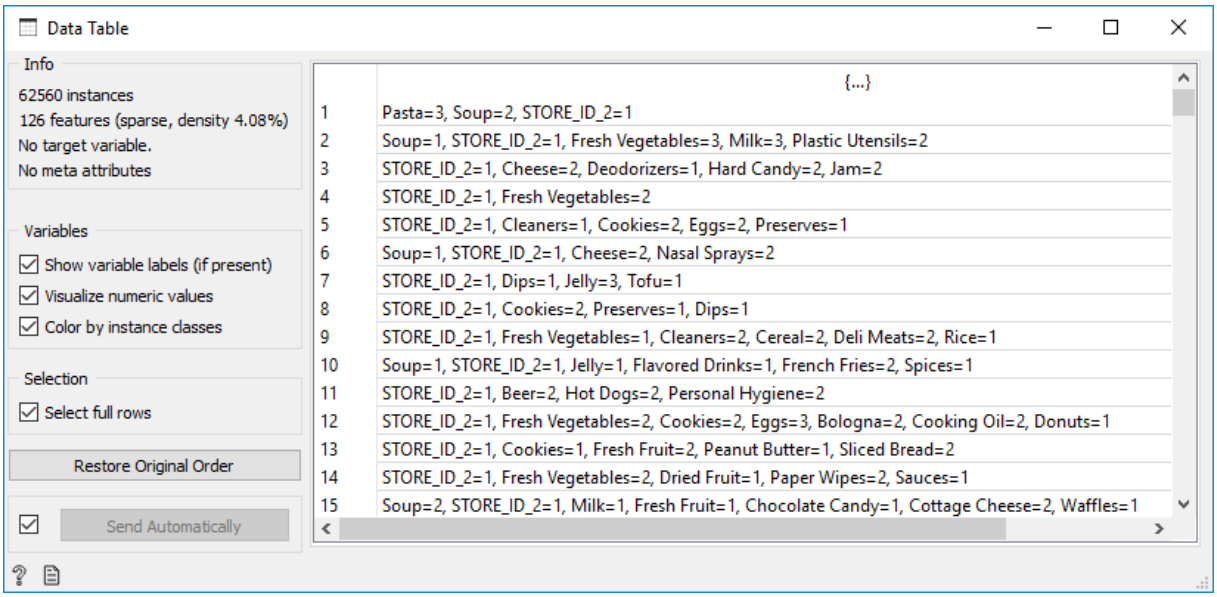
	A	B	C	D	E	F	G
1	**ID_House**	**Water_Dist=long**	**Noisy_Road=yes**	**Relief=slope**	**Price=high**	**Price=low**	**Price=medium**
2	1	1	1	0		1	
3	2	0	0	1			1
4	3	1	0	0	1		
5	4	0	1	0		1	
6	5	0	1	0		1	
7	6	0	1	0		1	
8	7	1	1	0	1		
9	8	0	1	0			1
10	9	1	0	1	1		

Obr. 136 Zdrojová dichotomická data o domech

Příklad 37

Volitelné cvičení

Pro další volitelné procvičení lze použít cvičná data **Foodmart 2000**, která jsou součástí instalace Orange. Představují typická data analýzy nákupního košíku (Market Basket Analysis) pomocí frekventovaných sad instancí a asociačních pravidel. Data obsahují 62 560 instancí řídkých dat o nákupu druhu zboží včetně počtu kusů od daného druhu. Navíc každý nákup obsahuje identifikátor prodejny STORE_ID.



Obr. 137 Vstupní data Foodmart 2000

Při vyhodnocení frekventovaných sad instancí má největší podporu 28 % nákup **Fresh Vegetables**, dále 17 % má podporu **Fresh Fruit** a nákupy dalších potravin uvedených ve výsledku se pohybují s podporou nad 11 %. Je zajímavé, že čerstvá zelenina a ovoce se velice často objevují v nákupech.

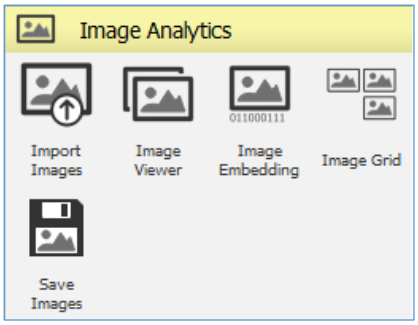
Itemsets	Support	%
Soup	7447	11.9
Dried Fruit	7304	11.68
Fresh Vegetables	17684	28.27
Cheese	7354	11.76
Fresh Fruit	10926	17.46

Obr. 138 Frekventované sady instancí nákupu potravin

8 NEURONOVÉ SÍTĚ A DOPLNĚK IMAGE ANALYTICS

Neuronová síť je algoritmus, který si bere za vzor činnost lidského mozku. Nervová soustava člověka je velmi složitý systém, který je stále předmětem zkoumání. Mozek je tvořen velkým množstvím vzájemně propojených neuronů, které komunikují pomocí elektrických impulsů. Neuron je základní stavební prvek nervové soustavy určený k přenosu, zpracování a uchování informací. Velmi zjednodušené neurofyziologické principy slouží k formulaci matematického modelu neuronové sítě. Princip neuronových sítí je implementován v řadě dostupných analytických a rozhodovacích programových nástrojích. Neuronové sítě podávají extrémně dobré výsledky ve srovnání se „standardními" typy rozhodovacích algoritmů a z toho důvodu jsou také implementovány v nástrojích pro data mining (Petr 2014a).

Pro analýzu obrazu je nutné doinstalovat samostatný doplněk **Image Analytics**. Tento doplněk se doinstaluje z menu volbou **Options** a **Add-ons…** Po instalaci je vlevo nová žlutá skupina **Image Analytics**, kde je dostupné pět různých widgetů. Po instalaci je nutný následný restart aplikace Orange.



Obr. 139 Uzly doplňku Image Analytics

Uzel **Image Embedding** obsahuje několik natrénovaných konvolučních neuronových sítí (pre-trained convolutional neural network), jako například Google Inception v3 (Godec *et al.* 2019). Neuronové sítě pro hluboké učení (deep neural networks) jsou většinou trénovány za určitým účelem na vybrané databance obrázků. Síť Inception v3 je určena pro klasifikace obrázků do celkem 1000 různých tříd (Stanford Vision Lab 2020) (Russakovsky et al. 2015). Účelem natrénované sítě je použití v režimu vybavování.

V uzlu Image Embeding jsou k dispozici tyto natrénované neuronové sítě k rozpoznávání obrazu:

- **SqueezeNet** – malý a rychlý model trénovaný na databázi ImageNet,
- **Inception v3** – Model Inception v3 od Googlu trénovaný na databázi ImageNet,
- **VGG-16** – 16-vrstvá síť, model trénován na databázi ImageNet (10 mil. tag images),
- **VGG-19** – 19-vrstvá síť, model trénován na databázi ImageNet,
- **Painters** – natrénováno na 79 433 obrazech různých malířů,
- **DeepLoc** – model natrénovaný na 21 882 obrázcích buněk.

> *Ukázka nalezení podobných obrazů od malířů Moneta a Maneta ukazují videa*
> (Orange Data Mining 2019a) *https://www.youtube.com/watch?v=R5uchDa_ba4*
> (Orange Data Mining 2019b) *https://www.youtube.com/watch?v=6srGs5w9x8w*

Pouze síť SqueezeNet zpracovává data lokálně na počítači. Ostatní sítě posílají data ke zpracování na vzdálený server.

8.1 Neuronová síť Painters

Data pro kapitolu 8 jsou samostatně v adresáři **Data\8_ImageAnalyticsData**, neboť se jedná o několik podadresářů s řadou obrázků.

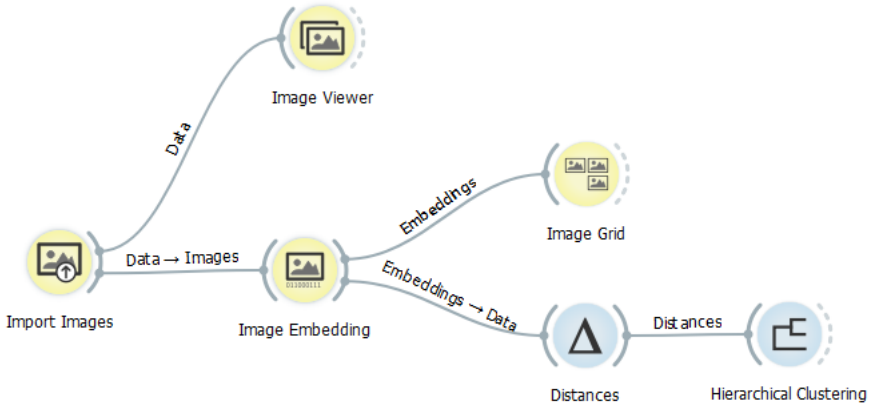
Příklad 38

Data adresář *8_ImageAnalyticsData\Domestic-animals,* **workflow** *8_Animals.ows*

Adresář *Domestic-animals* obsahuje 20 obrázků různých zvířat, která jsou dostupná na GitHub (https://github.com/ajdapretnar/datasets/blob/master/images/domestic-animals.zip).

Příklad je zpracován na základě článku na webu Towards Data Science (Foong 2019).

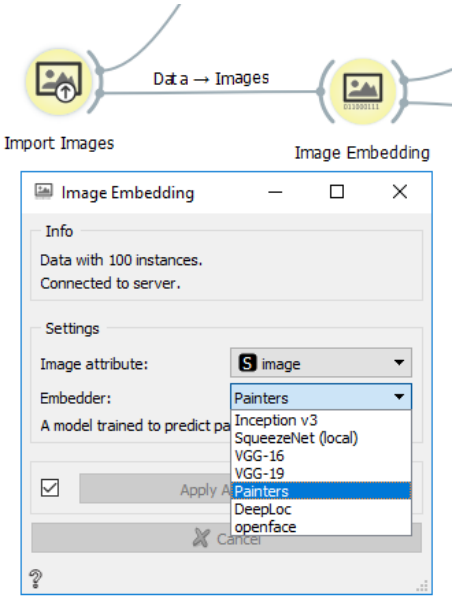
Workflow má následující sled uzlů (Obr. 140). Žlutý uzel **Import Images** slouží pro nahrání obrázků z vybraného adresáře. Pomocí uzlu **Image Viewer** lze prohlížet všechny nahrané obrázky. Za uzlem *Import Images* je uzel **Image Embedding**, kde se vybírá jedna z natrénovaných neuronových sítí, která zpracuje vstupní obrázky. Při velkém počtu vstupních obrázků trvá výpočet nějaký čas a zpracování je naznačeno procenty a animovaným kruhem kolem žlutého uzlu. Výsledek lze prohlížet v uzlu **Image Grid**. Uzly ze skupiny *Image Analytics* lze kombinovat s ostatními uzly Orange, jako jsou například modré uzly ze skupiny *Unsupervised*.



Obr. 140 Workflow pro hierarchické shlukování obrázků na základě klasifikace pomocí natrénované neuronové sítě v uzlu Image Embedding

V uzlu **Image Embedding** vybereme natrénovanou síť embedder **Painters** (Obr. 141). Jedná se o natrénovanou konvoluční neuronovou síť na obrazech malířů. Lze použít i jiný embedder z nabídky, ale ty vrací horší výsledky. Síť Painters byla trénovaná na celkem 79 433 různých výtvarných dílech od celkem 1 584 malířů (Kaggle 2016).

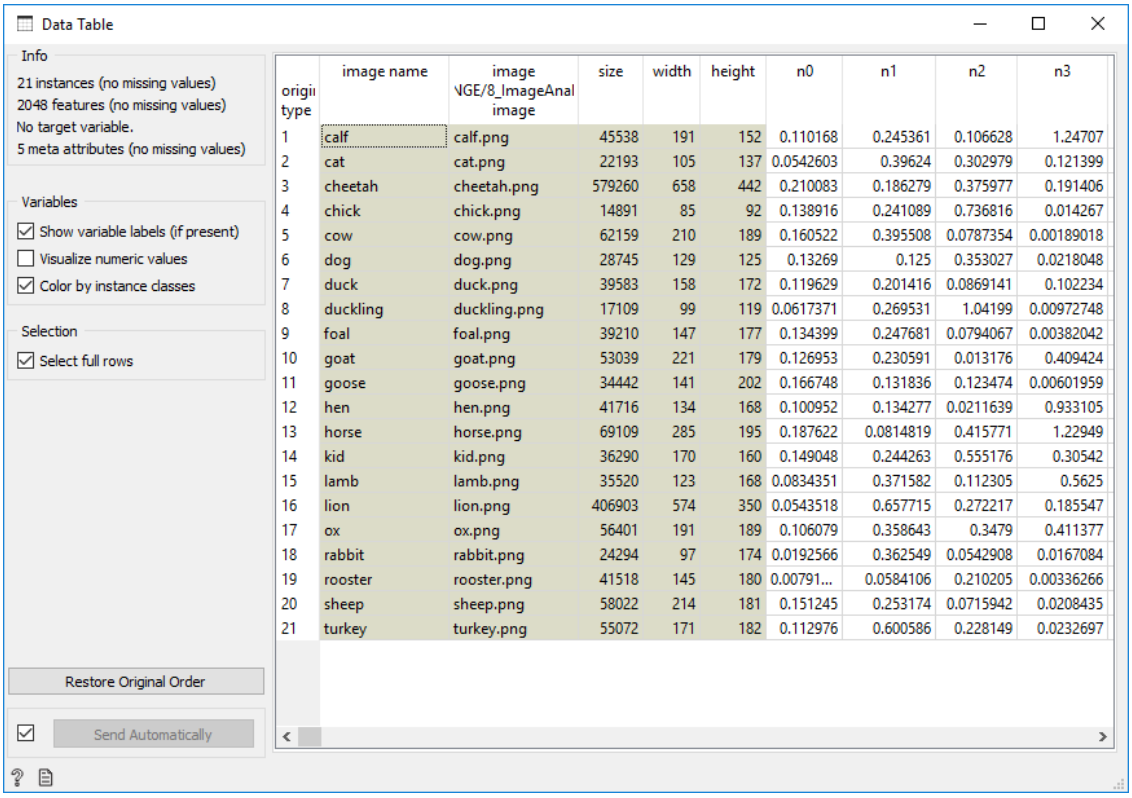
.



Obr. 141 Výběr embedderu Painters v uzlu Image Embedding

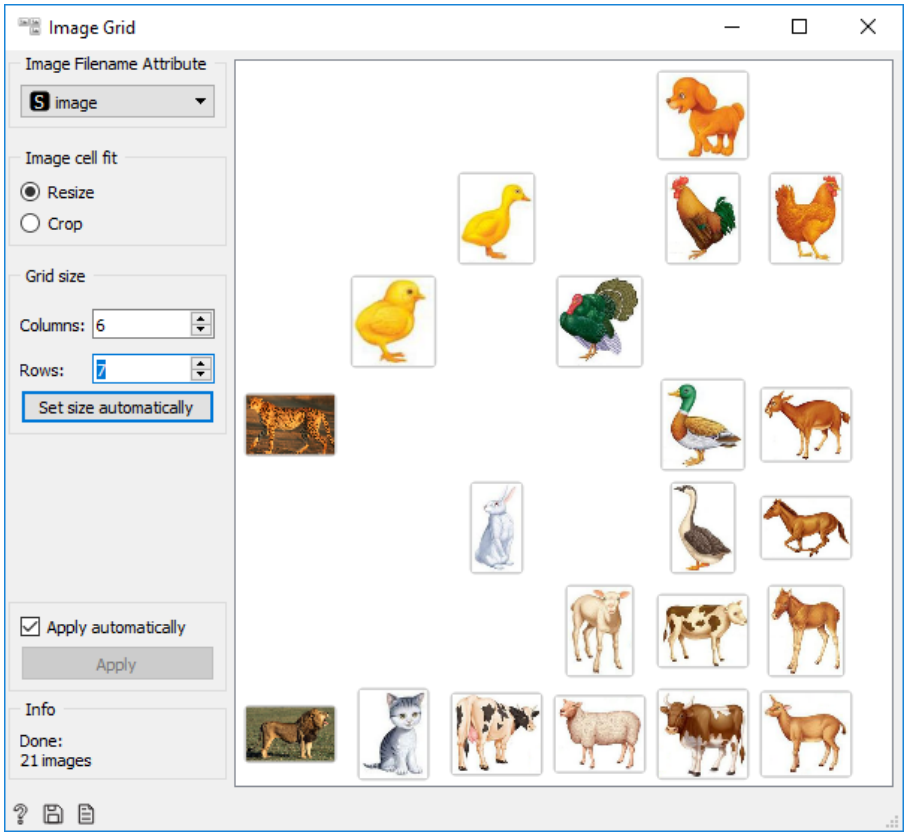
Uzel Image Embeding a embeder Painters vrací pro každý vstupní obraz jako výsledek vektorovou reprezentaci tzv. **feature vector** – 2048 čísel, které charakterizují právě jeden vstupní obraz (atributy n0, n1, n2, …, n2047). Tyto feature vektory lze zobrazit jednoduše v uzlu **Data Table** (na obrázku Obr. 142 nejsou viditelné všechny sloupce n*x*). Každá řádka popisuje jeden vstupní obrázek. Název obrázku je použit z názvu souboru.

> *Volitelně experimentujte i s jinými natrénovanými sítěmi. Například síť VGG-16 a VGG-19 vrací feature vector s vyšším počtem atributů a to s 4096 čísly. Naopak SqueezeNet vrací pouze 1000 čísel ve feature vektoru.*



Obr. 142 Feature vektory vytvořené neuronovou sítí pro 21 obrázků zvířat

Uzel **Image Grid** ukáže v mřížce matice shluky blízkých zvířat. V uzlu lze měnit počet řádků a sloupců matice, a tak docílit vizuálně lepšího zobrazení shluků podobných zvířat. Výchozí počet sloupců a řádků je volen automaticky a lze uživatelsky měnit za účelem názornějšího seskupení. Na obrázku je evidentní shluk vpravo dole, kde jsou čtyřnohá zvířata – skot, a také blízko k sobě mají obrázky drůbeže vpravo nahoře.

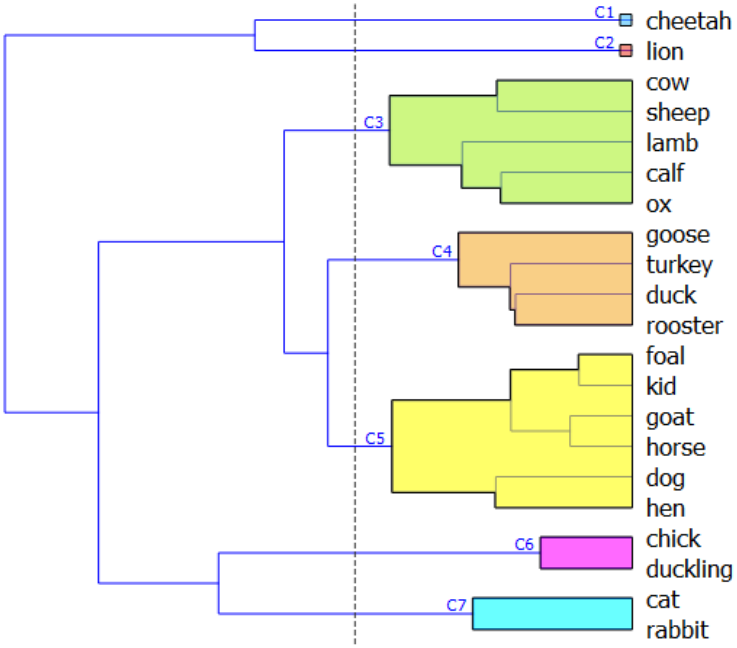


Obr. 143 Výsledek blízkých zvířat v matici Image Grid

Uzel **Image Grid** využívá vestavěnou metodu **t-SNE** k zobrazení obrázků do dvourozměrné roviny na základě jejich spočítaných popisných feature vektorů. Více podobných obrázků bude v ploše blíže. Algoritmus t-SNE vykresluje data pomocí metody stochastického vyhledávání sousedů s t-distribucí. t-SNE je technika redukce dimenzionality podobná MDS (multidimensional scaling), při níž jsou body mapovány do 2-D prostoru podle jejich pravděpodobnostního rozdělení.

