

Merjenje sreče

Seminarska naloga pri predmetu
Tehnologija upravljanja podatkov 2020/21

Študenta: Martin Štrekelj in Simon Babnik

Povezava do Github repozitorija: <https://github.com/MartinStrekelj/HappinessPredictorModel>

Kazalo vsebine

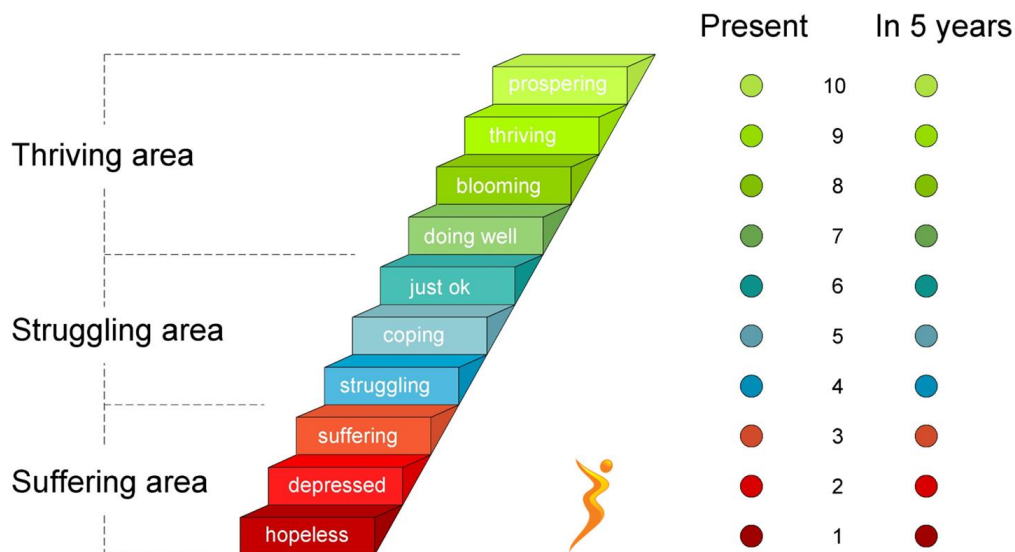
1. Uvod	1
2. Razumevanje problema in razumevanje podatkov	1
3. Priprava podatkov	3
3.1 Vnos podatkov s python skripto	4
3.2 Ročno pregledovanje podatkov	5
4. Modeliranje.....	6
4.1 Vnos podatkov v MindsDB scout	6
4.2 Kvaliteta podatkov	7
4.3 Gradnja modela.....	8
5. Vrednotenje.....	9
6. Uporaba.....	10
7. Zaključek	11
8. Viri in literatura.....	12

1. Uvod

Najina tema za seminarsko nalogo je uporaba MindsDB orodja na večji podatkovni zbirki. Seminarsko nalogo sva izdelala po standardu CRISP-DM, uporabila pa sva sistem za upravljanje podatkov PostgreSQL. Uporabljena podatkovna zbirka predstavlja indeks sreče v različnih državah po svetu ter zadostuje standardom in je primerna za učenje v MindsDB. Cilj seminarske naloge je s pomočjo nadzorovanega učenja izgraditi napovedni model indeksa sreče.

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.



Slika 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 spreminjala, kar sva upoštevala z normalizacijo parametra in tako normalizirala vrednosti med 0 in 1. Konkretno, če je parameter 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.

```
ddl.sql
1  create table if not exists happiness_schema.happiness
2  (
3      id serial not null
4          constraint happiness_pk
5              primary key,
6      country varchar(100),
7      region varchar(100),
8      year integer,
9      family double precision,
10     health double precision,
11     freedom double precision,
12     government_trust double precision,
13     economy double precision,
14     happiness_score double precision,
15     generosity numeric
16 );
17
18 comment on column happiness_schema.happiness.health is 'life expectancy';
19
20 comment on column happiness_schema.happiness.freedom is 'to act at your free will';
21
22 comment on column happiness_schema.happiness.government_trust is 'Corruption';
23
24 comment on column happiness_schema.happiness.economy is 'GDP per Capita';
25
26
27 alter table happiness_schema.happiness owner to postgres;
28
29 create unique index if not exists happiness_id_uindex
30 on happiness_schema.happiness (id);
```

