Univerza v Ljubljani

Fakulteta za računalništvo in informatiko

Merjenje sreče

Seminarska naloga Tehnologija upravljanja podatkov 2020/21

Študenta: Martin Štrekelj in Simon Babnik

Ljubljana, december 2020

# Kazalo

# 1. Uvod

Najina tema za seminarko nalogo je uporaba MindsDB orodja na večji podatkovni zbirki.

# 2. Razumevanje problema in razumevanje podatkov

Najina podatkovna zbirka je osnovana na informacijah *the Gallup World Survey* imenovanih "The World Happiness Report" iz let 2015-2020. V poročilu so ocene nacionalnih povprečij sreče ljudi. Ocene sreče so po skali *Centrilove lestvice (eng. Cantril ladder)* od 0, ki predstavlja dno lestvice in največje nezadovoljstvo oziroma nesrečo, do 10, ki predstavlja največjo razpoložljivo srečo.

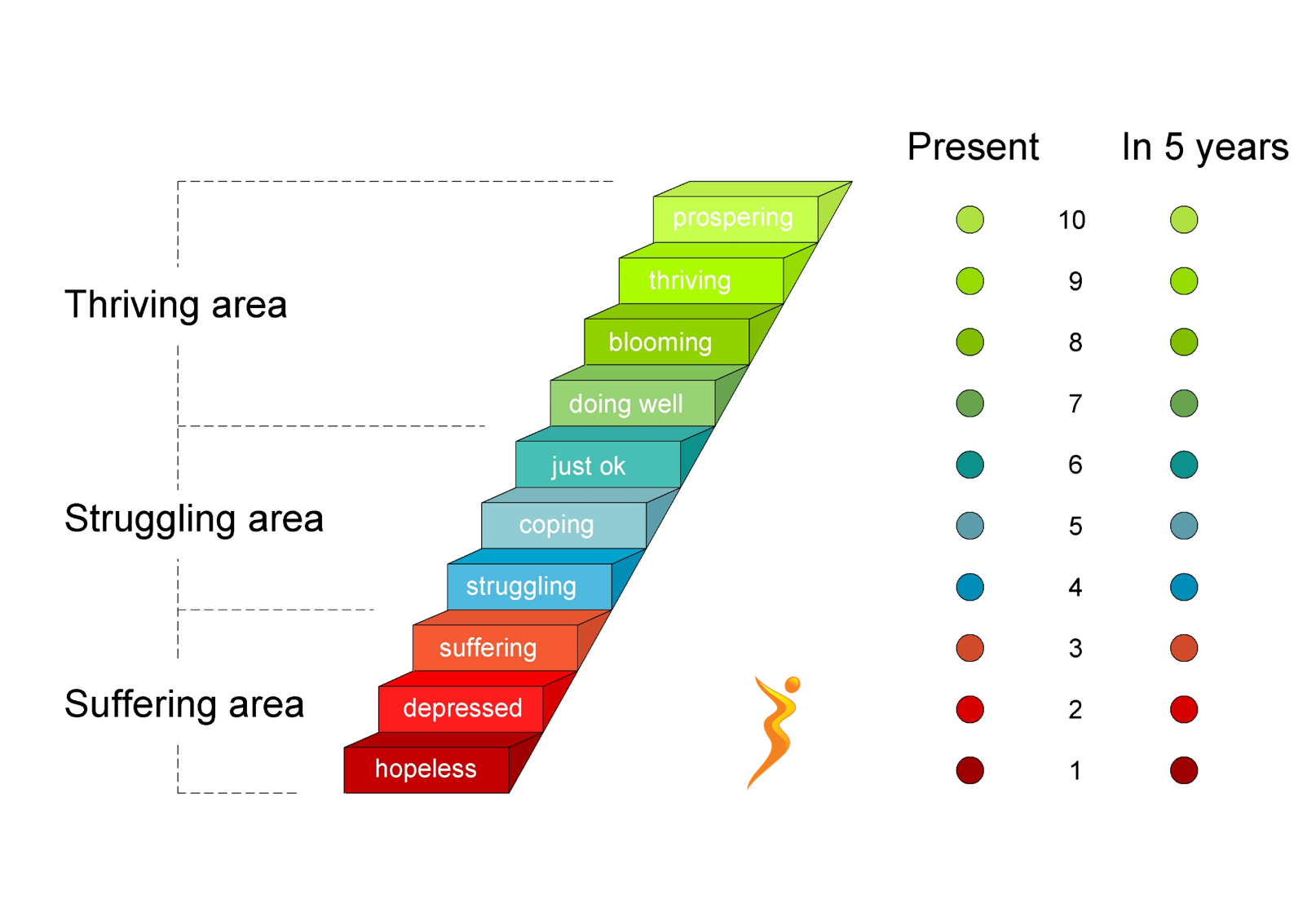


Figure 1: Centrilova lestvica (vir: https://innobatics.com/cantril-ladder/)

Točkovanje je na podlagi 6 parametrov, in sicer:

- **ekonomija** oziroma bruto domači proizvod na glavo

- nivo **zdravja** ter pričakovana dolžina zdravega življenja,

- pomen **družine**, socialna podpora in standarda življenja,

- nivo **svobode** do lastnih odločitev,

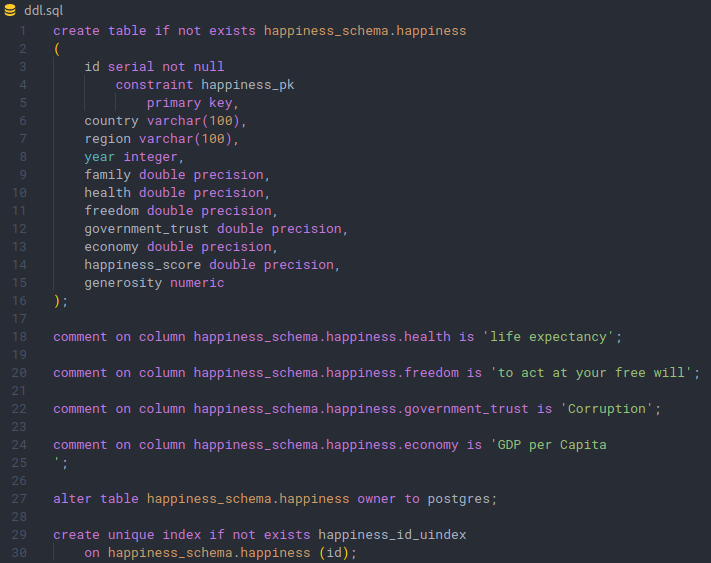
- nivo **dobrodelnosti** v državi,

