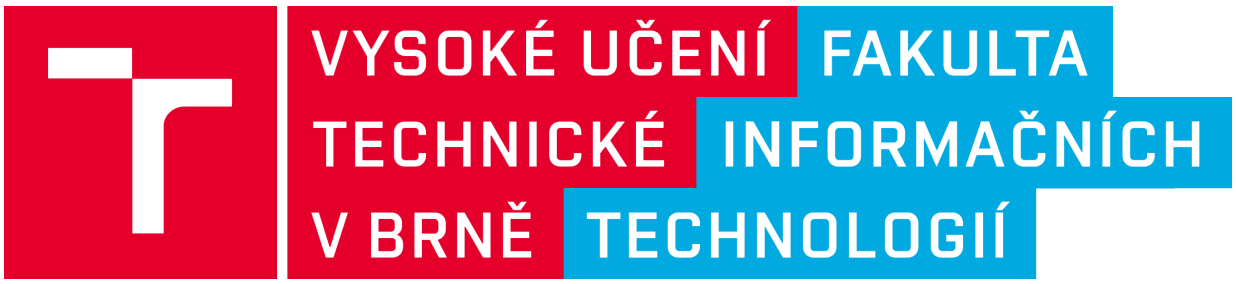
Fakulta informačních technologií
Vysoké učení technické v Brně



Získávání znalostí z databází

Databáze restaurací

2019/2020

Tomáš Aubrecht (xaubre02)
Stanislav Bartoš (xbarto87)

11. února 2020

1 Zadání projektu

Cílem projektu do předmětu ZZN bylo nejprve nastudovat strukturu dat, která byla k dispozici ke konkrétní variantě projektu. V našem případě se jednalo o data restaurací. Struktura dat je podrobněji popsána v kapitole 2. Po prostudování dat následovala volba úlohy, kterou budeme v rámci projektu zkoumat. Šlo o to zvolit takovou úlohu, která přinese o datech novou znalost a nebude zcela triviální, tedy například nebude nová znalost zjistitelná pouhým pohledem do tabulek. Úloha, kterou jsem se rozhodli v rámci projektu zpracovat, je popsána v kapitole 3.

2 Popis struktury dat

Data, která jsou k dispozici pro řešení varianty restaurací, se sestávají z devíti csv souborů (tabulek). Na základě analýzy dat, jež se v souborech nachazí, je možno rozdělit tyto soubory do tří kategorií:

1. **Informace o restauracích**
   - `geoplaces2.csv` – obsahuje základní informace o restauracích, tedy:
     ID_restaurace, zeměpisná šírka/délka, jméno, adresa, město, region, stát, fax, PSČ, podávání alkoholu, možnost kouření, úroveň oblečení, dostupnost, cenová kategorie, adresa webové stránky, franšíza, atmosféra, rahrádka, ostatní služby
   - `chefmozaccepts.csv` – obsahuje informaci o možných způsobech platby:
     ID_restaurace, typ platby
   - `chefmozcuisine.csv` – obsahuje informaci o typech kuchyní v nabídce:
     ID_restaurace, typ kuchyně
   - `chefmozhours4.csv` – obsahuje informaci o otevírací době:
     ID_restaurace, hodiny, dny
   - `chefmozparking.csv` – obsahuje informaci o možných způsobech parkování:
     ID_restaurace, způsob parkování
2. **Informace o zákaznících**
   - `userprofile.csv` – obsahuje základní informace o zákaznících:
     ID_zákazníka, zeměpisná šírka/délka, kuřák, úroveň pití alkoholu, oblíbený styl oblékání, oblíbená atmosféra, způsob dopravy, rodinný stav, děti, rok narození, zájmy, osobnost, náboženství, zaměstnání, oblíbená barva, váha, rozpočet, výška
   - `usercuisine.csv` – obsahuje informaci o oblíbených kuchyních jednotlivých zákazníků:
     ID_restaurace, oblíbená kuchyně
   - `userpayment.csv` – obsahuje informaci o možných způsobech platby zákazníků:
     ID_restaurace, způsob platby
3. **Hodnocení restaurací**
   - `rating_final.csv` – hodnocení restaurací zákazníky:
     ID_zákazníka, ID_restaurace, hodnocení, hodnocení jídla, hodnocení nabízených služeb