Slika 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.

```
import csv
import psycopg2

con = psycopg2.connect(database="tup_seminarska", user="postgres",
                        password="postgres", host="127.0.0.1", port="5432")

cursor = con.cursor()

INPUT_DATA = "input_data/"
def readFile2015():
    with open(f"{INPUT_DATA}2015.csv", newline='') as csvfile:
        spamreader = csv.reader(csvfile, delimiter=',', quotechar='"')
        i = 1
        for row in spamreader:
            if i > 1:
                country = row[0]
                region = row[1]
                happiness = row[3]
                economy = row[5]
                normalised_economy = (float(economy) - 0) / (1.69 - 0)
                family = row[6]
                normalised_family = (float(family) - 0) / (1.4 - 0)
                health = row[7]
                normalised_health = (float(health) - 0) / (1.03 - 0)
                freedom = row[8]
                normalised_freedom = (float(freedom) - 0) / (0.67 - 0)
                g_trust = row[9]
                normalised_trust = (float(g_trust) - 0) / (0.55 - 0)
                generosity = row[10]
                normalised_genero = (float(generosity) - 0) / (0.8 - 0)
                stmt = f"INSERT INTO happiness_schema.happiness (country, region, year, economy, family, health, freedom, government_trust, happiness_score, generosity) VALUES ('{country}', '{region}', {2015}, {normalised_economy}, {normalised_family}, {normalised_health}, {normalised_freedom}, {normalised_trust}, {float(happiness)}, {normalised_genero})"
                cursor.execute(stmt)
            i += 1
        con.commit()
```

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

Glava programa je na koncu izgledala sledeče.

```
def main():
    readFile2015()
    readFile2016()
    readFile2017()
    readFile2018()
    readFile2019()
    readFile2020()
    con.close()

main()
```

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

3.2 Ročno pregledovanje podatkov

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

```
SELECT * FROM happiness_schema.happiness WHERE health > 1 OR freedom > 1 OR family > 1 OR government_trust > 1 OR generosity > 1 OR happiness_score > 10;
```

Slika 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.

```
UPDATE happiness_schema.happiness SET health = 1, freedom = 1, economy = 1, happiness_score = 7.1 WHERE id = 386;
```

Slika 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: <https://github.com/MartinStrekeli/HappinessPredictorModel>

