POROČILO ZA 2. SEMINARSKO NALOGO PRI PREDMETU UMETNA INTELIGENCA

Šolsko leto 2020/21

Nik Prinčič (63190240)

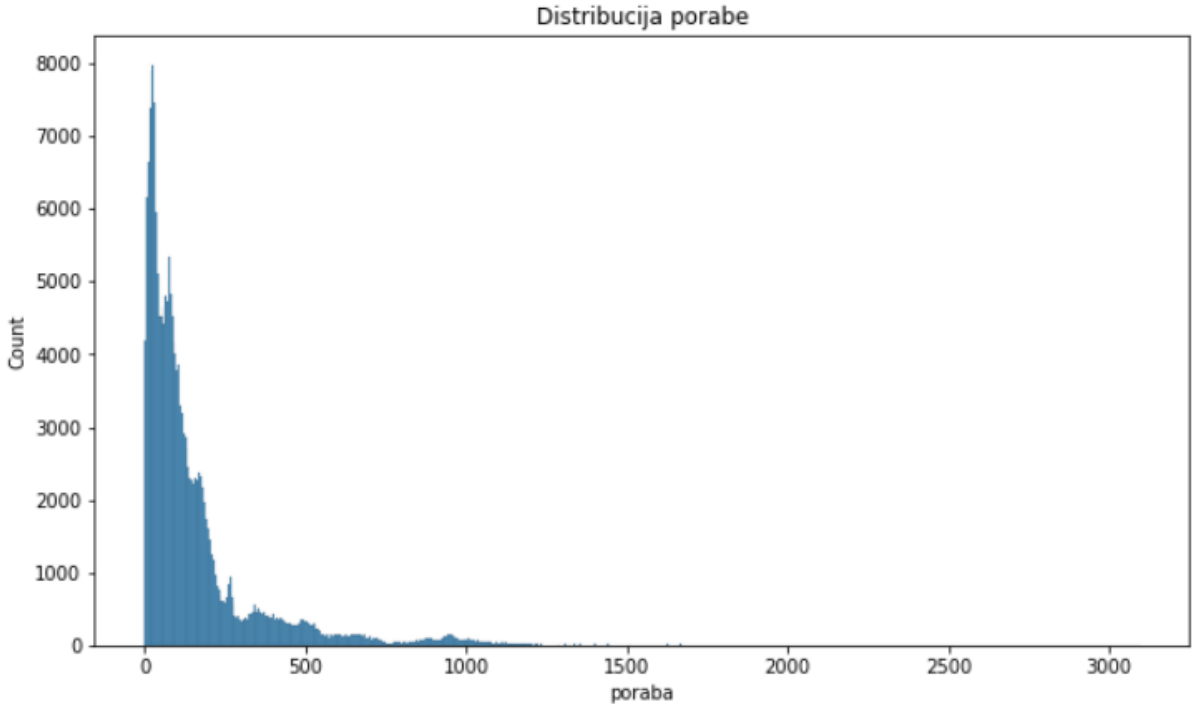
1. Priprava atributov

Dodal sem naslednje atribute:

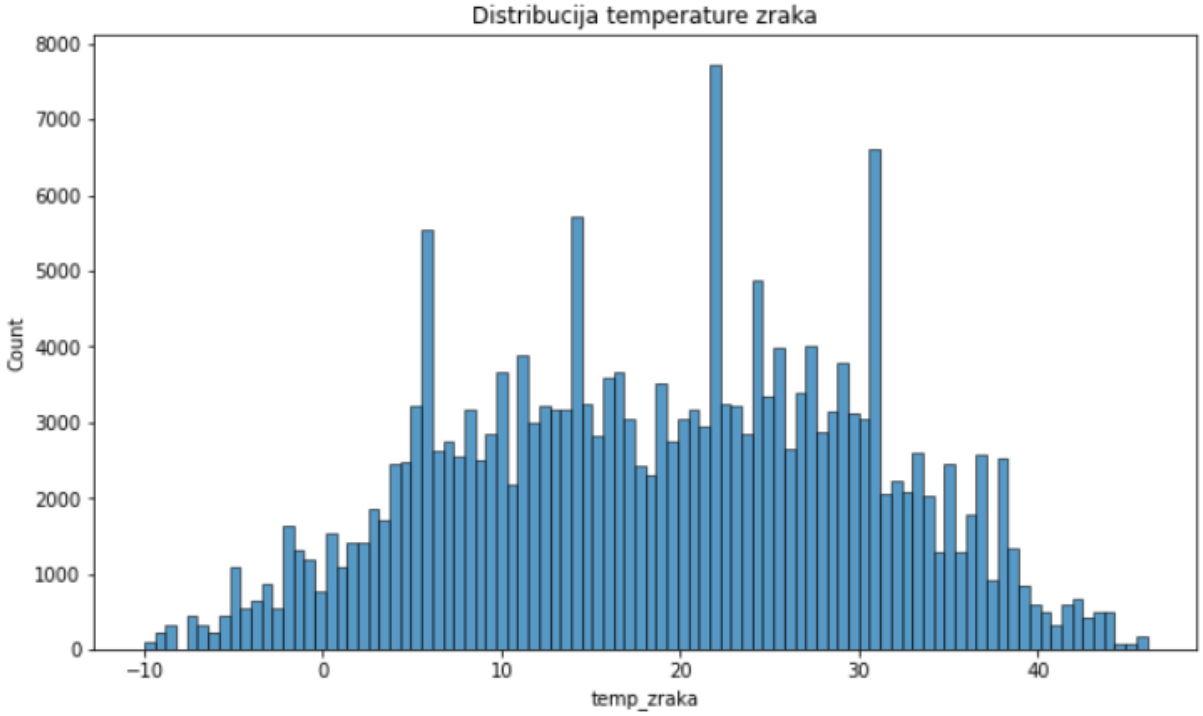
- datetime, ki hrani Datetime objekt
- vikend, binarna vrednost, kjer 1 predstavlja vikend (sobota, nedelja), 0 pa dan v tednu
- mesec, ki hrani zaporedno številko meseca v letu
- teden, ki hrani zaporedno številko tedna v letu
- dan, ki hrani zaporedno številko dneva v letu
- 7, 14, 21 in 28 dnevno premikajoče zaporedje (moving average) za vse zvezne atribute (temp_zraka, temp_rosisca, oblacnost, padavine, pritisk, hitrost_vetra)

2. Vizualizacija podatkov

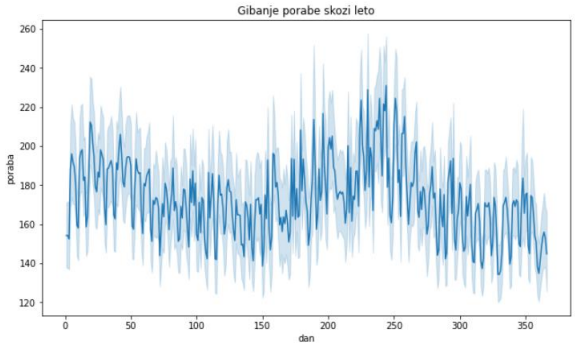
Kot prvi zanimiv podatek se mi je zdelo to, da poraba ni normalno porazdeljena.

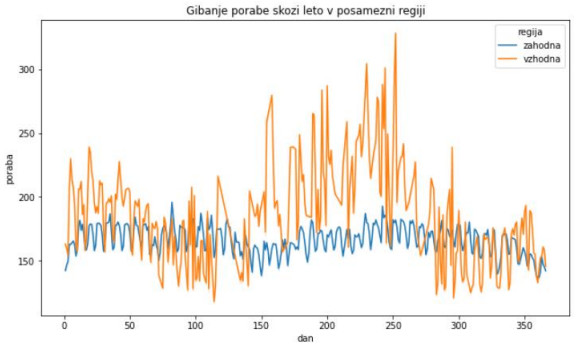


Nasprotno od porabe pa je temperatura zraka približno normalno porazdeljena.

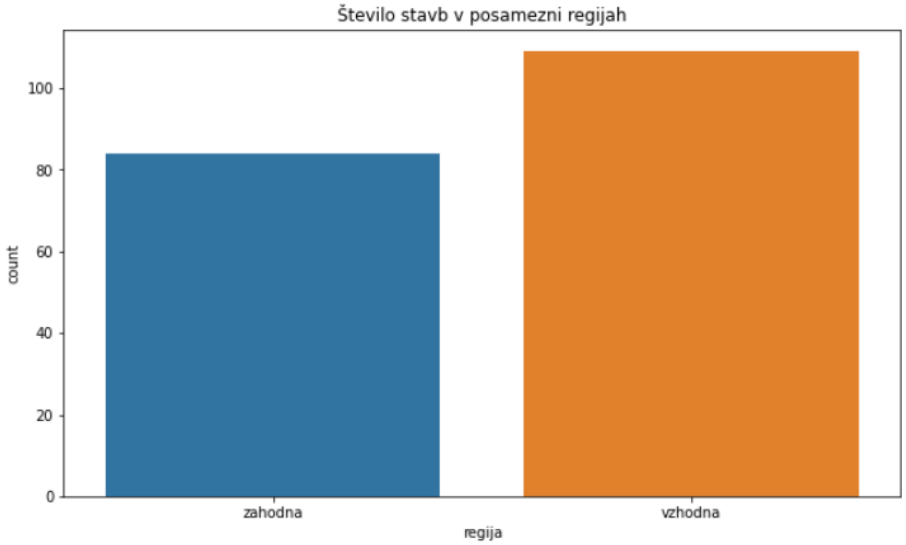


Gibanje porabe skozi leto pa najbolj narase nekje sredi jeseni, sicer pa je to naraščanje bolj opazno v vzhodni regiji.



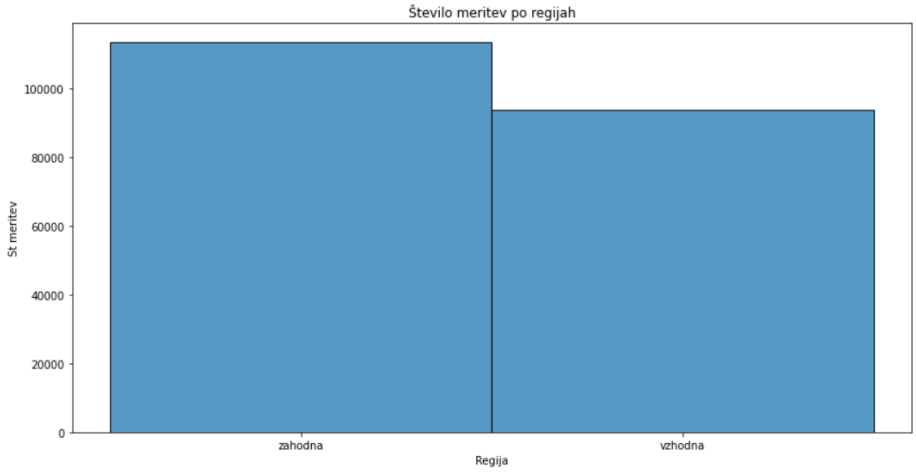


Zanimiv je tudi podatek, da stavbe v podatkovni množici niso enakomerno porazdeljene med regijami.