3 Formulace úlohy

Majitel nově budované restaurace se snaží zjistit, jakým směrem by se jeho nová restaurace měla ubírat, aby byla nejen konkurenceschopná, ale aby se také těšila velké oblibě. Z tohoto důvodu musí zajistit co nejvhodnější podmínky pro získání velké spokojenosti mezi zákazníky. Pokud je zákazník spokojen, dá se předpokládat, že udělí restauraci i vysoké hodnocení, které bude mít za následek větší navštěvovanost.

Snahou je tedy predikce spokojenosti zákazníka na základě dostupných informací o konkurenčích restauracích a informací o tom, jak různí zákazníci tyto restaurace hodnotí vzhledem k jejich nabízenému portfóliu. Na základě této predikce bude mít majitel možnost se rozhodnout, jakou kuchyni bude restaurace preferovat a jaké bude nabízet služby. Jinými slovy, do kterých služeb je vhodné investovat a které naopak budou mít minimální dopad na hodnocení restaurace.

4 Řešení

V této kapitole je podrobně popsán postup při řešení výše specifikované úlohy. Jejím cílem bylo vytvořit model predikující hodnocení restaurace, z něhož by bylo možné vyčíst informace o tom, jaké atributy a jejich kombinace mají největší vliv na dané hodnocení. Tento model by měl sloužit pro podporu rozhodování o tom, jaké služby zahrnout v nabízeném portfóliu při budování nové či rekonstrukci stávající restaurace za účelem co nejvyšší spokojenosti zákazníků.

Pro zadanou úlohu jsou důležité všechny dostupné soubory, které popisují vlastnosti jednotlivých restaurací a jejich hodnocení zákazníky. Úlohu lze rozdělit do několika postupně navazujících procesů, které jsou detailněji popsány v následujících kapitolách. První z nich je předzpracování dostupných dat, následně klasifikace s výběrem relevantních dat a reprezentace získaných výsledků v závěru.

4.1 Předzpracování dat

Je nutné předzpracovat obsah všech vstupních souborů, protože některá data jsou nekonzistentní, zašumělá či u nich chybí některé hodnoty. Jak již bylo zmíněno, jako zdroj dat byly zvoleny ty soubory, které se týkají restaurací či jejich nabízených služeb a hodnocení zákazníků. Jedná se o následující soubory:

- `geoplaces2.csv`
- `rating_final.csv`
- `chefmozaccepts.csv`
- `chefmozparking.csv`
- `chefmozcuisine.csv`
- `chefmozhours4.csv`

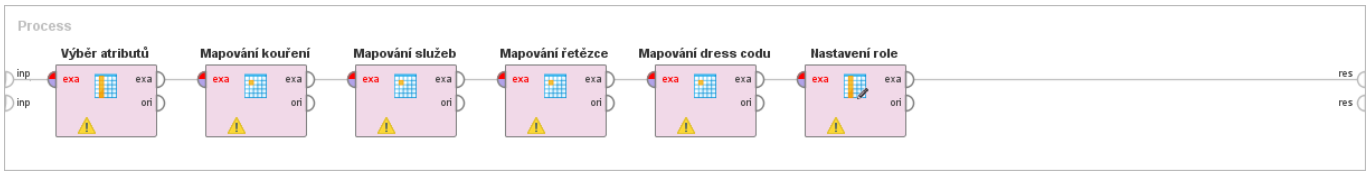
Pro každý soubor byl vytvořen samostatný proces, ve kterém je předzpracován obsah daného souboru. Tyto procesy jsou popsány v následujících podkapitolách. Data získaná z jednotlivých procesů předzpracování jsou spojena dohromady a jsou tak připravena pro dolování informací. Celkový proces předzpracování dat můžete vidět na obrázku 4 na straně 6.