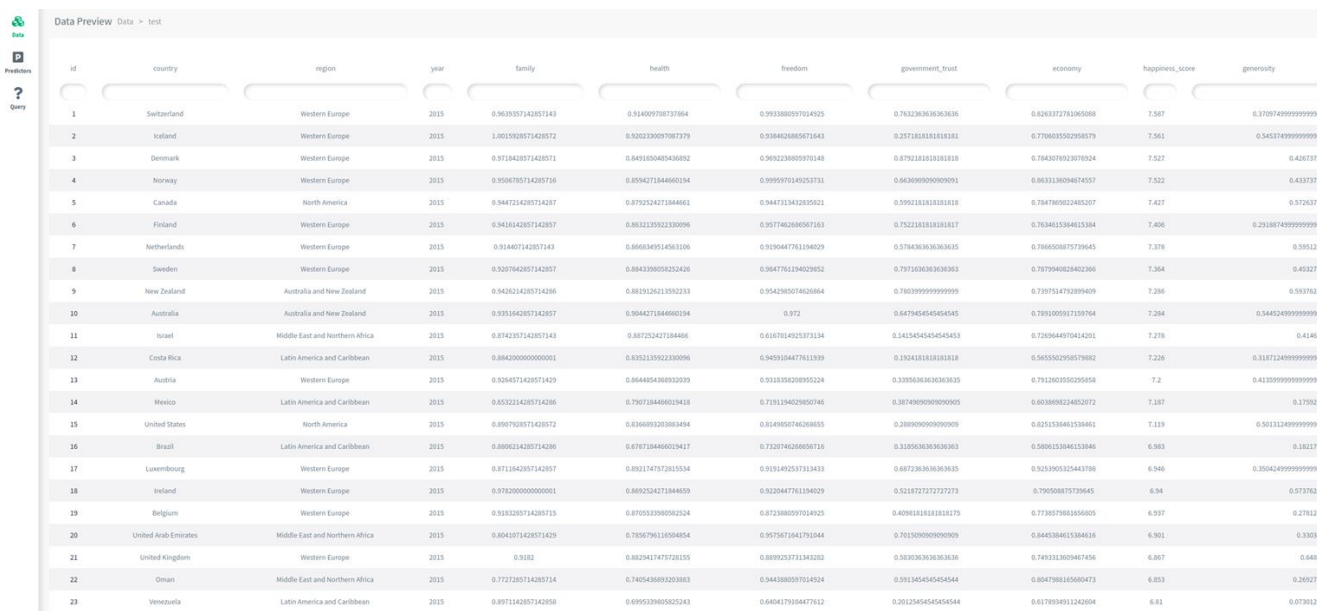
4. Modeliranje

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.



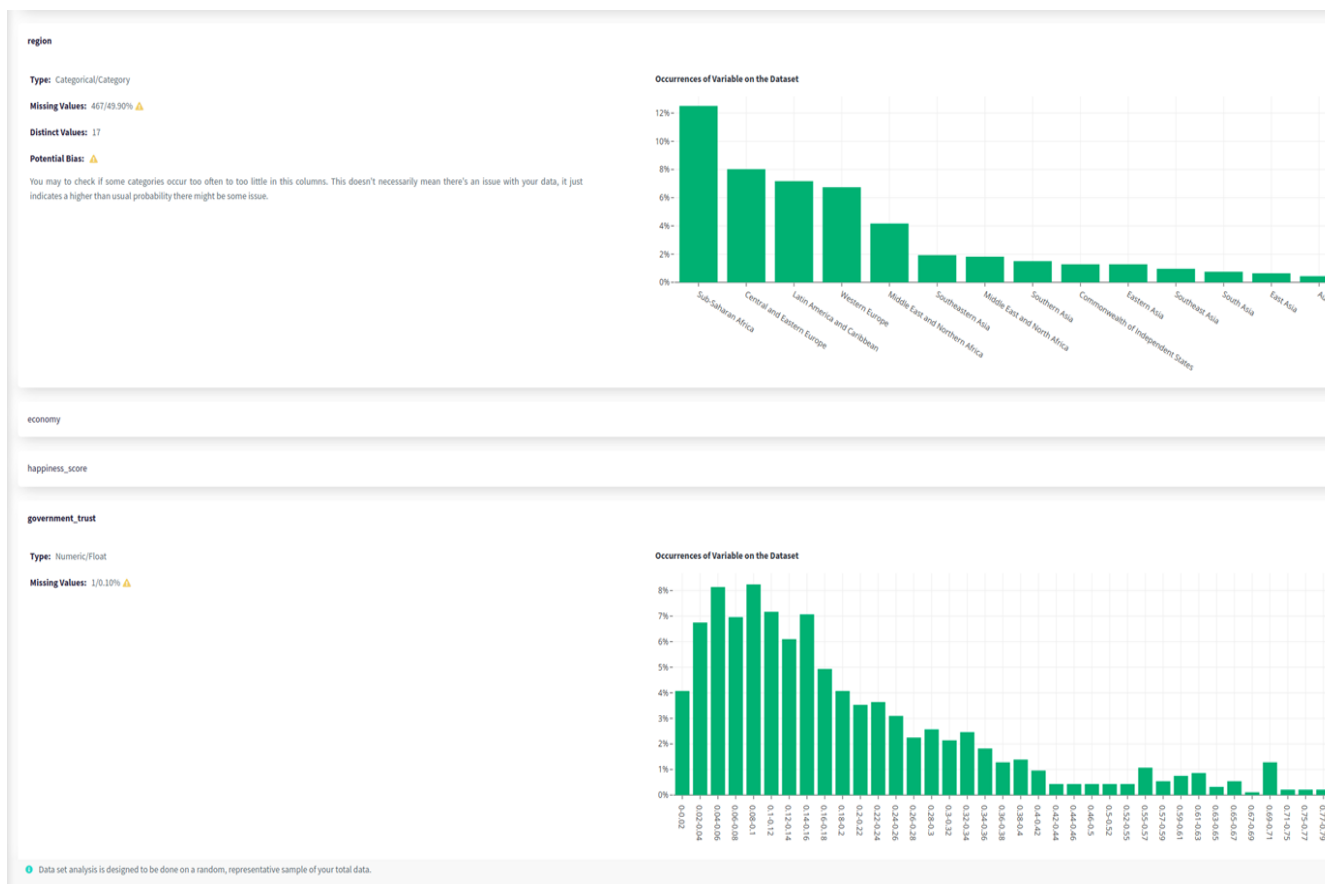
The screenshot displays the 'Data Preview' window of the MindsDB scout application. The interface includes a sidebar with icons for 'Data', 'Predictions', and 'Query'. The main area shows a table with 11 columns: id, country, region, year, family, health, freedom, government_trust, economy, happiness_score, and generosity. The table contains 23 rows of data, each representing a different country. The data is as follows:

