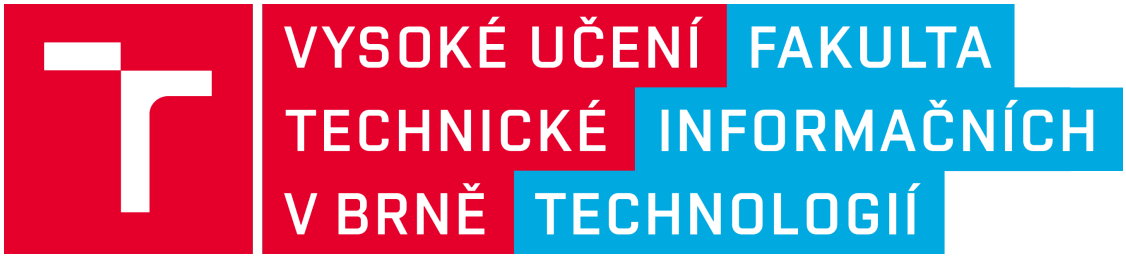

Získávání znalostí z databází

Databáze z výzkumu příčin obezity

Zadání

David Chocholatý (xchoch09)
Martin Baláž (xbalaz15)

Brno, 25. září 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.

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