4.1.1 Předzpracování restaurací

Soubor `geoplaces2.csv` obsahuje velké množství atributů týkající se jednotlivých restaurací, kde pro zadanou úlohu nejsou všechny atributy relevantní. Prvním krokem je tedy výběr podmnožiny atributů, se kterými se bude dále pracovat. Vybrané atributy josu: `accessibility`, `alcohol`, `area`, `dress_code`, `franchise`, `other_services`, `placeID`, `price`, `Rambience` a `smoking_area`.

Dalším krokem je namapování některých z těchto atributů na obecnější hodnoty za účelem redukce počtu hodnot, které jednotlivé atributy mohou nabývat. Kouření, služby navíc a obchodní řetězec jsou namapovány pouze na hodnoty `yes` a `no` a dress code je namapován na hodnoty `formal` a `informal`.

Posledním krokem je nastavení role `ID` pro atribut `placeID`, který se využije při spojování dat. Nastavení této role se děje v každém následujícím procesu předzpracování, proto to již nebude zmiňovat.

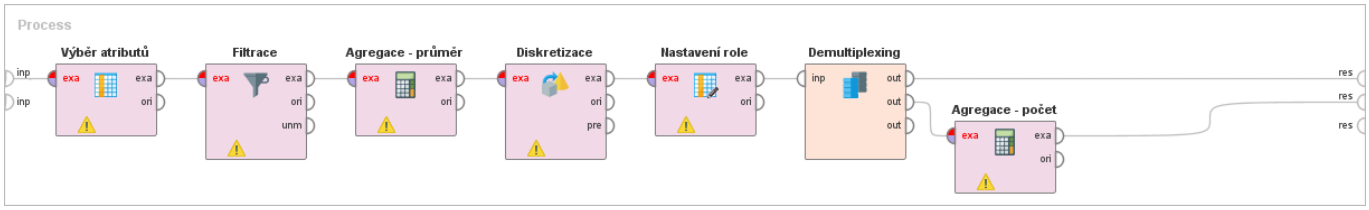


Obrázek 1: Proces předzpracování restaurací.

4.1.2 Předzpracování hodnocení

Vzhledem k povaze úlohy je nejdůležitějším atributem vytvářeného modelu již udělené hodnocení restaurací zákazníky. Nejprve se vyfiltrovaly chybějící hodnoty. Následně je potřeba toto hodnocení agregovat pro jednotlivé restaurace, kde vhodnou agregační funkcí pro daný případ je průměr. Získané hodnoty jsou dále diskretizovány z intervalu <0,2> do tříd *, ** a *** odpovídající danému hodnocení.

Po nastavení role je zde další agregace. Ta má spíše informativní účel, kde se využívá agregační funkce počet pro zjištění počtu restaurací, které získaly jednotlivá ohodnocení(*Celkem je 44 restaurací, které získaly 3 hvězdičky, 75 restaurací, které získaly 2 hvězdičky a 11 restaurací s jednou hvězdičkou.*). Tato agregace je využita i v následujících procesech, proto ji dále také nebudeme zmiňovat.

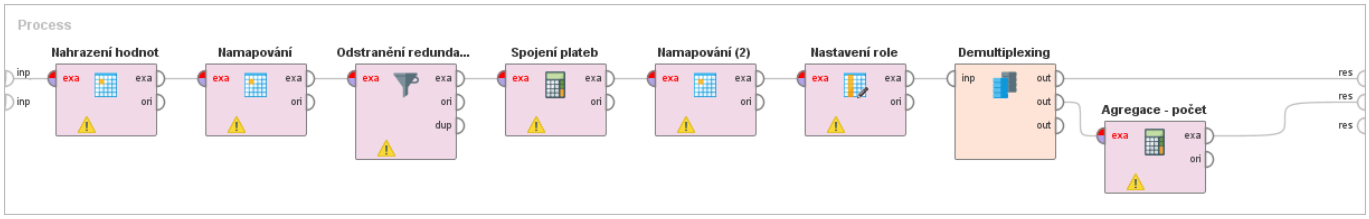


Obrázek 2: Proces předzpracování hodnocení restaurací.