id	country	region	year	family	health	freedom	government_trust	economy	happiness_score	generosity
1	Switzerland	Western Europe	2015	0.9639357142857143	0.914009708737864	0.9933880957014925	0.7632363636363636	0.8263372781065088	7.587	0.3709749999999999
2	Iceland	Western Europe	2015	1.0015928571428572	0.9202330097087379	0.938462685671643	0.2571831818181818	0.7706035562958579	7.561	0.5453749999999999
3	Denmark	Western Europe	2015	0.9718428571428571	0.8491850485436892	0.9893236859570148	0.8791218181818182	0.7843076923076924	7.527	0.4263737
4	Norway	Western Europe	2015	0.950678714285716	0.8594271844660194	0.9995976149253731	0.6638090909090901	0.8633136094614557	7.523	0.4337375
5	Canada	North America	2015	0.9447234285714287	0.8792524271844661	0.944731342859521	0.5992181818181818	0.784769822485207	7.427	0.5726375
6	Finland	Western Europe	2015	0.9416142857142857	0.8632135922330096	0.957746268567163	0.7522318181818182	0.7634615384615384	7.406	0.2918874999999999
7	Netherlands	Western Europe	2015	0.9144807142857143	0.8666349514563106	0.9190447761194029	0.9190447761194029	0.786605087579645	7.378	0.595125
8	Sweden	Western Europe	2015	0.9267642857142857	0.884336859252426	0.984776124028652	0.7971603636363636	0.7879540824602366	7.364	0.453275
9	New Zealand	Australia and New Zealand	2015	0.9426234285714286	0.8819126213592233	0.9542385074626864	0.7803999999999999	0.7397514792899409	7.286	0.5837625
10	Australia	Australia and New Zealand	2015	0.9351642857142857	0.9044271844660194	0.972	0.6478454545454545	0.7891005917159764	7.294	0.5445249999999999
11	Israel	Middle East and Northern Africa	2015	0.8742357142857143	0.887251427184466	0.6167614925373134	0.1413454545454545	0.7263444970414201	7.278	0.41465
12	Costa Rica	Latin America and Caribbean	2015	0.8842000000000001	0.8352135922330096	0.949304477611939	0.1924181818181818	0.5455502954857982	7.226	0.3187124999999999
13	Austria	Western Europe	2015	0.9264871428571429	0.8644854368922039	0.933838288955224	0.3395636363636363	0.7912603550295558	7.2	0.4139999999999999
14	Mexico	Latin America and Caribbean	2015	0.6532234285714286	0.7987184466019418	0.7319124023950746	0.38749090909090905	0.6038698224852072	7.187	0.179525
15	United States	North America	2015	0.8987928571428572	0.8366893203834904	0.8148805746288855	0.28880000000000009	0.8251153846153846	7.119	0.5011214999999999
16	Brazil	Latin America and Caribbean	2015	0.8896242857142856	0.8767184466019417	0.732074626856716	0.3189563636363636	0.5806153846153846	6.983	0.182175
17	Luxembourg	Western Europe	2015	0.8711642857142857	0.8921747512815534	0.9101492537313433	0.6872363636363635	0.9253905125443788	6.946	0.3504249999999999
18	Ireland	Western Europe	2015	0.9782000000000001	0.8692524271844659	0.9228447761194029	0.5218727272727273	0.790508875739645	6.94	0.5737625
19	Belgium	Western Europe	2015	0.9383285714285715	0.870553990562524	0.8723880957014925	0.4096181818181818	0.773857981656005	6.937	0.278125
20	United Arab Emirates	Middle East and Northern Africa	2015	0.8043071428571429	0.7856796116504854	0.9575471844718104	0.70150000000000009	0.8445384615384616	6.901	0.33035
21	United Kingdom	Western Europe	2015	0.9382	0.882941747572155	0.889525731343302	0.5836363636363636	0.7493313609461456	6.867	0.6489
22	Oman	Middle East and Northern Africa	2015	0.7727285714285714	0.7405436893203883	0.9443880957014924	0.5913454545454544	0.804788165600473	6.853	0.249275
23	Venezuela	Latin America and Caribbean	2015	0.897142857142858	0.699533805625243	0.4404179104477612	0.20125454545454544	0.6178934911242604	6.81	0.0738125

