Kalibracija meritev pametnih svetilk s strojnim učenjem

Marko Hostnik

Povzetek

V poročilu predstavimo projekt kalibracije meritev pametnih svetilk s strojnim učenjem. Analiziramo meritve treh starih in deset novejših pametnih svetilk. Osredotočimo se na meritve temperature in jih grafično predstavimo ter primerjamo. Na podatkih naučimo različne modele strojnega učenja in jih med sabo primerjamo.

Ključne besede
nadzorovano učenje, pametne svetilke, kalibracija meritev

1. Uvod

Za projekt pri predmetu Matematika z računalnikom rešujem problem kalibracije meritev pametnih svetilk. Pri tem je cilj uporabiti metode nadzorovanega in nenadzorovanega strojnega učenja. Pobudo za projekt je dalo podjetje Garex[1], ki v sodelovanju z drugimi podjetji razvijajo rešitev za *pametno* mestno razsvetljavo. S projektom sem se vključil tudi na tekmovanje, ki ga organizira Fakulteta za računalništvo in informatiko[2]. V ekipi z mano sodeluje Gregor Kikelj, študent 3. letnika dodiplomskega študija Matematike na FMF.

1.1 Opis problema

Pametne svetilke za mestno razsvetljavo niso le svetilke, temveč so opremljene z raznimi merilnimi napravami. Vsaka svetilka vsebuje merilnike temperature, tlaka, vlage, onesnaževanja zraka (CO2, PPM) ter radar za detekcijo pešcev in avtomobilov. Vendar vse meritve niso zanesljive, zato jih želimo kalibrirati. Ker so svetilke geografsko blizu (na isti ulici – glej Sliko 1) si morda lahko pomagamo z združenimi podatki vseh svetilk.

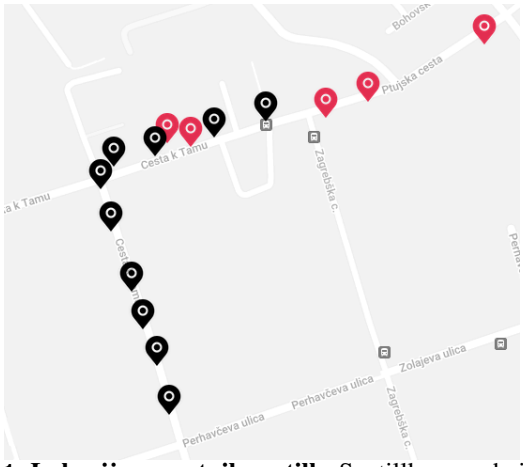
V času dela na projektu so bile svetilke zamenjane. Sredi marca 2022 je podjetje namreč posodobilo tri stare svetilke in jih nadomestilo z desetimi novimi svetilkami. Za vmesno poročilo smo delali s tremi starimi svetilkami. V tem končnem poročilu bomo primerjali rezultate pridobljene iz obeh množic svetilk. Naš cilj je uporabiti zbrane podatke meritev in jih obdelati z metodami strojnega učenja v namen približati se k resničnim vrednostim izmerjenih količin.

2. Uporabljene metode

Za delo uporabljamo programski jezik Python in programske knjižnice za strojno učenje ter analizo podatkov (*scikit-learn, matplotlib, pandas, ...*).

Projekt je razdeljen na dva dela – analiza podatkov in evalvacija modelov. V vsakem delu med sabo primerjamo stare in nove podatke.

[1] https://www.garex.si/
[2] https://datascience.fri.uni-lj.si/competition/



Slika 1. Lokacije pametnih svetilk. Svetillke se nahajajo v Mariboru. Črni markirji označujejo deset novih svetilk, rdeči prejšnje oziroma ostale neuporabljene svetilke.

2.1 Obdelava podatkov

Dobili smo dostop do podatkov svetilk in jih obdelali v obliko primerno za obdelavo. Ker ima vsaka meritev 68 spremenljivk, smo se odločili, da se osredotočimo najprej na vremenske podatke. Za primerjavo z resničnimi vremenskimi podatki smo si pomagali z arhivskimi vremenskimi podatki ARSO[3] iz najbližje samodejne postaje *Maribor – Vrbanski plato*. Vremenska postaja javno razpolaga z natančnimi meritvami temperature in vlage zato se v prihodnje osredotočamo predvsem na ti dve količini.