- nivo **zaupanja v oblast** oziroma v percepcija korupcije

Ker so standardi in vrednosti po državah, regijah ali kontinentih različni so avtorji raziskave utežili vrednosti parametrov na podlagi primerjave z distopijo. Če je parameter ocenjen z 0 je primerljiv z distopijo, torej absolutno dno lestvice. Nasprotno je največja vrednost utopija, ali maksimalno najboljša vrednost v posameznem segmentu. Maksimalna vrednost se je skozi leta raziskav spremenijala, kar sva upoštevala z normalizacijo parametra in tako normalizirala vrednosti med 0 in 1. Konkretno, če je paramater X označen z vrednostjo 0.5 to pomeni, da je približno v sredini v primerjavi med utopijo in distopijo.

# 3. Priprava podatkov

Podatke sva pridobila na spletni strani Kaggle, ki je ponujal že pripravljene podatke v csv formatu ločeno po letih raziskave (2015 – 2020). Čeprav bi bilo to že dovolj, da bi jih uvozila v MindsDB, sva se odločila, da prej podatke prečistiva, pogledava in vstaviva v lokalno PostgreSQL bazo. Opravila sva kratko analizo atributov in stolpcev csv formata in oblikovala DDL (data definition language) stavek, ki je ustrezal najinim potrebam. Odločila sva se, da bo zadostovala ena tabela, saj so podatki že normalizirani do tretje normalne oblike.