4.1.3 Předzpracování plateb

Hlavním problémem zpracování plateb byla nekonzistence hodnot `card` a `visa`, proto jsme ji vyřešili nahrazením hodnot pomocí regulárního výrazu. Posléze došlo k namapovaní dosavadních hodnot na `cash` nebo `card`, které vyjadřují, zdali se dá v dané restaraci platit hotově nebo kartou. Dalším krokem bylo odstranění redundantních dat a spojení těch zbylých, které se nyní namapovaly na hodnoty `cash`, `card` nebo `both`.

Tento proces můžete vidět na obrázku 3. Zbylé procesy jsou velmi podobné tomuto nebo některému z předcházejících procesů, proto u nich již nebudeme uvádět obrázek.



Obrázek 3: Proces předzpracování plateb.

4.1.4 Předzpracování parkování

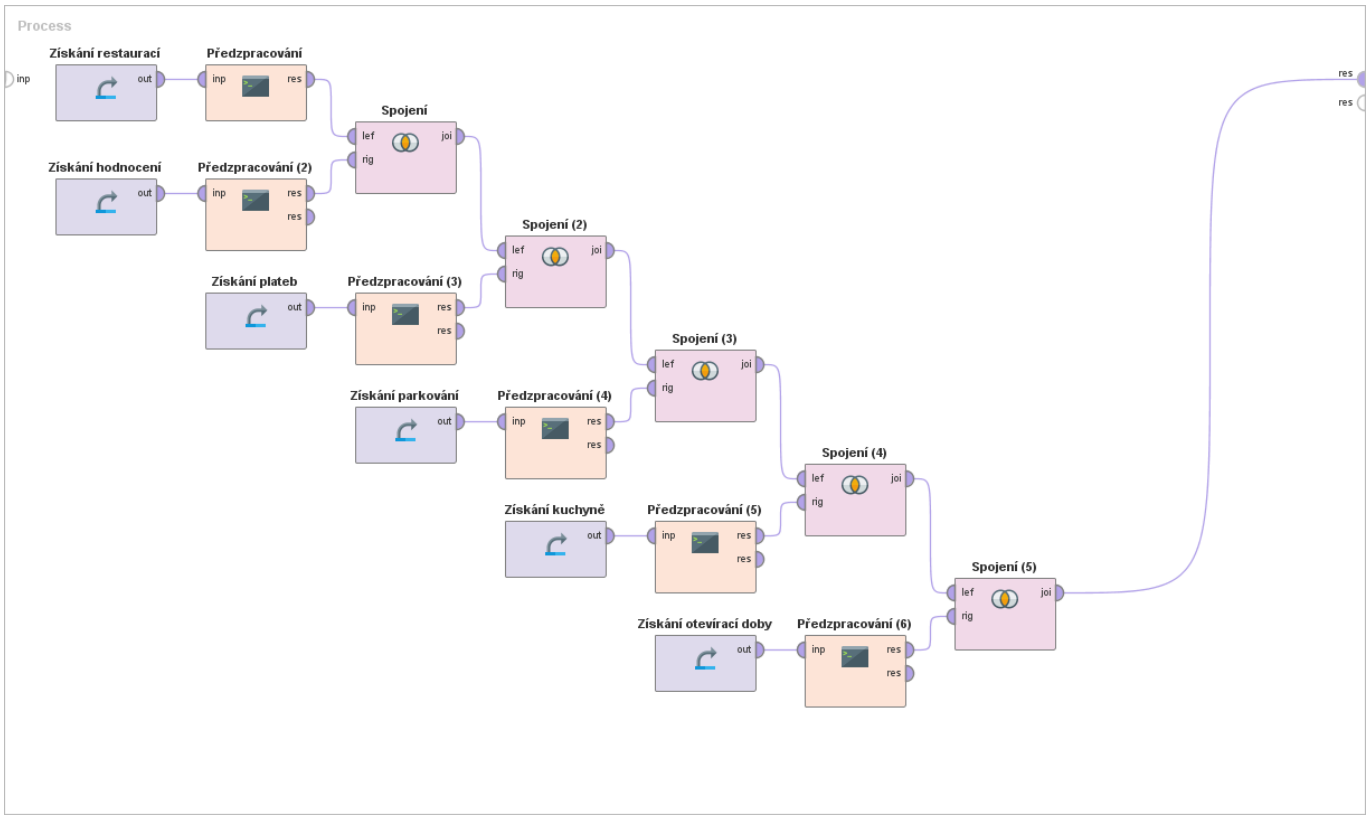
Hlavní činností při zpracování dat o možnosti parkování bylo namapování všech možností parkování pouze na hodnoty `yes` a `no` vyjadřující, zdali se dá poblíž dané restaurace parkovat či nikoliv. Nakonec se ze získaných výsledků odstranily duplicity.