> *Metoda t-SNE je dostupná v Orange i jako samostatný uzel v sekci Unsupervised. Zájemci si mohou vyzkoušet tento uzel samostatně včetně experimentování s jeho parametry jako je Perplexity (obdoba určení počtu nejbližších sousedů), počet komponent v metodě PCA (t-SNE se vždy počítá nad komponentami získanými PCA), volba normalizace (standardizace sloupců) atd. V tomto případě exportujte data získaných feature vektorů a zpracujte jako nové workflow, kdy uvidíte výsledky shlukování se zohledněním různé kompaktnosti shluků (důraz na lokální nebo globální strukturu).*

Pomocí uzlu **Distance** (vzdálenost **Cosine**) a uzlu **Hierarchical Clustering** (linkage **Ward**) můžeme najít skupiny podobných zvířat. V dendrogramu na Obr. 144 je viditelné, že jsou seskupeny podobné skupiny zvířat, jako je drůbež, skot a čtyřnohá zvířata. Gepard a lev jsou samostatně v jednoprvkových shlucích patrně z důvodu barevného pozadí obrázku, následně je z nich v dendrogramu vytvořena jedna skupina.



Obr. 144 Hierarchické shlukování obrázků zvířat

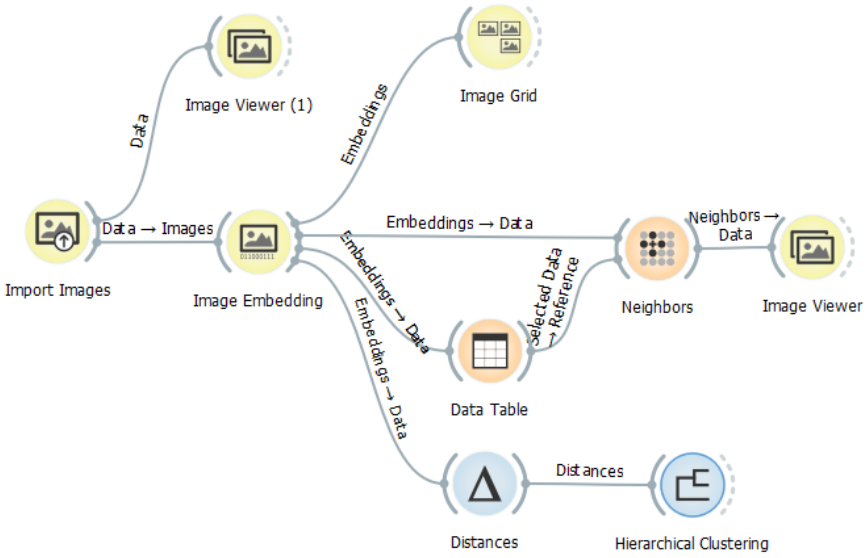
8.2 Podobnost map

Příklad 39

Data adresář *8_ImageAnalyticsData\AllMaps,* **workflow** *8_MapyCZ.ows*

V adresáři **AllMaps** jsou nachystány výřezy map z českého portálu *mapy.cz* jako obrázky. Byly připraveny jednak různé výřezy základní mapy, výřezy historických map, výřezy map leteckých snímků, a navíc všechny základní mapy byly převedeny do černo-bílé podoby. Území byla vybrána pestře – krajská města a středně velká města s okolím, horské oblasti, zemědělské oblasti, oblasti s vodními plochami a to jak pro historické mapy, tak pro letecké snímky a základní mapy. Celkem adresář obsahuje 65 výřezů map. Při pořizování výřezů bylo voleno přibližně stejné měřítko. Velikosti výřezu mají různé rozměry a nebylo ani zamýšleno dosáhnout stejného rozměru výřezů. Natrénovaná neuronová síť nebere ohled na velikost obrázků, ale na jejich obsah.

Otázkou této úlohy je, zda natrénovaná sít Painters, která je natrénovaná na dílech malířů, dobře popíše feature vektorem nachystané mapy. Workflow s natrénovanou neuronovou sítí je podobné jako v předchozím příkladu. Po importu obrázků se v uzlu **Image Embedding** nastaví síť **Painters**.

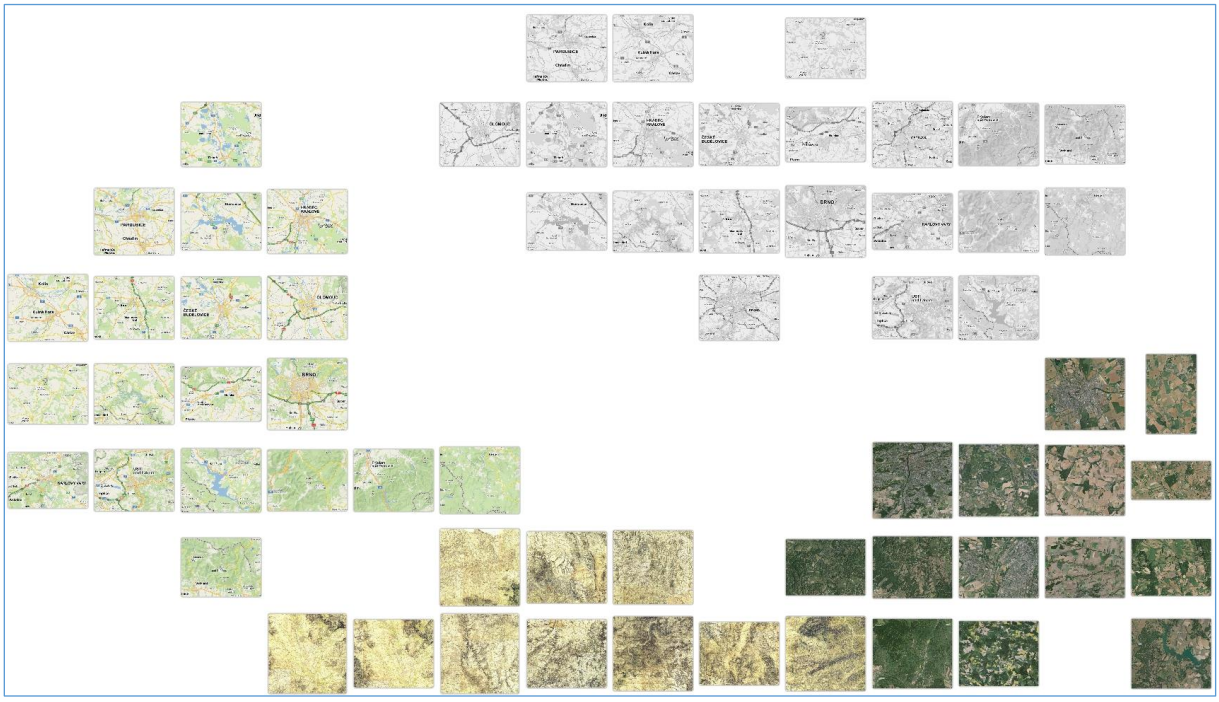


Obr. 145 Workflow hledání podobných mapových výřezů

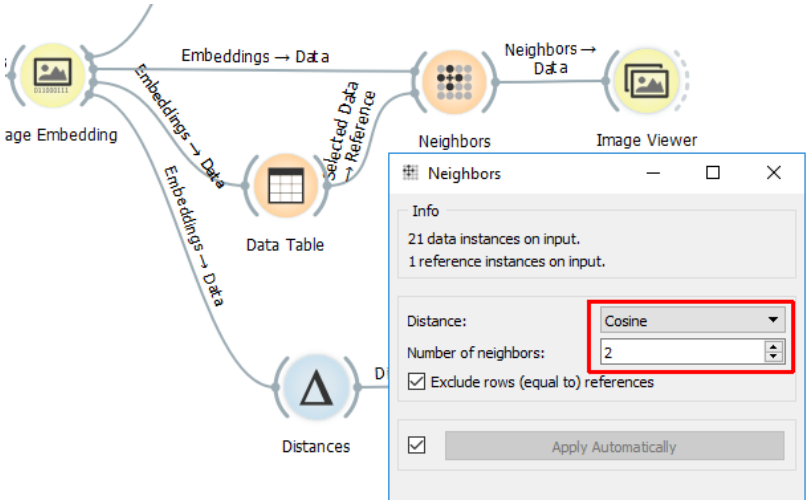
V uzlu **Image Grid** vidíme zobrazení map v matici, kde jsou mapy uspořádány podle blízkosti (Obr. 146). Je evidentní, že je zde vlevo shluk základních map, nahoře jsou si blízké černobílé mapy, dole uprostřed jsou mapy historické a vpravo dole jsou letecké snímky. Je evidentní, že čtyři druhy map jsou seskupeny do čtyř oblastí.

Zajímavé je hledání podobných měst pomocí uzlu **Neighbors** (Obr. 147) (Zupan 2020). Vzdálenost je zvolena *Cosine* z toho důvodu, že vstupní popisný feature vektor obrázků, který je výstupem natrénované neuronové sítě, má vysokou dimenzi a to 2048 atributů.

Pro hledání podobných map je nejprve nutné vybrat v **Data Table** jeden řádek jako referenci – tzn. jednu mapu a k té se hledají podobné mapy. V uzlu se nastaví počet hledaných podobných map ve volbě *Number of neighbors*. Pokud bude odebráno zatržítko *Exclude rows (equal to references)*, bude na výstupu zobrazena i referenční mapa spolu s blízkými mapami. Podobné mapy zobrazí následně uzel **Image Viewer.**



Obr. 146 Matice map v uzlu Image Grid



Obr. 147 Nastavení uzlu Neighbors pro hledání podobných map

Obr. 148 Výsledek nalezení podobné mapy k vzorové mapě Aer_Pole

Obr. 148 ukazuje podobné výřezy leteckých snímků, kde převládá mozaika zemědělských ploch, polí, luk a lesů. Při prohledávání základních map jsou nalezeny různé zajímavé dvojice podobných základních map (Dobesova 2020a). Evidentní je, že byly dobře nalezeny dvě mapy, které obsahují vodní plochy – rybníky v okolí Třeboně a mapa se soustavou vodních nádrží Nové Mlýny (Obr. 149). Druhá dvojice podobných map ukazuje výřez s městy Kolín a Pardubice a sítí silnic první a druhé třídy s minimem zelených lesních ploch a převahou zemědělské půdy (Obr. 150).

Obr. 149 Dvojice podobných map s vodními plochami

Obr. 150 Dvojice podobných základních map

8.3 Kategorizace mapy

Příklad 40

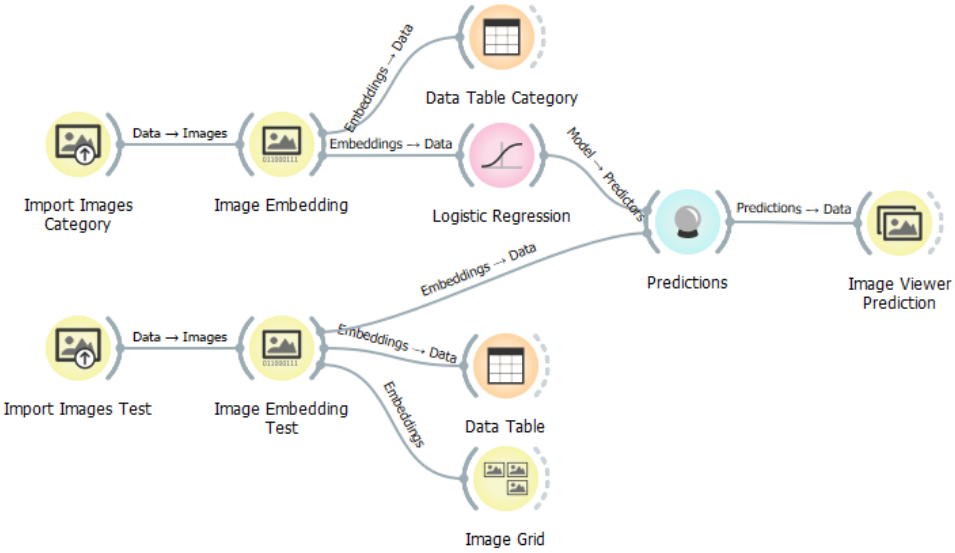
Data adresář *8_ImageAnalyticsData\MapCategory* a adresář *Test,* **workflow** *8_PredictMap.ows*

V této úloze budeme opět pracovat se stejnými mapovými výřezy jako v předchozím příkladu. Tyto mapy jsou ale uloženy ve čtyřech samostatných podadresářích podle druhu mapy (adresáře Aerial, Base, CB, Hist). Název adresáře je potom automaticky přiřazen jako **kategorie k obrázku**, což je viditelné v uzlu Data Table jako nový samostatný sloupec. Tyto mapy můžeme chápat jako trénovací data se známou kategorií, která jsou následně využita v logistické regresi pro určení kategorie čtyř testovacích map o neznámém přiřazení do kategorie.

Ve workflow jsou dvě samostatné větvě. V první větvi se naimportuje 65 map z adresáře *MapCategory* pomocí uzlu *Import Images*. Ve druhé větvi se naimportují 4 mapy z adresáře *Test.* Za oba vstupní uzly je připojen uzel *Image Embedding* (embedder *Painters*).

Dále je v horní větvi s trénovacími daty připojen růžový uzel **Logistic Regression** (nastavení uzlu logistické regrese ponechat výchozí). Obě větvě workflow spojuje uzel **Predictions**. V tomto uzlu vidíme výsledek predikce kategorie pro čtyři mapy podle kategorií trénovacích map (Obr. 152).

Ve výsledkovém okně predikce můžeme zapnout zobrazení pravděpodobnosti predikce (*Show probabilities for*) výběrem myší. Je vidět, že první dvě mapy se zaklasifikovaly správně s maximální pravděpodobností 1 do kategorie Aerial – tj. leteckých map. Třetí (i čtvrtá) mapa je také zařazena správně do kategorie Hist – historických map s poměrem pravděpodobností jednotlivých kategorií 0.01 : 0.02 : 0.00 : 0.97. Navíc je pravděpodobnost znárorněna barevnou čárou podle kategorie. Výslednou predikci kategorie můžeme zobrazit v uzlu *Image Viewer*, kde nastavíme popis volbou výsledné kategorie označené jako *Logistic Regression*. Predikce přinesla překvapivě dobré výsledky určení kategorie map (Dobesova 2019b).



Obr. 151 Workflow pro určení druhu mapy na základě testovacích dat

Predictions

Show probabilities for

Aerial
Base
CB
Hist

	Logistic Regression	image name	image	size	width	height	n0
1	1.00 : 0.00 : 0.00 : 0.00 → Aerial	Aer1	Aer1.jpg	143534	451	356	0.0801169
2	1.00 : 0.00 : 0.00 : 0.00 → Aerial	Aer2	Aer2.jpg	162454	475	339	0.0538663
3	0.01 : 0.02 : 0.00 : 0.97 → Hist	Hist1	Hist1.jpg	172759	478	389	0.167575
4	0.00 : 0.00 : 0.00 : 0.99 → Hist	Hist2	Hist2.jpg	223540	494	453	0.058509

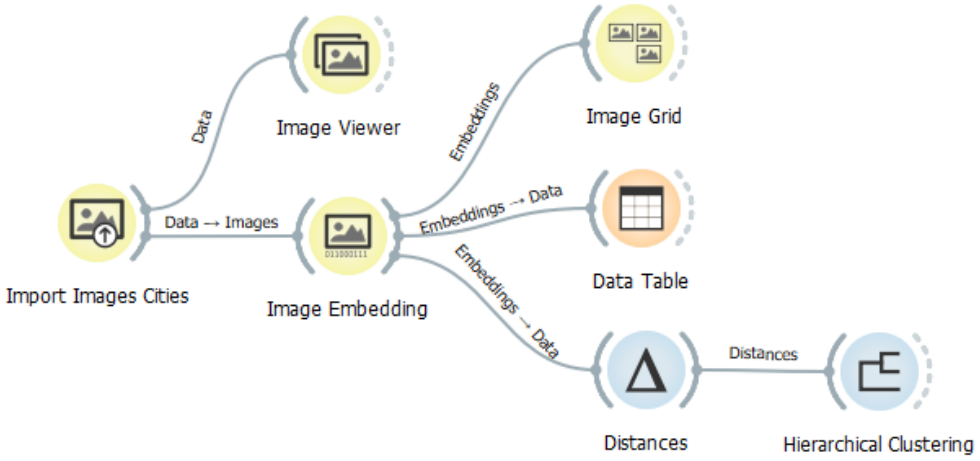
Obr. 152 Výsledek správné predikce typu mapy pomocí logistické regrese

8.4 Hledání podobných evropských měst

Příklad 41

Data: adresář *8_ImageAnalyticsData/100Cities,* **workflow** *8_Landuse.ows*

Natrénovanou síť *Painters* lze použít i pro mapy, které zobrazují typy využití půdy – landuse měst. Zdroj dat je z datové sady Copernicus Urban Atlas (Copernicus Programme, 2020). Zde jsou dostupné celé funkční urbanistické celky FUA (Functional Urban Area), které zahrnují centrální město a širší okolí (vysvětlení viz Eurostat, 2020). Nejprve byly nachystány kruhové mapové výřezy 100 měst podle barevné legendy Urban Atlasu (Janoušek 2019). Data byla zpracována obdobným workflow jako v příkladu hledání podobných map. V dendrogramu byly hledány dvojice, které byly spojeny na nejnižší úrovni, tudíž jsou si nejpodobnější.

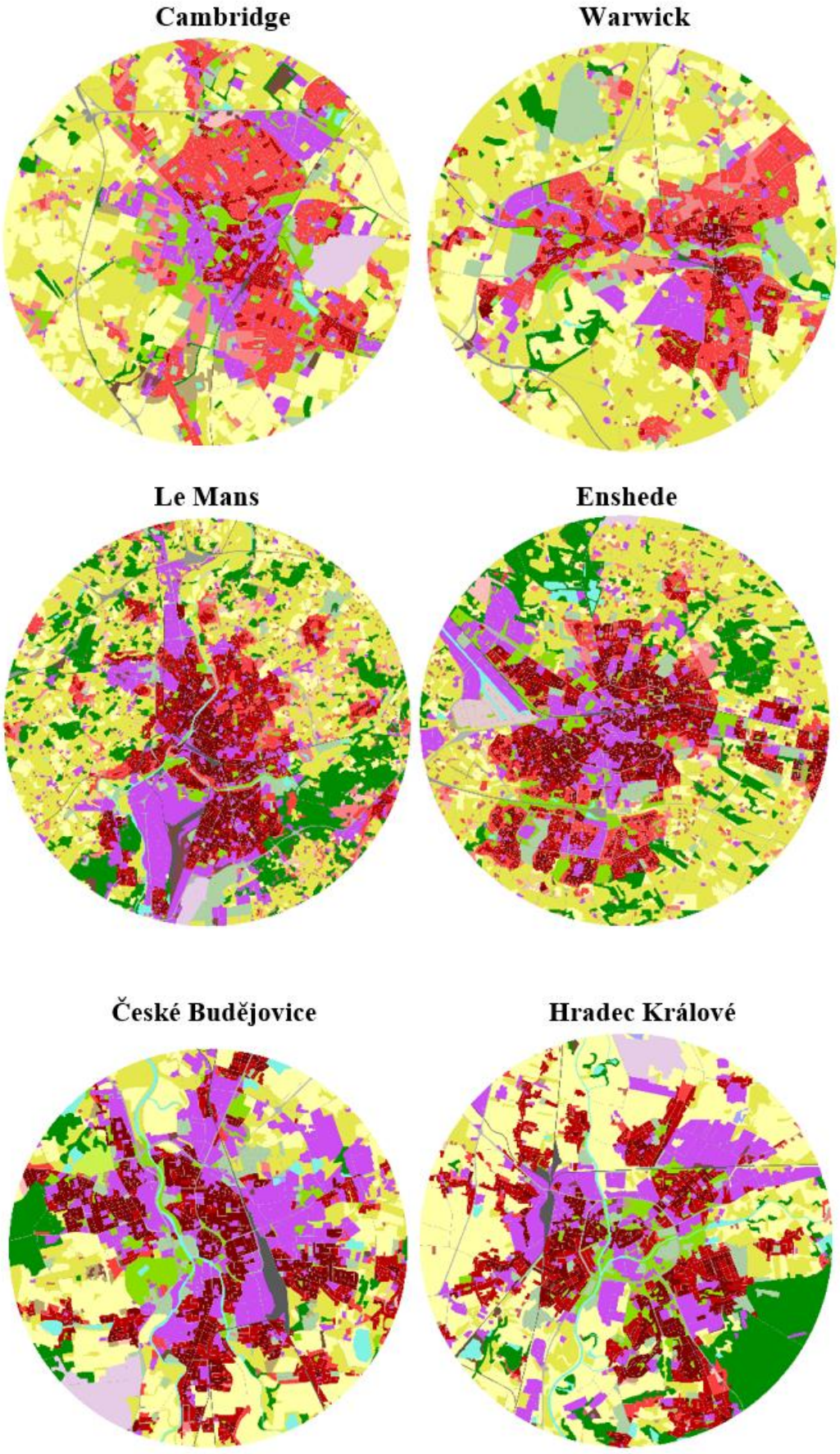


Obr. 153 Workflow pro nalezení podobných měst podle landuse

Výsledky přinesly zajímavé dvojice podobných evropských měst (Dobesova, 2019). Příklady výřezů landuse dvojic měst jsou uvedeny na Obr. 154. První dvojice měst Cambridge a Warwick si je podobná chybějící hustou zástavbou v centru (tmavě červená barva), kdy převažuje řidší zástavba (světle červená barva). Hustá zástavba se vyskytuje sporadicky. Většina měst právě obsahuje v centru hustou zástavbu, která je většinou starší historickou zástavbou. Charakteristické je i podobné okolí obou měst, kdy převládají pastviny (světle zelená barva). U jiných Evropských měst často převažuje v okolí měst orná půda a louky. Pastviny se vyskytují sporadicky.

