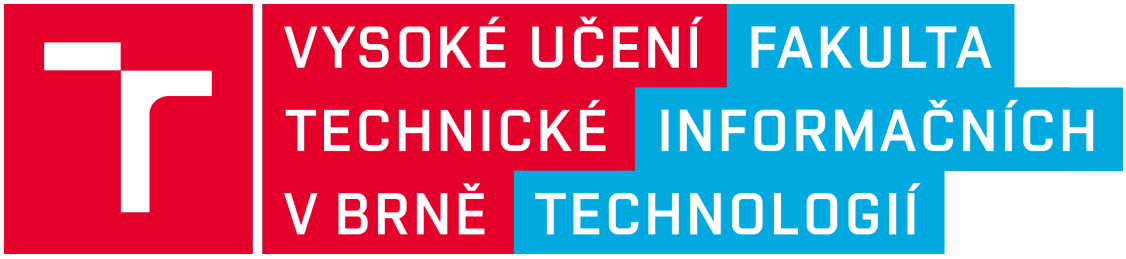


Získávání znalostí z databází

Databáze z výzkumu příčin obezity

Řešení

David Chocholatý (xchoch09)
Martin Baláž (xbalaz15)

Brno, 14. října 2024

1 Informace o datasetu

Vytvoření datasetu bylo motivováno cílem odhadu úrovní obezity u jedinců na základě jejich fyzického stavu a stravovacích návyků. Dataset obsahuje údaje ze tří zemí: Mexika, Peru a Kolumbie. Byl sestaven v roce 2020 s hlavním cílem shromáždit data pro výzkum faktorů ovlivňujících obezitu, která patří mezi rostoucí globální zdravotní problémy. Analýza těchto dat umožňuje predikovat hlavní faktory spojené s obezitou, čímž může přispět k lepšímu porozumění této celosvětové epidemii.

1.1 Struktura databázových dat

Dataset obsahuje celkem 17 atributů a 2111 záznamů. Záznamy jsou rozděleny do kategorií podle třídy, označované jako *NObesity*, která určuje úroveň obezity. Na základě této třídy jsou data klasifikována do 7 kategorií: *Nedostatečná hmotnost*, *Normální hmotnost*, *Nadváha I. stupně*, *Nadváha II. stupně*, *Obezita I. stupně*, *Obezita II. stupně* a *Obezita III. stupně*. Z datasetu tvoří ze 77 % data vygenerovaná synteticky pomocí nástroje Weka a filtru SMOTE. Zbylých 23 % dat bylo přímo získáno od uživatelů prostřednictvím webové platformy [1].

Atribut	**Typ**	**Popis**
Gender	Kategorický	Pohlaví
Age	Kontinuální	Věk
Height	Kontinuální	Výška
Weight	Kontinuální	Váha
family_history_with_overweight	Binární	Má člen rodiny problémy s nadváhou?
FAVC	Binární	Často konzumujete vysoce kalorické jídlo?
FCVC	Kontinuální	Obvykle jíte zeleninu při jídlech?
NCP	Kontinuální	Kolik hlavních jídel denně máte?
CAEC	Kategorický	Jíte nějaké jídlo mezi hlavními jídly?
SMOKE	Binární	Kouříte?
CH2O	Kontinuální	Kolik vody denně pijete?
SCC	Binární	Sledujete množství kalorií, které denně přijímáte?
FAF	Kontinuální	Jak často máte fyzickou aktivitu?
TUE	Kontinuální	Kolik času používáte technologická zařízení (mobil, videohry, atd.)?
CALC	Kategorický	Jak často pijete alkohol?
MTRANS	Kategorický	Jaký druh dopravy obvykle používáte?
NObeyesdad	Kategorický	Úroveň obezity

Tabulka 1: Popis atributů datasetu.

Informace o respondentech, které jsou v databázi obsaženy, zachycuje tabulka 1.

2 Zadání vlastních úloh

Pro účely pokrytí co možná největší škály přístupů budou zvoleny 3 úlohy, a to úloha *klasifikace*, *klastrování* a *regrese*. Pro každou z úloh budou vybrány vhodné modely, mezi kterými bude provedeno závěrečné srovnání. Zároveň pomocí vhodných modelů bude možné určit důležitost jednotlivých atributů a jejich vliv na výsledné vyhodnocení.

2.1 Úloha klasifikace — stanovení úrovně obezity

V rámci této úlohy se budeme zaměřovat na predikci různých úrovní obezity na základě cílového atributu *NObeyesdad*. Daný atribut představuje různé kategorie obezity, což umožňuje určit úrovně obezity jednotlivců na základě jejich dalších charakteristik, jako jsou věk, pohlaví, hmotnost a další faktory.

Stanovení úrovně obezity je vhodné, protože obezita se neurčuje přímo pomocí váhy (výpočet tzv. *indexu tělesné hmotnosti*), ale na základě podílu tělesného tuku [2]. Navíc s využitím vhodných modelů můžeme lépe porozumět vztahu mezi různými atributy a úrovní obezity, což může vést k účinnějším strategiím pro prevenci a léčbu obezity.

2.2 Úloha klastrování — určení skupin v populaci

Cílem této úlohy je nalézt přirozeně se vyskytující skupiny v populaci na základě atributů souvisejících s obezitou, jako jsou stravovací návyky, fyzická aktivita a další životní faktory. Klastrování umožní identifikovat skupiny jedinců, kteří sdílejí podobné charakteristiky a vzorce chování, aniž by byla předem stanovena kritéria pro rozdělení.

Tento přístup odhalí různé skupiny s podobným rizikem vzniku obezity nebo se zdravými návyky, což pomůže k lepšímu pochopení struktury populace a případné identifikaci vzorců chování, které mohou souviset s obezitou. Získané informace mohou být užitečné pro cílené preventivní programy a personalizované zdravotní osvěty.

2.3 Úloha regrese — predikce váhy

Tato úloha si klade za cíl předpovědět váhu osoby na základě dostupných informací, jako je věk, výška, rodinná anamnéza, fyzická aktivita, a dalších faktorů. Vytvořený regresní model lze využít k odhadu budoucí váhy osoby v horizontu 5, 10 či 20 let, za předpokladu, že si udrží stejné návyky a životní styl. K tomu stačí použít stejné vstupní údaje o osobě, s jedinou výjimkou — změnou věku.

I přesto, že váha přímo neurčuje úroveň obezity, která se vyhodnocuje podle podílu tuku v těle [2], lze pomocí nárůstu váhy v budoucích letech identifikovat potenciálně rizikové osoby. U těchto jedinců je možné blíže určit důvody nárůstu váhy. Může to být například nárůst svalové hmoty nebo zvýšení podílu tuku v důsledku nevhodných zdravotních návyků.

3 Příprava a čištění dat

Před samotným dolováním dat jsme provedli důkladnou přípravu, která zahrnovala následující kroky:

- **Odstranění duplikovaných dat**: Nejprve jsme zkontrolovali, zda v datasetu neexistují duplikované záznamy. Po provedení analýzy jsme nalezli celkem 24 duplikátů, které jsme následně odstranili, aby nedošlo k ovlivnění výsledků.
- **Kontrola chybějících hodnot**: Zkontrolovali jsme, zda se v datasetu nevyskytují chybějící hodnoty, a potvrdili jsme, že žádné hodnoty nechybí.
- **Kontrola odlehlých hodnot**: Identifikovali jsme odlehlé hodnoty pomocí statistických metod, jako je interkvartilové rozpětí (IQR) a vizualizací dat pomocí boxplotů. Několik odlehlých hodnot jsme ověřili a rozhodli se je ponechat, protože odrážely skutečné reálné jevy.
- **Kontrola korelovaných atributů**: Zkontrolovali jsme vzájemné korelace mezi atributy pomocí korelační matice, avšak nenašli jsme žádné silně korelované atributy, které by bylo nutné odstranit.
- **Odstranění atributů s nízkou variabilitou**: Na základě analýzy distribuce hodnot jsme odstranili dva atributy – *SMOKE* a *SCC*, protože více než 95% hodnot v těchto atributech bylo stejných, což znamená, že tyto atributy neposkytují dostatečně rozmanité informace pro další analýzu.
- **Převod nominálních atributů na numerické**: Pro úlohu regrese je zapotřebí převést veškeré nominální atributy na numerické. Tato část byla provedena pouze pro úlohu regrese, viz blokové schéma 16 v sekci 4.3.
- **Další úpravy**: Další úpravy dat, jako je standardizace nebo normalizace, nebyly v tomto případě potřeba, protože všechny atributy již byly v odpovídajícím rozsahu.