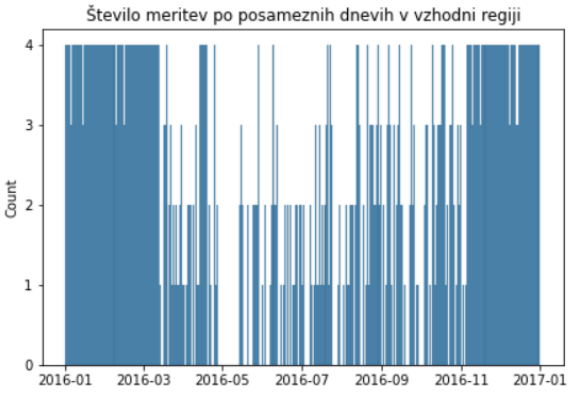


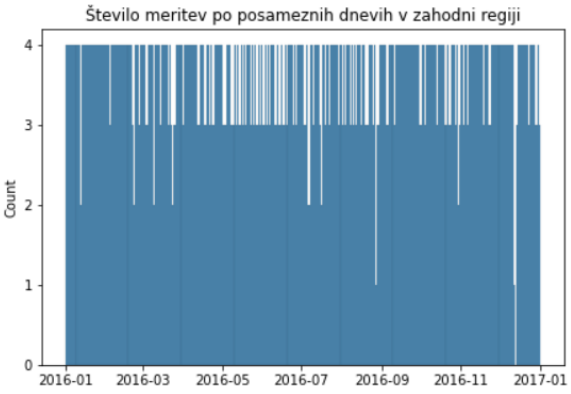
Podobno je tudi kar nekaj več izobraževalnih stavb kot pa drugih.

Tudi število meritev med regijama ni enako, v zahodni je bilo izvedenih več meritev.

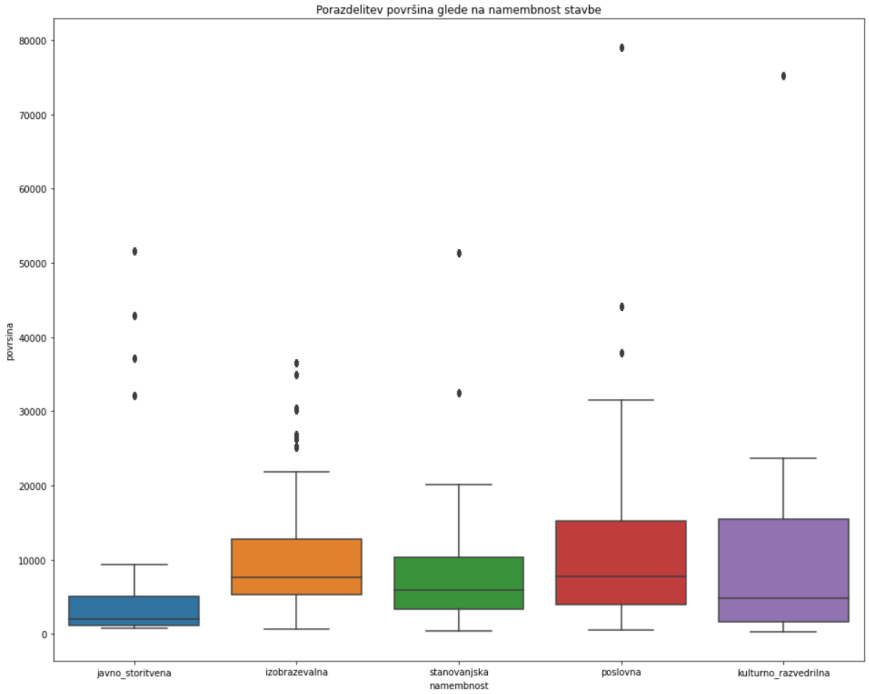


V podatkovni množici so vsi zapisi popolni (ni vrstice, ki bi imela NAN/NA), so pa podatki rahlo ne konsistentni, saj so obstajajo časovni intervali, ko se meritve niso izvajale štirikrat na dan, ali pa se sploh nekaj dni niso izvajale, sklepam, da bi v realnem svetu to lahko bila posledica nepravilnega delovanja senzorjev, omrežja, itd. Iz spodnjih histogramov lahko razberemo, da je na vzhodu od sredine marca pa do sredine novembra kar veliko manjkajočih meritev, posledica tega je tudi razvidna v zgornjem grafu, ki prikazuje število meritev v posamezni regiji.

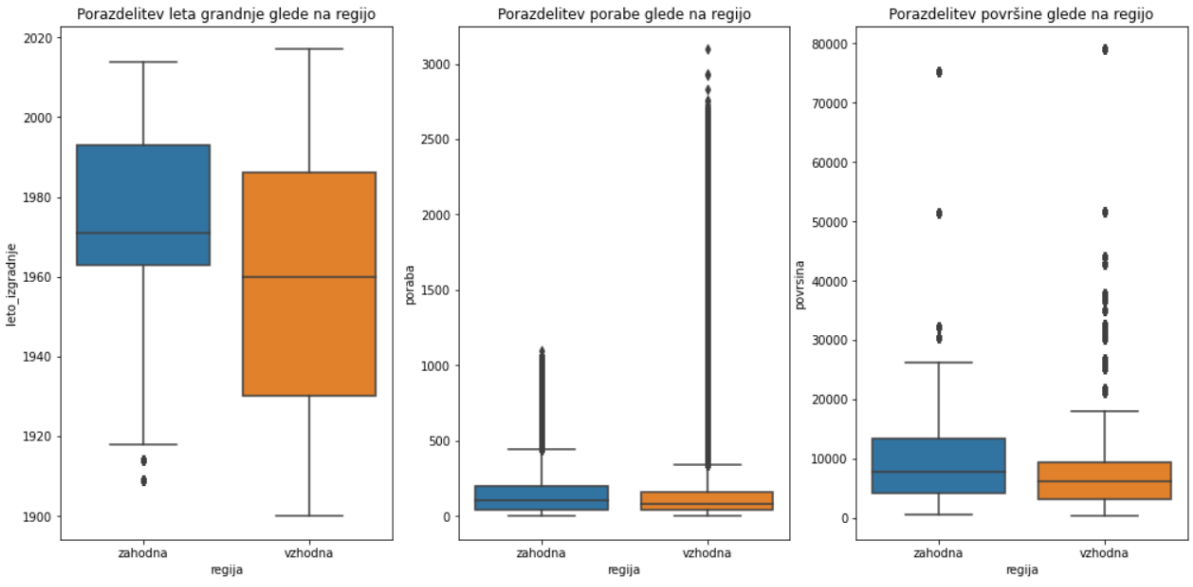




Nekako predvidljiva pa je distribucija površin stavb glede na namembnost, kjer so poslovne in izobraževalne stavbe v povprečju največje, javno storitvene pa najmanjše.



Zanimivo je tudi kako regija vpliva na leto izgradnje in površino stavb ter porabo.



3. Analiza atributov

3.1. Apriorne verjetnosti razredov

	norm_poraba	verjetnost razreda
1	SREDNJA	0.369635
2	NIZKA	0.233745
3	VISOKA	0.186957
0	ZELOVISOKA	0.132553
4	ZELONIZKA	0.077109

3.2. Ocena atributov z metriko Chi^2

	Atribut	ocena
0	povrsina	9.138277e+07
1	stavba	6.955992e+04
2	dan	2.865281e+04
3	ura	1.930063e+04
4	smer_vetra	1.105394e+04
5	namembnost	5.704203e+03
6	leto_izgradnje	5.043999e+03
7	teden	4.613801e+03
8	vikend	1.362082e+03
9	temp_rosisca	1.156765e+03
10	temp_zraka	8.698973e+02
11	temp_rosisca_MA_7	8.652343e+02
12	mesec	8.444787e+02
13	temp_rosisca_MA_14	8.203148e+02
14	regija	7.684409e+02
15	temp_rosisca_MA_21	6.726223e+02
16	temp_zraka_MA_14	6.151689e+02
17	temp_zraka_MA_7	6.038999e+02
18	temp_zraka_MA_21	5.831537e+02
19	temp_zraka_MA_28	5.439896e+02
20	temp_rosisca_MA_28	5.196205e+02
21	hitrost_vetra_MA_28	1.380060e+02
22	hitrost_vetra_MA_21	1.370714e+02
23	hitrost_vetra_MA_7	1.329736e+02
24	hitrost_vetra_MA_14	1.286485e+02
25	oblacnost_MA_28	1.143028e+02
26	oblacnost_MA_21	1.037772e+02
27	oblacnost_MA_7	9.526975e+01
28	oblacnost_MA_14	9.284265e+01
29	padavine	7.373481e+01
30	padavine_MA_14	7.273259e+01
31	padavine_MA_21	7.141134e+01
32	padavine_MA_7	7.096048e+01
33	padavine_MA_28	6.882838e+01
34	oblacnost	6.278009e+01
35	hitrost_vetra	3.285671e+01
36	pritisk	2.102491e+01
37	pritisk_MA_28	1.218917e+01
38	pritisk_MA_21	1.143751e+01
39	pritisk_MA_14	1.079086e+01
40	pritisk_MA_7	9.886802e+00

3.3. Ocena atributov z metriko Relieff

	Atribut	ocena
0	povrsina	15.486930
1	smer_vetra	10.200000
2	ura	5.520000
3	dan	5.510000
4	leto_izgradnje	4.370000
5	temp_zraka	3.844000
6	stavba	3.400000
7	temp_rosisca	3.257000
8	pritisk	2.684000
9	temp_rosisca_MA_7	1.712107
10	oblacnost	1.560000
11	hitrost_vetra	1.414000
12	temp_rosisca_MA_14	1.362399
13	temp_zraka_MA_7	1.294607
14	temp_zraka_MA_14	1.096705
15	temp_rosisca_MA_21	1.058123
16	pritisk_MA_7	0.973393
17	temp_zraka_MA_21	0.960891
18	temp_zraka_MA_28	0.901109
19	temp_rosisca_MA_28	0.876530
20	teden	0.770000
21	pritisk_MA_14	0.736118
22	pritisk_MA_21	0.559929
23	pritisk_MA_28	0.501043
24	oblacnost_MA_7	0.430000
25	vikend	0.370000
26	hitrost_vetra_MA_7	0.346679
27	oblacnost_MA_14	0.302258
28	namembnost	0.250000
29	oblacnost_MA_21	0.229774
30	oblacnost_MA_28	0.183398
31	hitrost_vetra_MA_14	0.172152
32	padavine	0.150000
33	mesec	0.140000
34	hitrost_vetra_MA_21	0.118676
35	padavine_MA_7	0.115000
36	hitrost_vetra_MA_28	0.103323
37	padavine_MA_14	0.089643
38	padavine_MA_21	0.060630
39	padavine_MA_28	0.057885
40	regija	0.020000

4. Predstavitev dobljenih modelov

4.1. Klasifikacija

Pri klasifikacijskih problemih sem najprej sestavil modele na osnovi ocene Relieff in nato še z uporabo ocene Chi^2 in pri vseh modelih dobil boljšo natančnost z uporabo slednje. Vse modele sem gradil na način, da sem model najprej zgradil na 5 do n najbolje ocenjenih atributih in nato kot končni model izbral tistega ki je imel najboljšo natančnost.