2.2 Evalvacija modelov

Ključen del modeliranja podatkov je preverjanje kako se bo model obnesel v praksi. Zato smo implementirali programsko ogrodje za testiranje napovedi modelov. Pri tem je pomembna skrb za pravilno metodologijo testiranja, kajti podatki so časovne narave. To pomeni, da ko testiramo model na nekem časovnem odseku, moramo model učiti samo na

[3] https://meteo.arso.gov.si/

podatkih, ki so kronološko pred časom vseh meritev v testnem odseku.

Natančneje, časovno urejene podatke razdelimo na 10 kosov. Pri i-tem kosu, model učimo na vseh podatkih iz prejšnjih kosev in model testiramo na naslednjem kosu. Tako so časovno urejeni podatki za učenje ločeni od podatkov za testiranje.

Rezultate napovedi modelov ovrednotimo z izračunom dveh mer napake:

- povprečno kvadratno odstopanje MSE

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{1}$$

  kjer $\hat{y}_i$ označuje napovedano meritev, y_i pa resnično vrednost.

- mediana absolutnega odstopanja MedAE

$$MedAE = \text{median}(|\hat{y_1} - y_1|, ..., |\hat{y_n} - y_n|) \tag{2}$$

Za razliko od MSE, je MedAE robustna mera, ki ni tako občutljiva na izjeme v podatkih (angl. *outliers*). Poleg izračunanih napak vključimo mero negotovosti – standardno napako. Standardno napako izračunamo z metodo stremena (angl. *bootstrapping*).

2.3 Napovedni modeli

Med sabo primerjamo različne modele, od zelo preprostih do kompleksnejših. V naslednjih opisih privzamemo, da napovedujemo temperaturo. Podobno bi bilo za poljubno drugo količino. Cilj modelov je torej napovedati dejansko temperaturo na podlagi neke izmerjene temperature v prihodnosti.

- **Identiteta.** Napoved temperature je kar izmerjena temperatura.
- **Povprečje.** Napoved temperature je povprečje vseh izmerjenih temperatur na učnih podatkih.
- **Mediana.** Tako kot povprečje, vendar z mediano.
- **Drseče povprečje.** Napoved temperature je povprečje zadnjih 30 minut meritev temperature.
- **Povprečno odstopanje.** Model na učnih podatkih izračuna povprečno odstopanje izmerjene od resnične temperature. Na novem podatku od izmerjene temperature odšteje izračunano odstopanje.
- **Drseče povprečno odstopanje.** V model povprečnega odstopanja damo kot vhodne podatke povprečja zadnjih 30 minut meritev.
- **Ridge.** Ridge je linearna regresija z L2 regularizacijo. Napoved modela je

$$\hat{y}(x) = \beta_0 + \sum_{i=1}^{n}\beta_i x_i = \beta^T x \tag{3}$$

  kjer so β koeficienti, ki minimizirajo cenovno funkcijo

$$J(\beta) = \sum_{i=1}^{m}(\hat{y}_i - y_i)^2 + \lambda\sum_{i=1}^{n}\beta_i^2 \tag{4}$$

- **Naključni gozdovi.** *Random Forest (**RF**)* je model, ki na učnih podatkih zgradi množico odločitvenih dreves. Napoved temperature je povprečje napovedi posameznih dreves.

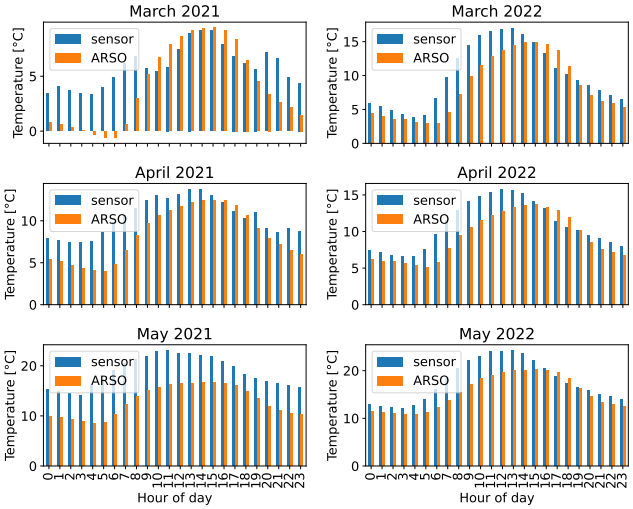
Pri tem izpostavimo, da gre pri modelih z drsečim povprečenjem bolj za različno predprocesiranje podatkov kot za bistveno različne modele. Isti postopek lahko torej uporabimo za nadgradnjo ostalih modelov, kar v prihodnje tudi naredimo z Ridge modelom.