  Zároveň nebyla využita technika *one-hot encoding* na kategorické atributy, jelikož při jejímž využitím dosahovaly modely horších výsledků, a to řádově 2% pro hodnotu *Accuracy*. Rovněž to stejné platí pro diskretizaci (*binning*) atributu Věk (*Age*) s rozdílem opět 2% pro stejnou metriku.

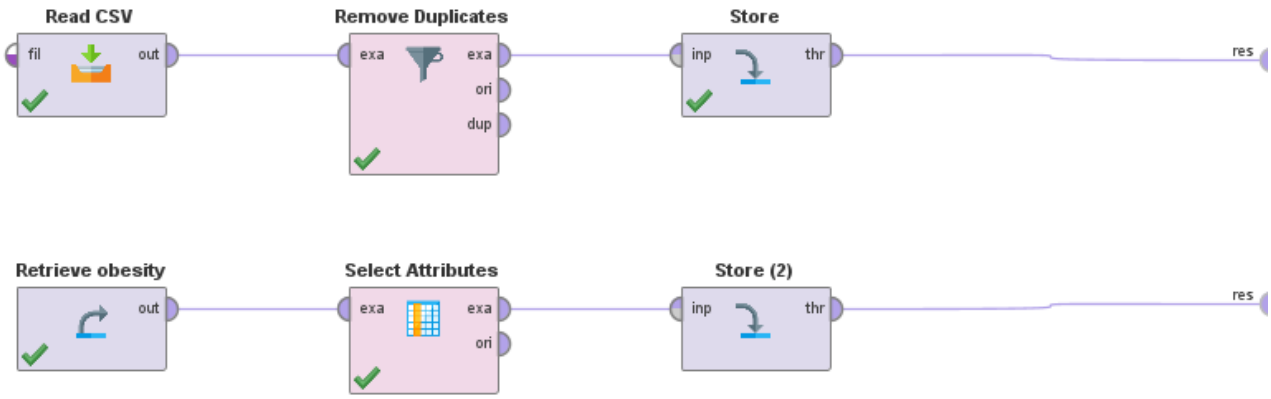
Tyto kroky zajistily, že data jsou připravena pro proces dolování, a že odstranění atributů s nízkou variabilitou zlepší přesnost a výkon následných modelů. Blokové schéma přípravy dat je vyobrazeno na obrázku 1.

4 Implementace úloh

Tato sekce popisuje implementaci jednotlivých úloh, a to úlohy klasifikace, klastrování a regrese. Pro každou z úloh je popsán výběr atributů, samotné dolování a následné vyhodnocení.

4.1 Úloha klasifikace — stanovení úrovně obezity

Tato úloha se zaměřuje na predikci různých úrovní obezity na základě cílového atributu *NObeyesdad*. Následující sekce popisuje jednotlivé dílčí kroky, a to od výběru atributů, přes parametry dolování, až po konečné vyhodnocení.



Obrázek 1: Blokové schéma přípravy dat.

4.1.1 Výběr atributů

Pro klasifikační úlohu byly vybrány veškeré dostupné atributy s výjimkou atributů *SMOKE* a *SCC*, které byly odstraněny na základě nízké variability (viz sekce 3). Výjimku tvoří také predikovaný atribut *NObeyesdad*.

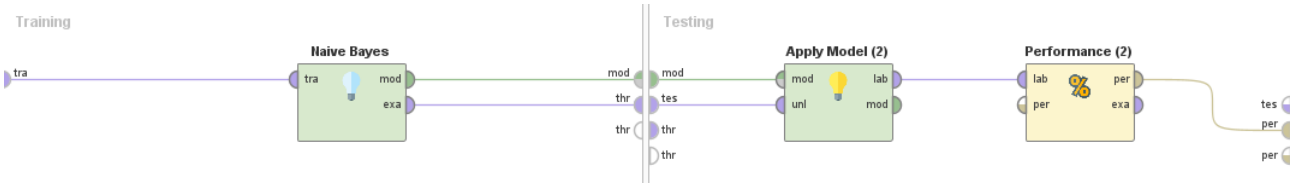
4.1.2 Dolování dat vybranými modely

Pro zvolenou dolovací úlohu bylo pro účely srovnání vybráno celkem 5 modelů, a to

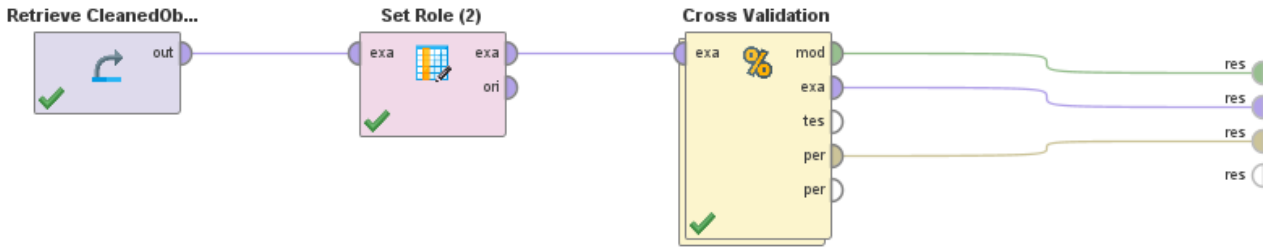
- Naive Bayes,
- Decision Tree,
- Deep Learning,
- Random Forest,
- Gradient Boosted Trees.

V našem průzkumu byly porovnány dva způsoby rozdělení datové sady, a to konkrétně bloky *Split Validation* a *Cross Validation*. Po srovnání dvou uvedených přístupů bylo vyhodnoceno, že technika křížové validace (*Cross Validation*) dosahuje lepších výsledků, a to řádově okolo 2.5% metriky *Accuracy*.

Veškeré modely využívají blokové schéma uvedené na obrázku 2. Dílčí schéma validačního bloku je uvedené na obrázku 3, a to konkrétně s využitím modelu *Naive Bayes*.



Obrázek 3: Blokové schéma samotného validačního bloku.



Obrázek 2: Blokové schéma úlohy klasifikace.

Naivní bayesovský klasifikátor (Naive Bayes)

Vyhodnocení modelu *Naive Bayes* je uvedeno na obrázku 4. Model dosahuje $64.88\% \pm 3.82\%$ v metrice *Accuracy*, 64.10% v průměrné hodnotě metriky *Precision*, v průměru 64.76% pro metriku *Recall* a *F1-score* 64.43%.

accuracy: 64.88% +/- 3.82% (micro average: 64.88%)

	true Normal_Weight	true Overweight_Level_I	true Overweight_Level...	true Obesity_Type_I	true Insufficient_Weight	true Obesity_Type_II	true Obesity_Type_III	class precision
pred. Normal_Weight	142	34	36	1	24	1	0	59.66%
pred. Overweight_Lev...	32	134	19	26	5	0	0	62.04%
pred. Overweight_Lev...	16	53	131	55	0	0	0	51.37%
pred. Obesity_Type_I	0	24	69	183	0	44	1	57.01%
pred. Insufficient_Wei...	80	10	0	0	237	0	0	72.48%
pred. Obesity_Type_II	0	0	9	54	0	202	0	76.23%
pred. Obesity_Type_III	12	21	24	32	0	50	323	69.91%
class recall	50.35%	48.55%	45.49%	52.14%	89.10%	68.01%	99.69%	

Obrázek 4: Vyhodnocení modelu Naive Bayes.