Figure 2: Kreiranje sheme za vnos podatkov (vir: lasten)

## 3.1 Vnos podatkov s python skripto

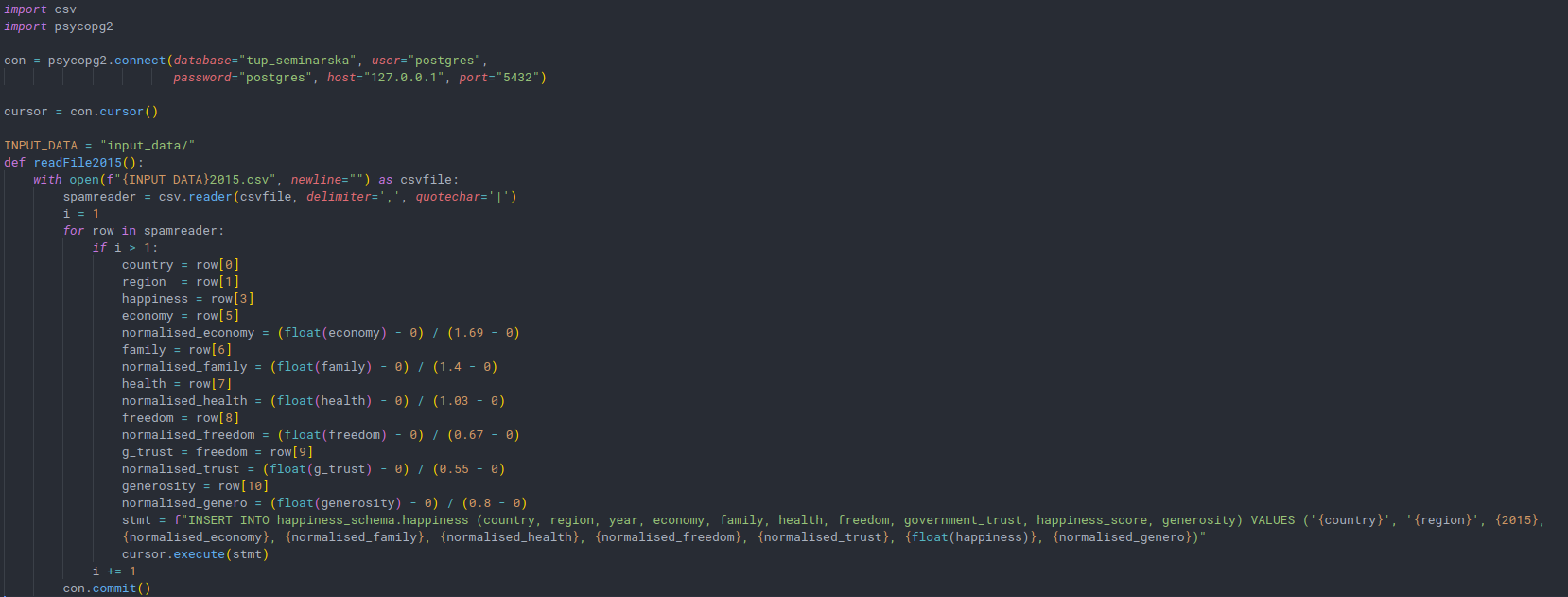
Nato sva z uporabo python skripte prečistila, normalizirala in vstavila podatke v prej kreirano podatkovno bazo oziroma tabelo. Za vnos podatkov sva uporabila dve knjižnici, in sicer **csv** za branje csv datotek ter **psycopg2**, ki opravlja vlogo pyodbc povezovalca med aplikacijsko in podatkovnim nivojem oziroma vnos podatkov mimo sistema za upravljanje z podatki. Ker so bili podatki različno strukturirani sva za vsako poročilo napisala funkcijo, ki je pravilno prebrala podatke, jih normalizirala (nastavila na interval med 0 in 1) ter nato pravilno vstavila v tabelo.

Figure 3: uvozi, povezava z bazo, funkcija za branje poročila 2015 (vir: lasten)

Glava programa je na koncu izgledala sledeče.