Druhá dvojice měst Le Mans a Enschede se vyznačují pestrou mozaikou husté zástavby s hustou mozaikou malých ploch průmyslových, komerčních a vojenských areálu v okolí obou měst (fialová barva). V okolí obou měst se objevují i meší plochy lesů.

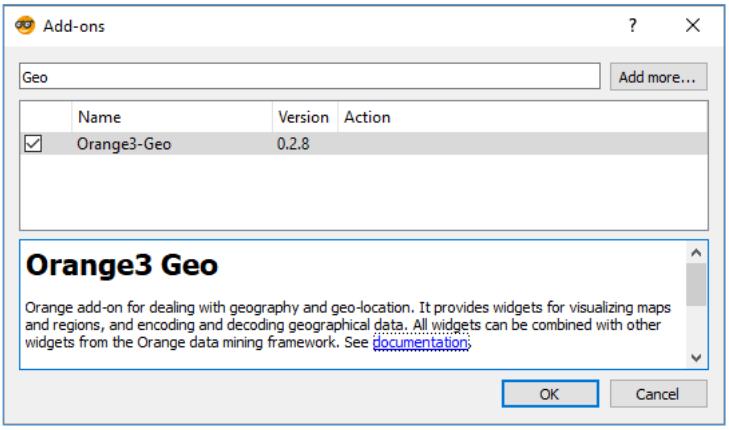
Dvě česká krajská města Hradec Králové a České Budějovice jsou si podobná díky několika odděleným malým centrům husté zástavby (původně několik vesnice) a velkým průmyslovým a komerčním areálům na území města (Dobesova 2019b). Několik lesní porostů tvoří spíše souvislé větší plochy. Podobný je tvar i vodních ploch – protékajících řek oběma městy.



Obr. 154 Dvojice podobných měst ze sady 100 evropských měst podle landuse (Dobesova, 2019)

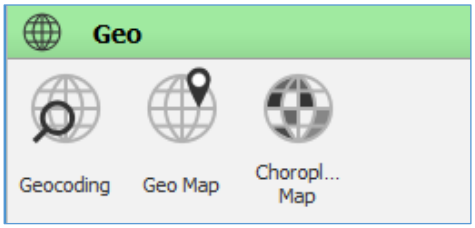
9 PROSTOROVÁ DATA A DOPLNĚK GEO

Doplněk **Geo** umožňuje kódování a dekódování geografických dat, dále zobrazení dat v mapách a tvorbu kartogramů (choropleth map). Všechny uzly mohou být kombinovány s ostatními uzly Orange. Je tak možné před zobrazením dat v mapě provést předzpracování vstupních dat pomocí jiných uzlů. Doplněk Geo je nutné doinstalovat při spuštění Orange s administrátorským oprávněním. Následně je nutný restart Orange.



Obr. 155 Instalace doplňku Geo

Sekce Geo obsahuje tři zelené uzly: *Geocoding*, *Geo Map* a *Choropleth Map*.



Obr. 156 Nabídka widgetů (uzlů) ve skupine Geo

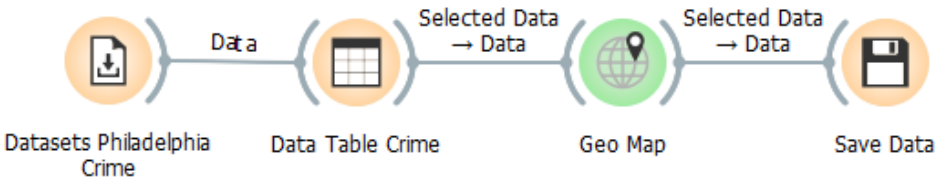
9.1 Zobrazení bodových dat v mapě pomocí Geo Map

Uzel **Geo Map** umí zobrazit vstupní data na základě údajů o poloze latitude a longitude, které jsou součástí dat. Předpokládá se, že jsou souřadnice v souřadnicovém systému WGS 84 (EPSG:4326).

Příklad 42

Data *Philadelphia Crime* z nabízených cvičných datasetů Orange nebo *9_Philadelphia-crime.xlsx*

Workflow 9_Geo.ows



Obr. 157 Workflow s použitím uzlu Geo Map pro zobrazení poloh kriminálních činů v mapě

Dataset *Philadelphia Crime* obsahuje bodovou vrstvu kriminálních činů v jedné čtvrti města Philadelphia z rozmezí let 2006 až 2012 v počtu 9 666 činů. Tabulka kriminálních činů obsahuje kromě data činu, jeho souřadnic *Lat* a *Lon* také typ kriminálního činu (*Liquor Law Violations, Public Drunkenness, Gambling, Homicide, Family Abuse*, *Prostitution*). Ve workflow je uzel *Data Table*, kde si lze data prohlédnout a vybrat záznamy pro další zpracování v uzlu **Geo Map**. Pro další zpracování jsou vybrány všechny záznamy.

V uzlu **Geo Map** se nastavují v dialogovém okně tyto parametry:

- *Map* – podkladová mapa (na výběr je *Open Street Map, Black and White, Topographic, Satellite, Print, Dark*)
- *Lat, Lon* – názvy atributů, které obsahují informaci o souřadnicích. Limit na rozsah souřadnic Latitude je v rozmezí -85.0511(S) do 85.0511(N), Longitude má limit v rozsahu -180(W) až 180(E). Pokud jsou atributy pojmenovány Lat a Lon, tak jsou tyto atributy automaticky rozpoznány a navoleny. Jinak je nutné zdrojové atributy nastavit.

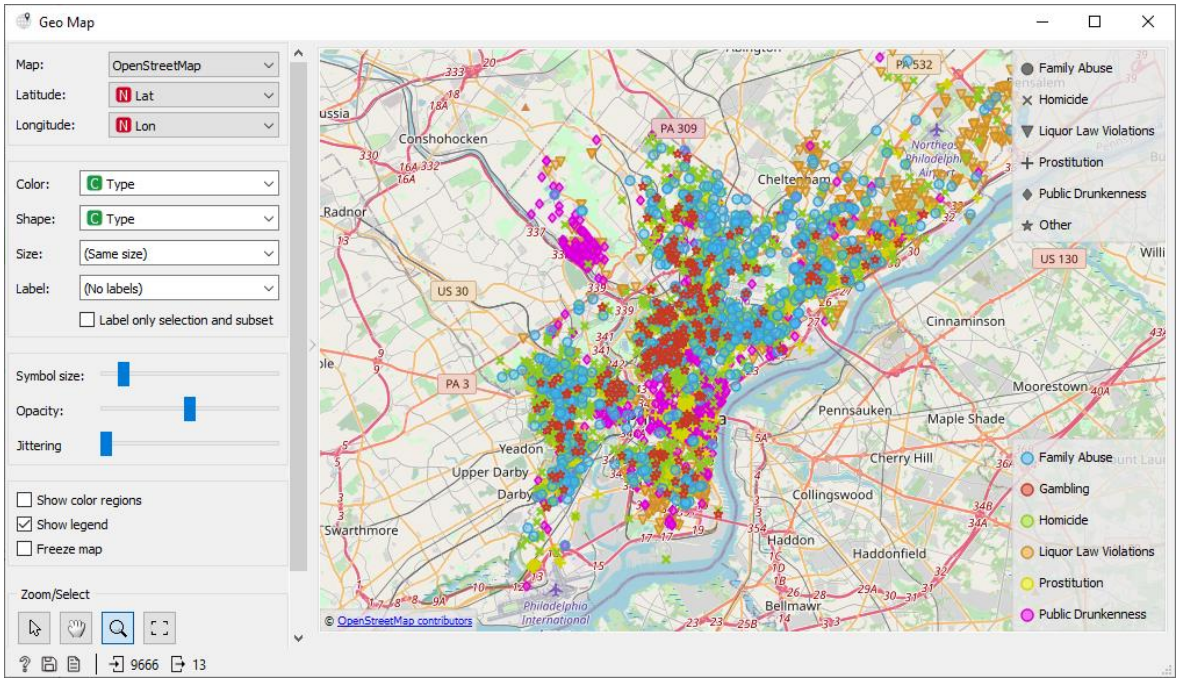
- *Color, Shape* – barva a tvar bodového znaku je nastavena podle počtu kategorií zvoleného atributu, zde je to typ kriminálního činu. Volba různých barev a tvaru bodových znaků se provede automaticky, nelze podle uživatelsky změnit.
- *Size, Label* – lze navázat na atributy, pokud existují vhodné atributy ve zdrojových datech. Zde lze popsat body datem nebo druhem kriminálního činu.

- *Symbol size* – nastavení velikosti bodového znaku
- *Opacity* – zde znamená míra výplně bodového znaku
- *Jittering* – míra odsunutí (disperse) překrývajících bodů z původních pozic

- *Show color regions* – vykreslí barevně oblasti, podle převládajících činů, kdy barvy souhlasí s barvou odpovídajících bodů. Shlukování blízkých bodů do oblastí je pomocí metody naive greedy clustering (funguje s omezením na 600 bodů v okně náhledu).

- *Show legend* – zobrazí v pravém rohu legendu pro použité barvy a bodové znaky
- *Freeze map* – nedojte k překreslení mapy při změně vstupních dat

- *Zoom, pan, zoom to fit* tlačítka – jsou určené pro přiblížení a posun mapy

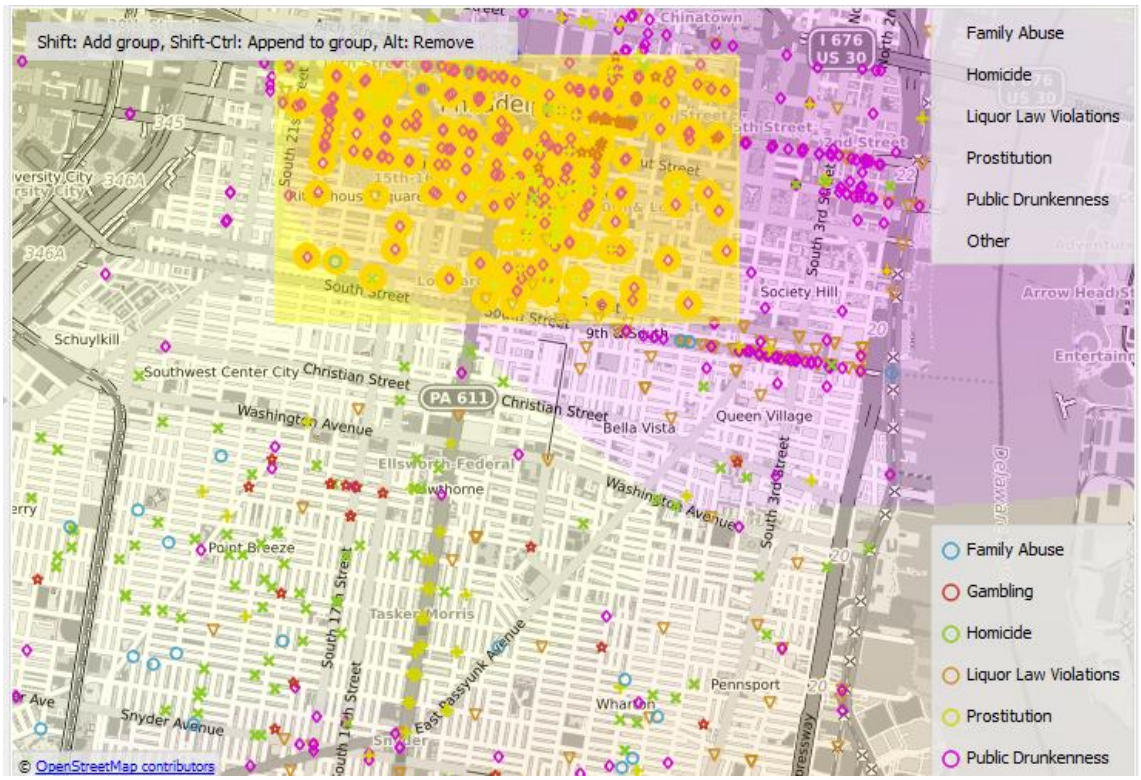


Obr. 158 Okno Geo Map se zobrazením bodů kriminálních činů

Vyzkoušejte nastavení *Jittering*, který různě odsune body z původní polohy. Body se potom tolik nepřekrývají. Je pak lépe zřetelné, jak hustý je výskyt některých činů na určitých ulicích.

Obr. 159 Stejné území bez odsunu bodů pomocí Jittering (vlevo) a s nastaveným Jittering (vpravo) pro čin Prostitution

Pomocí tlačítka *Select* lze v mapě obdélníkem vybrat určité body a ty potom uložit pomocí uzlu *Save Data.*



Obr. 160 Výběr obdélníkové oblasti pro následnou možnost uložení výběru dat do tabulky. Fialově je vykreslena oblast, kde převládá Public Drunkenness (zapnuta volba Show color region)

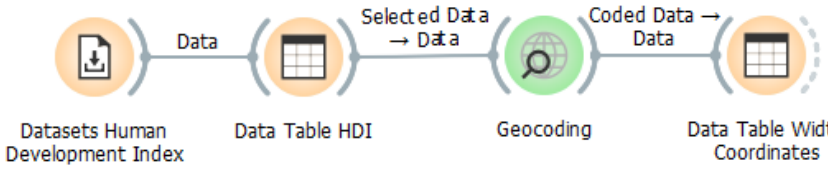
> *Další dokumentace je zde* (Orange Data Mining 2015)*:*
> *https://orange3.readthedocs.io/en/3.5.0/widgets/visualize/geomap.html*

9.2 Geokódování pomocí uzlu Geocoding

Uzel **Geocoding** umí podle názvů regionů přiřadit (geokódovat) těmto bodům nebo regionům geografické souřadnice. Pokud se jedná o celé státy nebo regiony, tak je přiřazen centroid polygonu. Stejně je možné použít i dekódování, kdy se ze souřadnic *Latitude* a *Longitude* se vygenerují názvy regionů nebo států. Doplněná data lze uložit jako novou tabulku uzlem *Save Data*.

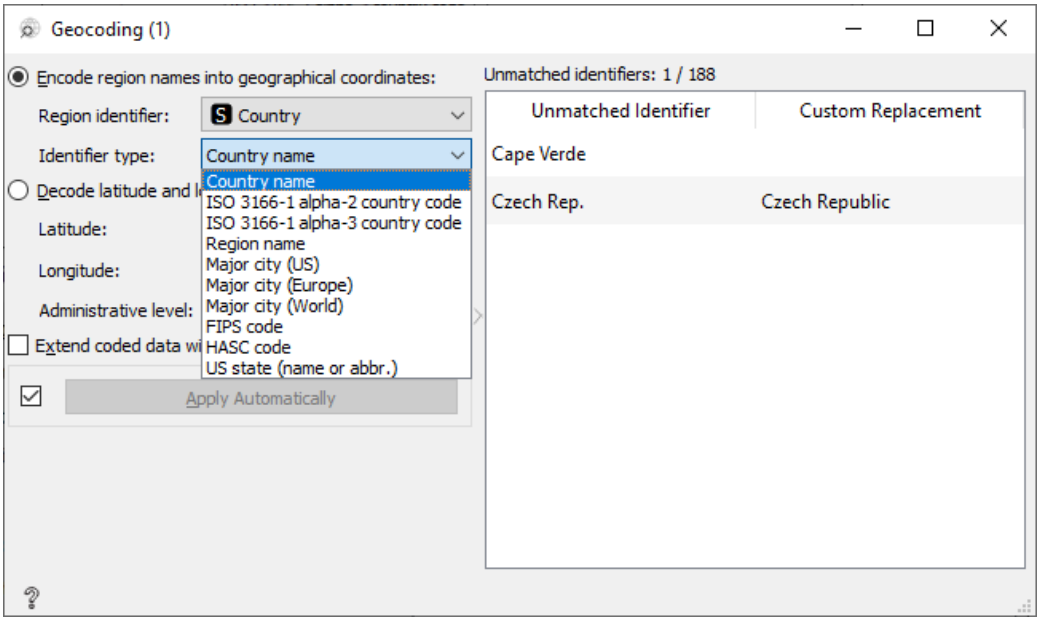
Příklad 43

Data *HDI* z nabízených cvičných datasetů Orange (nebo *9_HDI.xlsx*), ***workflow*** *9_Geo.ows*



Obr. 161 Workflow pro geokódování států

Nejprve pomocí horních voleb vyzkoušíme kódování – volba **Encode**. Vstupní dataset HDI obsahuje sloupec *Country* s názvy států celého světa, který zvolíme jako *Region identifier*. Dále lze zvolit typ identifikátoru, kde je na výběr Country name, ... , Region name atd. Nenalezené identifikátory jsou vypsané v pravém okně a lze je manuálně opravit doplněním správného názvu v pravém sloupci.



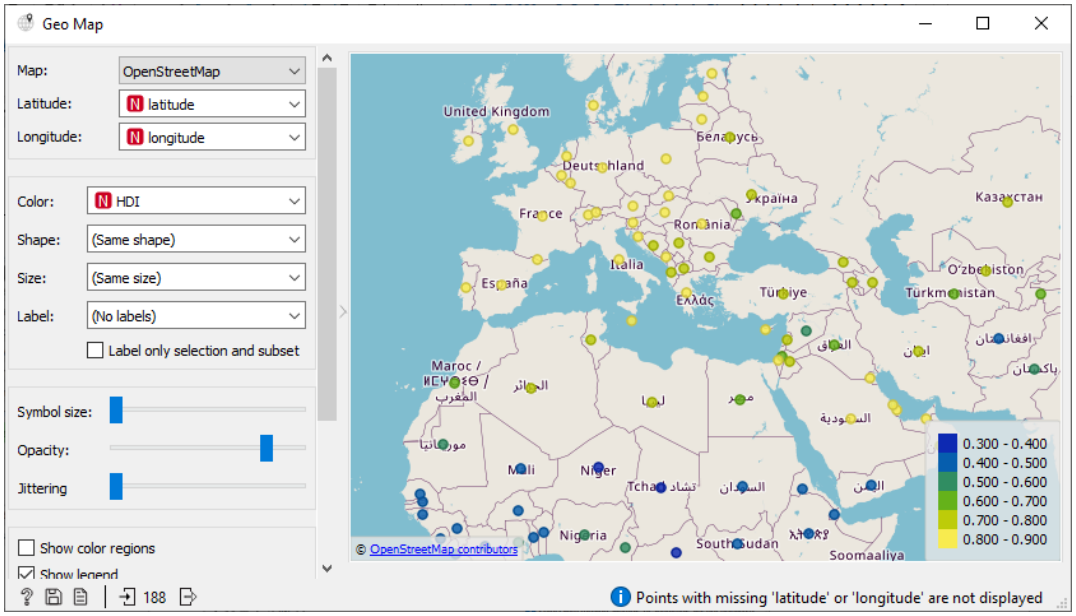
Obr. 162 Nastavení pro doplnění souřadnic – geokódování

Do tabulky jsou doplněny dva nové sloupce s **latitude** a **longitude**. Obohacenou tabulku si zobrazíme uzlem *Data Table* a uložíme ji pro testování druhé možnosti a to určení států pomocí souřadnic (název souboru *9_HDI_Coordinates.xlsx*).

	HDI	Country	latitude	longitude	Life expectancy
1	0.949	Norway	79.8486	22.6903	81.7
2	0.939	Australia	-24.915	133.076	82.5
3	0.939	Switzerland	46.8119	8.427	83.1
4	0.926	Germany	51.1135	10.5197	81.1

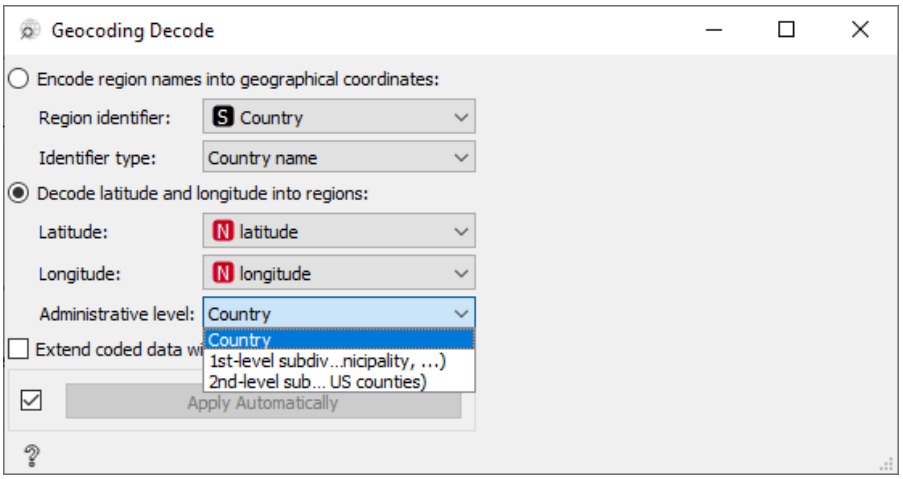
Obr. 163 Ukázka doplněných dat o dva nové sloupce latitude a longitude

Polohu vygenerovaných centroidů států můžeme zkontrolovat zobrazením pomocí přidání uzlu uzlu *Geo Map*.