Rozhodovací strom (Decision Tree)

Vyhodnocení modelu *Decision Tree* je uvedeno na obrázku 5. Model dosahuje $87.05\% \pm 4.9\%$ v metrice *Accuracy*, 87.43% v průměrné hodnotě metriky *Precision*, v průměru 86.89% pro metriku *Recall* a *F1-score* 87.16%.

accuracy: 87.05% +/- 4.90% (micro average: 87.04%)

	true Normal_Weight	true Overweight_Level_I	true Overweight_Level...	true Obesity_Type_I	true Insufficient_Weight	true Obesity_Type_II	true Obesity_Type_III	class precision
pred. Normal_Weight	200	13	0	0	9	0	0	90.09%
pred. Overweight_Lev...	53	209	19	2	0	1	0	73.59%
pred. Overweight_Lev...	6	54	247	49	0	1	0	69.19%
pred. Obesity_Type_I	0	0	22	292	0	8	1	90.40%
pred. Insufficient_Wei...	23	0	0	0	257	0	0	91.79%
pred. Obesity_Type_II	0	0	0	8	0	286	0	97.28%
pred. Obesity_Type_III	0	0	0	0	0	1	323	99.69%
class recall	70.92%	75.72%	85.76%	83.19%	96.62%	96.30%	99.69%	

Obrázek 5: Vyhodnocení modelu Decision Tree.

Hluboké učení (Deep Learning)

Vyhodnocení modelu *Deep Learning* je uvedeno na obrázku 6. Model dosahuje 83.45% ± 7.44% v metrice *Accuracy*, 84.43% v průměrné hodnotě metriky *Precision*, v průměru 82.83% pro metriku *Recall* a *F1-score* 83.62%.

accuracy: 83.45% +/- 7.44% (micro average: 83.45%)

	true Normal_Weight	true Overweight_Level_I	true Overweight_Level...	true Obesity_Type_I	true Insufficient_Weight	true Obesity_Type_II	true Obesity_Type_III	class precision
pred. Normal_Weight	217	51	0	0	38	0	0	70.92%
pred. Overweight_Lev...	14	128	8	0	0	0	0	85.33%
pred. Overweight_Lev...	12	94	273	37	0	0	0	65.62%
pred. Obesity_Type_I	0	1	7	302	0	28	0	89.35%
pred. Insufficient_Wei...	39	2	0	0	228	0	0	84.76%
pred. Obesity_Type_II	0	0	0	12	0	268	1	95.37%
pred. Obesity_Type_III	0	0	0	0	0	1	323	99.69%
class recall	76.95%	46.38%	94.79%	86.04%	85.71%	90.24%	99.69%	

Obrázek 6: Vyhodnocení modelu Deep Learning.

Náhodný les (Random Forest)

Vyhodnocení modelu *Random Forest* je uvedeno na obrázku 7. Model dosahuje 90.60% ± 1.85% v metrice *Accuracy*, 90.59% v průměrné hodnotě metriky *Precision*, v průměru 90.23% pro metriku *Recall* a *F1-score* 90.41%.

accuracy: 90.60% +/- 1.85% (micro average: 90.60%)

	true Normal_Weight	true Overweight_Level_I	true Overweight_Level...	true Obesity_Type_I	true Insufficient_Weight	true Obesity_Type_II	true Obesity_Type_III	class precision
pred. Normal_Weight	240	28	10	2	15	0	0	81.36%
pred. Overweight_Lev...	13	202	5	1	0	0	0	91.40%
pred. Overweight_Lev...	9	45	251	15	0	0	0	78.44%
pred. Obesity_Type_I	0	1	21	328	0	3	0	92.92%
pred. Insufficient_Wei...	20	0	0	0	251	0	0	92.62%
pred. Obesity_Type_II	0	0	1	5	0	293	1	97.67%
pred. Obesity_Type_III	0	0	0	0	0	1	323	99.69%
class recall	85.11%	73.19%	87.15%	93.45%	94.36%	98.65%	99.69%	

Obrázek 7: Vyhodnocení modelu Random Forest.

Rozhodovací stromy zesílené gradientem (Gradient Boosted Trees)

Vyhodnocení modelu *Gradient Boosted Trees* je uvedeno na obrázku 8. Model dosahuje 89.59%± 2.45% v metrice *Accuracy*, 89.62% v průměrné hodnotě metriky *Precision*, v průměru 89.43% pro metriku *Recall* a *F1-score* 89.52%.

accuracy: 89.59% +/- 2.45% (micro average: 89.59%)

	true Normal_Weight	true Overweight_Level_I	true Overweight_Level...	true Obesity_Type_I	true Insufficient_Weight	true Obesity_Type_II	true Obesity_Type_III	class precision
pred. Normal_Weight	230	25	0	1	11	1	0	85.82%
pred. Overweight_Lev...	39	239	37	4	1	0	0	74.69%
pred. Overweight_Lev...	2	10	222	25	0	0	0	85.71%
pred. Obesity_Type_I	0	2	29	311	0	8	1	88.60%
pred. Insufficient_Wei...	11	0	0	0	254	0	0	95.85%
pred. Obesity_Type_II	0	0	0	10	0	288	0	96.64%
pred. Obesity_Type_III	0	0	0	0	0	0	323	100.00%
class recall	81.56%	86.59%	77.08%	88.60%	95.49%	96.97%	99.69%	

Obrázek 8: Vyhodnocení modelu Gradient Boosted Trees.

Celkové vyhodnocení jednotlivých modelů a veškerých změřených či vypočítaných metrik shrnuje tabulka 2.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
Naive Bayes	64.88 ± 3.82	64.10	64.76	64.43
Decision Tree	87.05 ± 4.9	87.43	86.89	87.16
Deep Learning	83.45 ± 7.44	84.43	82.83	83.62
Random Forest	$\mathbf{90.60 \pm 1.85}$	**90.59**	**90.23**	**90.41**
Gradient Boosted Trees	$\mathbf{89.59 \pm 2.45}$	**89.62**	**89.43**	**89.52**

Tabulka 2: Výkonové metriky různých modelů. Pro každý model je uvedena hodnota celkové přesnosti (*Accuracy*), dále pak průměr jednotlivých dílčích hodnot *Precision*, průměrná hodnota *Recall*, a nakonec hodnota *F1-score* vypočítaná s využitím průměrných hodnot *Precision* a *Recall*.

4.1.3 Vyhodnocení dolování

Na základě stanovených hodnot v tabulce 2 vyšel z testovanych modelů nejlépe model *Random Forest*, který má nejlepší výkon ve všech vyhodnocovaných parametrech (*Accuracy*, *Precision*, *Recall* a *F1-score*). Jako druhý skončil v našem umístění model *Gradient Boosted Trees* s lehce horším výkonem.

Naopak nejhůře skončil základní přístup *Naive Bayes* a přístup *Deep Learning*. To může být především způsobeno tím, že první z uvedených má příliš základní a jednoduchou charakteristiku vzhledem k řešenému problému. Výkon druhého modelu, *Deep Learning*, může být ovlivněn celkovým množstvím trénovacích dat, což způsobí zhoršenou generalizaci modelu.

4.2 Úloha klastrování — určení skupin v populaci

Tato úloha se zaměřuje na nalezení přirozeně se vyskytujících skupin v populaci na základě atributů souvisejících s obezitou. Následující sekce popisuje jednotlivé dílčí kroky, a to od výběru atributů, přes samotné dolování, až po výsledné vyhodnocení.

4.2.1 Výběr atributů

Opět jako pro úlohu klasifikace, blíže popsané v sekci 4.1, i nyní byly využity veškeré atributy s výjimkou atributů *SMOKE* a *SCC*, které nebyly použity na základě nízké variability (viz

sekce 3). Nyní ovšem využíváme i dříve predikovaný atribut *NObeyesdad*.