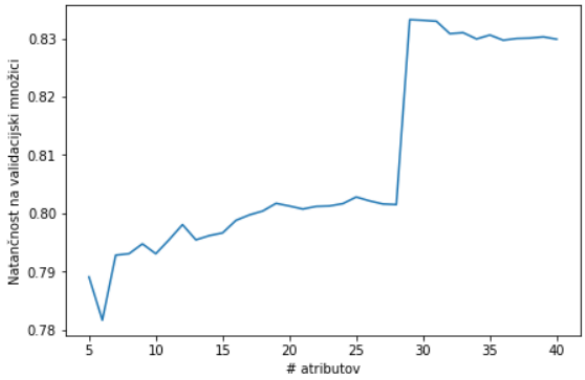
Pri nekaterih vrstah modelov (tisti, ki za učenje potrebujejo veliko časa), pa sem uporabil samo oceno Chi^2, saj mi je v vseh prejšnjih modelih dala boljšo natančnost.

Ko sem uporabljal metodo ocenjevanja atributov Chi^2 sem vse atribute, ki so lahko potencialno negativni transformiral tako, da sem jim prištel nek teoretičen maksimum. Za padavine sem uporabil +1 ter za temp_zraka in temp_rosisca sem uporabil +100.

4.1.1. Odločitveno drevo

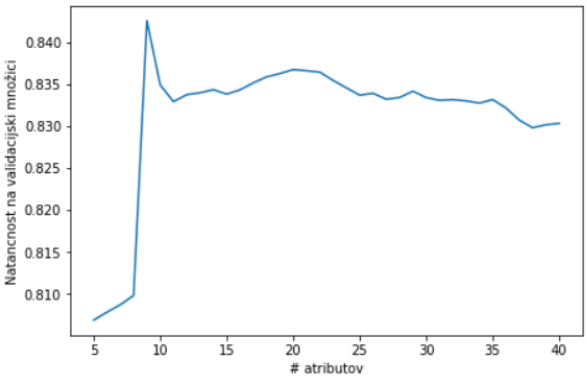
4.1.1.1. Odločitveno drevo (atributi izbrani z metriko Relieff)

```
Najboljša natančnost na validacijski množici: 0.8332354206446769 (# atributov: 29)
Natančnost na testni množici: 0.832726159004449
Izbrani atributi:
['ura', 'stavba', 'povrsina', 'leto_izgradnje', 'temp_zraka', 'temp_rosisca', 'oblacnost', 'padavine', 'pritisk', 'smer_vetra', 'hitrost_vetra',
'teden', 'dan', 'vikend', 'temp_zraka_MA_7', 'temp_zraka_MA_14', 'temp_zraka_MA_21', 'temp_zraka_MA_28', 'temp_rosisca_MA_7', 'temp_rosisca_MA_1
4', 'temp_rosisca_MA_21', 'temp_rosisca_MA_28', 'oblacnost_MA_7', 'oblacnost_MA_14', 'pritisk_MA_7', 'pritisk_MA_14', 'pritisk_MA_21', 'pritisk_MA
_28', 'hitrost_vetra_MA_7']
```



4.1.1.2. Odločitveno drevo (atributi izbrani z metriko Chi^2)

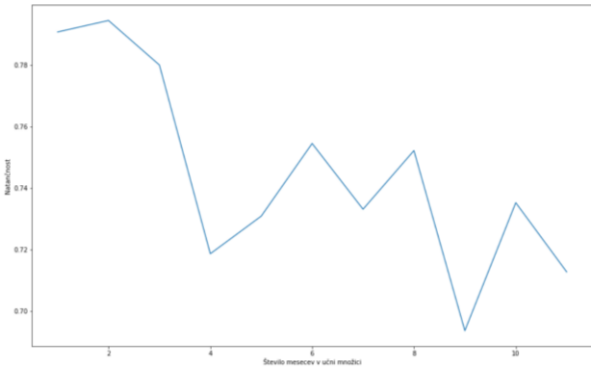
```
Najboljša natancnost na validacijski množici: 0.8425619572037097 (# atributov: 9)
Natancnost na testni množici: 0.845766974015088
Izbrani atributi:
['ura', 'stavba', 'namembnost', 'povrsina', 'leto_izgradnje', 'smer_vetra', 'teden', 'dan', 'vikend']
```



4.1.1.3. Ovrednotenje modela na mesečnih podatkih ({januar} proti {februar}, {januar, februar} proti {marec}, ...)

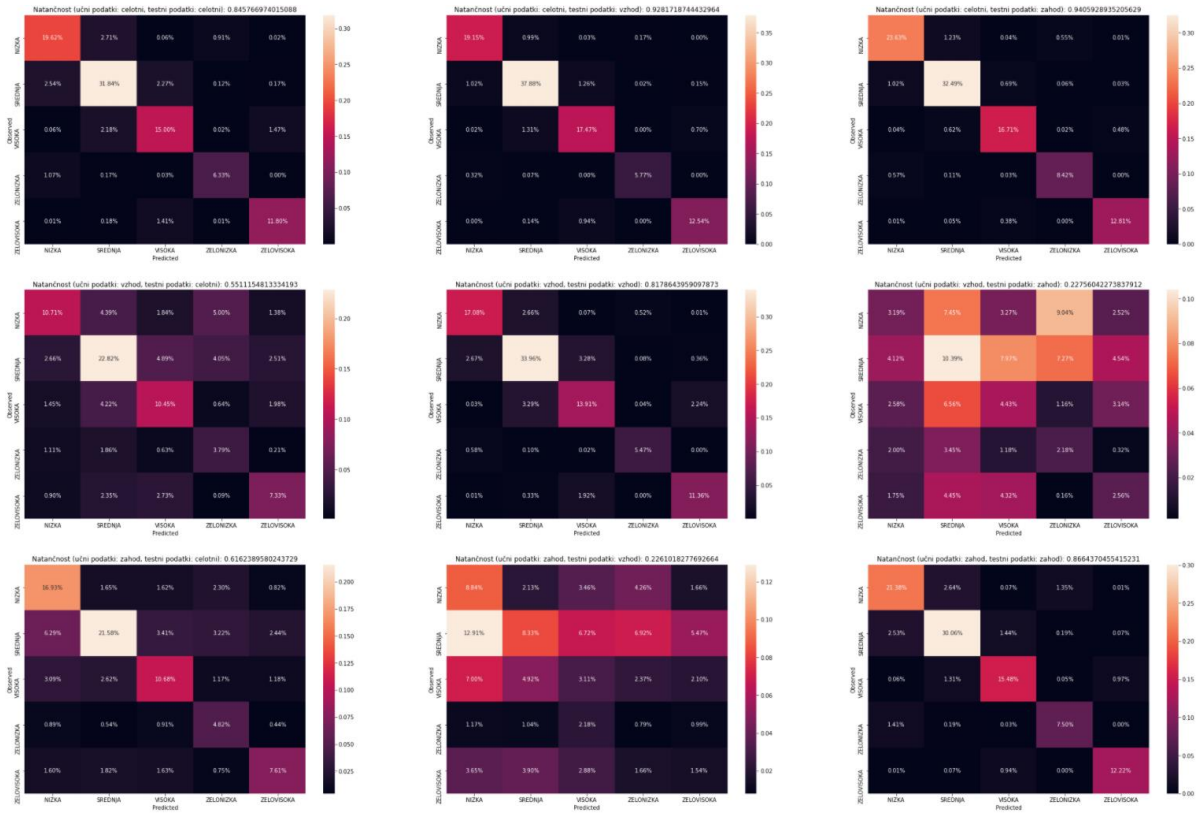
Za ovrednotenje sem izbral model, ki mi je pri navadnem ovrednotenju dal boljše rezultate.

```
Število mesecev v učni množici: 1, natančnost: 0.7908848417954378
Število mesecev v učni množici: 2, natančnost: 0.7946196762967463
Število mesecev v učni množici: 3, natančnost: 0.7801206861482134
Število mesecev v učni množici: 4, natančnost: 0.7186514886164623
Število mesecev v učni množici: 5, natančnost: 0.73090105872496
Število mesecev v učni množici: 6, natančnost: 0.7545653761869978
Število mesecev v učni množici: 7, natančnost: 0.7331579622738355
Število mesecev v učni množici: 8, natančnost: 0.7522525620744205
Število mesecev v učni množici: 9, natančnost: 0.69361163057248
Število mesecev v učni množici: 10, natančnost: 0.7352760009476428
Število mesecev v učni množici: 11, natančnost: 0.7127494800194716
Povprečna natančnost: 0.7451627966960608
```



4.1.1.4. Primerjava modelov naučenih na podatkih posameznih regij

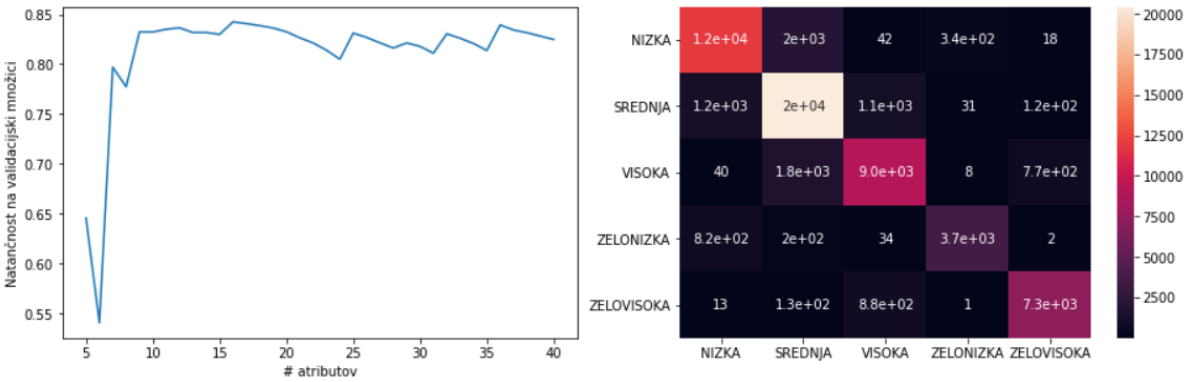
Iz spodnjih matrik zmot lahko opazimo, da je do največ napačnih napovedi prišlo ko se je model učil na podatkih iz vzhodne regije in napovedoval na podatkih iz zahodne ter ko so ti podatki bili v obratnih vlogah (testni/učni), najmanj napak pa ko se je učil na celotnih podatkih in napovedoval na podatkih iz zahodne regije.