Obr. 164 Zobrazení vygenerovaných souřadnic centroidů států.

Pro dekódování sestavíme další workflow s použitím stejného uzlu **Geocoding** a zvolíme druhý přepínač **Decode**. Atributy se souřadnicemi jsou automaticky rozpoznány, pokud se jmenují *latitude* a *longitude*, jinak je nutné je ručně nastavit. Dále se zvolí administrativní úroveň. Do dat je doplněn nový sloupec s názvem *name.* Název sloupce nelze volitelně zadat.



Obr. 165 Nastavení dekódování

Data zobrazíme pomocí uzlu *Data Table*. Můžeme porovnat hodnoty obou sloupců *Country* a *name*, zda se shodují. Nová data můžeme uložit.

	Country	name	Life expectancy
1	Norway	Norway	81.7
2	Australia	Australia	82.5
3	Switzerland	Switzerland	83.1
4	Germany	Germany	81.1

Obr. 166 Obsah dat po dekódovaní

> *Poznámka: Vyzkoušejte geokódovat data s názvy krajů, ORP nebo okresů ČR s použitím anglických názvů.*

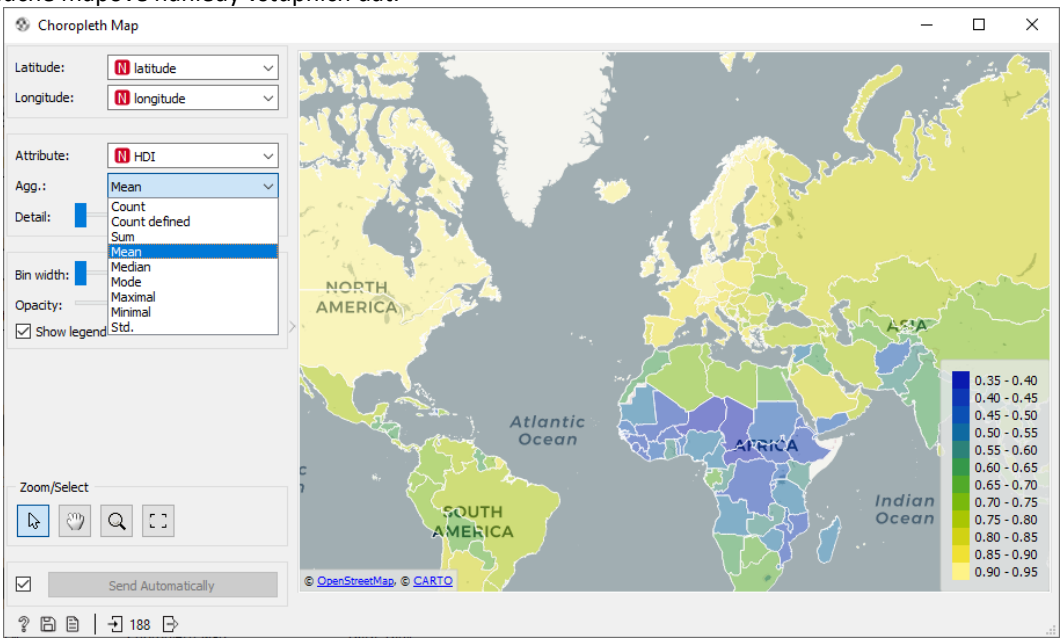
9.3 Tvorba kartogramu pomocí uzlu Choropleth map

Příklad 44

Data *HDI* z nabízených cvičných datasetů Orange (nebo *9_HDI.xlsx*)

Workflow *9_Geo.ows*

Uzel **Choropleth Map** umí vykreslit plošný kartodiagram na základě souřadnic *lat, lon*. Při vstupních datech, která nejsou relativní, vzniká psuodokartogram. O relativní data se jedná, pokud jsou přepočtena na jednotku plochy zobrazovaného území, nebo se jedná o přepočet na 1000 obyvatel apod. Nastavení barevné stupnice a dalších vlastností kartogramu a mapy v uzlu Choropleth Map má jen omezené možnosti. Pro tvorbu plnohodnotných kartografických výstupů je lepší použít GIS software nebo grafický editor. Tento uzel je vhodné použít jen pro jednoduché mapové náhledy vstupních dat.



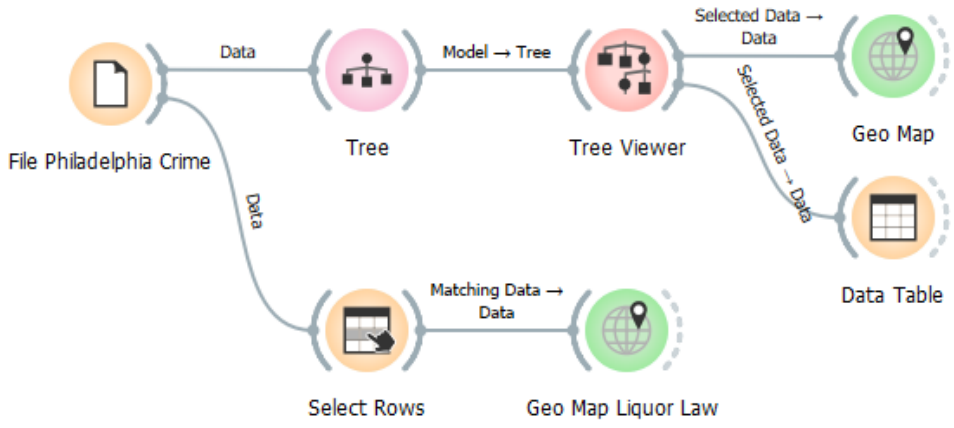
Obr. 167 Dialog uzlu Choropleth Map a vytvořený kartodiagram pro Human Development Index

9.4 Vymezení převahy jevů v území pomocí rozhodovacího stromu

Uzly pro práci s prostorovými daty lze kombinovat s ostatními uzly Orange. Následující příklad ukáže, jak lze pomocí rozhodovacího stromu vymezit oblasti, kde se shlukují a převládají konkrétní druhy trestného činu v určité lokalitě města Philadelphia. Bude použit klasifikační rozhodovací strom.

Příklad 45

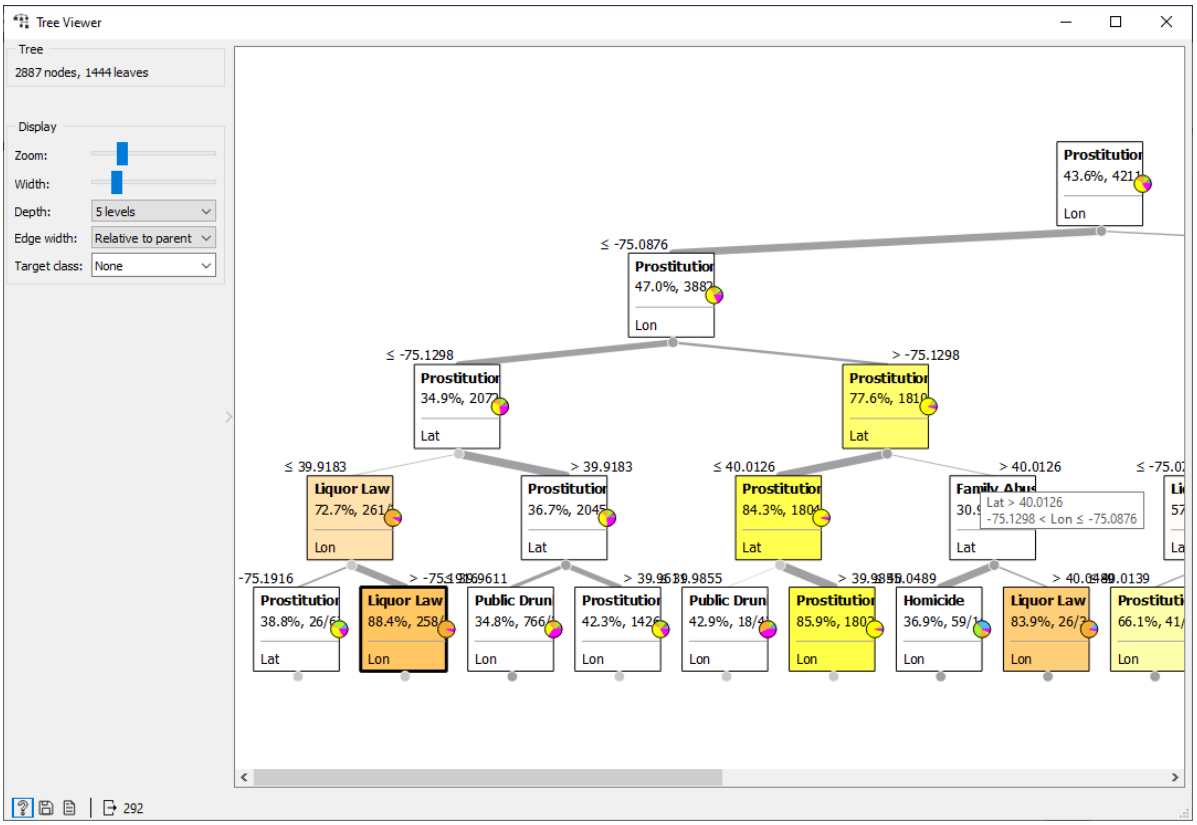
Data *9_Philadelphia-crime.xlsx*, **workflow** *9_PredictCrime.ows*



Obr. 168 Workflow s rozhodovacím stromem pro zobrazení převládajících trestných činů

Nejprve je nutné nastavit v uzlu *File* atribut **Type** na roli *target,* neboť v následujícím uzlu **Tree** je chápán jako predikovaná veličina v rozhodovacím stromě. Pomocí uzlu **Tree Viewer** si zobrazíme poměrně rozsáhlý klasifikační rozhodovací strom, který zobrazíme pouze do hloubky páté úrovně. Vidíme, že na rozhodování mají vliv rozsahy Lon a Lat, tj. umístění činů.

Jednotlivé barevné listy stromu se sytější barvou představují oblasti, kde je převládající výskyt jedné kategorie činu. Vyberte myší uzel druhý uzel zleva v páté úrovni (viz Obr. 169). Vybraný uzel se orámuje silnou černou čárou. Tento list stromu představuje oblast, kde převládají činy *Liquor Law Violation* (88,4 %), kterých se tam nachází 258 z celkového počtu 292 trestných činů ve čtvrti města Philadelphia. Větev rozhodovacího stromu lze přepsat ve formě pravidel pro souřadnice *Lat* a *Lon* a získat tak souřadnice oblasti s převažujícím výskytem činů pro vybraný list.



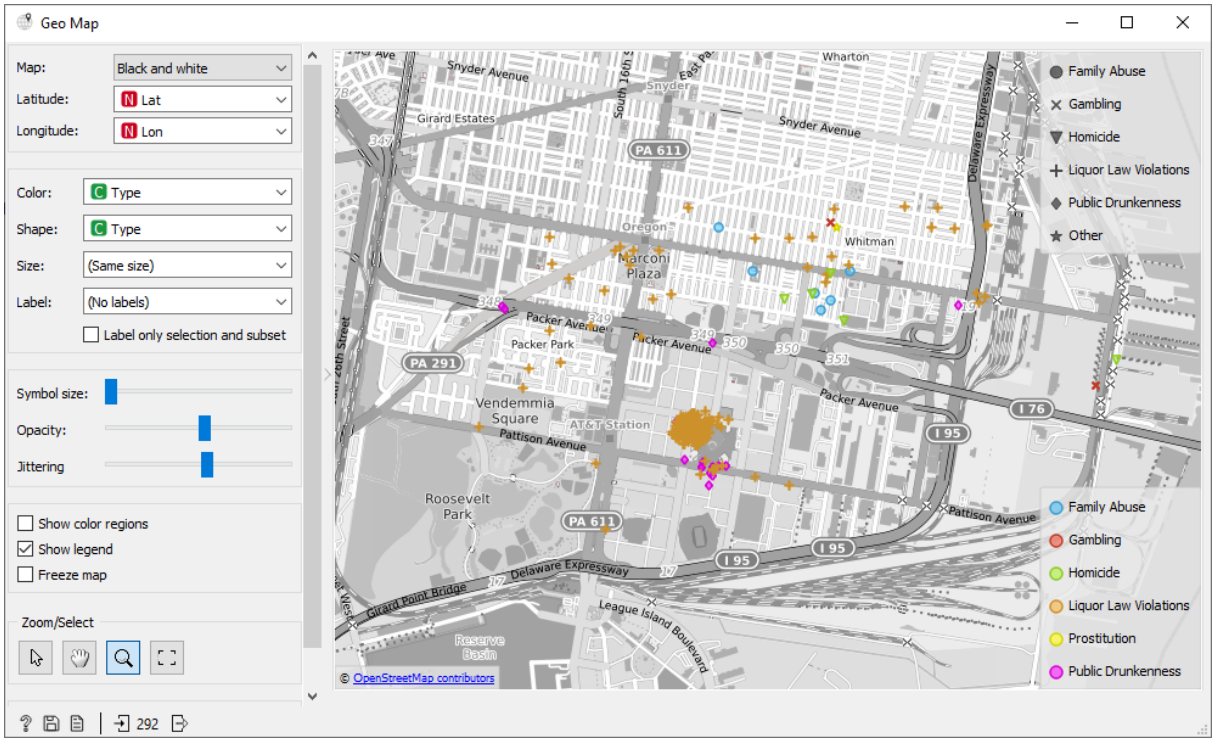
Obr. 169 Rozhodovací strom kriminálních činů

Oblast odpovídající vybranému listu si zobrazíme pomocí následujícího uzlu **Geo Map**, kdy je oblast automaticky přiblížena v mapě (Obr. 170). Pro lepší čitelnost zvolte podkladovou mapu *Black and white*. Také vyšší *Jittering* pomůže prozkoumat jedno místo, kde je velmi vysoký výskyt typu činu *Liquor Law Violations*.

Jednotlivé listy s vysokým zastoupením jednoho konkrétního činu (sytě vybarvené) lze tak postupně použít a lokalizovat a predikovat opakování výskytu jednotlivých typů kriminálních činů. Zkuste postupně vybírat listy a zobrazit odpovídající oblast v mapě pomocí uzlu **Geo Map**.

Záznamy spadající pod vybraný list stromu lze zobrazit také v tabelární formě následujícím listem **Data Table** za uzlem **Tree Viewer.**

V druhé části workflow je proveden výběr ze zdrojových dat pomocí uzlu *Select Rows*, kdy jsou vybrány pouze činy *Liquor Law Violations*. Následující uzel *Geo Map* zobrazuje jen tyto činy na celém území. Lze takto volitelně vybírat různé typy činů a zobrazovat pouze tento výběr v území.



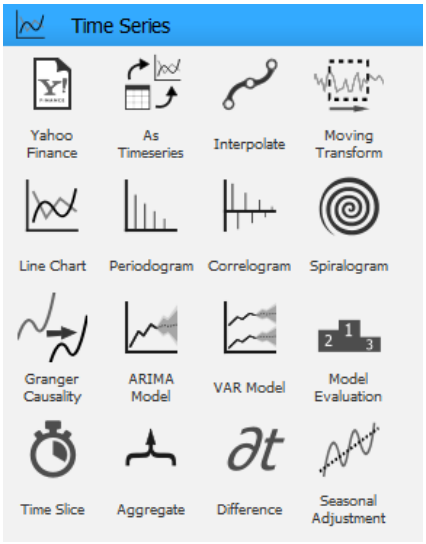
Obr. 170 Mapa části města s body kriminálních činů, kde převládjí činy Liquor Law Violations

> *Poznámka: Zdroj inspirace je na*
> *https://orange3.readthedocs.io/en/3.5.0/widgets/visualize/geomap.html*

> *Vyzkoušejte geokódování a zobrazení dat v mapě pro Evropu, soubor 9_EuropeBaseData.xlsx, workflow 9_GeoEurope.ows.*

10 ČASOVÉ ŘADY

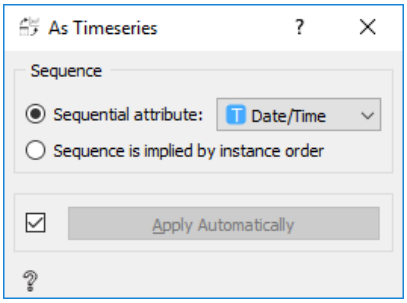
Pro práci s časovými řadami je nutné doinstalovat doplněk Add-on **Time Series**. V levém pruhu se objeví modrá skupina uzlů. Dostupná cvičná data lze otevírat prvním uzlem *Yahoo Finance*, ale ten ve cvičení nebude použit.



Obr. 171 Uzly v doplňku Time Series

Skupina obsahuje uzly pro základní úpravu a analýzu časových řad. Uzel **As Timeseries** převádí vstupní data na časovou řadu. V tomto uzlu vybíráme atribut s údajem času nebo určíme, že sekvence je dána pouze pořadím dat (Obr. 172). Uzel *As Timeseries* je vhodné zařadit vždy za uzel *File* před dalším zpracováním.

Uzel **Interpolate** dopočítává chybějící hodnoty. Uzel **Aggregate** poskytuje možnost spočítat pro řadu různé týdenní, měsíční, či roční agregace.



Obr. 172 Uzel As Timeseries

Analýza časových řad řeší dvě základní úlohy:

- **rozklad** časové řady na složky, kdy se provádí popis historických dat,
- časovou **predikci** vývoje hodnot do budoucnosti.

10.1 Korelace dvou časových řad

Časové řady mohou vykazovat navzájem časovou korelaci. Tu jednoduše zjistíme pomocí uzlu **Correlations** ze základní skupiny uzlů *Data*.

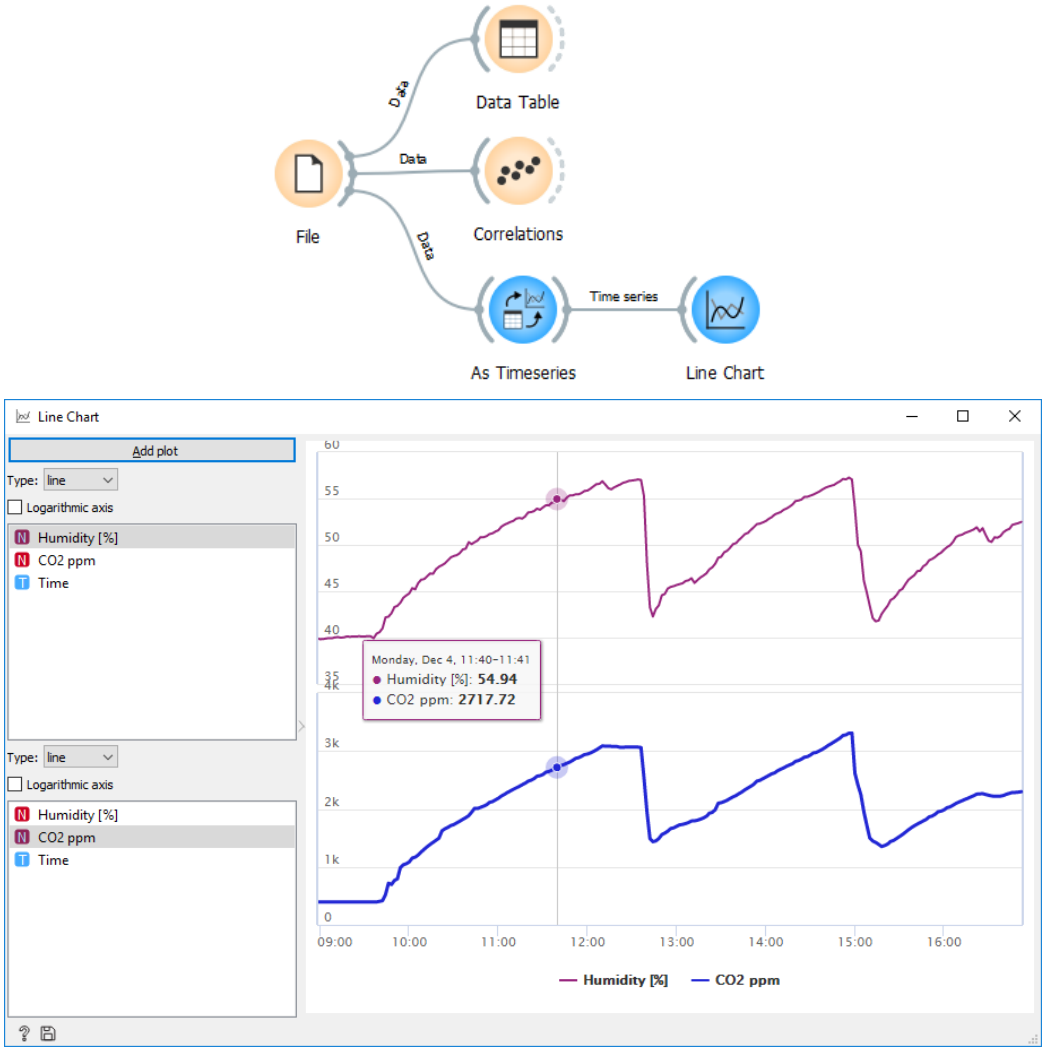
Příklad 46

Data *10_Ucebna.xlsx,* **workflow** *10_Ucebna.ows*

Data obsahují naměřenou koncentraci CO_2 a vlhkost (Humidity) v průběhu jednoho dne na učebně. Průběh hodnot je vykreslen pomocí uzlu **Line Chart**. V uzlu *Line Chart* jsou zobrazeny dva grafy z důvodu různého rozsahu a měřítka hodnot na ose Y pro každou zobrazovanou veličinu. Další grafy se přidávají tlačítkem **Add plot**. V levém okně se vybírá veličina k zobrazení.