4.2.2 Dolování dat — metoda k-means

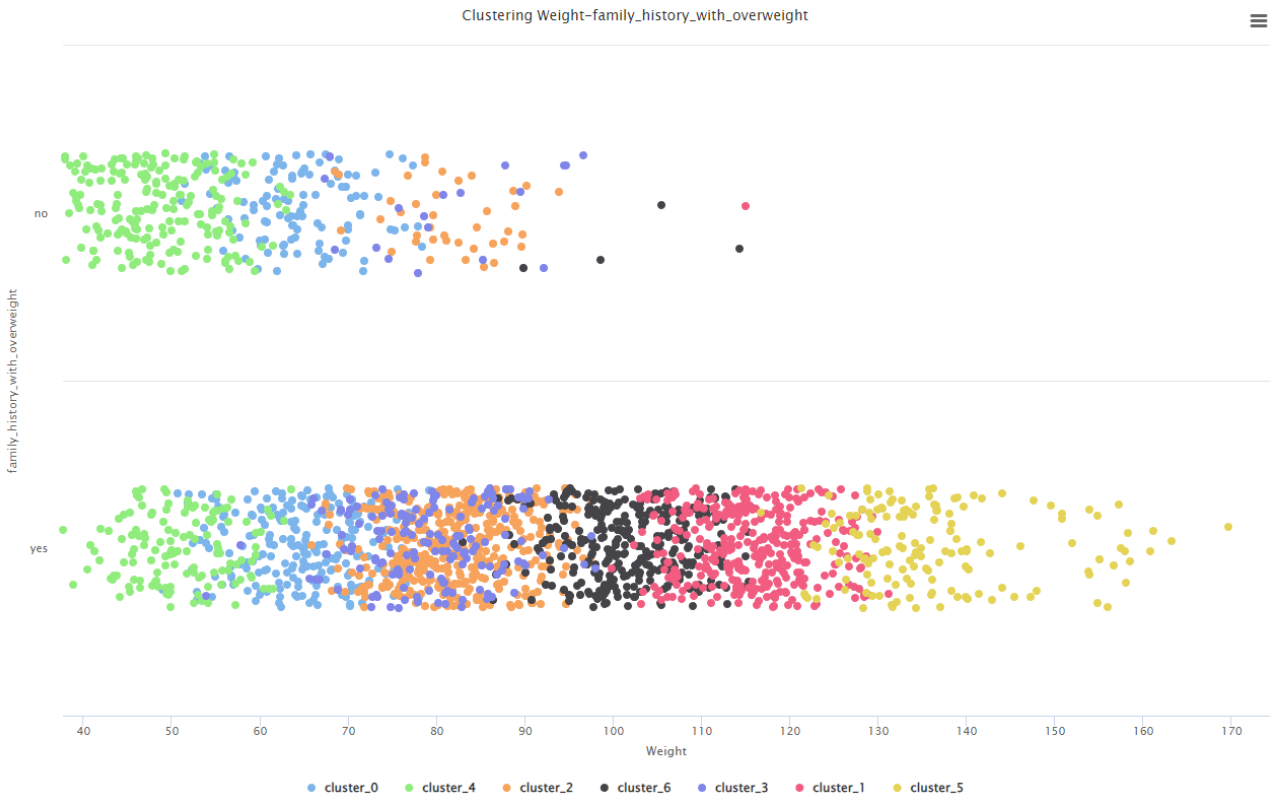
Pro úlohu klastrování byla zvolena jedna z nejrozšířenějších metod, a to metoda *k-means*. Pomocí této metody byly postupně vyhodnocovány jednotlivé vztahy mezi páry atributů a blíže určeny vzory pozorované v datech.

Obrázek 9: Blokové schéma úlohy klastrování.

Schéma zapojení pro úlohu regrese je uvedené na obrázku 9. Opět jako pro úlohu klasifikace samotnému dolování předchází předzpracování dat blíže popsané v sekci 3.

Pro veškeré porovnávané dvojice bylo uvažováno celkem 7 skupin, což koreluje s počtem skupin v atributu *NObeyesdad*.

Vztah váhy a přítomnosti člena rodiny s nadváhou
Následující graf, zachycený na obrázku 10, vyobrazuje vztah atributů *Weight* a *family_history_with_overweight* a skupiny získané pomocí zvoleného algoritmu. První z atributů zastupuje váhu a druhý zodpovídá na otázku, zda má nějaký člen rodiny dotazované osoby problémy s nadváhou.



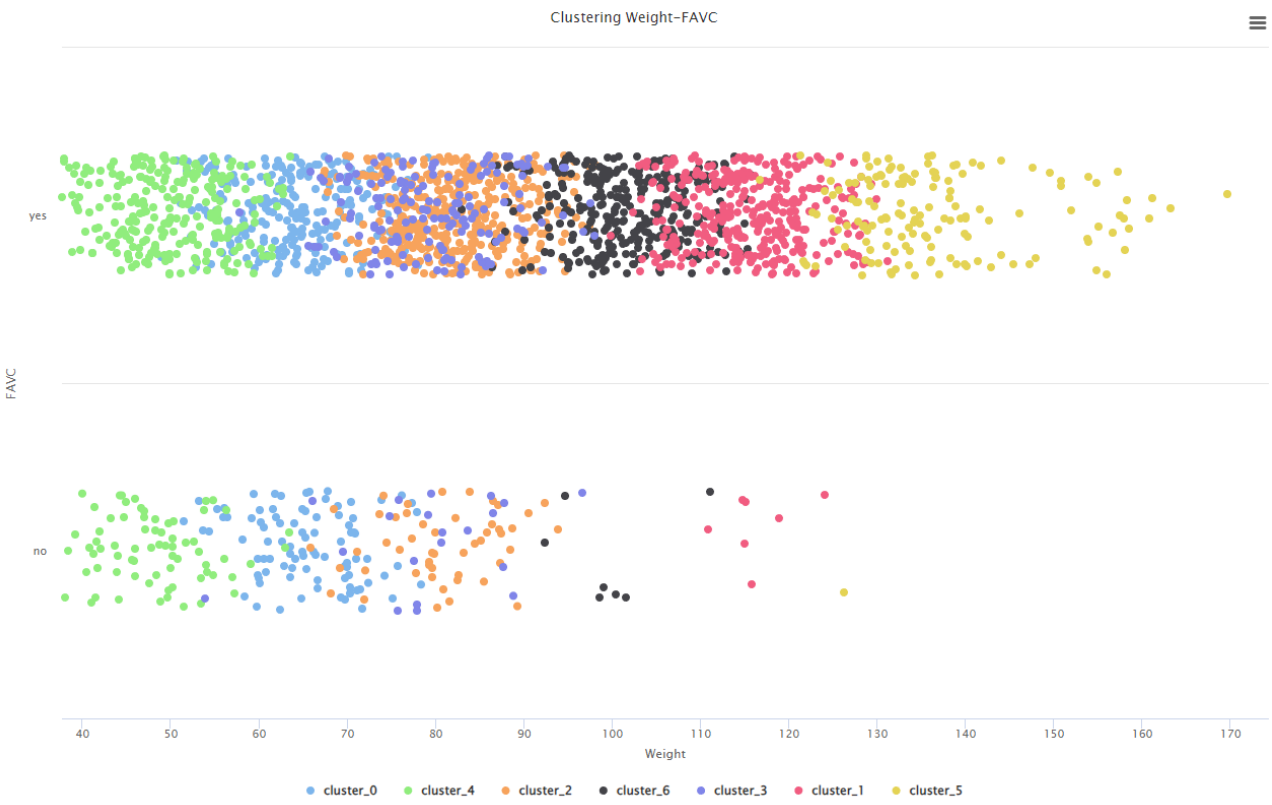
Obrázek 10: Vztah atributů *Váha* (*Weight*) a *Má člen rodiny problémy s nadváhou?* (*family_history_with_overweight*).

Z vyhodnoceného grafu je přímo patrné, že všechny osoby s váhou nad 120 kilogramů mají v rodině osobu s nadváhou. I když váha přímo nedefinuje obezitu, tak přesto s ní stále úzce souvisí. Na základě seskupení dat taky velmi málo osob má váhu v rozsahu $90 - 120$ kilogramů bez přítomnosti druhé obézní osoby v rodině.

Samozřejmě, že přítomnost druhé osoby v rodině nemusí mít přímý vliv na obezitu, ovšem může to inklinovat k určení špatných stravovacích návyků rodiny či vzájemnému ovlivnění.