Implementirali smo tudi druge modele, vendar jih izpustimo iz poročila zaradi podobnih rezultatov z opisanimi modeli. Modeli bi lahko bili tudi kompleksnejši, vendar bolj kot točne napovedi si želimo modele robustne na napake zato se nismo ukvarjali z modeli kot so nevronske mreže.

3. Rezultati

3.1 Analiza podatkov

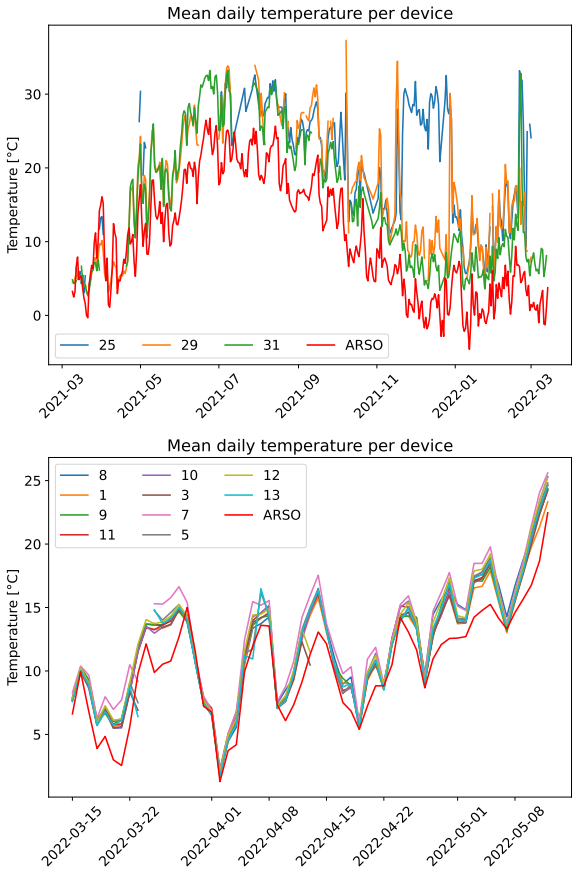
Stare svetilke so podatke zbirale približno eno leto od marca 2021 do marca 2022, ko so bile zamenjane z novimi. Nove svetilke zbirajo podatke od sredine marca 2022 do zdaj (mesec maj 2022). Med starimi in novimi podatki ni preseka, da bi lahko primerjali delovanje svetilk v istem časovnem obdobju.



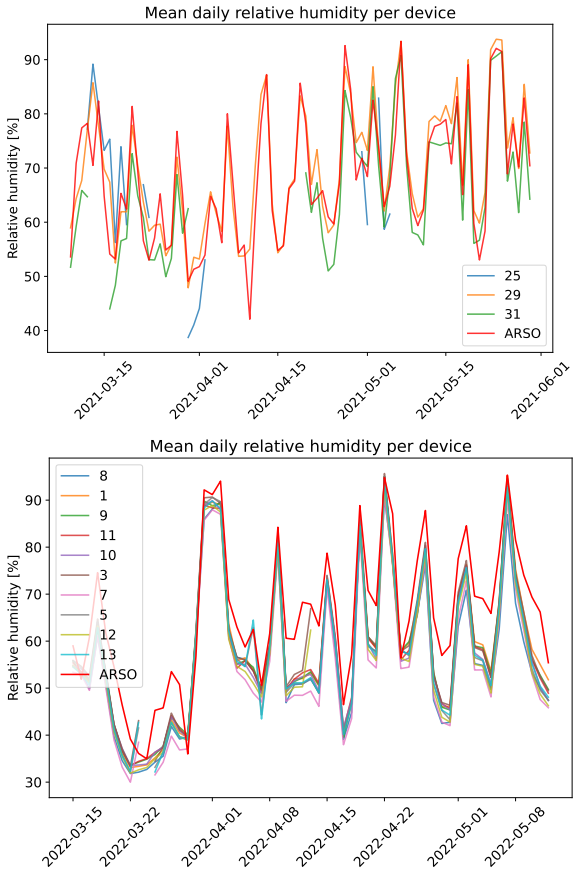
Slika 2. Povprečna temperatura po urah v mesecih. Levi stolpec je za stare podatke, desni za nove. Poudarek je na odstopanju izmerjenih meritev od ARSO meritev.