Z průběhu hodnot je patrné, že hodnoty obou veličin stoupaly mezi 9:45 až 12 hodinou, kdy probíhala výuka. Poté následuje prudký pokles, kdy se místnost vyvětrala a přišla na výuku jiná skupina studentů. Poté následuje opět nárůst hodnot přibližně od 13 do 15 hodin a následný rychlý pokles. V závěru je viditelný opětný mírný nárůst hodnot.

Korelace je spočítána pomocí uzlu **Correlations** a vykazuje hodnotu 0,971.



Obr. 173 Workflow a časový průběh hodnot koncentrace CO_2 a vlhkosti

10.2 Zjištění stacionarity časové řady

Před rozkladem časové řady na složky a případnou predikcí je nutné zjistit stacionaritu časové řady. Data stacionární řady nezávisí na čase, naopak hodnoty **nestacionární** řady se s časem mění. Časové řady jsou **stacionární**, pokud nemají trendovou ani sezónní složku. Souhrnné statistiky jako je průměr nebo rozptyl, které jsou vypočtené z časové řady, jsou v průběhu času konzistentní (Brownlee 2018). Stacionaritu lze nejprve zjišťovat pomocí přímého vykreslení hodnot, potom je možné spočítat souhrnné statistiky a poslední možností je použití statistického testu Augmented Dickey Fuller (také nazývaný Unit root test).

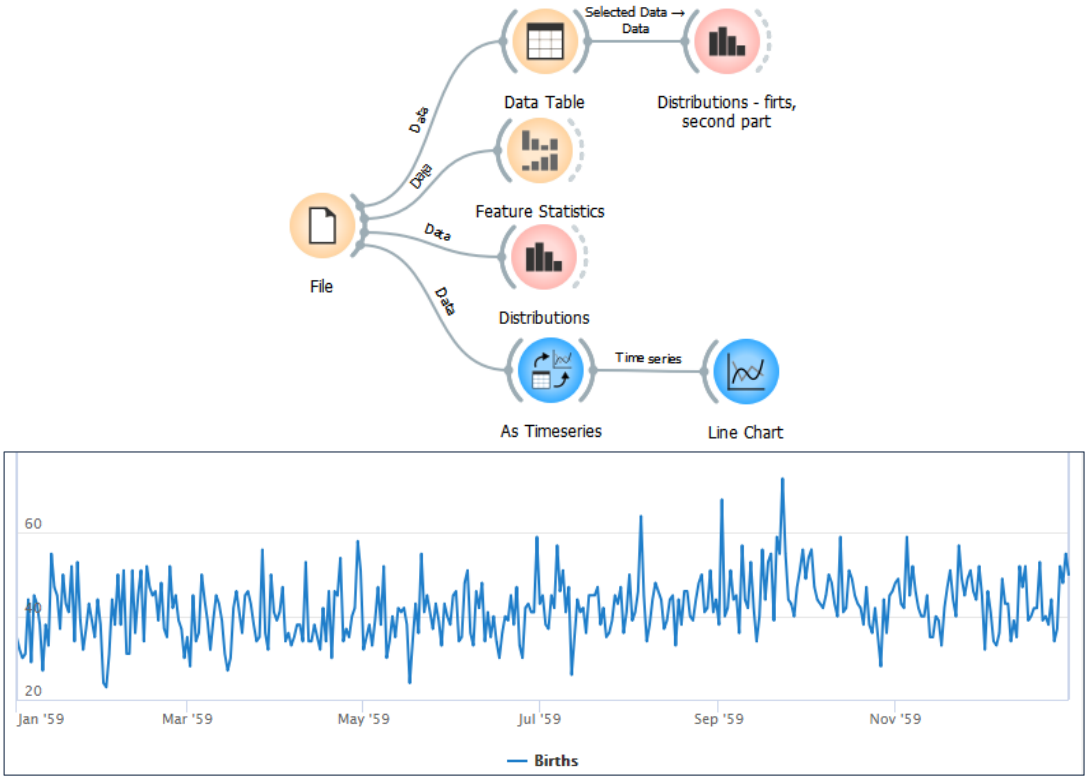
Více o testu na https://machinelearningmastery.com/time-series-data-stationary-python/ (Brownlee 2018)

Příklad 47

Data *10_FemaleBirth.xlsx,* **workflow** *10_FemaleBirth.ows*

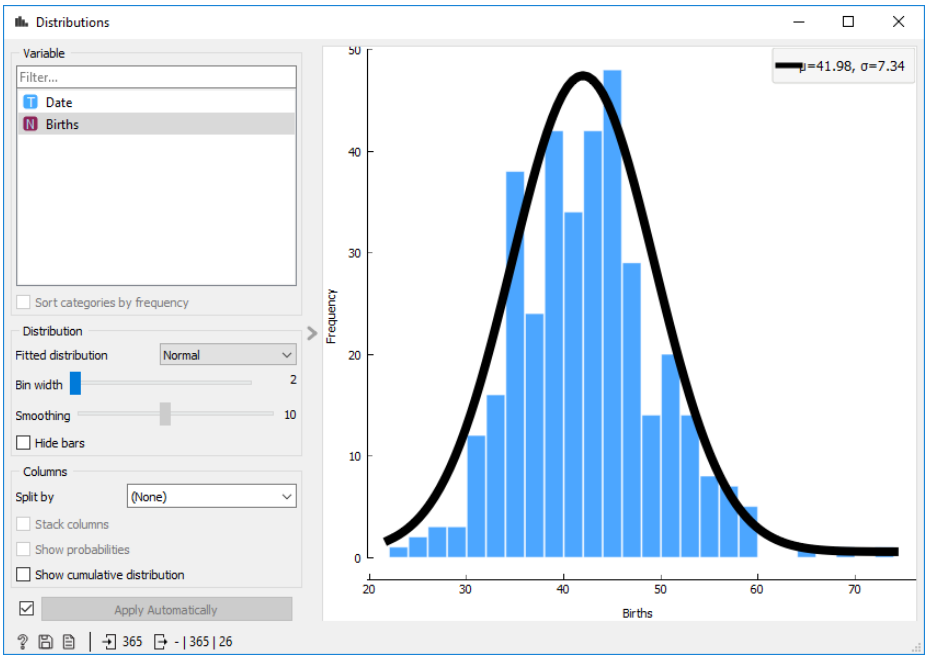
Data *10_FemaleBirth* představují ukázku stacionární řady. Data obsahují denní úhrny počtu narozených žen od ledna do prosince v roce 1959.

Nejprve vykreslíme graf průběhu hodnot. Je patrné, že počet narozených žen v průběhu roku kolísá, ale nevykazuje trend nebo sezónní složku. Ve workflow je použit modrý uzel **Line Chart** (ze skupiny Time Series) pro základní vykreslení průběhu hodnot. Pozor nezaměnit tento uzel s uzlem *Line Plot* ze skupiny Visualize.



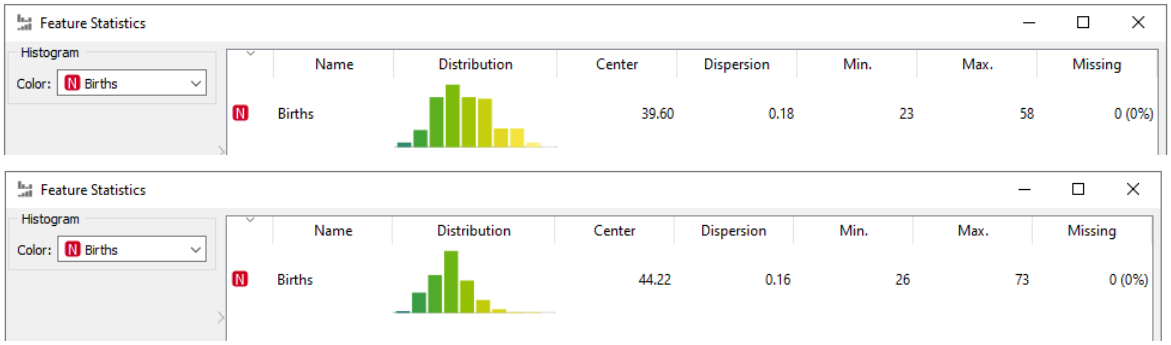
Obr. 174 Workflow a časová řada počtu narozených žen v průběhu jednoho roku zobrazená uzlem Line Chart

Uzel **Distribution** vykreslí četnosti jednotlivých hodnot. Tento uzel dobře poslouží pro rychlé a hrubé zjištění stacionarity. Histogram hodnot má zvonovitý tvar normálního Gausova rozdělení s delším pravým ocasem. Pro porovnání histogramu s teoretickým rozdělením zvolte v dialogu uzlu *Distribution* volbu *Fitted distribution–Normal*. Pro porovnání se vykreslí černá souvislá čára teoretického rozdělení. Gausovo rozdělení skutečných hodnot naznačuje, že se jedná v případě počtu narozených žen o stacionární řadu.



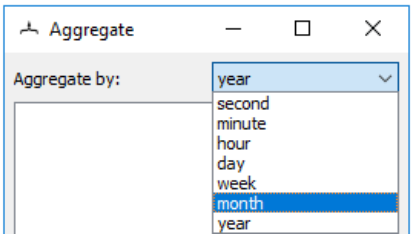
Obr. 175 Distribuce hodnot počtu narozených žen ve dnech

Dále lze stacionaritu zjistit porovnáním průměru a rozptylu hodnot první části (leden až červen) a druhé části souboru (červenec až prosinec) pomocí uzlu *Feature Statistics* (viz kapitola 3). Vybereme v datové tabulce vždy příslušný rozsah měsíců. První část souboru má průměrný počet narozených žen 39,6 a koeficient variace (Dispersion) 0,18. Druhá část souboru má průměrnou hodnotu 44,22 a dispersion 0,16 (Obr. 176). Průměrné hodnoty a dispersion si zhruba odpovídají. Také distribuce hodnot dílčích částí souboru má Gausovo rozdělení.



Obr. 176 Distribuce hodnot počtu narozených v první (nahoře) a druhé půlce roku (dole)

> *Samostatně použijte uzel **Aggregate** pro výpočet měsíčních agregací počtu narozených žen. Tyto agregace si zobrazte v tabulce a v grafu Line Chart. Pozor nelze jednotlivé měsíce porovnávat mezi sebou, neboť měsíce v roce mají různý počet dní. Pro porovnání je nutné očistit data a přepočítat je na stejný počet dní v měsíci.*

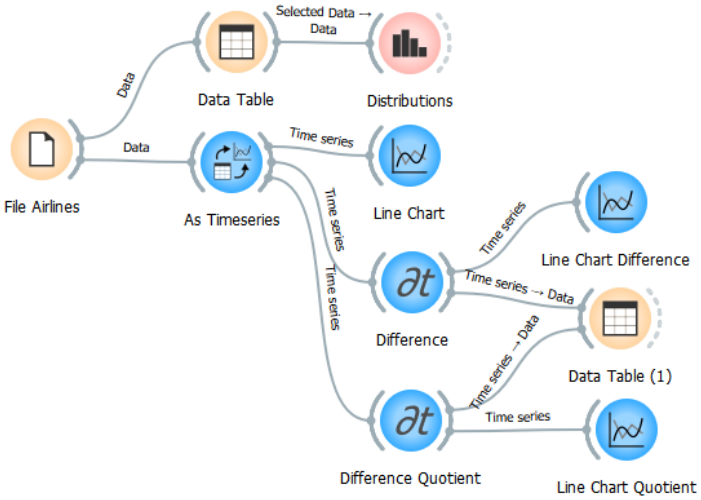


Obr. 177 Možnosti uzlu Aggregate

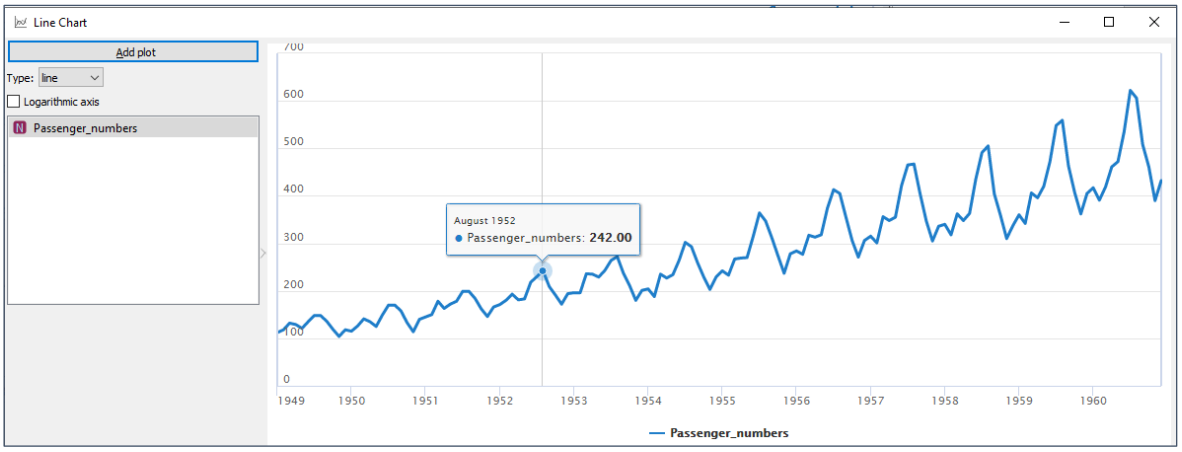
Příklad 48

Data *10_ Airlines.xlsx,* **workflow** *10_ Airlines.ows*

Ukázku **nestacionární časové řady** obsahuje soubor 10_Airlines.xlsx. Tato data obsahují počty pasažérů přepravených letadly v USA. Soubor obsahuje 144 záznamů počtu pasažérů od roku 1949 do roku 1961 v agregacích za jednotlivé měsíce. Již ze základního vykreslení dat je evidentní, že data obsahují rostoucí trend hodnot s výraznými výkyvy do vyšších hodnot v červenci a srpnu každého roku (období cestování na letní dovolené). Další menší kladné výkyvy hodnot v průběhu roku jsou opakovaně v prosinci každého roku (cestování na Vánoce a Nový rok) a v březnu (Velikonoce) následovaný částečným poklesem.

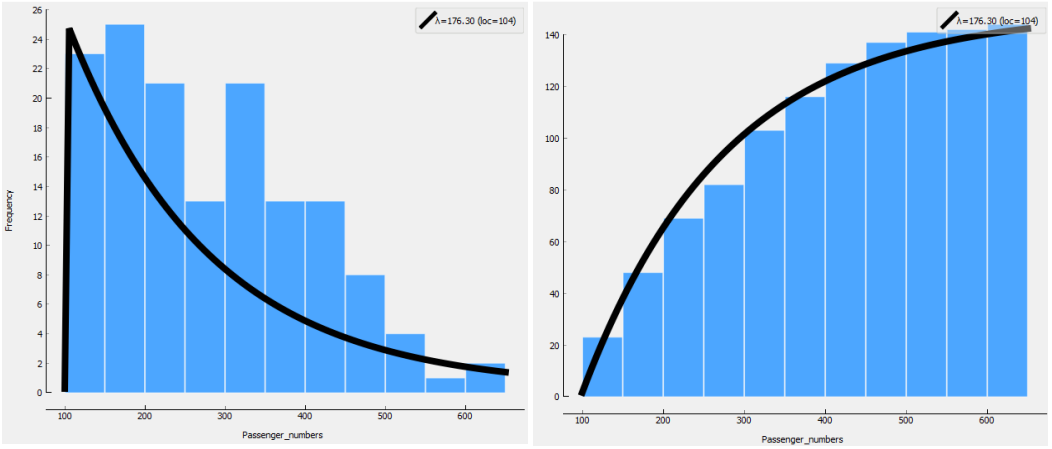


Obr. 178 Workflow pro časovou řadu počtu pasažérů



Obr. 179 Nestacionární časová řada počtu přepravených pasažérů letadly v období 1949 až 1961 v USA

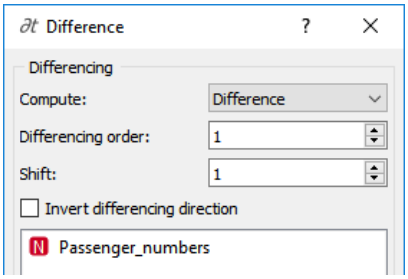
Histogram četnosti hodnot nemá tvar normálního rozdělení jako v předchozím případě. Naopak naznačuje exponenciální růst hodnot v čase. Pro porovnání zvolte v uzlu *Distribution* zobrazení *Fitted distribution–Exponential*. Po zaškrtnutí volby *Show cumulative distribution* se zobrazí kumulativní distribuce, kterou opět můžeme porovnat s křivkou exponenciálního rozdělení (Obr. 180 b).



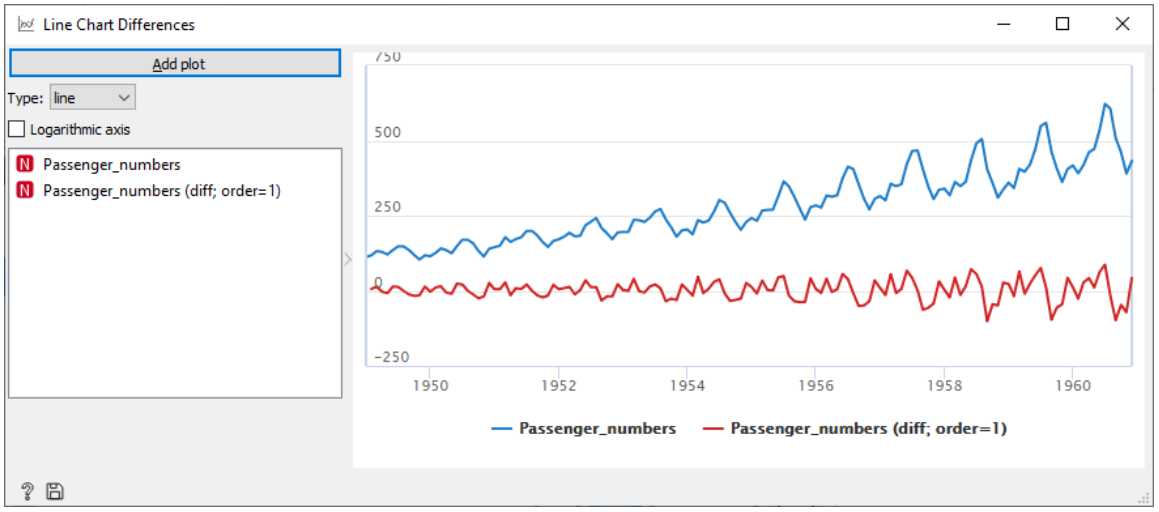
Obr. 180 Histogram hodnot počtu pasažérů za celé časové období (a), kumulativní distribuce (b)

Pomocí uzlu **Difference** lze spočítat rozdíly sousedních hodnot neboli absolutní přírůstek (1. diference), tzv. delta. Velké změny v diferencích prvního řádu poukazují na odlehlé hodnoty v původních datech. Lze nastavit i výpočet druhé diference. Volbou *Shift* se nastavuje velikost diference, tj. vzdálenost odečítaných hodnot, pro *Shift* 2 se počítají rozdíly hodnot v čase t a čase t-2. Diference lze spočítat v absolutních hodnotách nebo v procentech *(*volba *Percentage change).*

Nahrazení původní řady řadou prvních diferencí lze převést původní řadu na řadu bez trendu. Odečtením sezónních diferencí lze řadu zbavit sezónních vlivů. Zde má sezónní diference délku (*Shift*) 12. Hodnoty diferencí jsou doplněny do tabulky dat. První hodnota diference není spočítaná, neboť neexistuje předchozí hodnota Y_{t-1} pro první údaj. Pomocí uzlu **Line Chart** lze vykreslit společně původní hodnoty řady a řadu prvních diferencí. U prvních diferencí stoupá rozkmit, což naznačuje nestacionární řadu. **Průměrná hodnota prvních diferencí** je 0,25, což je průměrný nárůst časové řady za jeden měsíc (lze zjistit uzlem *Feature Statistics*).