Vztah váhy a konzumace vysoce kalorického jídla

Následující graf, zachycený na obrázku 11, vyobrazuje vztah atributů *Weight* a *FAVC* a skupiny získané pomocí zvoleného algoritmu. První z atributů zastupuje váhu a druhý zodpovídá na otázku, zda dotazovaná osoba konzumuje vysoce kalorická jídla.



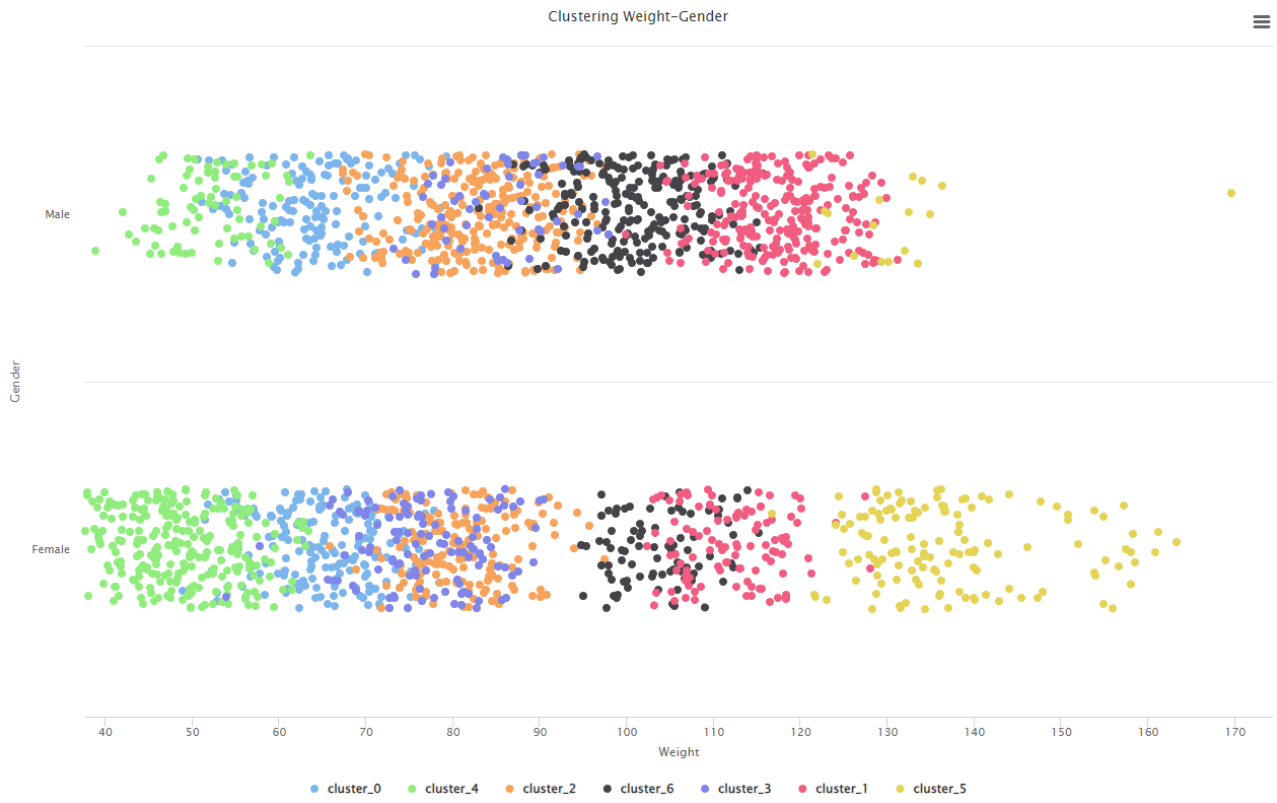
Obrázek 11: Vztah atributů *Váha* (*Weight*) a *Často konzumujete vysoce kalorické jídlo?* (*FAVC*).

Opět jako v předchozím případě i nyní je jasně patrné, že všechny osoby s váhou nad 130 kilogramů konzumují vysoce kalorická jídla. I když váha nedefinuje přímo obezitu, je s ní úzce svázána. Také skupiny s váhou v rozmezí $100 - 130$ výrazněji konzumují vysoce kalorická jídla.

I když i osoby s nízkou váhou konzumují vysoce kalorická jídla, v datech se nenachází osoby, které by měly stejné stravovací návyky a přitom netrpěly výraznou obezitou.

Vztah váhy a pohlaví

Následující graf, zachycený na obrázku 12, vyobrazuje vztah atributů *Weight* a *Gender* a skupiny získané pomocí zvoleného algoritmu. První z atributů zastupuje váhu a druhý pohlaví.



Obrázek 12: Vztah atributů *Váha* (*Weight*) a *Pohlaví* (*Gender*).

Z dat je patrné, že rozptyl váhy mužů je menší než rozptyl váhy žen. I když většina mužů dosahuje větší váhy, než ženy, přesto osoby s nejvyšší váhou a typy obezity jsou ženského pohlaví. Na druhou stranu ženy také dosahují nižší váhy než muži.

Otázkou zde může být, zda vyšší váha u žen souvisí s počtem dětí a tudíž s těhotenstvím a především s následnou zvýšenou váhou po porodu a návratem k původní váze před otěhotněním. K tomuto určení ovšem v současnosti nemáme dostupná potřebná data.

Vztah váhy a počtu hlavních jídel za den

Následující graf, zachycený na obrázku 13, vyobrazuje vztah atributů *Weight* a *NCP* a skupiny získané pomocí zvoleného algoritmu. První z atributů zastupuje váhu a druhý počet hlavních jídel za den u dotazované osoby.

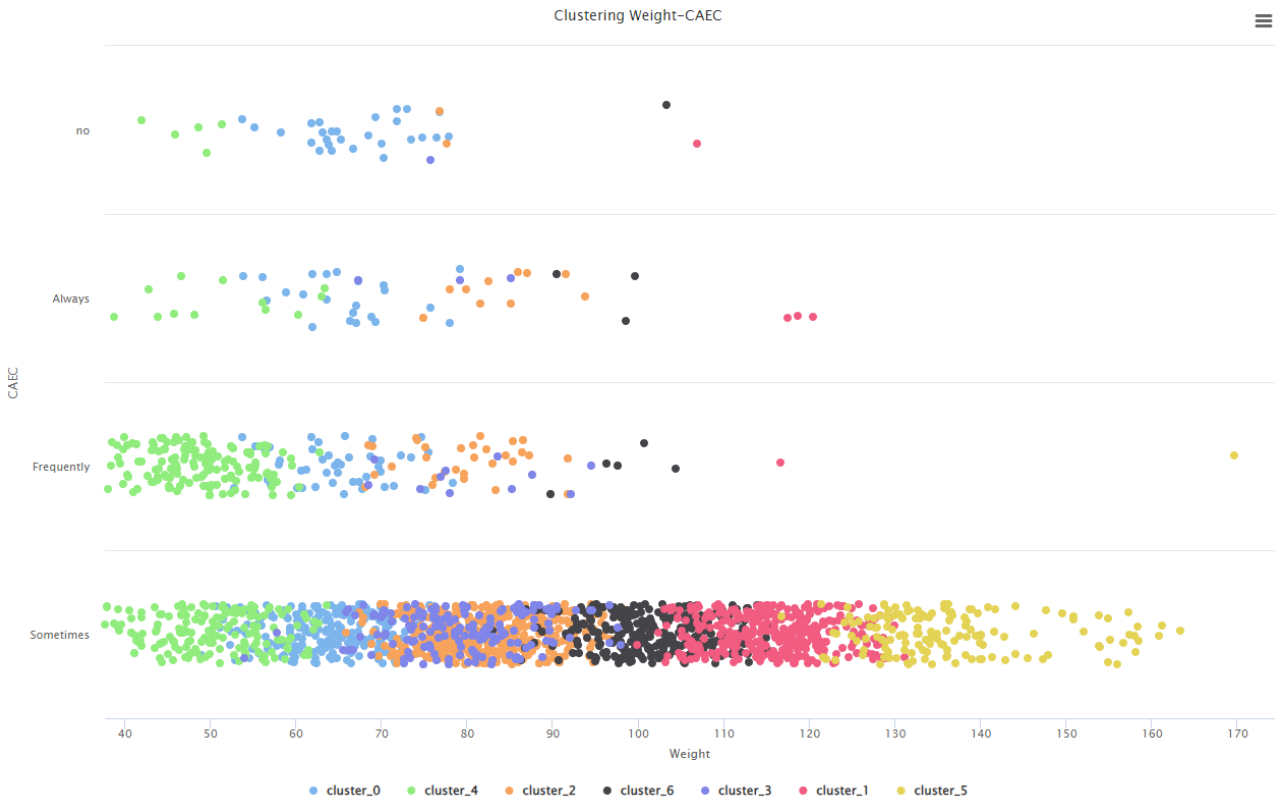
Obrázek 13: Vztah atributů *Váha* (*Weight*) a *Kolik hlavních jídel denně máte?* (*NCP*).