4.1.5 Předzpracování kuchyně

Vzhledem k množství hodnot, jaké mohou být kuchyně, jsme provedli spojení a namapování na daleko obecnější hodnoty. Těmito hodnotami jsou: `restaurant`, `fast_food`, `pub_or_bar`, `pizzeria` a `cafe`.

4.1.6 Předzpracování otevírací doby

U otevírací doby jsme pracovali pouze s hodinami. Dny, kdy má daná restaurace otevřeno, jsme nevyužili. Tyto hodiny jsme pak namapovali na hodnoty, které vyjadřují hlavní čas, kdy má daná restaurace otevřeno. Tyto hodnoty jsou: `nonstop`, `whole_day`, `morning`, `noon`, `afternoon`, `evening` a `night`.



Obrázek 4: Celkový proces předzpracování dat.

4.2 Klasifikace

V této kapitole jsou uvedeny kromě zvolených klasifikačních metod i vybrané atributy a rovněž podrobné výsledky vykázané jednotivými metodami.

4.2.1 Výběr atributů

Pro řešení výše uvedené úlohy jsme experimentovali s různými kombinacemi atributů, které se nám zdály relevantní vzhledem k povaze této úlohy a které by mohly mít zásadní vliv na hodnocení restaurací. Jmenovitě se jednalo o tyto atributy:

`rating, Rambience, accesibility, alcohol, franchise, other_services, parking_lot, dress_code, price, area, hours, smoking_area, Rpayment, Rcuisine.`