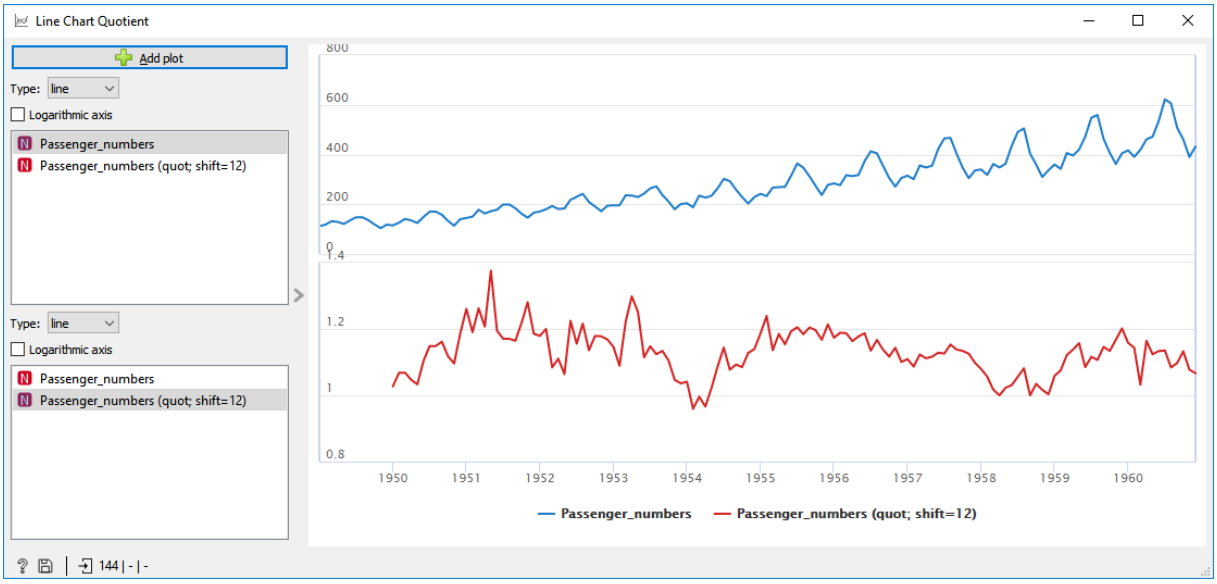


Obr. 181 Uzel Difference



Obr. 182 Graf původní řady a prvních diferencí

Uzel **Difference** umožňuje spočítání i **tempa růstu *k***, což je poměr dvou hodnot v různém čase. Tempo růstu se spočítá volbou *Compute: Quotient.* V případě těchto dat je zajímavé počítat meziroční tempo růstu a porovnávat stejné měsíce dvou po sobě jdoucích roků. Z toho důvodu byla nastavena volba *Shift* na hodnotu 12. Při intepretaci křivky tempa růstu můžeme konstatovat, že tempo růstu nabývá většinou hodnot vyšších než 1, a to mezi 1,1 až 1,4 tzn., že dochází k růstu. Pokles tempa půstu (méně než 1) byl pouze na počátku roku 1954, kdy poklesl počet přepravených osob oproti měsícům předchozího roku. V některých měsících roku 1958 a 1959 koeficient byl roven 1. To znamená, že hodnoty počtu cestujících nerostly, ale byly stejné jako v předchozím roce.



Obr. 183 Původní řada a řada meziročního tempa růstu po měsících

*Samostatně použijte uzel **Aggregate** pro výpočet ročních agregací počtu pasažérů. Tyto agregace si zobrazte v tabulce a v grafu Line Chart.*

Napočítaná data prvních diferencí nebo tempa růstu lze zobrazit pomocí uzlu Data Table a následně také uložit pomocí uzlu Save Data. Stejně tak napočítané agregace.

Uložením se získají výstupní data pro případné další zpracování mimo Orange.

10.3 Rozklad časové řady

Nestacionární řadu lze rozložit aditivním nebo multiplikativním rozkladem (Hančlová a Tvrdý 2003) (Křivý 2012).

Multiplikativní časová řada má se rozkládá podle vzorce

$$y_t = T_t * S_t * C_t * R_t \quad (8)$$

Aditivní dekompozice se rozkládá podle následujícího vzorce

$$y_t = T_t + S_t + C_t + R_t \quad (9)$$

kde y_t jsou hodnoty časové řady, T_t je trend, S_t je sezónní složka, C_t je cyklická složka a R_t je náhodná (reziduální) složka v čase t.

Trend je obecná tendence vývoje zkoumaného jevu za dlouhé období. Je výsledkem dlouhodobých a stálých procesů. Trend může být rostoucí, klesající nebo může existovat řada bez trendu.

Sezónní složka je pravidelně se opakující odchylka od trendové složky. Perioda této složky je menší než celková velikost sledovaného období.

Cyklická složka udává kolísání okolo trendu v důsledku dlouhodobého cyklického vývoje (požíváno např. v makroekonomických úvahách).

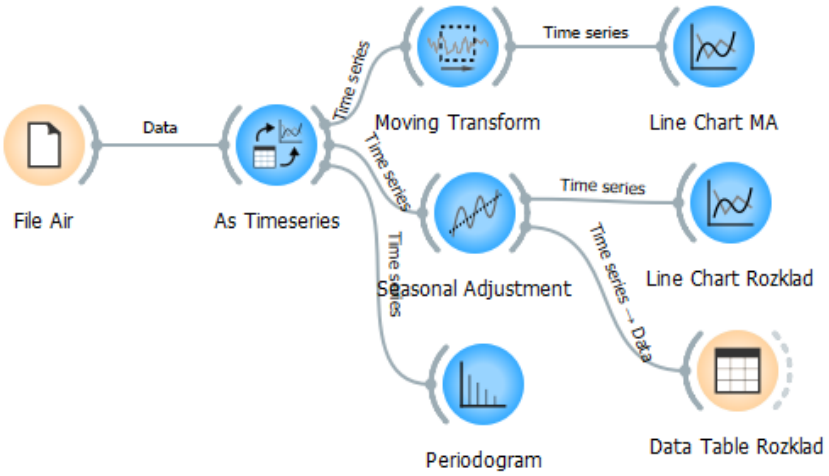
Sezónní a cyklické složce se dohromady říká **periodická složka** (Dvořáková 2015).

Náhodná složka se nedá popsat žádnou funkcí času. „Zbývá“ po vyloučení trendu, sezónní a cyklické složky.

Příklad 49

Data *10_Airlines.xlsx,* **workflow** *10_Airline_MA.ows*

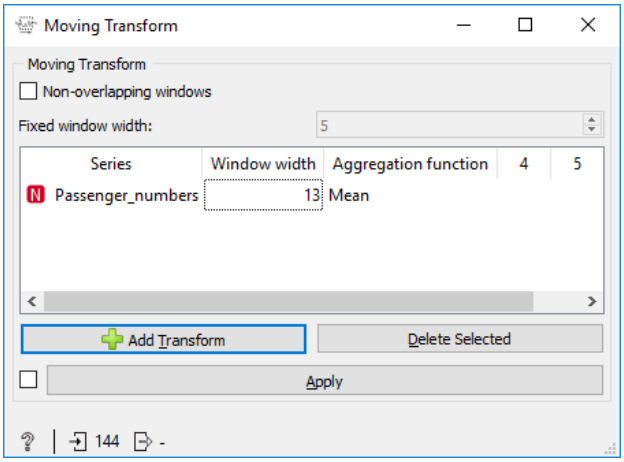
Jedná se o nestacionární časovou řadu vhodnou pro **multiplikativní rozklad.** Multiplikativní dekompozice se používá v případě, že variabilita časové řady roste v čase, nebo se v čase mění (Arlt et al. 2002). V případě této řady roste trend a roste i variabilita hodnot.



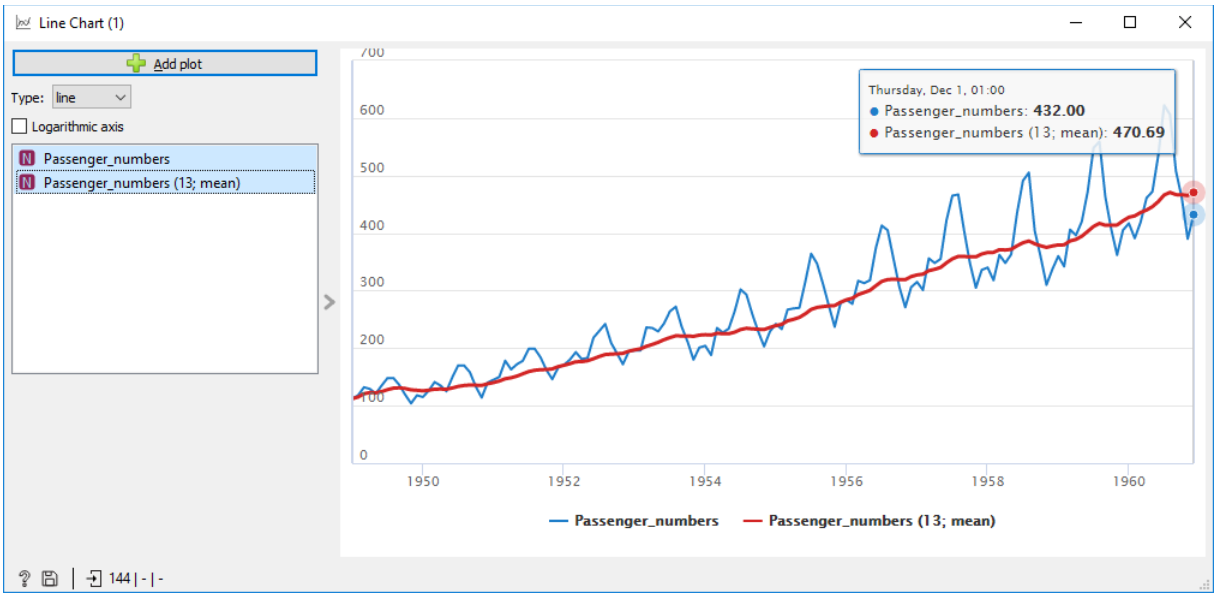
Obr. 184 Workflow pro zracování časové řady

Nejprve zjistíme trend pomocí **metody klouzavých průměrů** (Litschmannová 2010). Uzel **Moving Transform** spočítá postupně průměrnou hodnotu z okolních hodnot. Ve sloupci **Window width** lze nastavit šířku zhlazovacího okna. Volíme hodnotu **13**, protože se jedná o měsíční agregace z několika roků a je evidentní, že jsou v měsíční časové řadě sezonní výkyvy. V létě je patrný nárůst přepravy cestujících na dovolenou a po podzimním poklesu následuje nárůst koncem roku v období Vánoc. V nabídce agregační funkce jsou kromě průměru k dispozici i další funkce jako medián, maximum atd. Náhled vypočtených shlazených dat lze zobrazit v uzlu **Data Table** – obsahuje nový sloupec s vypočítanou shlazenou hodnotou. Volitelně lze přidat i transformaci s délkou okna 25. Okno s lichým počtem hodnot je symetrické na obě strany, kdy prostřední hodnota je právě hodnota, pro kterou se počítá průměr.

Následně lze z tabulky zobrazit data pomocí uzlu **Line Chart**, kde vybereme původní hodnoty a spočítané hodnoty klouzavých průměrů pro zobrazení v grafu. Lze měnit typ grafu na liniový (line), sloupcový, atd. Taktéž je možnost změnit svislou osu na logaritmickou. Metoda klouzavých průměrů se řadí mezi adaptivní techniky rozkladu a používá se tam, kdy se parametry trendu mění v čase (není např. stále lineární rostoucí trend).

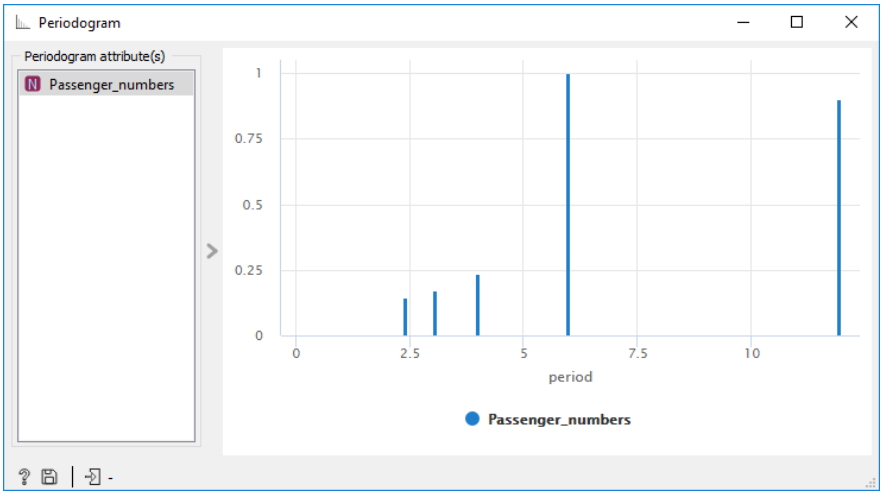


Obr. 185 Nastavení okna pro transformaci pomocí klouzavých průměrů



Obr. 186 Vykreslení zdrojové časové řady a klouzavých průměrů

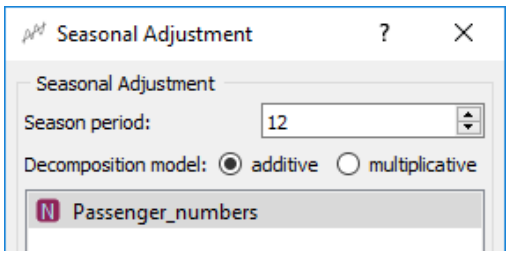
Periodogram je užitečným vizuálním prostředkem analýzy sezónních časových řad. Používá se pro vyhledávání významných periodických složek v časových řadách (Arltová a Arlt 1995). Délku sezónní periody lze vyšetřit pomocí uzlu **Periodogram**. Tento diagram zobrazuje dvě významné periody a to přibližně v délce 6 a 12 měsíců. Tento uzel je dobré použít pro zjištění periody před provedením rozkladu pomocí uzlu *Seasonal Adjustment*.



Obr. 187 Zjištění významných period sezónní složky pomocí periodogramu

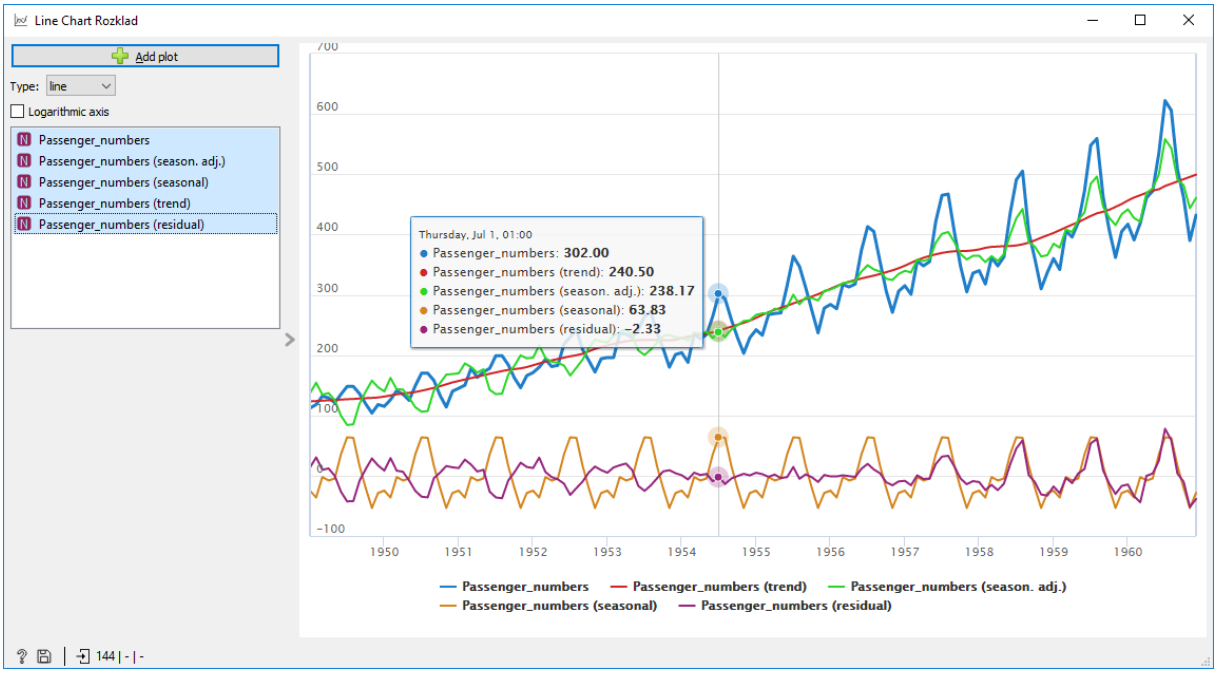
Rozklad časové řady se provádí pomocí uzlu **Seasonal Adjustment**. Provede se očištění časové řady od sezónní složky. Z vizuálního vyšetření řady a z periodogramu je evidentní, že **sezónní perioda je 12** a tak je hodnota zadána v dialogu uzlu *Seasonal Adjustment.*

V nastavení uzlu se volí, zda je časová řada **aditivní** nebo **multiplikativní**. Následně lze rozklad časové řady znázornit v uzlu **Line Chart**, kde v levém okně vybereme všech pět hodnot – původní data, data očištěná od sezónní složky (season. adj.), sezónní složku, trend a reziduální složku.



Obr. 188 Nastavení uzlu Seasonal Adjustment

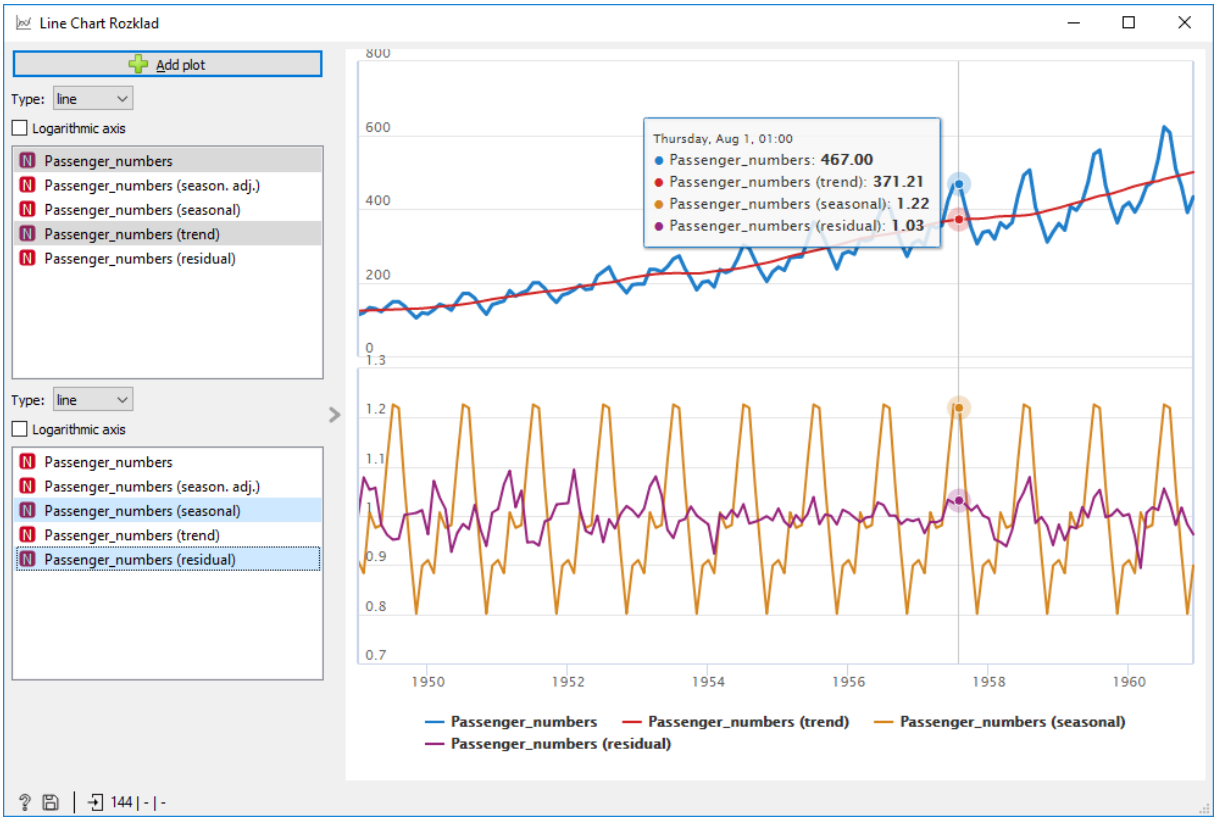
Můžeme experimentovat a zvolit nejprve aditivní rozklad *Decomposition model: additive*. Výsledek aditivního rozkladu obsahuje vysokou reziduální složku (hodnoty až 77) rostoucí s časem, což není správně a je evidentní, že aditivní rozklad je nevhodný (Obr. 189). V porovnání se sezónní složkou jsou sezónní maxima stejná jako náhodná složka, což není správné.