4.1.2. Naključni gozdovi

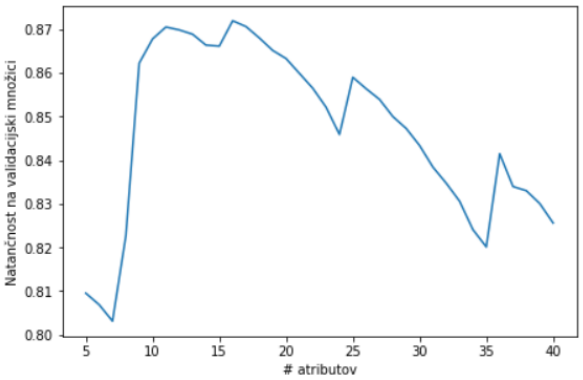
4.1.2.1. Naključni gozdovi (atributi izbrani z metriko relieff)

```
Najboljša natančnost na validacijski množici: 0.8424306610889205 (# atributov: 16)
Natančnost na testni množici: 0.8456541363079503
Izbrani atributi:
['ura', 'stavba', 'povrsina', 'leto_izgradnje', 'temp_zraka', 'temp_rosisca', 'oblacnost', 'pritisk', 'smer_vetra', 'hit
rost_vetra', 'teden', 'dan', 'temp_zraka_MA_7', 'temp_zraka_MA_14', 'temp_rosisca_MA_7', 'pritisk_MA_7']
```



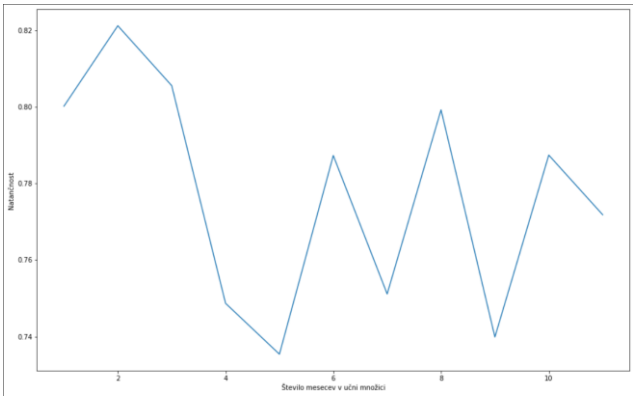
4.1.2.2. Naključni gozdovi (atributi izbrani z metriko Chi^2)

```
Najboljša natančnost na validacijski množici: 0.8719161958683346 (# atributov: 16)
Natančnost na testni množici: 0.8747179057321556
Izbrani atributi:
['ura', 'regija', 'stavba', 'namembnost', 'povrsina', 'leto_izgradnje', 'temp_zraka', 'temp_rosisca', 'smer_vetra', 'mesec', 'teden', 'dan', 'vikend', 'temp_rosisca_MA_7', 'temp_rosisca_MA_14', 'temp_rosisca_MA_21']
```



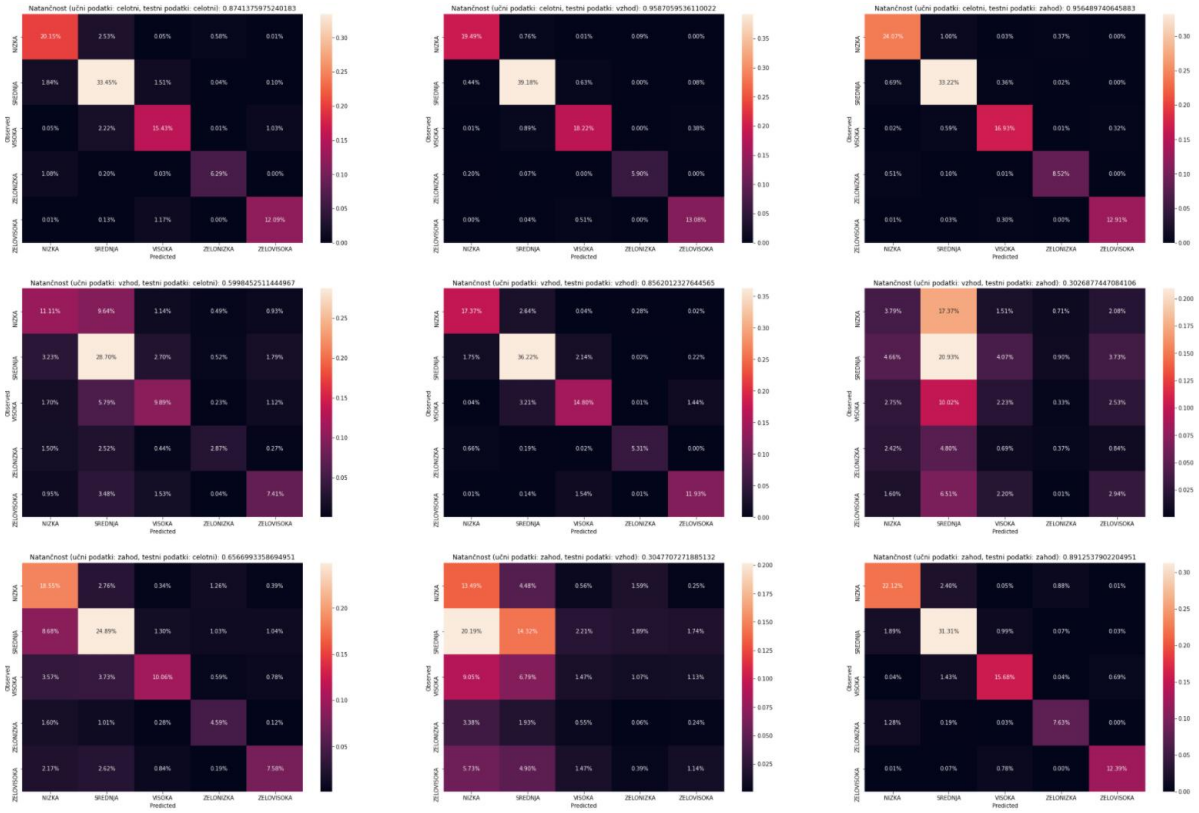
4.1.2.3. Ovrednotenje modela na mesečnih podatkih ({januar} proti {februar}, {januar, februar} proti {marec}, ...)

```
Število mesecev v učni množici: 1, natančnost: 0.8001747608535688
Število mesecev v učni množici: 2, natančnost: 0.8211898580345799
Število mesecev v učni množici: 3, natančnost: 0.8055461387212692
Število mesecev v učni množici: 4, natančnost: 0.7486865148861647
Število mesecev v učni množici: 5, natančnost: 0.7353949272602636
Število mesecev v učni množici: 6, natančnost: 0.7872899926953981
Število mesecev v učni množici: 7, natančnost: 0.7511228025150777
Število mesecev v učni množici: 8, natančnost: 0.7992296581608088
Število mesecev v učni množici: 9, natančnost: 0.7398805573988055
Število mesecev v učni množici: 10, natančnost: 0.7873963515754561
Število mesecev v učni množici: 11, natančnost: 0.7718281187768288
Povprečna natančnost: 0.777067243716202
```



4.1.2.4. Primerjava modelov naučenih na podatkih posameznih regij

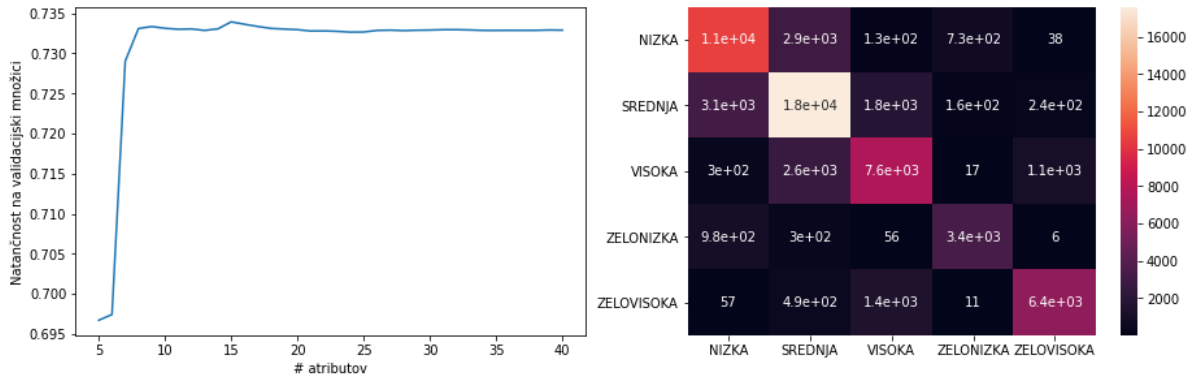
Iz spodnjih matrik zmot lahko opazimo, da je do največ napačnih napovedi prišlo ko se je model učil na podatkih iz vzhodne regije in napovedoval na podatkih iz zahodne ter ko so ti podatki bili v obratnih vlogah (testni/učni), najmanj napak pa ko se je učil na celotnih podatkih in napovedoval na podatkih iz zahodne ali vzhodne regije.



4.1.3. KNN

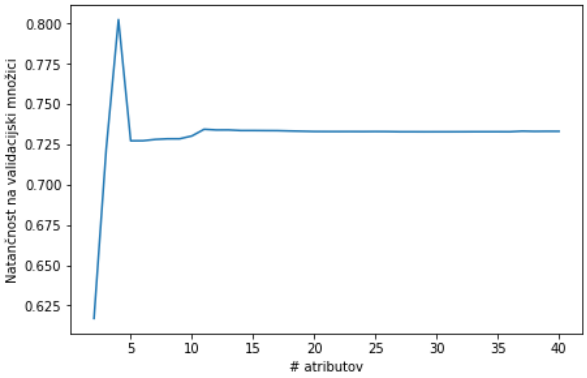
4.1.3.1. KNN (atributi izbrani z metriko Relieff)

```
Najboljša natačnost na validacijski množici: 0.7339394256582525 (# atributov: 15)
Natančnost na testni množici: 0.7342510800180541
Izbrani atributi:
['ura', 'stavba', 'povrsina', 'leto_izgradnje', 'temp_zraka', 'temp_rosisca', 'oblacnost', 'pritisk', 'smer_vetra', 'hitrost_vetra', 'teden',
'dan', 'temp_zraka_MA_7', 'temp_rosisca_MA_7', 'pritisk_MA_7']
```