Prvním zjištěným faktem je, že většina osob má za den celkem 3 hlavní jídla, což je podle mnohých odborníků správný přístup ke stravování. Co ovšem by se dalo přirozeně očekávat je to, že s vyšším počtem hlavních jídel za den bude vzrůstat i váha osob jednotlivých skupin dle počtu jídel. Tomu ovšem tak není.

Na druhou stranu může taky s počtem hlavních jídel ovlivňovat váhu velikost porce vzhledem k váze osoby, nebo zda osoba jí i menší jídla mezi hlavními jídly a v jakém množství.

Vztah váhy a snězení jídla mezi hlavními jídly

Následující graf, zachycený na obrázku 14, vyobrazuje vztah atributů *Weight* a *CAEC* a skupiny získané pomocí zvoleného algoritmu. První z atributů zastupuje váhu a druhý reprezentuje, zda dotazované osoby jedí nějaké jídlo mezi hlavními jídly.



Obrázek 14: Vztah atributů *Váha* (*Weight*) a *Jíte nějaké jídlo mezi hlavními jídly?* (*CAEC*).

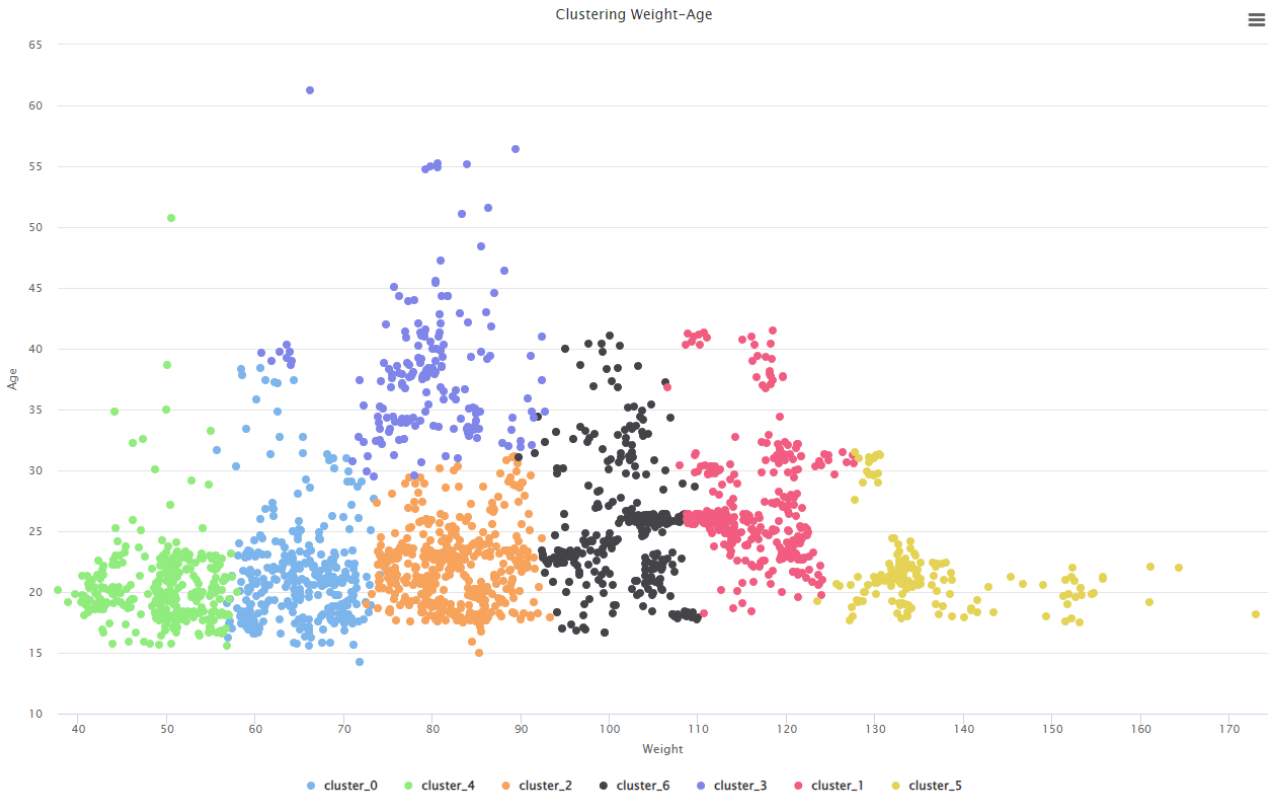
Na základě předchozí domněnky lze vyvrátit, že by s vyšším počtem snězených menších jídel mezi hlavními jídly vzrůstala váha dotazovaných osob. Naopak většina dotazovaných osob uvedla, že někdy sní menší jídlo mezi hlavními jídly. Ovšem mezi těmy osobami, které uvedly, že často nebo vždy, se nevyskytují osoby s vyšší váhou.

Nakonec osoby, které nejedí žádné jídlo mezi hlavními jídly, zastupují nejmenší část dotazované skupiny a jejich váha je nižší.

Uvedené fakty ovšem nevyvracejí, že kromě samotného počtu jídel může výrazně ovlivňovat váhu také velikost porce vzhledem k váze osoby a její fyzické aktivitě. Pro další vyhodnocení ovšem nejsou v datové sadě dostupné potřebné informace.

Vztah váhy a věku

Následující graf, zachycený na obrázku 15, vyobrazuje vztah atributů *Weight* a *Age* a skupiny získané pomocí zvoleného algoritmu. První z atributů zastupuje váhu a druhý věk.



Obrázek 15: Vztah atributů *Váha* (*Weight*) a *Věk* (*Age*).

Z dat je patrné, že osoby s vyšším věkem dosahují vyšší váhy. To může být také způsobeno snížením pohybové aktivity starších osob i v důsledku zdravotních omezení. Na druhou stranu osoby s nejvyšší váhou nepatří mezi nejstarší dotazované osoby, ale mezi ty nejmladší, přičemž se jedná o zcela alarmující fakt a může se jednat o problém v celé populaci a aktuálním přístupu vytváření zdravých stravovacích či pohybových návyků u mladých lidí.

4.2.3 Vyhodnocení dolování

Z vyhodnocení dolování byly zjištěny následující pozoruhodné fakty:

- Osoby s vyšší váhou, a tudíž pravděpodobně nadváhou či obezitou, mají v rodině alespoň jednu osobu, která má nadváhu.
- Vyšší váha úzce souvisí s konzumací vysoce kalorického jídla.
- Ženy dosahují jak nižší, tak ovšem i vyšší váhy než muži.
- Počet hlavních jídel za den ani menších jídel mezi hlavními jídly není hlavním faktorem způsobujícím vyšší tělesnou hmotnost.
- Osoby s nejvyšší hmotností jsou ve věku v rozsahu $15 - 20$ let. Osoby s vyšší hmotností se výrazně nacházejí ve věku $20 - 35$ let.

Souhrnně nejrizikovější osobou může být osoba ženského pohlaví ve věku v rozsahu $15-20$, která konzumuje vysoce kalorická jídla a má v rodině osobu s nadváhou.

Zároveň bychom chtěli uvést, že zvýšená váha sice velmi souvisí s nadváhou a obezitou, ale je nutné si uvědomit, že na určení obezity je nutné stanovit podíl tuku v lidském těle.