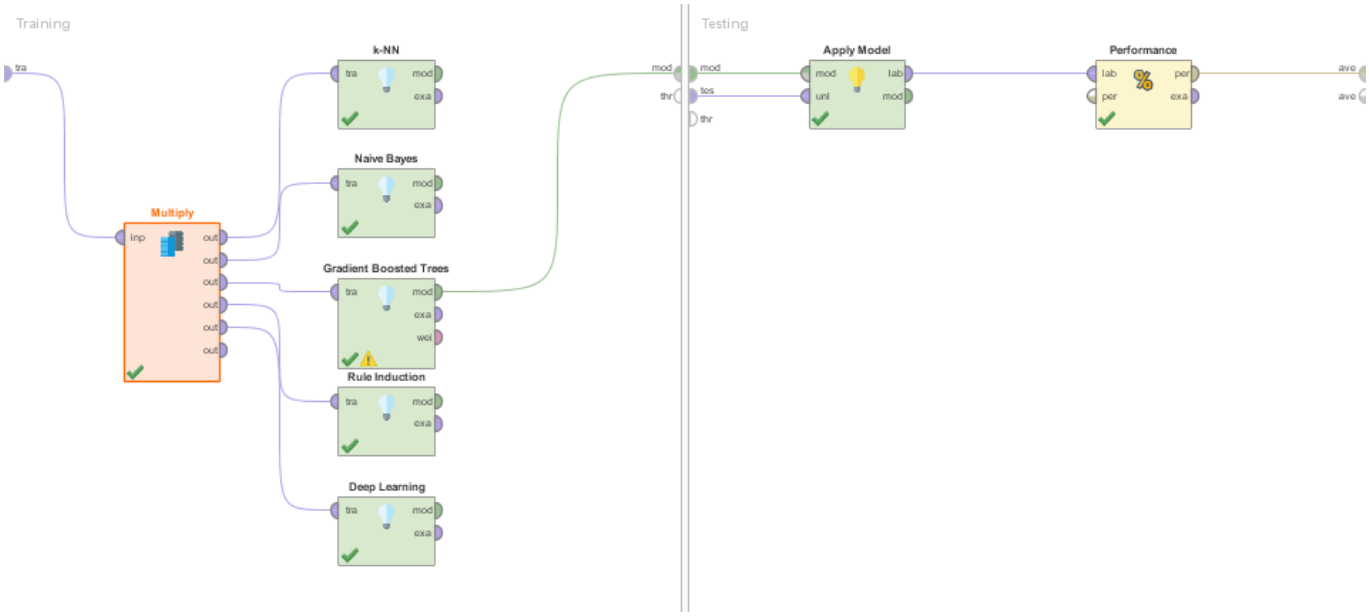
4.2.2 Výběr klasifikačních metod

Z metod pro klasifikaci, které RapidMiner nabízí, jsme zvolili z každé aplikovatelné kategorie (např. *Trees*) vždy tu, jež dosahovala nejlepších výsledků. Zvolené metody jsou tyto: `k-NN`, `Naive Bayes`, `Gradient Boosted Tree`, `Rule Induction` a `Deep Learning`. Celý proces klasifikace a porovnávání jednotlivých metod je znázorněn na obrázku 5.

Data pro učení a testování byla rozdělena stratifikovaným vzorkováním v poměru 80 % pro učení a 20 % pro testování.

Zároveň jsme zkoušeli i upravovat parametry některých metod. Zatímco na většinu metod to nemělo žádný vliv, nebo se výsledky zhoršily, tak u metody Deep Learning se nám osvědčilo použít jako aktivační funkci *Maxout* a počet epoch zvýšit na 100 – to sice prodloužilo dobu učení, ale zároveň se znatelně zvýšila přesnost.

V tabulkách jsou pak uvedeny přesnosti jednotlivých metod nad konkrétními třídami. Počet hvězdiček označuje třídu podle hodnocení – způsob jejich vytvoření ze zadaných dat byl již popsán výše v sekci o předzpracování dat. Pro zopakování platí následující: čím více hvězdiček, tím vyšší průměrné hodnocení.



Obrázek 5: Proces validace klasifikačních metod.

accuracy: 53.85%

	true *	true **	true ***	class precision
pred. *	1	2	0	33.33%
pred. **	0	11	7	61.11%
pred. ***	1	2	2	40.00%
class recall	50.00%	73.33%	22.22%	

Obrázek 6: k-NN.

accuracy: 50.00%

	true *	true **	true ***	class precision
pred. *	1	2	0	33.33%
pred. **	1	9	6	56.25%
pred. ***	0	4	3	42.86%
class recall	50.00%	60.00%	33.33%	

Obrázek 7: Naive Bayes.

accuracy: 61.54%

	true *	true **	true ***	class precision
pred. *	0	0	0	0.00%
pred. **	2	15	8	60.00%
pred. ***	0	0	1	100.00%
class recall	0.00%	100.00%	11.11%	

Obrázek 8: Gradient Boosted Tree.

accuracy: 42.31%

	true *	true **	true ***	class precision
pred. *	0	0	0	0.00%
pred. **	1	9	7	52.94%
pred. ***	1	6	2	22.22%
class recall	0.00%	60.00%	22.22%	

Obrázek 9: Rule Induction.

accuracy: 69.23%

	true *	true **	true ***	class precision
pred. *	1	0	0	100.00%
pred. **	1	13	5	68.42%
pred. ***	0	2	4	66.67%
class recall	50.00%	86.67%	44.44%	