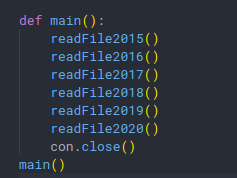


Figure 4: Glava programa za vnos podatkov v podatkovno bazo (vir: lasten)

## 3.2 Ročno pregledovanje podatkov

Ko so bili podatki vnešeni v podatkovno bazo, sva s pomočjo ročnih poizvedb nad bazo iskala anomalije v podatkih, ki so se morebito zgodile med branjev csv datoteke in vnosom v podatkovno bazo.

Figure 5: poizvedba za iskanje anomalij v podatkovni bazi (vir: lasten)

Na najino srečo so bile csv datoteke dobro pripravljene in tako nama je izstopal samo en vnos, ki sva ga ročno popravila na primerne vrednosti.



Figure 6: Popravljanje anomalij (vir: lasten)

S tem korakom je bila najina podatkovna baza pripravljena na modeliranje.

Celotna skripta, vhodni podatki in sql datoteke so na voljo preko *Githuba* na povezavi: povezava.

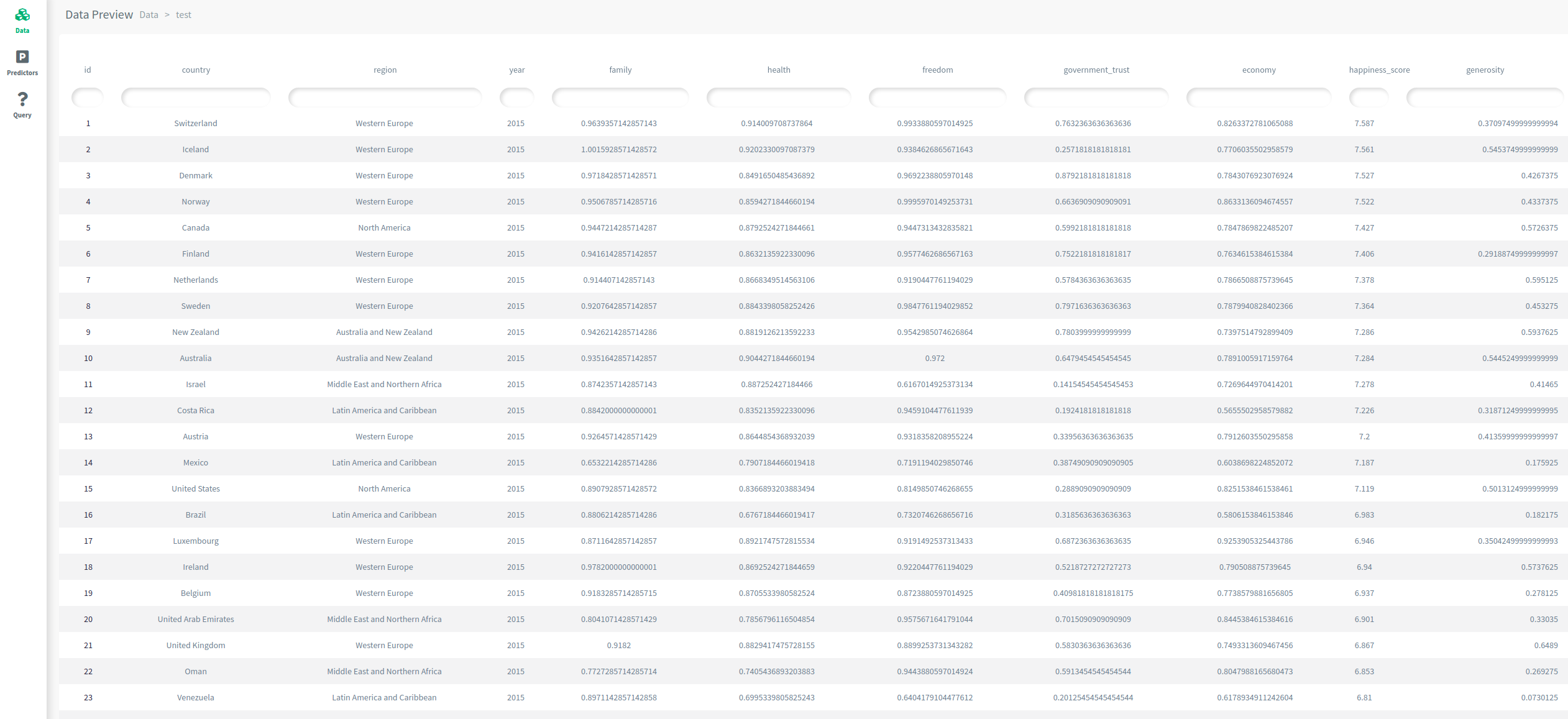
# 4. Modeliranje

Cilj modeliranja je bil kreirati model, ki bi deloval napovedovalno oziroma meril srečo na podlagi danih parametrov. Za doseg tega sva model učila nadzorovano. Model sva učila neposredno na podatkovni plasti oziroma na podatkovni bazi, kar je tudi prednost uporabe MindsDB

Uporabila sva orodja *MindsDB* in *MindsDB scout*, ki je grafični vmesnik za delanje z MindsDB.