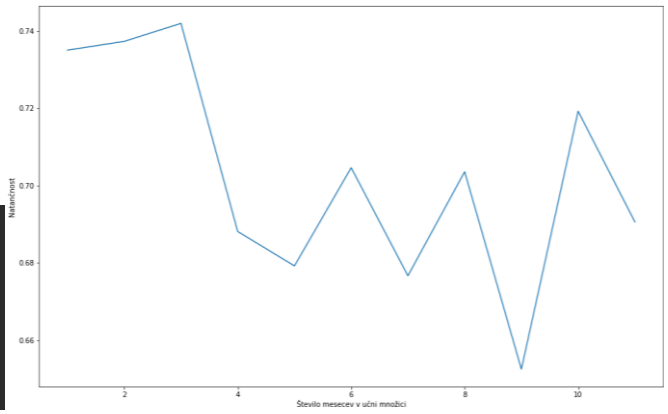
4.1.3.2. KNN (atributi izbrani z metriko Chi^2)

```
Najboljša natačnost na validacijski množici: 0.8024442073846666 (# atributov: 4)
Natančnost na testni množici: 0.8013733960925914
Izbrani atributi:
['ura', 'stavba', 'povrsina', 'dan']
```



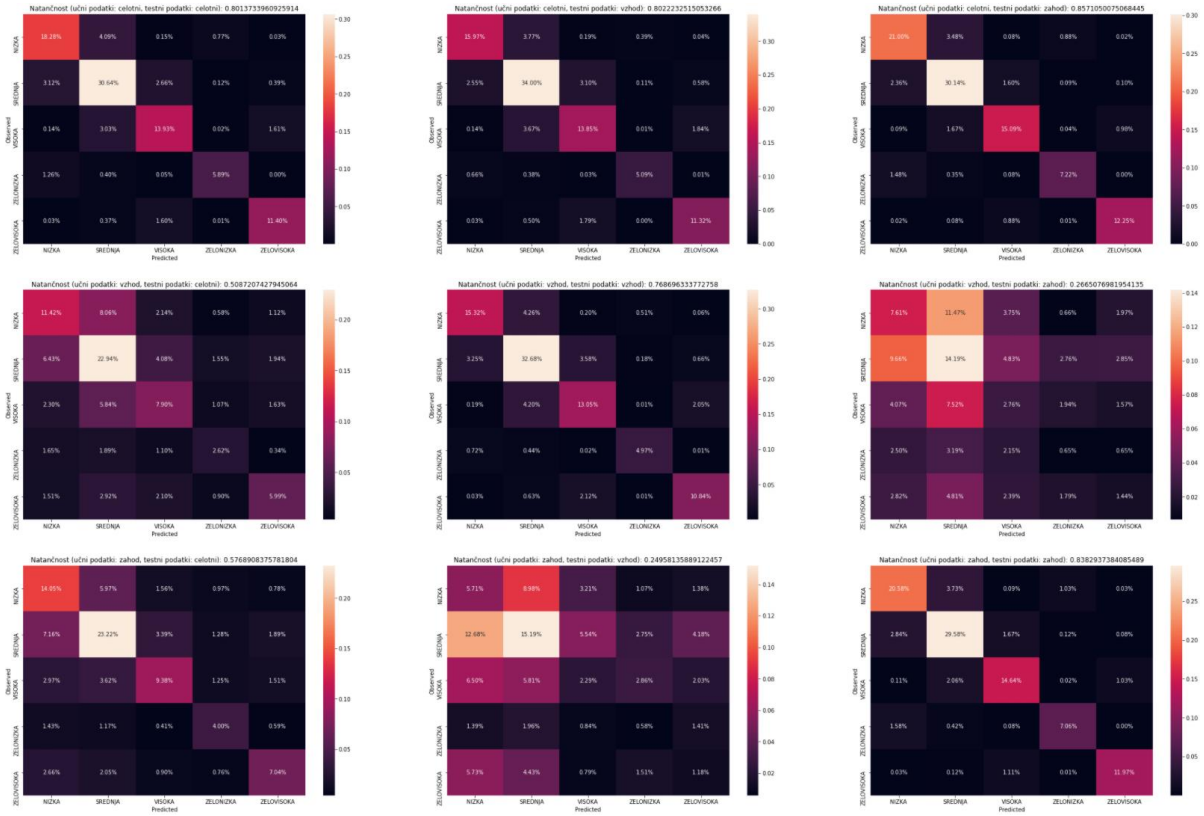
4.1.3.3. Ovrednotenje modela na mesečnih podatkih ({januar} proti {februar}, {januar, februar} proti {marec}, ...)

```
Število mesecev v učni množici: 1, natančnost: 0.7350073583517293
Število mesecev v učni množici: 2, natančnost: 0.7372811136275755
Število mesecev v učni množici: 3, natančnost: 0.741948606685199
Število mesecev v učni množici: 4, natančnost: 0.6880910683012259
Število mesecev v učni množici: 5, natančnost: 0.6791834869373143
Število mesecev v učni množici: 6, natančnost: 0.7046018991964937
Število mesecev v učni množici: 7, natančnost: 0.6766328756576415
Število mesecev v učni množici: 8, natančnost: 0.7035559529541234
Število mesecev v učni množici: 9, natančnost: 0.6524702901610665
Število mesecev v učni množici: 10, natančnost: 0.7192134565268894
Število mesecev v učni množici: 11, natančnost: 0.6905783953622162
Povprečna natančnost: 0.702596773069225
```



4.1.3.4. Primerjava modelov naučenih na podatkih posameznih regij

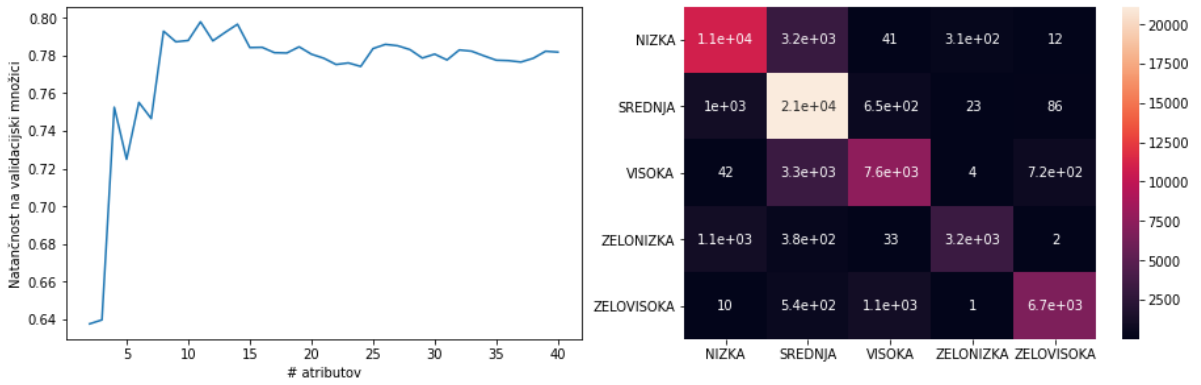
Iz spodnjih matrik zmot lahko opazimo, da je do največ napačnih napovedi prišlo ko se je model učil na podatkih iz vzhodne regije in napovedoval na podatkih iz zahodne ter ko so ti podatki bili v obratnih vlogah (testni/učni), najmanj napak pa ko se je učil na celotnih podatkih in napovedoval na podatkih iz zahodne regije.



4.1.4. Bagging

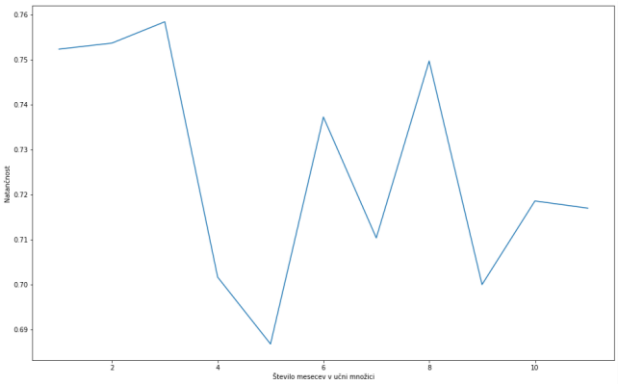
4.1.4.1. Bagging (atributi izbrani z metriko Chi^2)

```
Najboljša natančnost na validacijski množici: 0.7980089193343648 (# atributov: 11)
Natančnost na testni množici: 0.7989715648978013
Izbrani atributi:
['ura', 'stavba', 'namembnost', 'povrsina', 'leto_izgradnje', 'temp_zraka', 'temp_rosisca', 'smer_vetra', 'teden', 'dan', 'vikend']
```



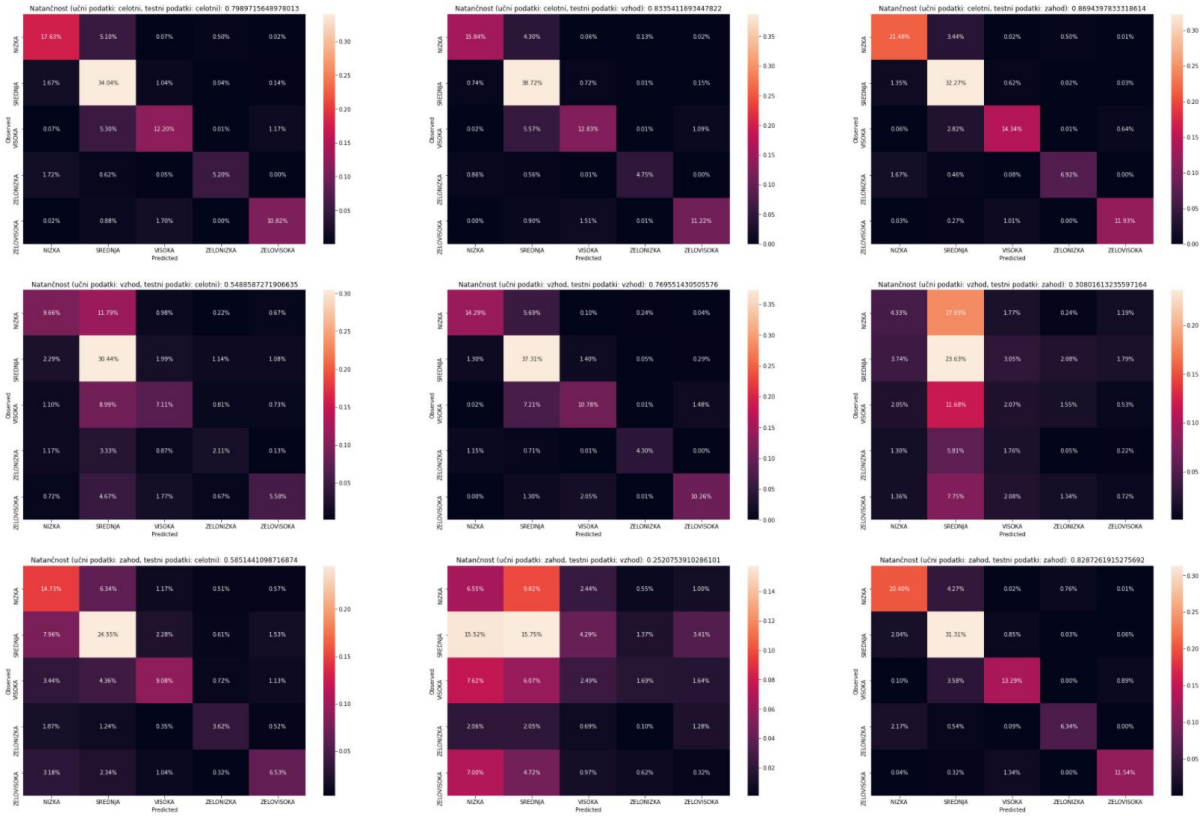
4.1.4.2. Ovrednotenje modela na mesečnih podatkih ({januar} proti {februar}, {januar, februar} proti {marec}, ….)

```
Število mesecev v učni množici: 1, natančnost: 0.7523454746136865
Število mesecev v učni množici: 2, natančnost: 0.753687234160084
Število mesecev v učni množici: 3, natančnost: 0.7584242999525391
Število mesecev v učni množici: 4, natančnost: 0.7016637478108582
Število mesecev v učni množici: 5, natančnost: 0.6868002132683373
Število mesecev v učni množici: 6, natančnost: 0.7372534696859021
Število mesecev v učni množici: 7, natančnost: 0.7103811112536892
Število mesecev v učni množici: 8, natančnost: 0.7497076827842355
Število mesecev v učni množici: 9, natančnost: 0.7000060324546058
Število mesecev v učni množici: 10, natančnost: 0.7185974887467425
Število mesecev v učni množici: 11, natančnost: 0.7169978315705625
Povprečna natačnost: 0.7259876896637493
```