Obr. 189 Aditivní rozklad řady nestacionární řady

Pro data počtu pasažérů je vhodnější multiplikativní rozklad. V uzlu **Seasonal Adjustment** vyberte volbu *Decomposition model: multiplicative*.

Složky lze vykreslit do samostatného grafu pomocí tlačítka **Add plot** a volby požadované hodnoty pro zobrazení. V případě multiplikativního rozkladu je v absolutních hodnotách pouze trend. Ostatní složky násobí trend a tak jsou vyjádřeny v relativních hodnotách. Sezónní a reziduální složku, je z tohoto důvodu lepší vykreslit v samostatném grafu. Z grafu, který je na Obr. 190 je vidět, že reziduální složka dosahuje nižších hodnot, a to kolem hodnoty 1. Sezónní složka se pohybuje v intervalu 0,8 až 1,2.



Obr. 190 Multiplikativní rozklad řady nestacionární řady

Nevýhodou zobrazení grafu v Orange je, že barvy linií jsou přiřazeny automaticky programem a nelze je uživatelsky měnit. Pro různá zpracování může tak např. trend mít různou barvu.

Výsledné hodnoty rozkladu lze uložit pomocí uzlu **Save Data** do XLSX souboru a použít je k dalšímu zpracování či prezentaci mimo Orange. V MS Excel lze podle potřeby upravit barvy, rozsahy os, popisy os, legendu apod.

10.4 Autokorelace a predikce hodnot časové řady

Hodnoty jedné časové řady mohou vykazovat vzájemný vztah, hovoří se potom o **autokorelaci** hodnot časové řady. **Autokorelační koeficient** je relativní míra proměnlivosti členů časové řady posunutých v čase o hodnotu *k*. Posun *k* se z angličtiny označuje jako *lag*. **Autokorelační funkce** (ACF) je potom závislost mezi hodnotami autokorelačního koeficientu a hodnotami posunu *k*. Graf ACF se nazývá **korelogram**. Hodnoty autokorelační funkce se pohybují v intervalu <-1, 1>. ACF je vhodným nástrojem k posouzení, zda časová řada obsahuje cyklickou nebo periodickou složku a také zda je či není řadou náhodných čísel – tedy do jaké míry je možné ji extrapolovat (predikovat). Korelogram se používá pro posouzení, zda řada reziduí má charakter tzv. bílého šumu. Jednotlivé sloupce autokorelační funkce vyjadřují sílu lineární závislosti mezi hodnotami časové řady (Arltová a Arlt 1995). Statistická významnost korelogramu se doplňuje intervalem spolehlivosti 95 %. Ten lze s dostatečnou přesností určit ze vztahu

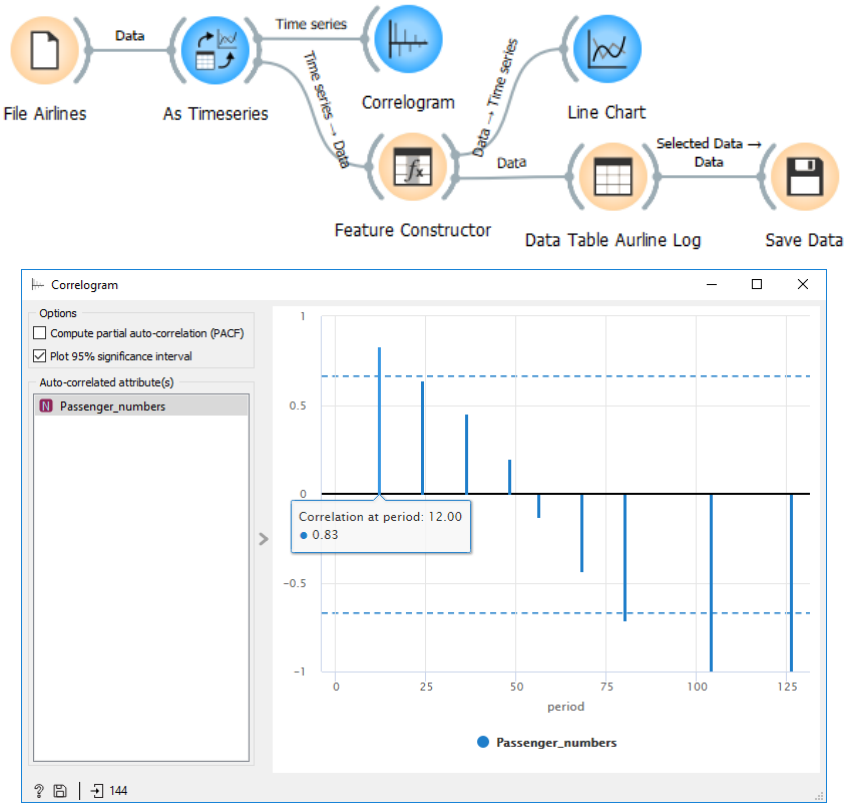
$$\frac{\pm 2}{\sqrt{N}} \qquad \text{kde N je délka časové řady} \qquad (10)$$

Příklad 50

Data *10_Airlines.xlsx,* **workflow** *10_Airline_Log.ows*

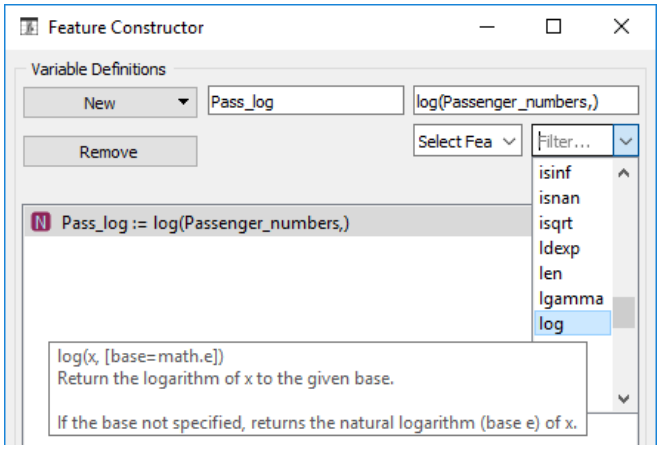
Z korelogramu v uzlu **Correlogram** zjistíme, že nejvyšší korelace je pro časový posun 12 měsíců. První modrá čára je nad vodorovnou čárkovanou čárou naznačující 95% hranici intervalu významnosti (signifikace). Další následující kladné autokorelační koeficienty jsou pod touto hranicí.

Zatržení první volby *Compute partial auto-correlation (PACF)* se spočítá parciální autokorelační funkce.



Obr. 191 Workflow a korelogram – graf autokorelační funkce

Vzhledem k tomu, že časová řada počtu přepravených pasažérů je nestacionární řadou, je nutné nejprve řadu převést na stacionární řadu. Původní data počtu pasažérů přepočítáme pomocí logaritmické funkce. Stacionarizovat lze i pomocí dalších funkcí jako je druhá a třetí mocnina nebo druhá a třetí odmocnina, časový posun-diference nebo exponenciální rozklad (Sunaysawant 2021). Využijeme uzel **Feature Constructor** ze základní sady *Data* a přepočítané hodnoty se uloží do nového atributu *Pass_log*. Upravená data uložíme pomocí uzlu *Save Data* do souboru *10_AirlinesLog.xlsx*. Data použijme jako vstupní data do nového workflow.



Obr. 192 Logaritmování vstupních hodnot

Příklad 51

Data *10_AirlinesLog.xlsx,* **workflow** *10_AirlinePrediction.ows*

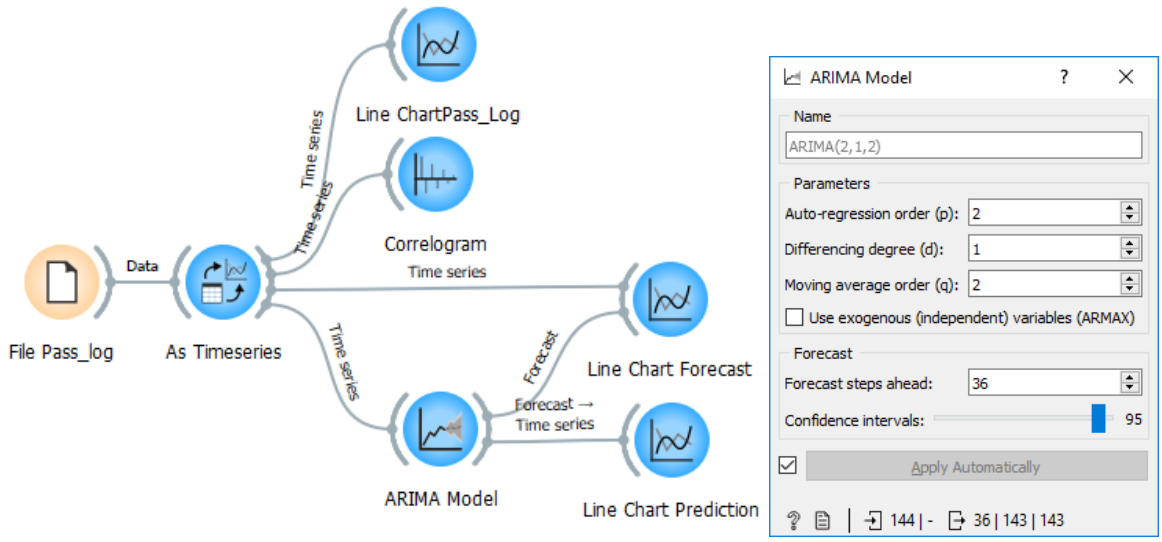
ARIMA model je klasický model časových řad, který modeluje auto-regresivní vlastnosti časových dat. Auto-regresivní znamená něco, co závisí na minulých hodnotách sebe sama. Časová řada může záviset:

- na své bezprostředně minulé hodnotě nebo
- může vykazovat sezónní chování a opakovat se po uplynutí určitého počtu časových bodů.

Boxova-Jenkinsova metodologie bere v úvahu při konstrukci modelu časové řady reziduální složku, která může být tvořena korelovanými (závislými) náhodnými veličinami. Boxova-Jenkinsova metodologie tedy nejen může zpracovávat časové řady s navzájem závislými pozorováními, ale dokonce těžiště jejich postupů spočívá právě ve vyšetřování těchto závislostí neboli tzv. korelační analýze. Kombinují se autoregresivní modely **AR(p)** s modely klouzavých průměrů reziduální složky **MA(q)**. V případě nestacionární časové řady se provádí stacionarizace např. diferencováním (I integration) a zjišťuje se řád s parametrem *d*. Výsledný model se potom označuje jako **ARIMA (p,d,q).** V případě sezónních vlivů se označují jako SARIMA modely (Hančlová a Tvrdý 2003).

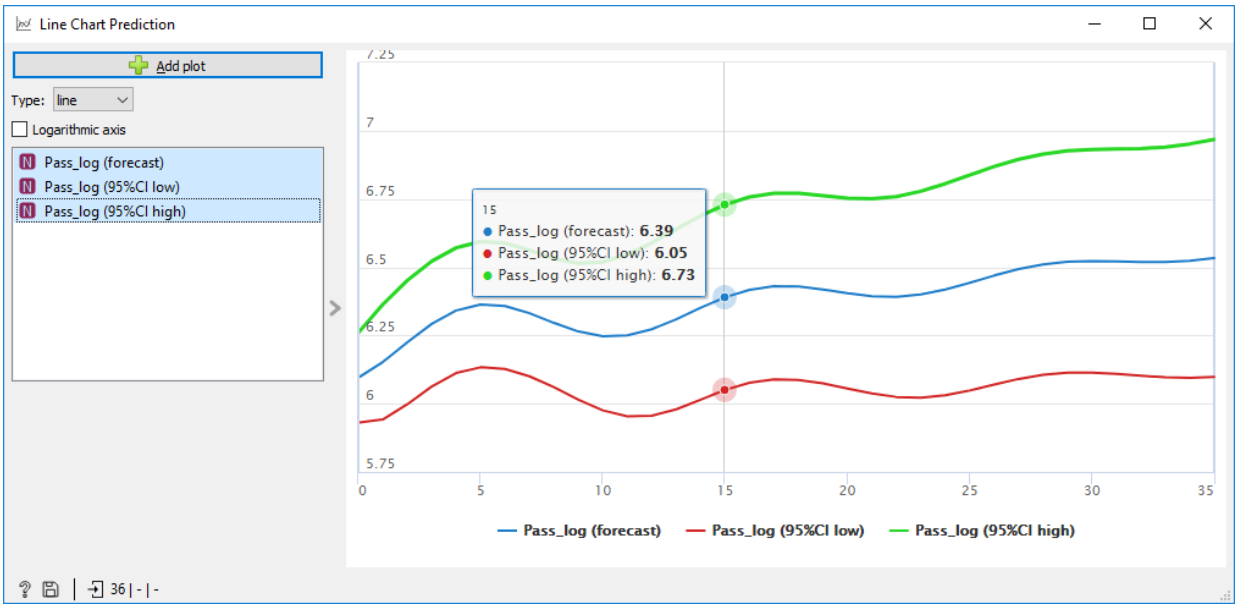
Ve workflow *10_AirlinePrediction.ows* je nutné z důvodu predikce nastavit vstupní atribut *Pass_log* do role *target*. Lze vykreslit logaritmovanou časovou řadu. V uzlu **ARIMA** nastavíme hodnoty koeficientů p, d, g na hodnoty 2, 1, 2. Hodnota d = 1 určuje první diferenci, která odstraňuje trend. Postup řešení predikce této časové řady v jazyce Python, testování stacionarizace a vysvětlení koeficientů je možné nalézt v článku „Air Passengers-Time Series-ARIMA“ (Sunaysawant 2021).

Počet predikovaných kroků může nastavit libovolně na 12 či více měsíců. Hodnota 36 měsíců predikuje na 3 roky dopředu.



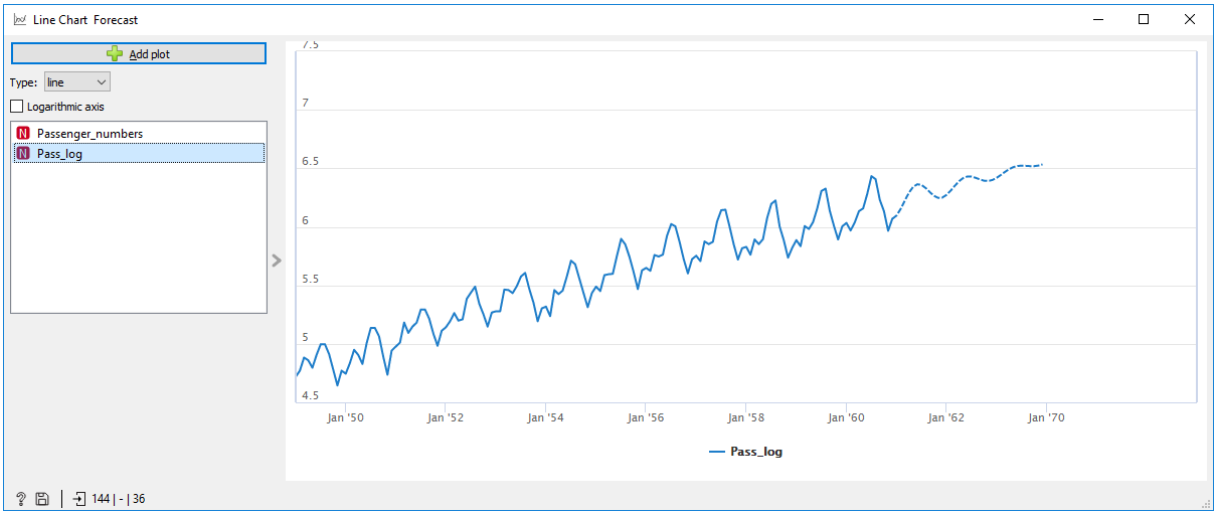
Obr. 193 Workflow s ARIMA modelem

Výsledek samotné predikce včetně 95% intervalu spolehlivosti lze zobrazit uzlem *Line Chart* (Obr. 194). Zobrazené hodnoty jsou v logaritmických jednotkách.



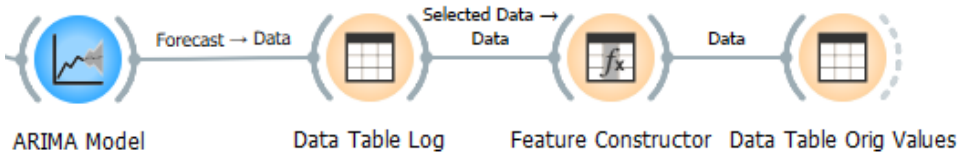
Obr. 194 Predikce na následující tři roky v logaritmických hodnotách

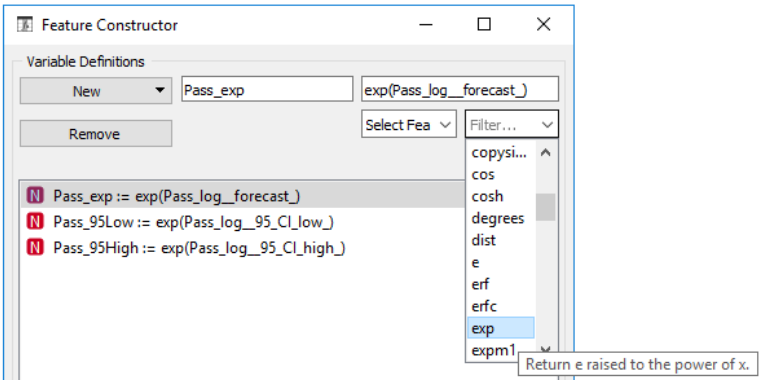
Lze zobrazit zároveň původní časovou řadu i predikované hodnoty. V tomto případě je nutné správně nastavit spojné čáry v dialogu *Edit Links*. Původní data jsou propojena jako *Time Series-Time Series* a výstup z uzlu ARIMA modelu je spojen jako *Forecast-Forecast.* V grafu jsou predikované hodnoty znázorněny čárkovanou čarou (Obr. 195). Hodnoty počtu přepravených pasažérů mají stále stoupající trend se sezónní složkou.



Obr. 195 Zobrazení původní časové řady a predikce ARIMA modelem v logaretmickýcj jednotkách

Hodnoty predikce do původních jednotek lze získat zpětným přepočtem exponenciální funkcí. Uzel *Feature Constructor* umožňuje zadat rovnice pro přepočet do nových tří atributů. Zpětně transformované hodnoty lze znázornit v grafu *Line chart*.