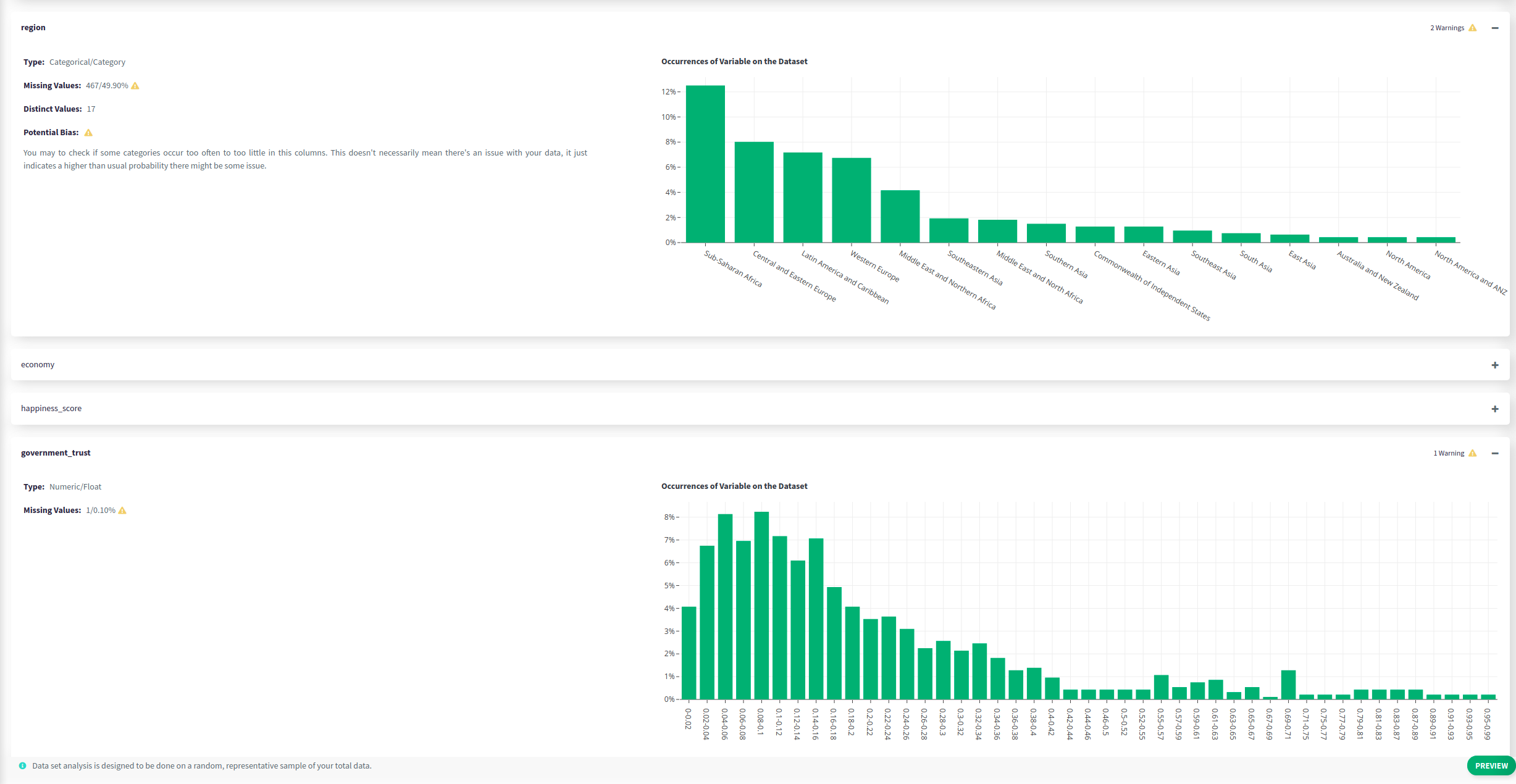
## 4.1 Vnos podatkov v MindsDB scout

Vnos v MindsDB preko MindsDB scout je bil preprost, saj se enostavno poveže z lokalno PostgreSQL bazo preko grafičnega uporabniškega vmesnika.

Figure 7: MindsDB scount vnos podatkov (vir:lasten)