4.3 Úloha regrese — predikce váhy

Tato úloha si klade za cíl předpovědět váhu osoby na základě dostupných informací, jako je věk, výška, rodinná anamnéza, fyzická aktivita, a dalších faktorů. Výstupem této části bude vytvořený regresní model, který lze využít k odhadu budoucí váhy osoby v horizontu několika let za předpokladu udržení stejných návyků a životního stylu. Následující sekce popisuje jednotlivé dílčí kroky, a to od výběru atributů, přes samotné vytvoření modelu, až po výsledné vyhodnocení.

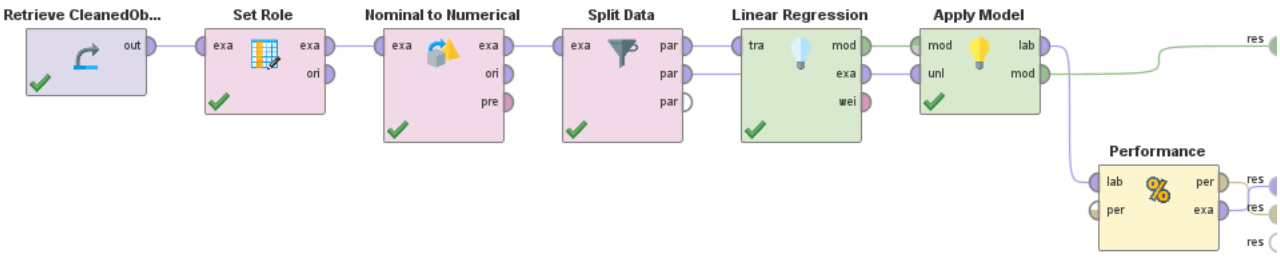
4.3.1 Výběr atributů

Opět jako v předchozích dvou úlohách i nyní byly využity veškeré atributy s výjimkou atributů *SMOKE* a *SCC*, které nebyly použity na základě nízké variability (viz sekce 3). Nyní ovšem využíváme i dříve predikovaný atribut *NObeyesdad* oproti úloze klasifikace. Predikovanou hodnotou bude atribut určující váhu — *Weight*.

Nutno uvést, že pro úlohu regrese je zapotřebí převést veškeré nominální hodnoty na numerické. Tento krok je také zachycen na obrázku 16 zobrazující celkové blokové schéma.

4.3.2 Vytvoření vybraných modelů

Pro úlohu regrese jsme pro srovnání vybrali 2 modely, a to lineární regresní model (*Linear Regression*) a neuronovou síť (*Neural Net*).



Obrázek 16: Blokové schéma úlohy regrese.

Schéma zapojení regresní úlohy (konkrétně pro model lineární regrese) zachycuje obrázek 16. Opět jako pro předchozí úlohy samotnému vytváření modelu předchází předzpracování dat blíže popsané v sekci 3.

Pro oba modely bylo zvoleno vyobrazení takové, kdy je zachycen vztah mezi skutečnou váhou (*Weight*) a skutečnou váhou současně s predikovanou váhou (*prediction(Weight)*). Zároveň kvalita modelů je vyhodnocena pomocí následujících zvolených metrik:

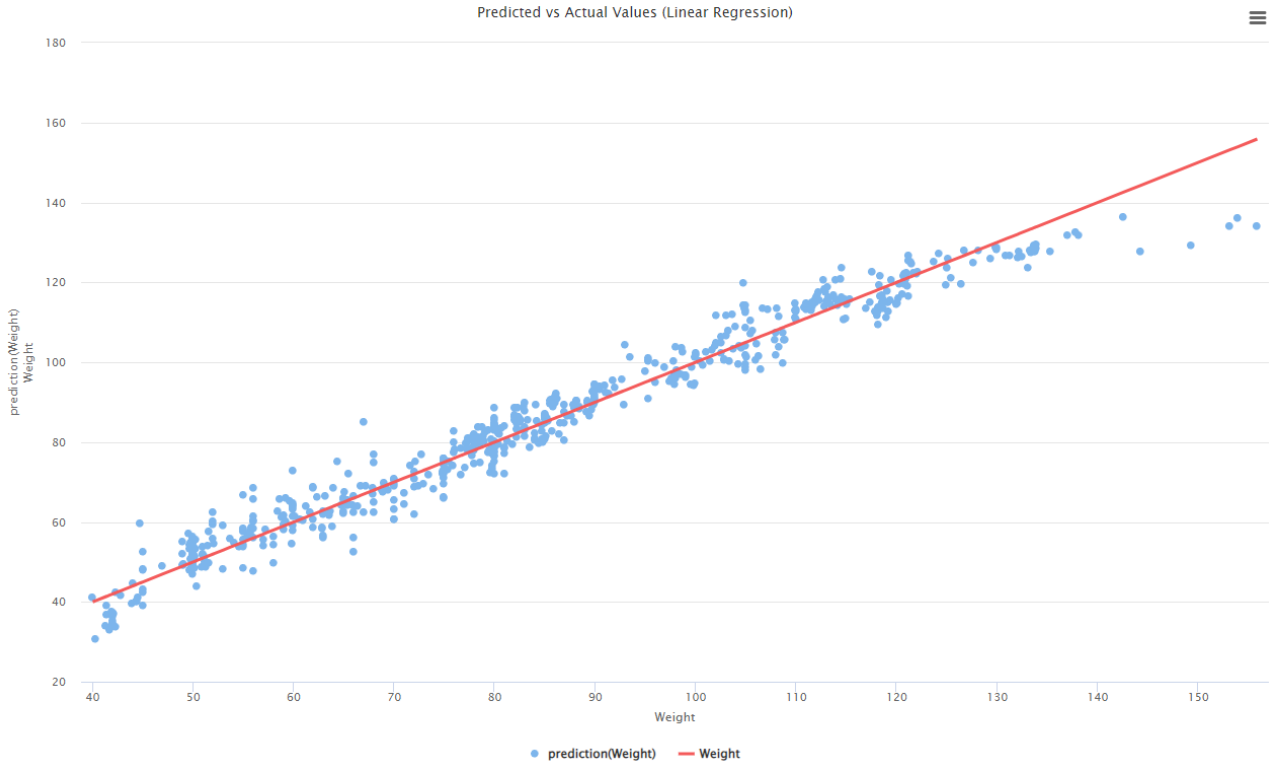
- Root Mean Squared Error (RMSE),
- Absolute Error,

- Relative Error,
- Squared Error,
- Squared Correlation,
- Prediction Average.

Dále jsou popsány jednotlivé modely.

Lineární regrese (Linear Regression)

Graf zachycující predikce lineární regrese vzhledem ke skutečné váze je vyobrazen na obrázku 17. Hodnoty zvolených metrik obsahuje tabulka 3.



Obrázek 17: Graf znázorňující vztah mezi skutečnou a predikovanou váhou pro model lineární regrese.