3.1.1 Stari podatki

Ugotovili smo, da meritve temperature (in drugih količin) niso stabilne. Včasih se izmerjena temperatura povzpne tudi do 600 °C. Zato smo očitno nesmiselne podatke odstranili iz zbirke meritev. Omenimo tudi, da se meritve izvajajo na približno 30 minut.



Slika 3. Dnevna povprečna temperatura starih (zgoraj) in novih (spodaj) svetilk. Za stare svetilke prikažemo celoletno temperaturno nihanje, za nove pa od marca do maja.



Slika 4. Dnevna povprečna vlaga starih (zgoraj) in novih (spodaj) svetilk v podobnem letnem času.

Za uvodno analizo smo združili podatke vseh svetilk in analizirali podatke za vsak mesec posebej. Ugotovili smo, da svetilke poročajo višjo temperaturo od resnične. Ta sprememba je bolj izrazita v zimskih mesecih kot v poletnih. Ker bi bila možna razlaga za odstopanje temperature v izpostavljenosti soncu, smo podatke združili tudi po uri (24 ur) in ugotovili, da je temperatura tudi ponoči previsoka. Primer je viden na Sliki 2.

3.1.2 Novi podatki

Novi podatki so bolj zanesljivih od starih, kot prikazujeta Sliki 2 in 3. Svetilke imajo bolj podobne si meritve, ki so bližje meritvam ARSO.

Na novih podatkih ni več tako izrazitega odstopanja temperature od ARSO meritev v nočnih urah, kot je bilo pri starih podatkih. Odstopanje je večje predvsem po dnevi, kar sumimo, da je povezano z izpostavljenostjo soncu.

3.1.3 Vlaga in zračni tlak

Podobno smo analizirali tudi vlago in tlak. Na Sliki 4 opazimo, da nove svetilke merijo vlago bolj podobno med sabo kot stare svetilke. Vendar pa nove svetilke bolj odstopajo od ARSO meritev pri merjenju vlage kot temperature.

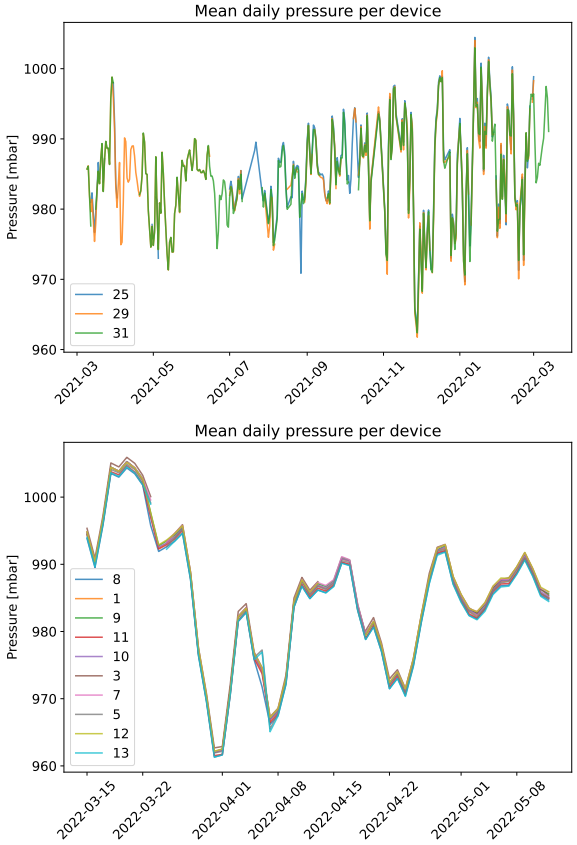
Slika 5 prikazuje meritve zračnega tlaka, ki so si med napravami bolj podobne kot meritve temperature ali vlage. Ker od ARSO postaje Vrbanski-Plato nismo dobili podatkov o tlaku, ne moremo preveriti ali so izmerjene vrednosti primerljive z dejanskim tlakom.