4.1.4.3. Primerjava modelov naučenih na podatkih posameznih regij

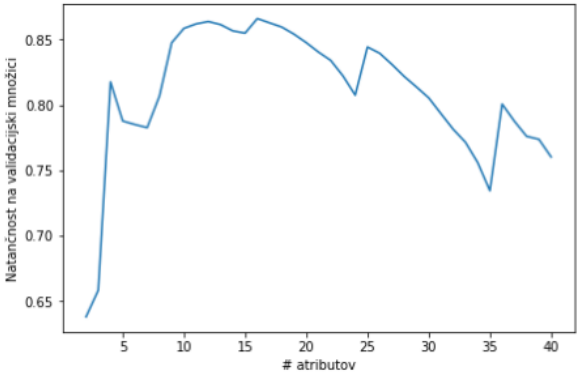
Iz spodnjih matrik zmot lahko opazimo, da je do največ napačnih napovedi prišlo ko se je model učil na podatkih iz vzhodne regije in napovedoval na podatkih iz zahodne ter ko so ti podatki bili zamenjani (testni/učni), najmanj napak pa ko se je učil na celotnih podatkih in napovedoval na podatkih iz zahodne regije



4.1.5. Extra Trees classifier (Extremely Randomized Trees)

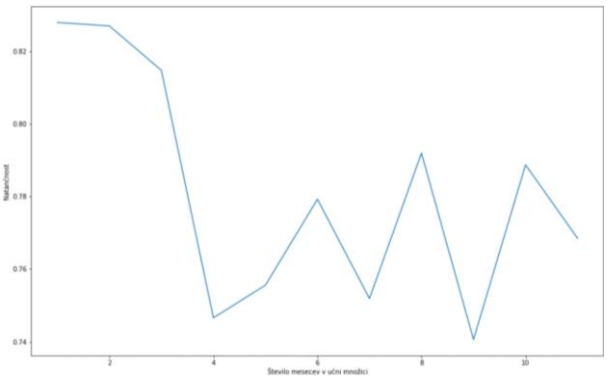
4.1.5.1. Extra Trees classifier (Extremely Randomized Trees) (atributi izbrani z metriko Chi^2)

```
Najboljša natačnost na validacijski množici: 0.8660715957122089 (# atributov: 16)
Natančnost na testni množici: 0.8680604810110258
Izbrani atributi:
['ura', 'regija', 'stavba', 'namembnost', 'povrsina', 'leto_izgradnje', 'temp_zraka', 'temp_rosisca', 'smer_v
etra', 'mesec', 'teden', 'dan', 'vikend', 'temp_rosisca_MA_7', 'temp_rosisca_MA_14', 'temp_rosisca_MA_21']
```



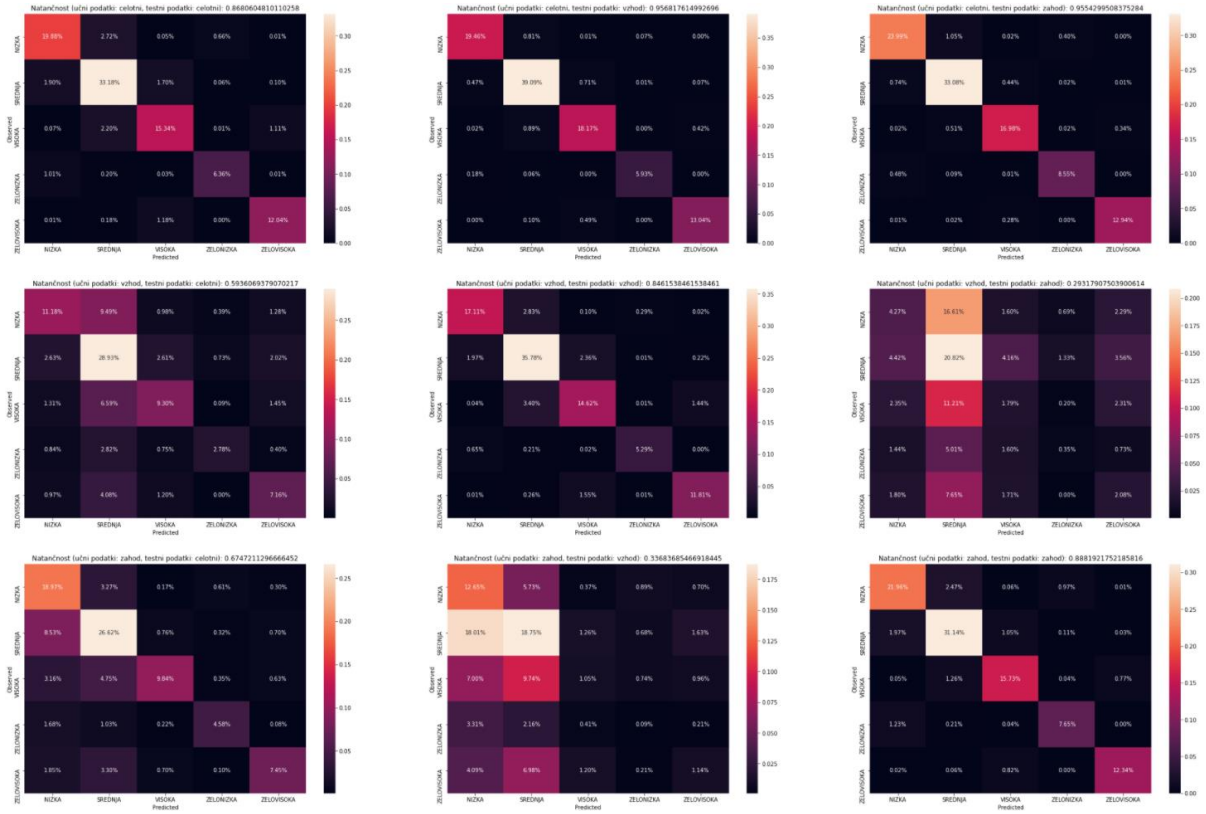
4.1.5.2. Ovrednotenje modela na mesečnih podatkih ({januar} proti {februar}, {januar, februar} proti {marec}, ...)

```
Število mesecev v učni množici: 1, natančnost: 0.8279525386313465
Število mesecev v učni množici: 2, natančnost: 0.8269900016571838
Število mesecev v učni množici: 3, natančnost: 0.8147671028544308
Število mesecev v učni množici: 4, natančnost: 0.746584938704028
Število mesecev v učni množici: 5, natančnost: 0.7555792520374743
Število mesecev v učni množici: 6, natančnost: 0.7792549306062819
Število mesecev v učni množici: 7, natančnost: 0.7518927242397023
Število mesecev v učni množici: 8, natančnost: 0.7919389228970356
Število mesecev v učni množici: 9, natančnost: 0.740604451951499
Število mesecev v učni množici: 10, natančnost: 0.7887230514096186
Število mesecev v učni množici: 11, natančnost: 0.7685090941275391
Povprečna natačnost: 0.7811633644651036
```



4.1.5.3. Primerjava modelov naučenih na podatkih posameznih regij

Iz spodnjih matrik zmot lahko opazimo, da je do največ napačnih napovedi prišlo ko se je model učil na podatkih iz vzhodne regije in napovedoval na podatkih iz zahodne ter ko so ti podatki bili v obratnih vlogah (testni/učni), najmanj napak pa ko se je učil na celotnih podatkih in napovedoval na podatkih iz zahodne ali vzhodne regije.

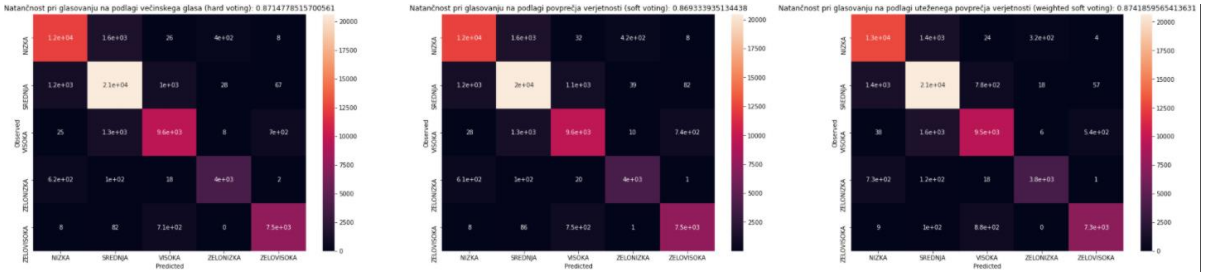


4.1.6. Klasifikacija z glasovanjem in uteženim glasovanjem

Klasifikator je sestavljen iz odločitvenega drevesa, naključnega gozda, KNN klasifikatorja in Extra tree klasifikatorja. Uteži za uteženo glasovanje na podlagi povprečja napovedanih verjetnosti sem izbral z večkratnim poizkušanjem in prišel do uteži (0.4, 0.2, 0.2, 0.3), ki dajejo malenkost boljšo točnost kot posamezni klasifikatorji.

4.1.6.1. Klasifikacija z glasovanjem in uteženim glasovanjem (atributi izbrani z metriko Chi^2)

```
Natančnost pri glasovanju na podlagi večinskega glasa (hard voting): 0.8714778515700561
Natančnost pri glasovanju na podlagi povprečja verjetnosti (soft voting): 0.869333935134438
Natančnost pri glasovanju na podlagi uteženega povprečja verjetnosti (weighted soft voting): 0.8741859565413631
```



4.1.7. Klasifikacija z globoko nevronsko mrežo

4.1.7.1. Klasifikacija z globoko nevronsko mrežo (atributi izbrani z metriko Chi^2)

Mreža je sestavljena iz petih skritih nivojev velikosti(256, 128, 128, 64, 32), do najboljšega rezultata pa sem prišel ko sem učenje izvajal 54 »epoch-ov«.