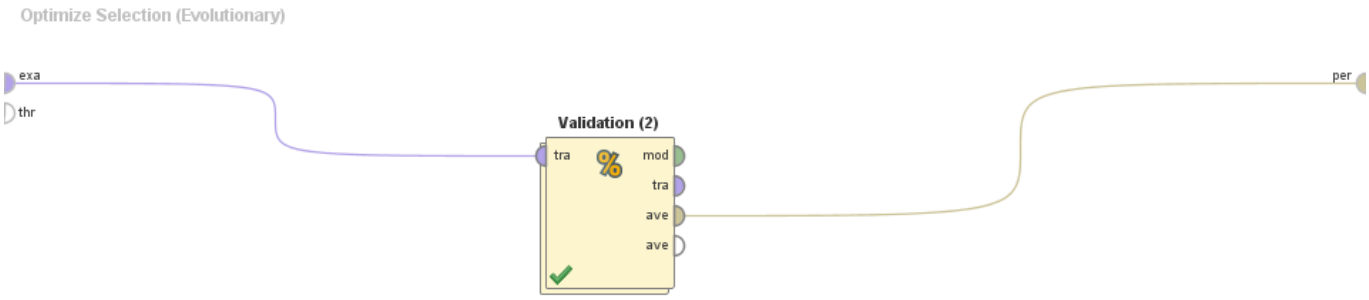
Obrázek 10: Deep Learning.

Jak je vidět, tak nejlepších výsledků bylo dosaženo pomocí metody Deep learning, ovšem přesnost této metody velmi kolísala (mezi 50 a 70 procenty), což je podle nás způsobeno zejména relativně malou množinou trénovacích dat.

4.3 Optimalizace

Jelikož cílem úlohy bylo zjistit, které atributy mají největší vliv na výsledné hodnocení, zkusili jsme výběr atributů optimalizovat. Nejprve jsme zkoušeli použít optimalizaci typu *Brute Force*, ale doba běhu této optimalizační metody se ukázala jako neúnosná. Z tohoto důvodu jsme se nakonec rozhodli použít optimalizaci pomocí evolučních alogoritmů.

Výběr klasifikačních metod je shodný s výběrem uvedeným v předchozí sekci. Proces optimalizace obsahuje jediný subproces, a to validaci viz obrázek 11. Tato validace je zcela totožná s tou popsanou na obrázku 5.



Obrázek 11: Proces optimalizace výběru atributů.

accuracy: 80.77%

	true *	true **	true ***	class precision
pred. *	0	0	0	0.00%
pred. **	1	15	3	78.95%
pred. ***	1	0	6	85.71%
class recall	0.00%	100.00%	66.67%	

Obrázek 12: k-NN.

accuracy: 73.08%

	true *	true **	true ***	class precision
pred. *	1	0	0	100.00%
pred. **	1	15	6	68.18%
pred. ***	0	0	3	100.00%
class recall	50.00%	100.00%	33.33%	

Obrázek 13: Naive Bayes.

accuracy: 80.77%

	true *	true **	true ***	class precision
pred. *	0	0	0	0.00%
pred. **	1	13	1	86.67%
pred. ***	1	2	8	72.73%
class recall	0.00%	86.67%	88.89%	

Obrázek 14: Gradient Boosted Tree.

accuracy: 76.92%

	true *	true **	true ***	class precision
pred. *	0	0	0	0.00%
pred. **	2	15	4	71.43%
pred. ***	0	0	5	100.00%
class recall	0.00%	100.00%	55.56%	

Obrázek 15: Rule Induction.

accuracy: 76.92%

	true *	true **	true ***	class precision
pred. *	1	0	0	100.00%
pred. **	1	15	5	71.43%
pred. ***	0	0	4	100.00%
class recall	50.00%	100.00%	44.44%	

Obrázek 16: Deep Learning – optimalizace výběru atributů.

Metoda	Přesnost	Zvolené atributy
k-NN	80.77%	dress_code, franchise, smoking_area, accessibility, price, area, hours
Naive Bayes	73.08%	dress_code, smoking_area, other_services, alcohol, accessibility, Rambience, area, parking_lot
Gradient Boosted Tree	80.77%	other_services, smoking_area, accessibility, Rambience, area, hours
Rule Induction	76.92%	franchise, other_services, accessibility, parking_lot
Deep Learning	76.92%	dress_code, other_services, accessibility, Rambience, area, parking_lot, hours

Dále jsme zkoušeli zjistit poměr, jakým jednotlivé atributy ovlivňují celkové hodnocení – tedy váhy. K tomu jsme se rozhodli využít optimalizaci vah opět pomocí evolučních algoritmů. S metodou Deep Learning se nám ovšem podařilo dosáhnout přesnosti úctyhodných 84.62 %.