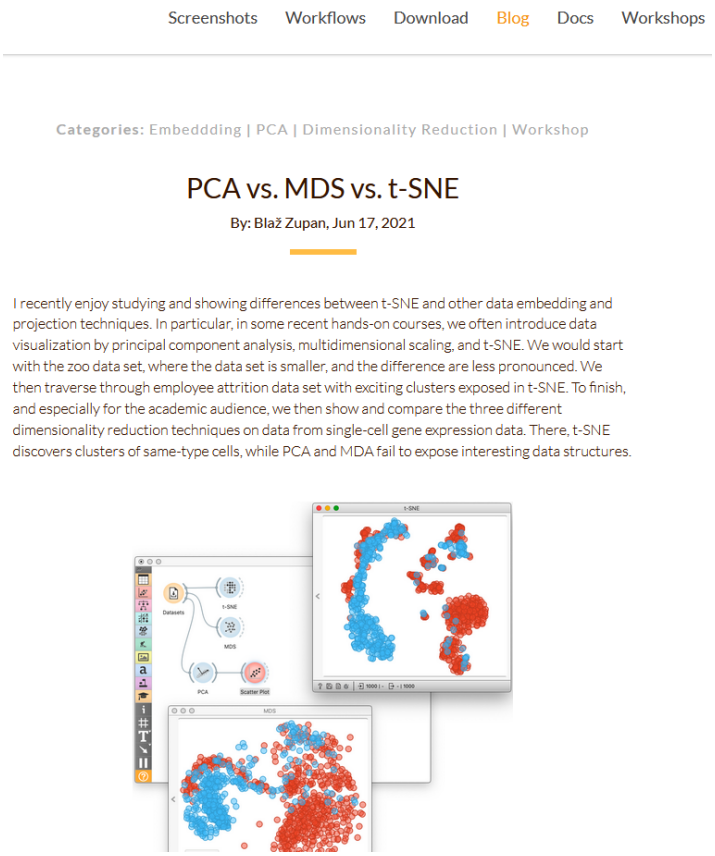
Obr. 196 Přepočet predikovaných hodnot do původních jednotek exponenciální funkcí

Samostatně vyšetřete časovou řadu 10_Rail_quartal_EUROSTAT.xlsx (list PassengerSelection), která obsahuje počty přepravených osob na železnici v Evropě z databáze EUROSTAT (Eurostat 2022) .
V časové řadě je viditelný pokles počtu pasažérů v roce 2020 a 2021 v době pandemie nemoci covid-19. Lze použít uzel ***Time Slice*** *a vybrat pouze data do konce roku 2019 a z nich udělat rozklad na složky.*

11 ZÁVĚR

Vážení studenti a čtenáři, dostali jste se úspěšně nakonec toho učebního textu. Oblast Data Maning a také software Orange skýtá mnoho dalších zajímavých oblastí využití. Na závěr uvedu několik inspirací do dalšího studia.

Doporučuji sledovat oficiální **Blog** software Orange na adrese https://orangedatamining.com/blog/. Zde se stále objevují nové zajímavé články, které mohou inspirovat k dalšímu zkoumání. Stejně tak historie Blogu obsahuje zajímavé čtení.



Obr. 197 Blog software Orange

Jednou z oblastí Data Mining je získávání znalostí z textových dokumentů. Software Orange disponuje doplňkem **Text Mining**, který obsahuje několik widgetů pro zpracování textu. Z těch základních je to word cloud, hledání klíčových slov v dokumentu, analýza sentimentu a třeba i zobrazení geografických jmen na mapě.

Na tento učební text navazuje v magisterském programu Geoinformatika a kartografie předmět *KGI/POGEO* **Pokročilé zpracování geodat**. Předmět se zaměřuje především na témata související s prostorovou statistikou – nejprve provádí studenty metodami pokročilé průzkumové analýzy, dále jsou představeny prostorově vážené metody, na které navazuje využití prostorových regresních modelů. Druhou část sylabu tvoří využití metod geocomputation, ve které se studenti seznámí s tématy fuzzy logiky, teorie informace a fraktální geometrie a jejich využití v prostoru. Pro cvičení jsou využívány regionální statistiky NUTS2 z databáze Eurostat, databáze OECD a další.

12 POUŽITÉ ZDROJE

AGARWAL, Ramesh, Charu AGGARWAL a V V V PRASAD, [b.r.]. *Depth First Generation of Long Patterns* [online]. Dostupné z: http://www.cs.tau.ac.il/~fiat/dmsem03/Depth First Generation of Long Patterns - 2000.pdf

AGRAWAL, Rakesh a Ramakrishnan SRIKANT, 1994. Fast Algorithms for Mining Association Rules in Large Databases. In: *VLDB conference* [online]. Dostupné z: http://www.vldb.org/conf/1994/P487.PDF

ARLT, Josef, Martina ARLTOVÁ a Eva RUBLÍKOVÁ, 2002. *Analýza ekonomických časových řad s příklady* [online]. Praha: Vysoká škole ekonomická, Fakulta informatiky a statistiky. Dostupné z: https://nb.vse.cz/~arltova/vyuka/crsbir02.pdf

ARLTOVÁ, Martina a Josef ARLT, 1995. Grafické metody analýzy ekonomických časových řad. *Statistika* [online]. **32**(11), 483–493 [vid. 2021-12-15]. ISSN 0322-788x. Dostupné z: Grafické metody analýzy ekonomických časových řad

BERKA, Petr, 2005. *Dobývání znalostí z databází*. Praha: Academia. ISBN 80-200-1062-9.

BIOLAB, 2016. *Association Rules* [online] [vid. 2021-07-09]. Dostupné z: https://orange3-associate.readthedocs.io/en/latest/widgets/associationrules.html

BREIMAN, Leo, 2001. Random Forests. *Machine Learning* [online]. **45**(1), 5–32. ISSN 1573-0565. Dostupné z: doi:10.1023/A:1010933404324

BROWNLEE, Jason, 2018. *How to Check if Time Series Data is Stationary with Python* [online]. Dostupné z: https://machinelearningmastery.com/time-series-data-stationary-python/

ČERVOVÁ, Lubomíra, 2020. *Bootstrapping aneb jak souvisí statistika s řemínky na botách* [online] [vid. 2021-09-03]. Dostupné z: https://acrea.cz/bootstrapping-aneb-jak-souvisi-statistika-s-reminky-na-botach/

CHANG, Chih-Chung a Chih-Jen LIN, 2011. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* [online]. **2**(3). ISSN 2157-6904. Dostupné z: doi:10.1145/1961189.1961199

CHATTAMVELLI, Rajan, 2011. *Data mining algorithms*. Oxford: Alpha Science International. ISBN 978-1-84265-684-6.

COPERNICUS PROGRAMME, 2020. *Urban Atlas* [online]. Dostupné z: https://land.copernicus.eu/local/urban-atlas

CORTES, Corinna a Vladimir VAPNIK, 1995. Support-Vector Networks. *Machine Learning* [online]. **20**(3), 273–297. ISSN 1573-0565. Dostupné z: doi:10.1023/A:1022627411411

DAWSON, Robert J. MacG., 1995. The „Unusual Episode" Data Revisited . *Journal of Statistics Education* [online]. **3**(3) [vid. 2021-07-22]. Dostupné z: http://jse.amstat.org/v3n3/datasets.dawson.html

DEMŠAR, Janez, Tomaž CURK, Aleš ERJAVEC, Črt GORUP, Tomaž HOČEVAR, Mitar MILUTINOVIČ, Martin MOŽINA, Matija POLAJNAR, Marko TOPLAK, Anže STARIČ, Miha ŠTAJDOHAR, Lan UMEK, Lan ŽAGAR, Jure ŽBONTAR, Marinka ŽITNIK a Blaž ZUPAN, 2013. Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research* [online]. **14**(35), 2349–2353. Dostupné z: http://jmlr.org/papers/v14/demsar13a.html

DOBESOVA, Zdena, 2019a. *Discovering association rules of information dissemination about geoinformatics university study* [online]. ISBN 9783319911885. Dostupné z: doi:10.1007/978-3-319-91189-2_32

DOBESOVA, Zdena, 2019b. The Similarity of European Cities Based on Image Analysis. In: Prokopova Z. SILHAVY R., SILHAVY P., ed. *Advances in Intelligent Systems and Computing* [online]. Cham: Springer, s. 341–348. ISBN 9783030303280. Dostupné z: doi:10.1007/978-3-030-30329-7_31

DOBESOVA, Zdena, 2020a. Experiment in Finding Look-Alike European Cities Using Urban Atlas Data. *ISPRS International Journal of Geo-Information* [online]. **9**(6), 20. ISSN 22209964. Dostupné z: doi:10.3390/ijgi9060406

DOBESOVA, Zdena, 2020b. Teaching decision tree using a practical example. In: R SILHAVY, ed. *Advances in Intelligent Systems and Computing* [online]. Cham: Springer, s. 247–256. ISBN 9783030519735. Dostupné z: doi:10.1007/978-3-030-51974-2_23

DOBESOVA, Zdena a Jan PINOS, 2019. *Using decision trees to predict the likelihood of high school students enrolling for university studies* [online]. 2019. Dostupné z: doi:10.1007/978-3-030-00211-4_12

DVOŘÁKOVÁ, Stanislava, 2015. *Statistická analýza a časové řady v příkladech*. Jihlava: Vysoká škola polytechnická Jihlava. ISBN 978-80-88064-18-3.

EUROSTAT, 2020. *Statistics explained, Glossary: Functional urban area* [online]. B.m.: Eurostat [vid. 2020-11-15]. Dostupné z: https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Functional_urban_area

EUROSTAT, 2021. *Eurostat database* [online]. Dostupné z: https://ec.europa.eu/eurostat/data/database

EUROSTAT, 2022. *Passengers transported (detailed reporting only) - (quarterly data)* [online] [vid. 2021-12-10]. Dostupné z: https://ec.europa.eu/eurostat/databrowser/product/page/RAIL_PA_QUARTAL

FOONG, Ng Wai, 2019. *Data Science Made Easy: Test and Evaluation using Orange* [online] [vid. 2020-10-10]. Dostupné z: https://towardsdatascience.com/data-science-made-easy-test-and-evaluation-using-orange-d74e554d9021

GODEC, Primož, Matjaž PANČUR, Nejc ILENIČ, Andrej ČOPAR, Martin STRAŽAR, Aleš ERJAVEC, Ajda PRETNAR, Janez DEMŠAR, Anže STARIČ, Marko TOPLAK, Lan ŽAGAR, Jan HARTMAN, Hamilton WANG, Riccardo BELLAZZI, Uroš PETROVIČ, Silvia GARAGNA, Maurizio ZUCCOTTI, Dongsu PARK, Gad SHAULSKY a Blaž ZUPAN, 2019. Democratized image analytics by visual programming through integration of deep models and small-scale machine learning. *Nature Communications* [online]. **10**(1), 4551. ISSN 2041-1723. Dostupné z: doi:10.1038/s41467-019-12397-x

HAN, Jiawei, Jian PEI, Yiwen YIN a Runying MAO, 2004. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. *Data Mining and Knowledge Discovery* [online]. **8**(1), 53–87. ISSN 1573-756X. Dostupné z: doi:10.1023/B:DAMI.0000005258.31418.83

HANČLOVÁ, Jana a Lubor TVRDÝ, 2003. *Úvod do analýzy časových řad*. Ostrava: Ekonomická fakulta, VŠB-TU.

HENDL, Jan, 2012. *Přehled statistických metod : analýza a metaanalýza dat*. 4., rozš. Praha: Portál. ISBN 978-80-262-0200-4.

JANOUŠEK, Matěj, 2019. *Porovnání urbánního prostoru pomocí kruhových výsečí*. magisterská práce, Olomouc, Česká republika. Univerzita Palackého.

JANOUŠOVÁ, E., J. HOLČÍK, D. HARUŠTIAKOVÁ, S. LITTNEROVÁ a J. JARKOVSKÝ, 2020a. Korespondenční analýza. *Analýza a hodnocení biologických dat, Institut biostatistiky a analýz Lékařské fakulty Masarykovy univerzity* [online] [vid. 2021-06-07]. Dostupné z: https://portal.matematickabiologie.cz/index.php?pg=analyza-a-hodnoceni-biologickych-dat--vicerozmerne-metody-pro-analyzu-dat--ordinacni-analyzy--korespondencni-analyza

JANOUŠOVÁ, E., J. HOLČÍK, D. HARUŠTIAKOVÁ, S. LITTNEROVÁ a J. JARKOVSKÝ, 2020b. Požadavky na data a omezení korespondenční analýzy. *Analýza a hodnocení biologických dat, Institut biostatistiky a analýz Lékařské fakulty Masarykovy univerzity* [online] [vid. 2021-06-07]. Dostupné z: https://portal.matematickabiologie.cz/index.php?pg=analyza-a-hodnoceni-biologickych-dat--vicerozmerne-metody-pro-analyzu-dat--ordinacni-analyzy--korespondencni-analyza--pozadavky-na-data-a-omezeni-korespondencni-analyzy

JOENSSEN, Dieter William a Udo BANKHOFER, 2012. Hot Deck Methods for Imputing Missing Data. In: Petra PERNER, ed. *Machine Learning and Data Mining in Pattern Recognition*. Berlin, Heidelberg: Springer Berlin Heidelberg, s. 63–75. ISBN 978-3-642-31537-4.

KAGGLE, 2016. *Painter by Numbers Competition, 1st Place Winner's Interview: Nejc Ilenič* [online]. Dostupné z: http://blog.kaggle.com/2016/11/17/painter-by-numbers-competition-1st-place-winners-interview-nejc-ilenic/

KEDRO, 2020. *Iris dataset example project* [online]. B.m.: QuantumBlack Visual Analytics Limited Revision. Dostupné z: https://kedro.readthedocs.io/en/stable/02_get_started/05_example_project.html#iris-dataset-example-project

KŘIVÝ, Ivan, 2012. *Analýza časových řad*. Ostrava: Univerzita Ostrava.

LITSCHMANNOVÁ, Martina, 2010. *Úvod do analýzy časových řad*. Ostrava: VŠB-TU, Fakulta elektrotechniky, Katedra aplikované matematiky.

LUKASOVÁ, Alena a Jana ŠARMANOVÁ, 1985. *Metody shlukové analýzy*. Praha: SNTL.

MBAABU, Onesmus, 2020. *Introduction to Random Forest in Machine Learning* [online] [vid. 2021-09-03]. Dostupné z: https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

MELOUN, Milan, Jiří MILITKÝ a Martin HILL, 2012. *Statistická analýza vícerozměrných dat v příkladech*. Gerstner. Praha: Academia. ISBN 978-80-200-2071-0.

ORANGE DATA MINING, 2015. *Geo Map* [online] [vid. 2021-07-09]. Dostupné z: https://orange3.readthedocs.io/en/3.5.0/widgets/visualize/geomap.html

ORANGE DATA MINING, 2019a. *Image Analytics: Clustering of Monet and Manet* [online]. Dostupné z: https://www.youtube.com/watch?v=R5uchDa_ba4

ORANGE DATA MINING, 2019b. *Image Analytics: Finding the Lost Monet* [online]. Dostupné z: https://www.youtube.com/watch?v=6srGs5w9x8w

ORANGE DATA MINING, 2021a. Orange. *Orange, Data Mining Fruitful and Fun* [online]. B.m.: University of Ljubljana. Dostupné z: https://orangedatamining.com

ORANGE DATA MINING, 2021b. *Orange Visual Programming* [online]. Dostupné z: https://orange3.readthedocs.io/projects/orange-visual-programming/en/master/

ORANGE DATA MINING, 2021c. *Orange Visual Programming Documentation* [online] [vid. 2021-07-09]. Dostupné z: https://buildmedia.readthedocs.org/media/pdf/orange-visual-programming/latest/orange-visual-programming.pdf

PETR, Pavel, 2014a. *Metody Data Miningu, část 1*. Pardubice: Univerzita Pardubice, Fakulta ekonomicko-správní. ISBN 978-80-7395-872-5.

PETR, Pavel, 2014b. *Metody Data Miningu, část 2*. Pardubice: Univerzita Pardubice, Fakulta ekonomicko-správní. ISBN 978-80-7395-873-2.

POLICIE ČR, 2020. *Majetkové trestné činy* [online]. Dostupné z: https://www.policie.cz/clanek/pomoc-obetem-tc-majetkove-trestne-ciny.aspx

PRETNAR, Ajda, 2016a. *All I See is Silhouette* [online] [vid. 2021-07-09]. Dostupné z: https://orangedatamining.com/blog/2016/03/23/all-i-see-is-silhouette/

PRETNAR, Ajda, 2016b. *Tips and Tricks for Data Preparation* [online] [vid. 2021-07-09]. Dostupné z: https://orangedatamining.com/blog/2016/01/29/tips-and-tricks-for-data-preparation/

PRETNAR, Ajda, 2019. *Explaining Models: Workshop in Belgrade* [online]. 2019. [vid. 2021-09-15]. Dostupné z: https://orangedatamining.com/blog/2019/2019-11-20-belgrade-workshop/

QUINLAN, J Ross, 1986. Induction of decision trees. *Machine Learning* [online]. **1**(1), 81–106. ISSN 1573-0565. Dostupné z: doi:10.1007/BF00116251

QUINLAN, J Ross, 1993. *C4.5: programs for machine learning*. B.m.: Morgan Kaufmann Publishers Inc. ISBN 1558602380.

RUSSAKOVSKY, Olga, Jia DENG, Hao SU, Jonathan KRAUSE, Sanjeev SATHEESH, Sean MA, Zhiheng HUANG, Andrej KARPATHY, Aditya KHOSLA, Michael BERNSTEIN, Alexander C BERG a Li FEI-FEI, 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* [online]. **115**(3), 211–252. Dostupné z: doi:10.1007/s11263-015-0816-y

ŠARMANOVÁ, Jana, 2012. *Metody analýzy dat* [online]. Ostrava: Vysoká škola báňská Technická univerzita Ostrava. Dostupné z: http://www.person.vsb.cz/archivcd/FEI/MAD/MAD.pdf

SAYAD, Saed, 2020a. *An Introduction to Data Science* [online]. Dostupné z: http://www.saedsayad.com/data_mining_map.htm

SAYAD, Saed, 2020b. *Support Vector Machine - Regression (SVR)* [online]. 2020. Dostupné z: http://www.saedsayad.com/support_vector_machine_reg.htm

STANFORD VISION LAB, 2020. *Imagenet: Large Scale Visual Recognition Challenge* [online] [vid. 2021-11-24]. Dostupné z: https://image-net.org/challenges/LSVRC/2014/browse-synsets

SUNAYSAWANT, 2021. *Air Passengers - Time Series - ARIMA* [online]. Dostupné z: https://www.kaggle.com/sunaysawant/air-passengers-time-series-arima

TAN, Edwin, 2021. Unsupervised Anomaly Detection in Python. *Towards Data Science* [online] [vid. 2021-12-14]. Dostupné z: https://towardsdatascience.com/unsupervised-anomaly-detection-in-python-f2e61be17c2b

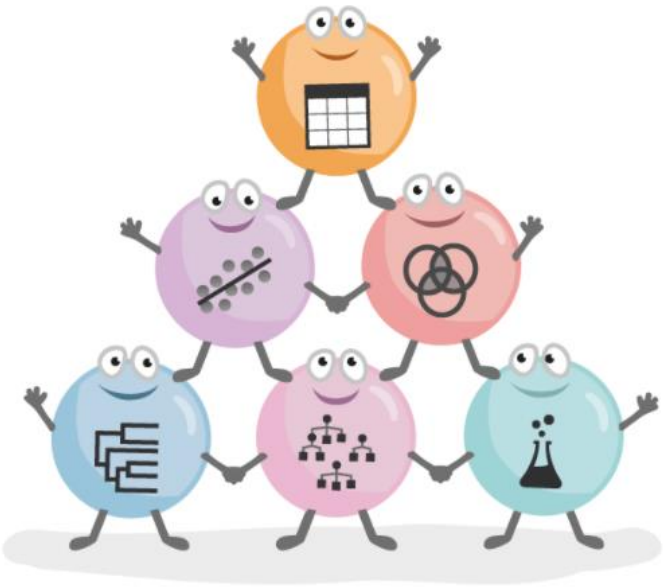
WARD, Joe H, 1963. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* [online]. **58**(301), 236–244. ISSN 0162-1459. Dostupné z: doi:10.1080/01621459.1963.10500845

WIKIPEDIA, 2020a. *Cosine similarity* [online]. Dostupné z: https://en.wikipedia.org/wiki/Cosine_similarity

WIKIPEDIA, 2020b. *DBSCAN* [online]. Dostupné z: https://en.wikipedia.org/wiki/DBSCAN

WIKIPEDIA, 2020c. *Lift (data mining)* [online]. Dostupné z: https://en.wikipedia.org/wiki/Lift_(data_mining)

ZUPAN, Blaž, 2020. *Look-alike Images* [online] [vid. 2021-05-15]. Dostupné z: https://orangedatamining.com/blog/2020/2020-01-08-neighbors-images/



ORANGE: Praktický návod do cvičení předmětu Data Mining

doc. Ing. Zdena Dobešová, Ph.D.

Odpovědný redaktor Otakar Loutocký
Předseda ediční komise PřF UP prof. RNDr. Jan Hlaváč, Ph.D.
Publikace neprošla redakční jazykovou úpravou
Loga software Orange Agnieszka Rovšnik
Sazbu provedla Zdena Dobešová

Vydala Univerzita Palackého v Olomouci, Křížkovského 8, 771 47 Olomouc
Vydáno pro Katedru geoinformatiky PřF UP jako její 94. publikaci

vydavatelstvi.upol.cz
1. vydání
Olomouc 2022
DOI: 10.5507/prf.22.2440864
ISBN 978-80-244-6086-4
VUP 2022/0058