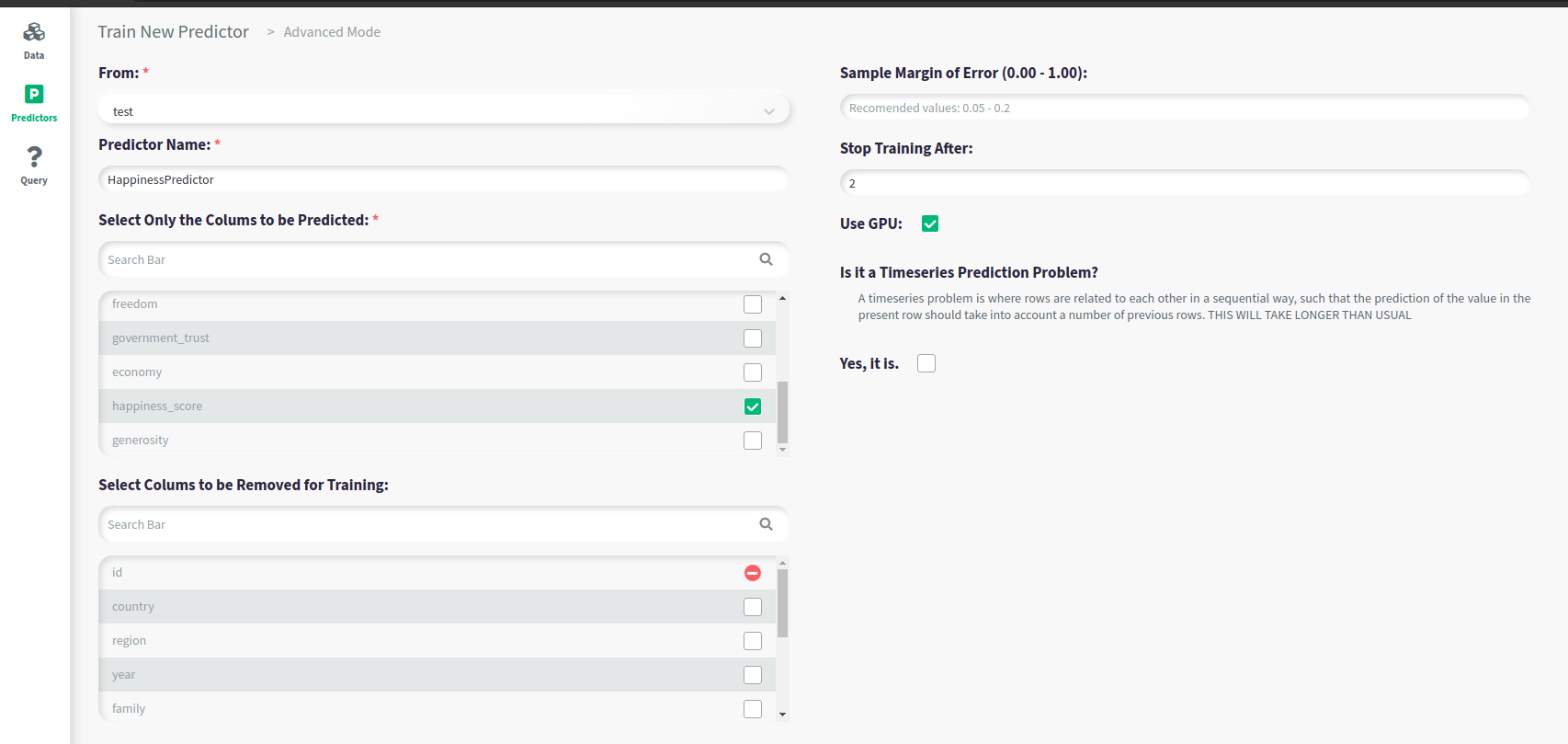
## 4.2 Kvaliteta podatkov

Na naslednjem koraku sva tudi preverila kvaliteto podatkov, ki je prikazala le nekaj opozoril. Opozorila niso resneje vplivala na pristranskost modela zato sva opozorila zanemarila. V nasprotnem primeru, bi pri gradnji modela odstranila problematične stolpce.

Figure 8: Opozorila glede pristranskosti podatkov (vir: lasten)

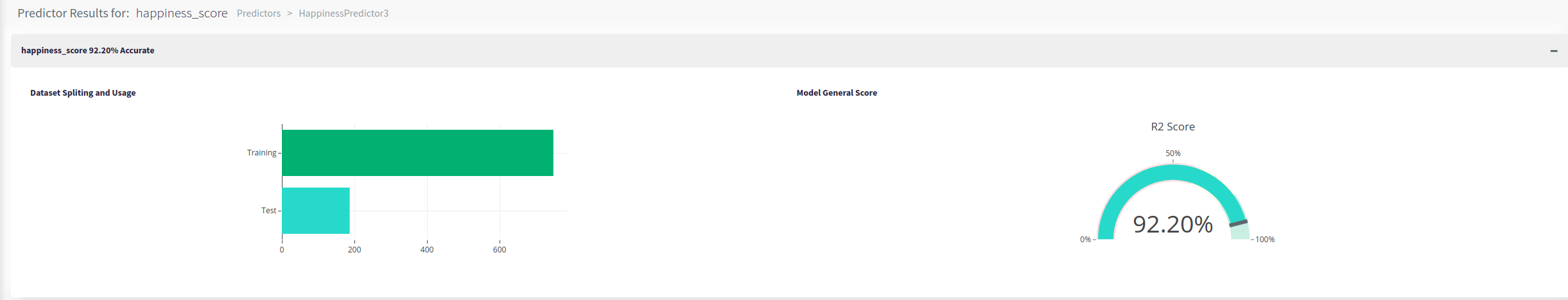
## 4.3 Gradnja modela

Model sva poimenovala *HappinessPredictor*, ki na podlagi