Slika 7: MindsDB scout vnos podatkov (vir: lasten)

4.2 Kvaliteta podatkov

Na naslednjem koraku sva 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.



Slika 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.

Train New Predictor > Advanced Mode

From: *

test

Predictor Name: *

HappinessPredictor

Select Only the Columns to be Predicted: *

Search Bar

freedom	<input type="checkbox"/>
government_trust	<input type="checkbox"/>
economy	<input type="checkbox"/>
happiness_score	<input checked="" type="checkbox"/>
generosity	<input type="checkbox"/>

Select Columns to be Removed for Training:

Search Bar

id	<input checked="" type="checkbox"/>
country	<input type="checkbox"/>
region	<input type="checkbox"/>
year	<input type="checkbox"/>
family	<input type="checkbox"/>

Sample Margin of Error (0.00 - 1.00):

Recommended values: 0.05 - 0.2

Stop Training After:

2

Use GPU: ☒

Is it a Timeseries Prediction Problem?

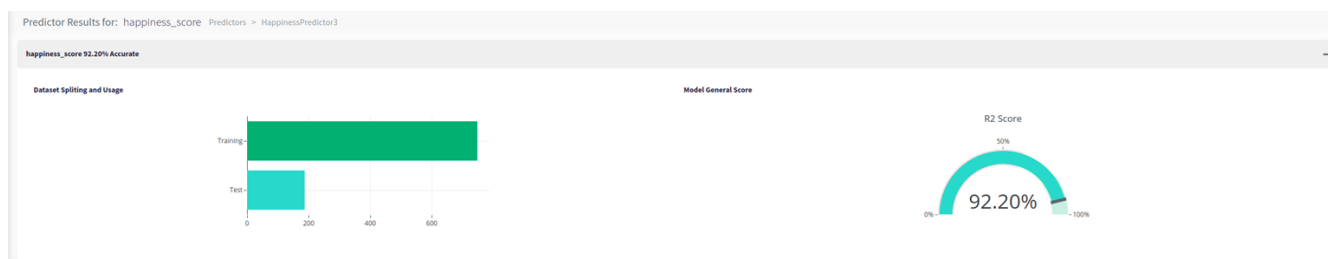
A timeseries problem is where rows are related to each other in a sequential way, such that the prediction of the value in the present row should take into account a number of previous rows. THIS WILL TAKE LONGER THAN USUAL.

Yes, it is. ☐

Slika 9: Treniranje modela (vir: lasten)

5. Vrednotenje

Po narejenem modelu sva pregledala vrednotenje modela. MindsDB je natančnost napovedovanja modela ocenil z 92.20% stopnje natančnosti, kar je glede na najine potrebe in dane pogoje razpoložljivih podatkov zadovoljivo glede na Brownlee, 2018.