accuracy: 84.62%

	true *	true **	true ***	class precision
pred. *	1	0	0	100.00%
pred. **	1	13	1	86.67%
pred. ***	0	2	8	80.00%
class recall	50.00%	86.67%	88.89%	

Obrázek 17: Deep Learning – optimalizace vah.

attribute	weight
dress_code	0.642
franchise	0.274
other_services	1
smoking_area	0.151
alcohol	0
accessibility	0.518
price	0.033
Rambience	0.530
area	0.259
concat(Rpayment)	0.504
parking_lot	0.013
concat(Rcuisine)	0.265
concat(hours)	0.221

Obrázek 18: Optimalizované váhy atributů.

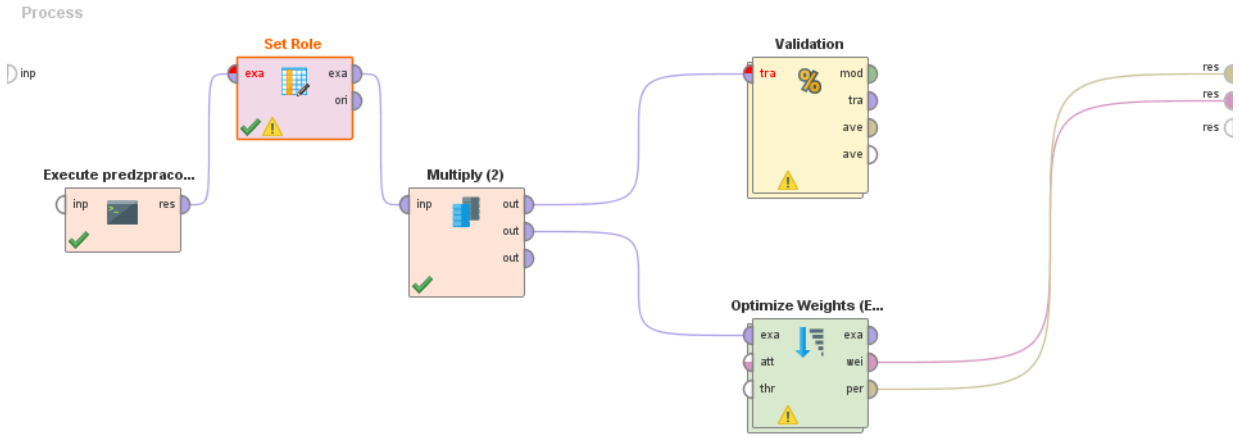
5 Závěr

Cílem tohoto projektu bylo vytvořit model, který predikuje hodnocení restaurace v závislosti na vlastnostech a službách, kterými daná restarace disponuje. Účelem tohoto modelu je podpora rozhodování při budování restarace zaměřená na maximální spokojenost zákazníků. Řešení bylo dosaženo v několika dílčích krocích, které zahrnují předzpracování vstupních dat, volba relevantních atributů, volba klasifikačních metod, samotná klasifikace a následně její optimilazace. Celé schéma tohoto procesu je znázorněno na obrázku 19.

Hodnocení bylo predikováno pomocí několika klasifikačních metod, které jsou dostupné v aplikaci RapidMiner. Při porovnání jednotlivých metod dostáváme nejlepší výsledky při použití metody Deep Learning. Její přesnost ale pro jednotlivá hodnocení znatelně kolísá.

Dalším krokem byla tedy optimalizace výběru atributů pro jednotlivé metody s cílem získat co nejpřesnější predikci. Tato optimalizace byla provedena pomocí evolučních algoritmů. Nejlépe opět vycházela metoda, která dokázala predikovat správný výsledek v 84.62 % případů.

Ukázalo se, že nejdůležitějšími parametry pro predikci hodnocení restaurace zákazníky jsou následující atributy: `accessibility`, `area`, `dress_code`, `smoking_area`, `other_services`, `parking_lot` a `hours`. Toto je informace, která představovala cíl našeho snažení. Díky ní se může majitel své restaurace snadněji rozhodnout, do kterých služeb je vhodné investovat.



Obrázek 19: Celé schéma řešení v RapidMineru.