vnesenih ostalih parametrov izračuna oziroma oceni nivo sreče. Gradnjo sva opravila preko vmesnika, kjer sva izbrala stolpec, ki ga želiva napovedovati ter parametre, ki naj vplivajo na napovedovanje.

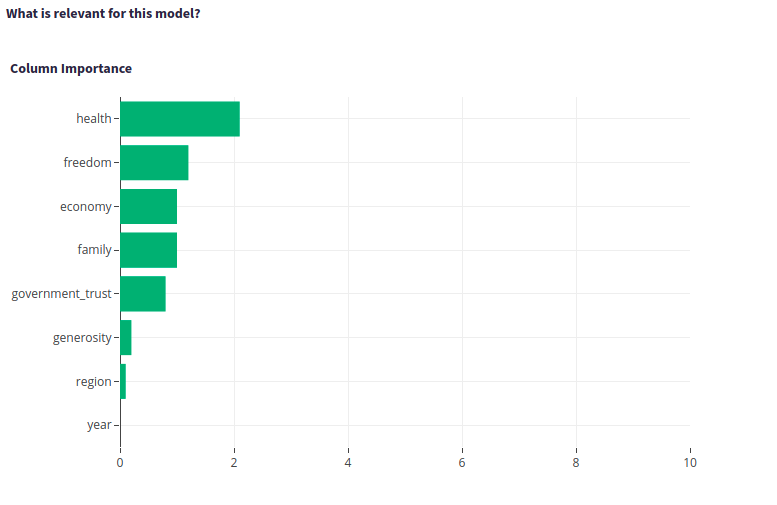
Figure 9: Treniranje modela (vir: lasten)

# 5. Vrednotenje

Po narejenem modelu sva pregledala vrednotenje modela. MindsDB je natačnost napovedovanja modela ocenil z 92.20% stopnje natačnosti, kar je glede na najine potrebe in dane pogoje razpoložljivih podatkov zadovoljivo.