3.2 Eksperimentalna evalvacija

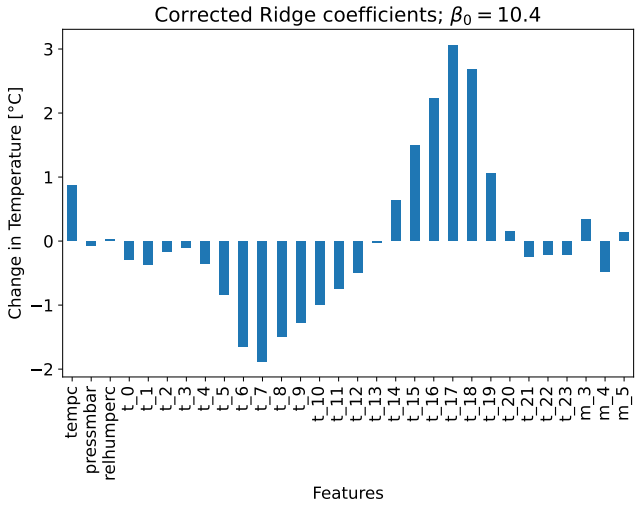
Rezultati evalviranja modelov niso odvisni le od izbire modela in podatkov, temveč tudi od hiper-parametrov modelov in pred-procesiranja podatkov. V sledečih rezultatih zato povzamemo rezultate le najboljših poskusov. Ostali podrobnejši rezultati vseh poskusov so dostopni v repozitoriju projekta.

Pri uporabi modela Ridge in Naključni gozdovi se je izkazala kot zelo uporabna transformacija podatka o času meritve, t.j. ura v dnevu in mesec v letu. Enodimenzionalen čas meritve lahko namreč predstavimo kot dvodimenzionalno točko na krožnici, kar modelu omogoča spoznati ciklično naravo časa (23:00 je le dve uri oddaljeno od 01:00, namesto 22 ur). Podobno storimo za dan oziroma mesec v letu (1.1. je blizu 31.12.).

Pri starih podatkih smo z modelom Ridge najboljše rezultate dobili predvsem z ustreznim pred-procesiranjem podatkov. Vzeli smo drseče povprečje zadnjih 30 minut meritev, transformirali čas kot opisano, uporabili pa smo vse vremenske spremenljivke, ki jih naprava izmeri (temperatura, vlaga, tlak).



Slika 5. Dnevni povprečni zračni tlak starih (zgoraj) in novih (spodaj) svetilk. Za stare svetilke prikažemo celoletno nihanje tlaka, za nove pa od marca do maja.



Slika 6. Koeficienti Ridge modela. Namesto prvih treh koeficientov prikažemo $\frac{\beta_i}{\sigma_i}$ zaradi standardizacije vhodnih spremenljivk.

Podatke smo tudi standardizirali. Parameter regularizacije smo nastavili na $\lambda = 1$.

Pri novih podatkih se je Ridge najbolje obnesel z drugačno predstavitvijo časa. Namesto, da je čas na krožnici smo čas (uro v dnevu in mesec v letu) predstavili z indikatorskimi spremenljivkami (angl. *one-hot encoding*). Menimo, da je razlog zakaj s tako predstavitvijo časa dobimo boljše napovedi v tem, da novi podatki vključujejo le tri mesece meritev zato letni čas ni prišel do izraza.

Pri naključnih gozdovih smo najboljše rezultate dobili z uporabo krožne transformacije časa meritve, uporabe vremenskih podatkov in poleg tega podatek o izvoru meritve (t.j. od katere naprave izvira meritev). Naključni gozdovi ne potrebujejo standardiziranih vhodnih podatkov. Vsa zgrajena drevesa so razvita do konca in v ansamblu jih je skupaj 100.

V Tabeli 1 so povzeti rezultati za stare svetilke in v Tabeli 2 za nove. Za lažjo interpretacijo rezultatov smo dodatno izračunali kolikšen delež vseh napovedi je odstopalo od resnične temperature za manj kot eno, tri ali pet stopinj.

3.2.1 Interpretacija rezultatov

Na starih podatkih (Tabela 1) so Naključni gozdovi (RF) jasen zmagovalec za model, ki najbolje napoveduje temperaturo. Na drugem mestu je Ridge. Pri obeh modelih bi rezultati lahko bili boljši, če bi izvedli optimizacijo hiper-parametrov modelov (npr. število dreves, globina dreves, parameter regularizacije, itd.) s prečnim preverjanjem. Za to se nismo odločili zaradi visoke računske zahtevnosti izvajanja poskusov in ker je to delo primerno za končno izbiro modela, ki se bo uporabil v produkciji.

Kljub rezultatski prednosti modela RF pred Ridge, uporaba Ridge omogoča določene prednosti. Učenje je pri Ridge bistveno hitrejše in Ridge je linearen model, zato je preprost in omogoča enostavnejšo interpretacijo naučenih koeficientov modela.