4.2. Regresija

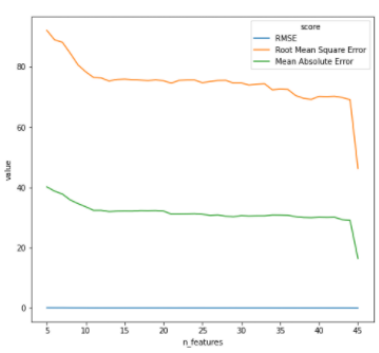
Pri regresijskih problemih sem najprej sestavil modele na osnovi ocene RReliefF in nato še z uporabo ocene f-regression in pri vseh modelih dobil nižji RMSE z uporabo slednje. Za vse modele sem gradil na način, da sem model najprej sestavil na 5 do n najbolje ocenjenih atributih in nato kot končni model izbral tistega ki je imel najnižji RMSE.

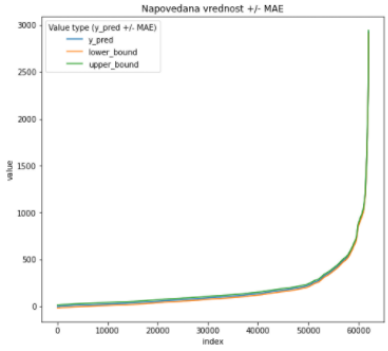
Pri nekaterih vrstah modelov (tisti, ki za učenje potrebujejo veliko časa), pa sem uporabil samo oceno f-regression, saj mi je pri vseh prejšnjih modelih dala nižji RMSE.

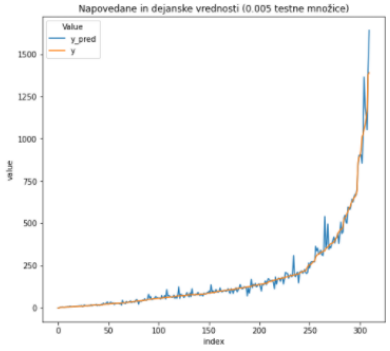
4.2.1. Regresijsko drevo

4.2.1.1. Regresijsko drevo (atributi izbrani z metriko RReliefF)

```
Najboljša RMSE na validacijski množici: 0.035514496017133 16 (# atributov: 45)
RMSE na testni množici: 0.03239550112785522
Root Mean Squared Error na testni množici: 45.758576554969494
Test Mean Absolute Error na testni množici: 16.20007104849632
Izbrani atributi:
['ura', 'stavba', 'povrsina', 'leto_izgradnje', 'temp_zraka', 'temp_rosisca', 'oblacnost', 'padavine', 'pritisk', 'smer_vetra', 'hitrost_
vetra', 'mesec', 'teden', 'vikend', 'temp_zraka_MA_7', 'temp_zraka_MA_14', 'temp_zraka_MA_21', 'temp_zraka_MA_28', 'temp_rosisca_MA_7',
'temp_rosisca_MA_14', 'temp_rosisca_MA_21', 'temp_rosisca_MA_28', 'oblacnost_MA_7', 'oblacnost_MA_14', 'oblacnost_MA_21', 'oblacnost_MA_2
8', 'padavine_MA_7', 'padavine_MA_14', 'padavine_MA_21', 'padavine_MA_28', 'pritisk_MA_7', 'pritisk_MA_14', 'pritisk_MA_21', 'pritisk_MA_
28', 'hitrost_vetra_MA_7', 'hitrost_vetra_MA_14', 'hitrost_vetra_MA_21', 'hitrost_vetra_MA_28', 'regija_vzhodna', 'regija_zahodna', 'name
mbnost_izobrazevalna', 'namembnost_javno_storitvena', 'namembnost_kulturno_razvedrilna', 'namembnost_poslovna', 'namembnost_stanovanjsk
a']
```

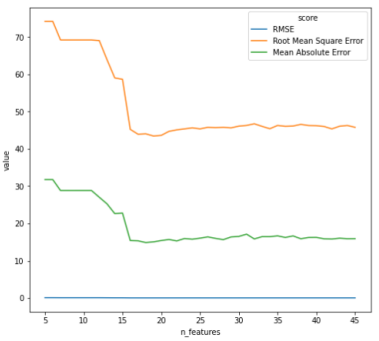


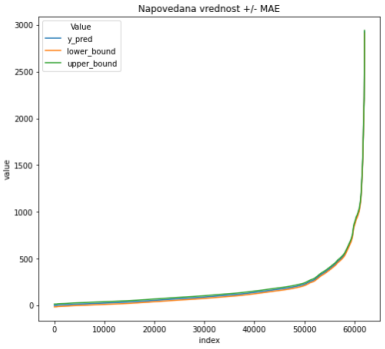


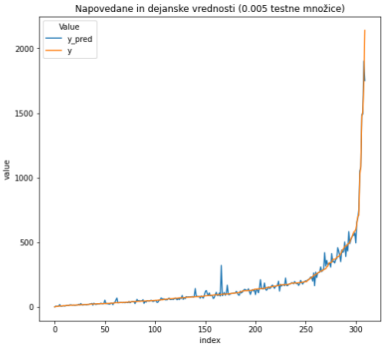


4.2.1.2. Regresijsko drevo (atributi izbrani z metriko f-regression)

```
Najboljša RMSE na validacijski množici: 0.03099734511558258 (# atributov: 18)
RMSE na testni množici: 0.02759171769855038
Root Mean Squared Error na testni množici: 42.22984854241357
Test Mean Absolute Error na testni množici: 14.703793668432422
Izbrani atributi:
['ura', 'stavba', 'povrsina', 'leto_izgradnje', 'temp_rosisca', 'mesec', 'teden', 'dan', 'vikend', 'temp_rosisca_MA_7', 'temp_rosisca_MA_14', 're
gija_vzhodna', 'regija_zahodna', 'namembnost_izobrazevalna', 'namembnost_javno_storitvena', 'namembnost_kulturno_razvedrilna', 'namembnost_poslov
na', 'namembnost_stanovanjska']
```

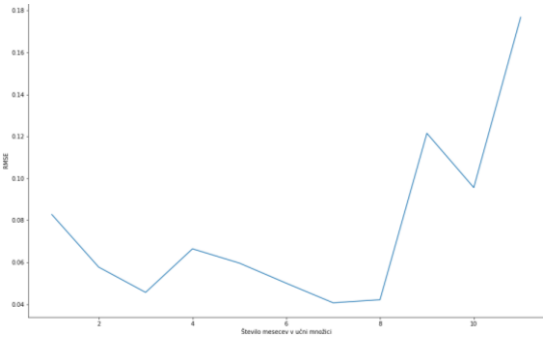






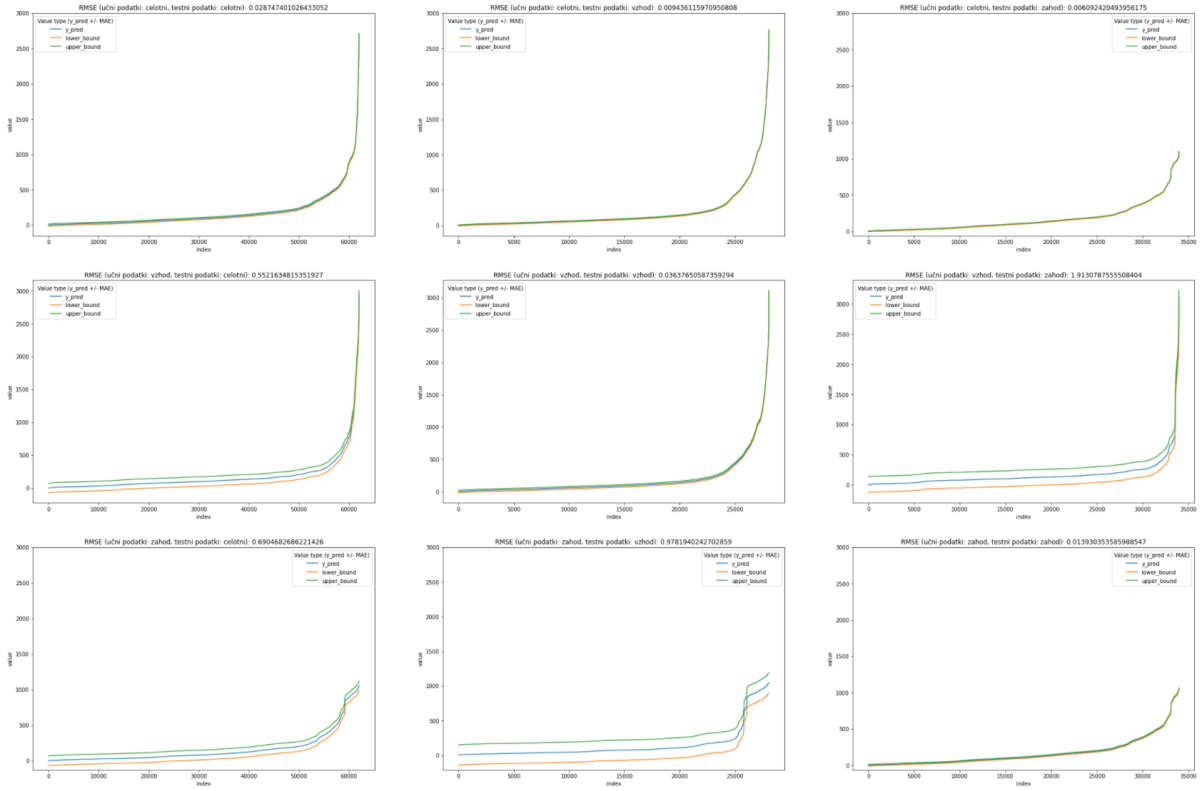
4.2.1.3. Ovrednotenje modela na mesečnih podatkih ({januar} proti {februar}, {januar, februar} proti {marec}, ...)

```
Število mesecev v učni množici: 1, RMSE: 0.08278323310669009
Število mesecev v učni množici: 2, RMSE: 0.0577107331531948
Število mesecev v učni množici: 3, RMSE: 0.04566540640854786
Število mesecev v učni množici: 4, RMSE: 0.06643814086437509
Število mesecev v učni množici: 5, RMSE: 0.059641423849880526
Število mesecev v učni množici: 6, RMSE: 0.05005173984847262
Število mesecev v učni množici: 7, RMSE: 0.04071076065789918
Število mesecev v učni množici: 8, RMSE: 0.04218477543260872
Število mesecev v učni množici: 9, RMSE: 0.12149651332433654
Število mesecev v učni množici: 10, RMSE: 0.09560459103714727
Število mesecev v učni množici: 11, RMSE: 0.17679597304704162
Povprečna RMSE: 0.0762802991572904
```



