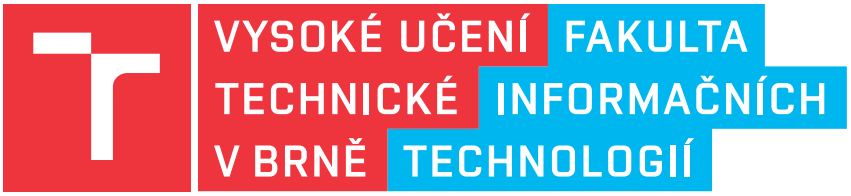

Projekt 2. část - dolovací data

Varianta 3: COVID-19

Ukládání a příprava dat

Šimon Galba, Bc. (xgalba03)
Gladiš Damián, Bc. (xgladi00)
Jeřábek František, Bc. (xjerab25)

16. prosince 2021

Obsah

1 Zadanie

Hledání skupin podobných měst z hlediska vývoje covidu a věkového složení obyvatel.
Atributy: počet nakažených za poslední 4 čtvrtletí, počet očkovaných za poslední 4 čtvrtletí, počet obyvatel ve věkové skupině 0..14 let, počet obyvatel ve věkové skupině 15 - 59, počet obyvatel nad 59 let.
Pro potřeby projektu vyberte libovolně 50 měst, pro které najdete potřebné hodnoty (můžete např. využít nějaký žebříček 50 nejlidnatějších měst v ČR).

2 Vytvorené CSV

Riadky:

- Ako objekt bol zvolený okres - každý riadok csv súboru zodpovedá jednému okresu v ČR. Počet očkovaných ľudí v jednotlivých mestách nie je verejne dostupný údaj a rozšírenie mesta na okres spôsobí minimálnu zmenu vo výsledku.

Stĺpce: Každý stĺpec tabuľky odpovedá jednému atribútu objektu

1. **LAU1** - kód okresu
2. **vaccination_count** - počet očkovaných ľudí v danom okrese
3. **infected_count** - počet infikovaných ľudí v danom okrese
4. **0-14** - počet obyvateľov podľa vo veku 0 až 14 rokov
5. **15-59** - počet obyvateľov podľa vo veku 15 až 59 rokov
6. **60+** - počet obyvateľov podľa vo veku 60 a vyššom
7. **název** - názov okresu
8. **infected_percentage** - percentuálny počet nakazených ľudí vzhľadom na počet obyvateľov okresu
9. **vaccinated_percentage** - percentuálny počet očkovaných ľudí vzhľadom na počet obyvateľov okresu
10. **kids_percentage** - percentuálny počet detí vo veku 0-14 rokov vzhľadom na počet obyvateľov okresu

3 Upravené CSV

3.1 Normalizácia

Pre potreby projektu bola použitá min-max normalizácia tak ako bola vysvetľovaná na prednáškach. Normalizované boli dáta udávajúce počty infikovaných v danom okrese, teda atribút infected_count.

- Min-max normalizace: na $[new\_min_A, new\_max_A]$

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Obrázek 1: Vzorec pre min-max normalizáciu.

3.2 Diskretizácia

Pre diskretizáciu bol zvolený postup diskretizácie do šírky, teda delenie dát na rovnako veľké intervaly použitím quantile-based rozdelenia. Pre potreby trénovania klasifikátoru boli normalizované atribúty **infected_percentage** a **vaccinated_percentage**. Na diskretizáciu bola využitá funkcia **qcut** z knižnice pandas.

3.3 Odľahlé hodnoty

3.3.1 Detekcia

Odľahlé hodnoty **outlier** boli detekované pomocou **z-score**. Odľahlé hodnoty boli hľadané na atribúte percentuálneho počtu detí, keďže tieto dáta môžu ovplyvniť prípadné výsledky dolovacieho algoritmu vzhľadom na naočkovanie detí. Ako parameter bola zvolená hodnota 3 určujúca signifikantný rozdiel dát a na výpočet použitá funkcia **stats** z knižnice **scipy**.

3.3.2 Náhrada

V tomto atribúte boli pre náš dataset odhalené dve odľahlé hodnoty. Tieto hodnoty boli potom podľa kvantilu 95% a 5% upravené na hodnotu bližšieho kvantilu a znovu vložené do datasetu.

3.4 Výsledné CSV:

Riadky:

- Okres ako objekt zostáva nezmenený

Stĺpce: Každý stĺpec tabuľky odpovedá jednému atribútu objektu. Stĺpce 1 až 7 zostávajú nezmenené.

11. **outlier** - odĺahlé hodnoty perpetuálneho počtu detí
12. **infection_category** - kategorizovaná hodnota počtu infikovaných v okrese
13. **vaccination_category** - kategorizovaná hodnota počtu očkovaných v okrese
14. **normalized_infected_count** - normalizovaná hodnota počtu infikovaných v okrese