MindsDB je poleg tega tudi ovrednotil pomembnost ostalih parametrov, ki so oziroma bodo služili za napovedovanje sreče. Glavni dejavnik je bil nivo zdravja oziroma dolga zdrava življenska doba.



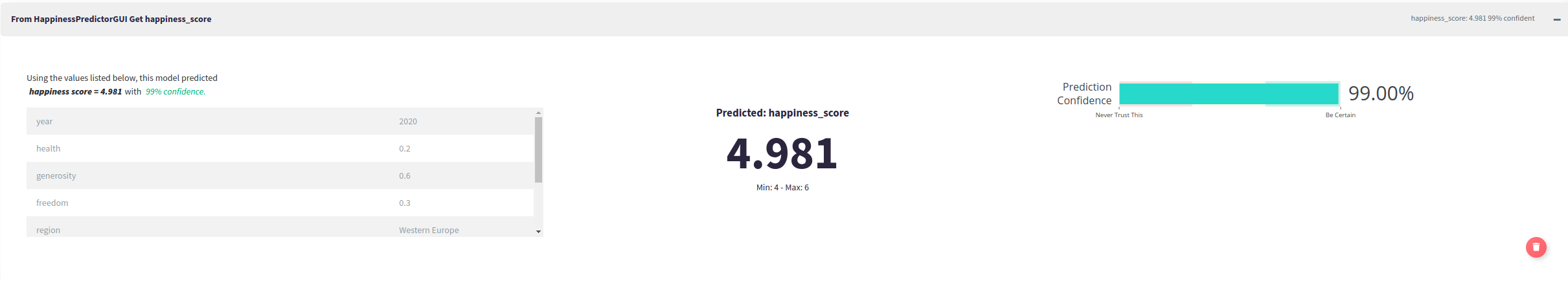
# 6. Uporaba

Na koncu sva preiskusila najin model s praktično uporabo. Osredotočila sva se na trenutne razmere v Sloveniji in na podlagi tega izpolnila parametre od modela pa pričakovala nivo sreče. Vhodni parametri so sledeči.

|  |  |
| --- | --- |
| PARAMETER | VREDNOST |
| year | 2020 |
| health | 0.2 |
| generosity | 0.6 |
| government\_trust | 0.2 |
| freedom | 0.3 |
| region | Western Europe |
| economy | 0.7 |
| family | 0.8 |

Nadpovprečno sva ocenila tri parametre, in sicer ekonomijo, socialno pomoč oziroma družino ter dobrodelnost. Zaradi trenutne korona situacije, ki so pokazali luknje v zdravstvenem sistemu in vzpostavile mehanizme za omejevanje svobode pa sta parametra ‘health’ in ‘freedom’ ocenjena podpovprečno. Poleg tega meniva, da si tudi parameter 'government\_trust', ki predstavlja zaupanje oblasti zasluži v Sloveniji leta 2020 podpovprečno oceno.

Model je napovedal, da smo v Sloveniji z 99% gotovostjo srečni 4.981 po Centrilovi lestvici. Glede na pretekla leta in ocene drugih držav je ta rezultat podpovprečen, kar pa je moč pripisati slabim ocenam najobčutljiveših dejavnikov, ki ju je prizadela korona kriza.



# Viri in literatura

Helliwell, John F., Richard Layard, Jeffrey Sachs, and Jan-Emmanuel De Neve, eds. 2020. World Happiness Report 2020. New York: Sustainable Development Solutions Network

https://machinelearningmastery.com/how-to-know-if-your-machine-learning-model-has-good-performance/