Slika 6 prikazuje koeficiente najboljšega Ridge modela učenega na novih podatkih, kjer je od koeficientov za temperaturo, tlak in vlago odstranjen vpliv standardizacije za lažje razumevanje. Vsak koeficient nam pove za koliko stopinj Celzija se bo napoved spremenila, če se vhodna spremenljivka spremeni za eno enoto (kar je ravno odvod funkcije napovedi po vhodnih spremenljivkah). Kot vidimo se je model uspešno naučil, da je temperatura čez dan višja kot ponoči.

Na novih podatkih (Tabela 2) je Ridge dosegel najboljše napovedi temperature. Naključni gozdovi so tokrat dosegli slabše rezultate od preprostejših modelov povprečnega odstopanja in drsečega povprečnega odstopanja. Razlog za to pripisujemo bolj natančnim vhodnim podatkom v primerjavi s starimi podatki. RF namreč dobro najde interakcije med vhodnimi spremenljivkami, kar sedaj ni tako pomembno, ker so meritve bolj zanesljive.

Novi podatki so bolj zanesljivi od starih, saj vidimo, da je Identiteta na novih podatkih dosegla bolj točne rezultate kot celo najbolj natančen model na starih podatkih. Dodatno, razlike med modeli so manjše na novih podatkih. Morda bi lahko razlog za napake v meritvah pripisali različni legi pametnih svetilk od ARSO vremenske postaje.

	MSE	MedAE	< 1 °C	< 3 °C	< 5 °C
Naključni gozdovi	14.144 ± 0.097	2.210 ± 0.011	0.245	0.621	0.828
Ridge	21.916 ± 0.157	2.624 ± 0.014	0.212	0.553	0.760
Drseče povp. odst.	31.586 ± 0.219	3.452 ± 0.017	0.155	0.439	0.658
Povp. odst.	50.713 ± 0.464	3.097 ± 0.019	0.179	0.488	0.682
Povprečje	80.277 ± 0.315	7.398 ± 0.026	0.065	0.201	0.333
Mediana	82.048 ± 0.328	7.400 ± 0.042	0.056	0.188	0.315
Drseče povp.	83.291 ± 0.405	7.034 ± 0.017	0.014	0.082	0.244
Identiteta	102.464 ± 0.755	6.330 ± 0.016	0.022	0.106	0.319

Tabela 1. Rezultati evalvacije modelov za napoved temperature [°C] na **starih** podatkih.

	MSE	MedAE	< 1 °C	< 3 °C	< 5 °C
Ridge	4.163 ± 0.016	1.214 ± 0.003	0.414	0.877	0.974
Drseče povp. odst.	5.938 ± 0.015	1.697 ± 0.003	0.268	0.778	0.964
Povp. odst.	7.891 ± 0.027	1.797 ± 0.004	0.262	0.745	0.922
Naključni gozdovi	8.547 ± 0.031	1.672 ± 0.004	0.322	0.737	0.910
Drseče povp.	9.262 ± 0.028	1.853 ± 0.007	0.326	0.675	0.874
Identiteta	11.229 ± 0.040	1.900 ± 0.007	0.318	0.680	0.861
Povprečje	39.532 ± 0.117	4.011 ± 0.009	0.135	0.381	0.575
Mediana	40.420 ± 0.118	4.100 ± 0.022	0.127	0.374	0.574

Tabela 2. Rezultati evalvacije modelov za napoved temperature [°C] na **novih** podatkih.

4. Zaključek

V projektu smo analizirali dve množici podatkov, od starih in od novih pametnih svetilk. Implementirali smo različne modele za napoved temperature in ugotovili, da so meritve novih svetilk bolj zanesljive od starih. Na starih svetilkah je bil najbolj natančen model Naključni gozdovi, na novih pa model Ridge.

4.1 Prihodnje delo

Tekmovanje še ni končano. V prihodnje bomo modele preizkusili na drugih količinah, kot sta tlak in vlaga. Ker nimamo na voljo resničnih podatkov za izvedbo nadzorovanega strojnega učenja na drugih meritvah, bomo preizkusili metodo napovedi izpada posamezne svetilke. Namesto, da bomo napovedali temperaturo na podlagi meritev desetih svetilk in rezultat primerjali s podatki vremenske postaje, bomo modele učili na meritvah devetih svetilk in z njimi skušali napovedati meritve izpuščene svetilke. Na tak način lahko zaznamo, ali nam uspe iz devetih svetilk rekonstruirati meritve desete.