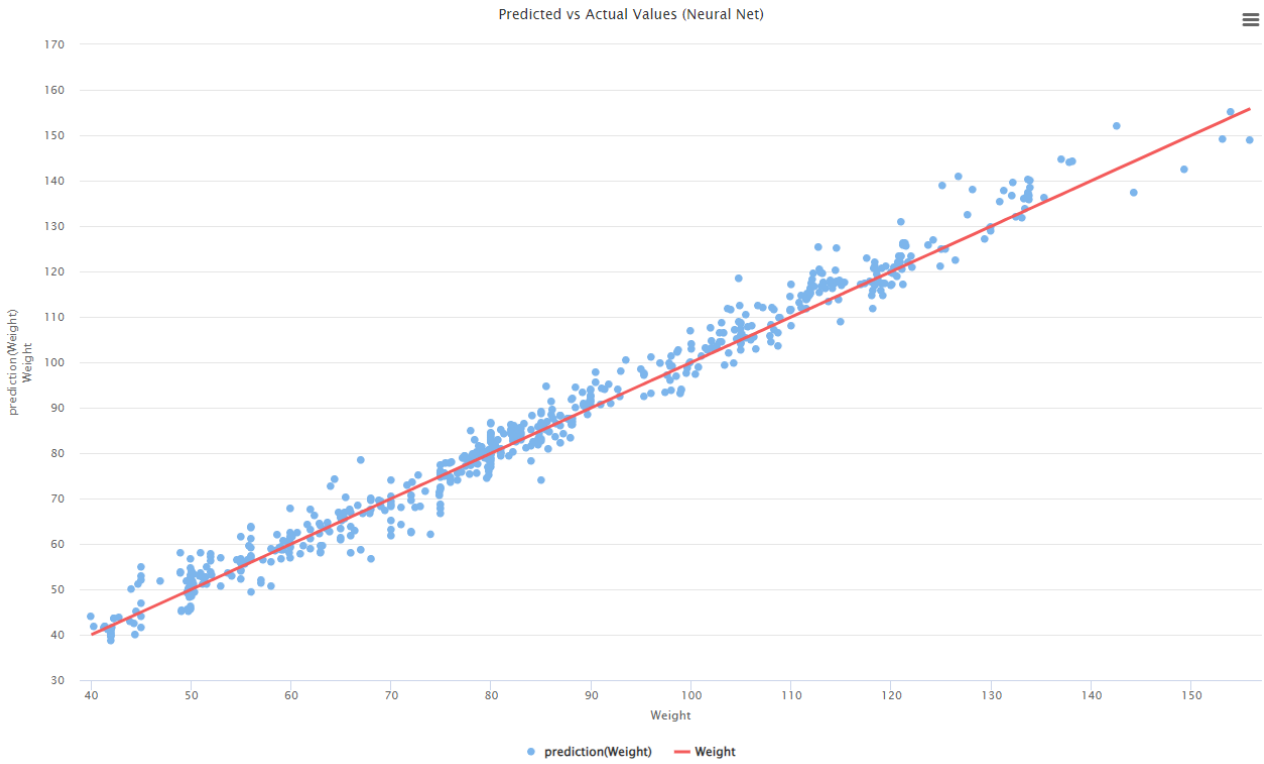
Metric	Value
Root Mean Squared Error (RMSE)	4.633 ± 0.000
Absolute Error	3.563 ± 2.962
Relative Error	4.60% ± 4.33%
Squared Error	21.469 ± 43.353
Squared Correlation	0.967
Prediction Average	85.053 ± 25.600

Tabulka 3: Metriky chyb modelu lineární regrese.

Vysoká hodnota metriky *Squared Correlation* (0, 967) naznačuje, že model vysvětluje většinu variability váhy, což je velmi pozitivní. Na druhou stranu variabilita v chybách (zejména vysoká směrodatná odchylka u kvadratické chyby (*Squared Error*) a absolutní chyby (*Absolute Error*)) naznačuje, že model nemusí být spolehlivý ve všech případech. Některé predikce mohou mít výrazně větší chyby než jiné, což by mohlo ukazovat na to, že model má problémy s některými vzory nebo odlehlými hodnotami v datech.

Neuronová síť (Neural Net)

Graf zachycující predikce neuronové sítě vzhledem ke skutečné váze je vyobrazen na obrázku 18. Hodnoty zvolených metrik obsahuje tabulka 4.



Obrázek 18: Graf znázorňující vztah mezi skutečnou a predikovanou váhou pro model neuronové sítě.

Metric	Value
Root Mean Squared Error (RMSE)	3.688 ± 0.000
Absolute Error	2.724 ± 2.486
Relative Error	3.44% ± 3.52%
Squared Error	13.603 ± 25.234
Squared Correlation	0.981
Prediction Average	88.054 ± 25.626

Tabulka 4: Metriky chyb modelu neuronové sítě.

Model neuronové sítě překonává lineární regresi téměř ve všech metrikách. Má nižší hodnotu metriky *Root Mean Squared Error (RMSE)*, absolutní chybu (*Absolute Error*) a relativní chybu (*Relative Error*), což naznačuje lepší přesnost predikcí. Mírně vyšší hodnota metriky *Squared Correlation* ukazuje, že neuronová síť lépe zachycuje vzory v datech než lineární regrese.

Model však stále vykazuje určitou variabilitu v chybách, jak je vidět na směrodatných odchylkách, což naznačuje, že může mít problémy s konkrétními případy nebo odlehlými hodnotami. Celkově model neuronové sítě dosahuje lepších výsledků.

4.3.3 Vyhodnocení modelů

Při porovnání lineární regrese a neuronové sítě je zřejmé, že neuronová síť poskytuje lepší výsledky ve většině sledovaných metrik. Dosahuje nižší hodnoty metriky *Root Mean Squared Error (RMSE)* oproti lineární regresi, což naznačuje přesnější predikce.

Neuronová síť má také nižší absolutní chybu (*Absolute Error*) a relativní chybu (*Relative Error*) ve srovnání s lineární regresí. Dále má neuronová síť nižší průměrnou kvadratickou chybu (*Squared Error*), čímž méně penalizuje větší chyby. Oba modely dosahují vysoké hodnoty metriky *Squared Correlation*, přičemž neuronová síť lépe zachycuje variabilitu v datech než lineární regrese.

Na základě uvedených odůvodnění by byl pro zvolenou regresní úlohu zvolen model neuronové sítě (*Neural Net*).

5 Závěr

V rámci této práce jsme se zabývali analýzou datasetu zaměřeného na příčiny obezity. Náš přístup zahrnoval tři hlavní úlohy: klasifikaci, klastrování a regresi.

Pro klasifikační úlohu jsme se zaměřili na predikci úrovně obezity. Testovali jsme pět různých modelů: *Naive Bayes*, *Decision Tree*, *Deep Learning*, *Random Forest* a *Gradient Boosted Trees*. Nejlepších výsledků dosáhl model *Random Forest* s přesností $90,60\%$ a F1-skóre $90,41\%$. Těsně za ním následoval model *Gradient Boosted Trees* s přesností $89,59\%$ a hodnotou *F1-score* $89,52\%$.

V úloze klastrování jsme použili metodu *k-means* k identifikaci přirozených skupin v populaci. Tato analýza odhalila několik zajímavých vztahů mezi různými atributy. Zjistili jsme například, že osoby s vyšší váhou často mají v rodině někoho s nadváhou a konzumují více vysoce kalorických jídel. Zajímavým zjištěním bylo také to, že osoby s nejvyšší hmotností se nacházejí ve věkové skupině $15 - 20$ let.

Pro regresní úlohu jsme se zaměřili na predikci váhy jedince. Porovnávali jsme dva modely: lineární regresi a neuronovou síť. Neuronová síť dosáhla lepších výsledků ve všech sledovaných metrikách, včetně nižší hodnoty metriky *Root Mean Squared Error (RMSE)* a vyšší hodnoty metriky *Squared Correlation*.

Celkově tato studie poskytla cenné poznatky o faktorech souvisejících s obezitou. Výsledky naznačují, že ke vzniku obezity přispívá kombinace genetických předpokladů, stravovacích návyků a životního stylu. Zvláště znepokojivé je zjištění týkající se vysoké míry obezity u mladých lidí, což podtrhuje potřebu cílených intervencí v této věkové skupině.

Tyto výsledky mohou být využity pro vývoj efektivnějších strategií prevence a léčby obezity, stejně jako pro personalizované zdravotní programy. Pro budoucí výzkum by bylo přínosné rozšířit dataset o další relevantní atributy, jako jsou například detailnější informace o stravovacích návycích nebo míra fyzické aktivity.

Reference

[1] Mehrparvar, F.: Obesity Levels. Online source: https://www.kaggle.com/datasets/fatemehmehrparvar/obesity-levels/data, 2024, accessed: 2024-09-25.

[2] Walsh, T. P.; Arnold, J. B.; Evans, A. M.; aj.: The association between body fat and musculoskeletal pain: a systematic review and meta-analysis. *BMC Musculoskeletal Disorders*, 2018: str. 233, ISSN 1471-2474, doi:10.1186/s12891-018-2137-0.
Dostupné z: https://doi.org/10.1186/s12891-018-2137-0