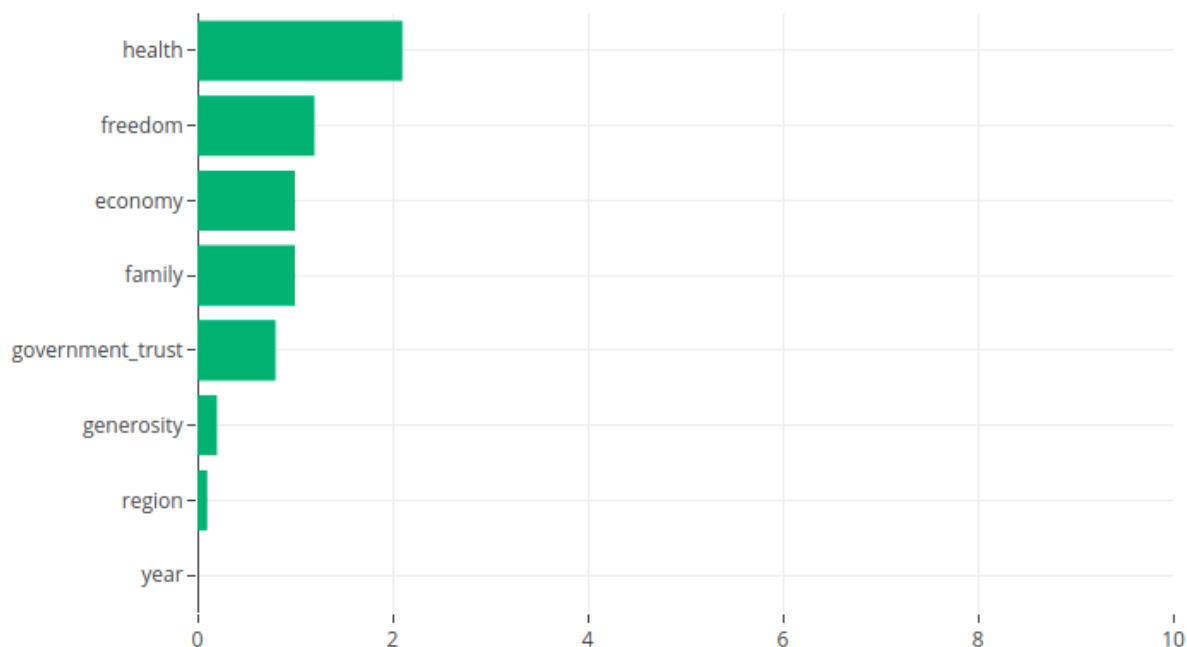


Slika 10: Natančnost napovedovanja modela (vir: lasten)

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 življenjska doba.

What is relevant for this model?

Column Importance



Slika 11: Uteži parametrov pri napovedovanju sreče (vir: lasten)

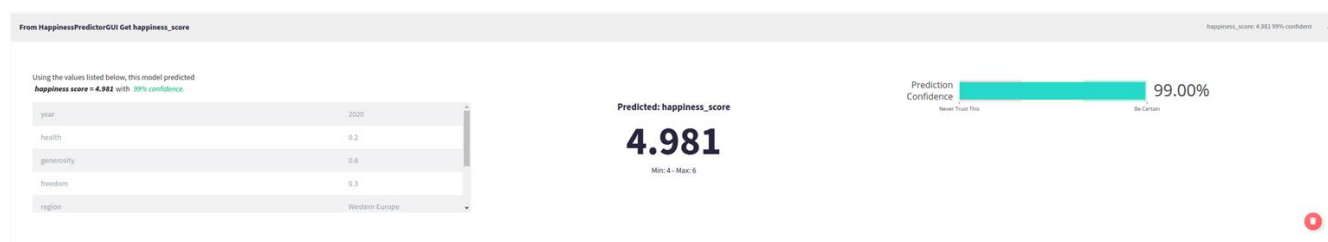
6. Uporaba

Na koncu sva preizkusila 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 najboljčutljivejših dejavnikov, ki ju je prizadela korona kriza ('health' in 'freedom').



Slika 12: Predviden indeks sreče glede na vnesene vrednosti parametrov (vir: lasten)

7. Zaključek

Uspelo nama je izdelati napovedovalni model indeksa sreče, ki glede na parametre zdravje, ekonomija, družina, svoboda, dobrodelnost in zaupanje v oblast z visoko natančnostjo predvidi indeks sreče. Poleg tega nama je model pokazal, kolikšno utež imajo posamezni parametri pri izračunu. Največjo utež imata parametra zdravje in svoboda, kar nam zelo dobro opiše trenutno situacijo po svetu, saj vsem primanjkuje ravno teh dveh stvari in smo posledično tudi manj srečni.

8. Viri in literatura

- [1] *AI Tables—MindsDB Documentation*. (b. d.). Pridobljeno 28. december 2020, s <https://docs.mindsdb.com/databases/>
- [2] Brownlee, J. (2018, april 19). How To Know if Your Machine Learning Model Has Good Performance. *Machine Learning Mastery*. <https://machinelearningmastery.com/how-to-know-if-your-machine-learning-model-has-good-performance/>
- [3] *World Happiness Report 2020*. (b. d.). Pridobljeno 28. december 2020, s <http://worldhappiness.report/>