4.2.1.4. Primerjava modelov naučenih na podatkih posameznih regiji

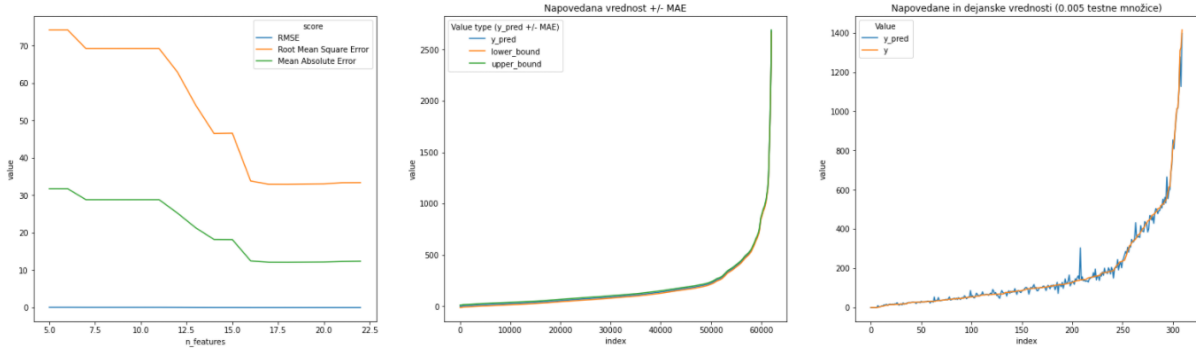
Podobno kot pri klasifikacijskih modelih so tudi tukaj najboljši rezultati ko se model uči na celotnih podatkih in napoveduje posamezno regijo, najslabši pa ko se model uči in napoveduje nasprotni regiji.



4.2.2. Regresijski naključni gozdovi

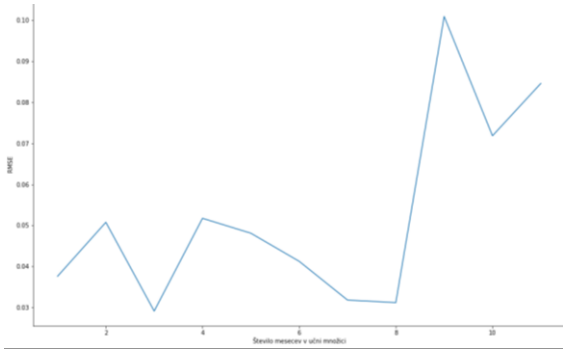
4.2.2.1. Regresijski naključni gozdovi (atributi izbrani z metriko f-regression)

```
Najboljša RMSE na validacijski množici: 0.018023799468503753 (# atributov: 19)
RMSE na testni množici: 0.01579708925597358
Root Mean Squared Error na testni množici: 31.95351881050272
Test Mean Absolute Error na testni množici: 11.922838675061064
Izbrani atributi:
['ura', 'stavba', 'povrsina', 'leto_izgradnje', 'temp_rosisca', 'mesec', 'teden', 'dan', 'vikend', 'temp_rosisca_MA_7', 'temp_rosisc
a_MA_14', 'temp_rosisca_MA_21', 'regija_vzhodna', 'regija_zahodna', 'namembnost_izobrazevalna', 'namembnost_javno_storitvena', 'name
mbnost_kulturno_razvedrilna', 'namembnost_poslovna', 'namembnost_stanovanjska']
```



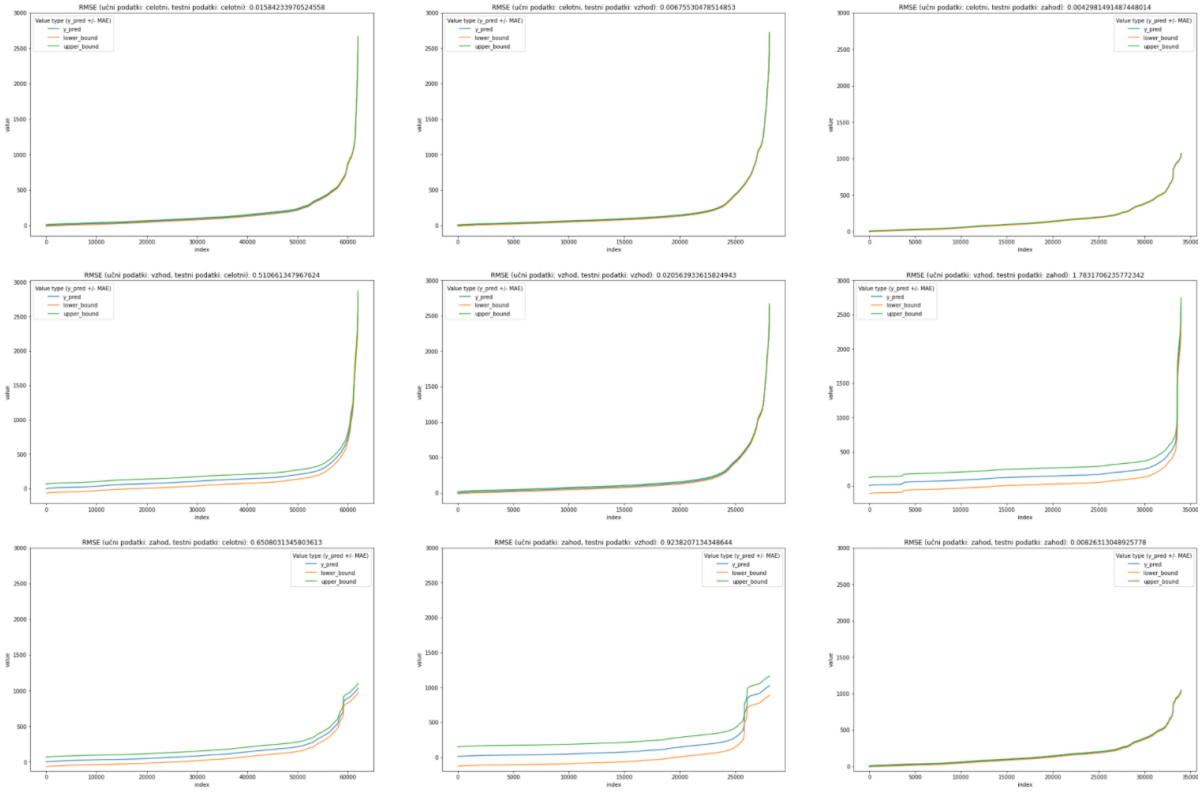
4.2.2.2. Ovrednotenje modela na mesečnih podatkih ({januar} proti {februar}, {januar, februar} proti {marec}, ...)

```
Število mesecev v učni množici: 1, RMSE: 0.037611347526648714
Število mesecev v učni množici: 2, RMSE: 0.05076325669992952
Število mesecev v učni množici: 3, RMSE: 0.029142868692872168
Število mesecev v učni množici: 4, RMSE: 0.051736281324047714
Število mesecev v učni množici: 5, RMSE: 0.04810609489249359
Število mesecev v učni množici: 6, RMSE: 0.04126426458654173
Število mesecev v učni množici: 7, RMSE: 0.03183426779276877
Število mesecev v učni množici: 8, RMSE: 0.03121236982045397
Število mesecev v učni množici: 9, RMSE: 0.10087912926908102
Število mesecev v učni množici: 10, RMSE: 0.07182432025460381
Število mesecev v učni množici: 11, RMSE: 0.08457207508351593
Povprečna RMSE: 0.05263147963117791
```



4.2.2.3. Primerjava modelov naučenih na podatkih posameznih regiji

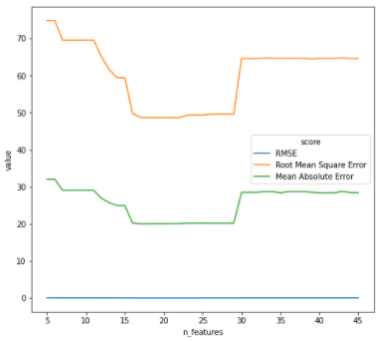
Iz spodnje matrike zmot lahko opazimo, da se tudi regresijski naključni gozdovi najslabše obnesejo ko jih učimo na vzhodnih podatkih napovedujemo pa na zahodnih in obratno, najboljše pa ko jih učimo na vseh podatkih in napovedujemo posamezno regijo, le za malenkost slabši rezultat pa dobimo ko pa jih učimo na zahodnih podatkih in na njih tudi napovedujemo.

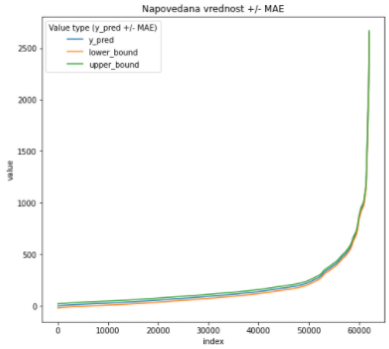


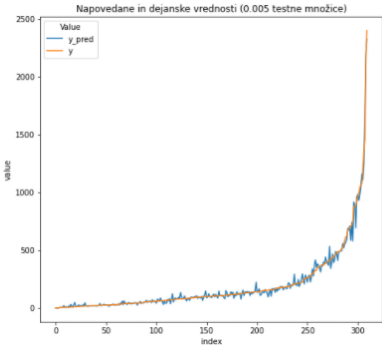
4.2.3. KNN regresija

4.2.3.1. KNN regresija (atributi izbrani z metriko f-regression)

```
Najboljša RMSE na validacijski množici: 0.039032476069527165 (# atributov: 17)
RMSE na testni množici: 0.03735721714102302
Root Mean Squared Error na testni množici: 49.13799290124059
Test Mean Absolute Error na testni množici: 19.96858449932297
Izbrani atributi:
['ura', 'stavba', 'povrsina', 'leto_izgradnje', 'temp_rosisca', 'teden', 'dan', 'vikend', 'temp_rosisca_MA_7', 'temp_rosisca_MA_14', 'regija_vzhodna', 'regija_zahodna', 'namembnost_izobrazevalna', 'namembnost_javno_storitvena', 'namembnost_kulturno_razvedrilna', 'namembnost_poslovna', 'namembnost_stanovanjska']
```

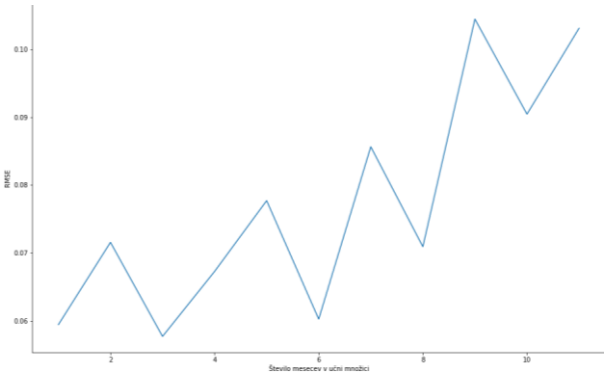




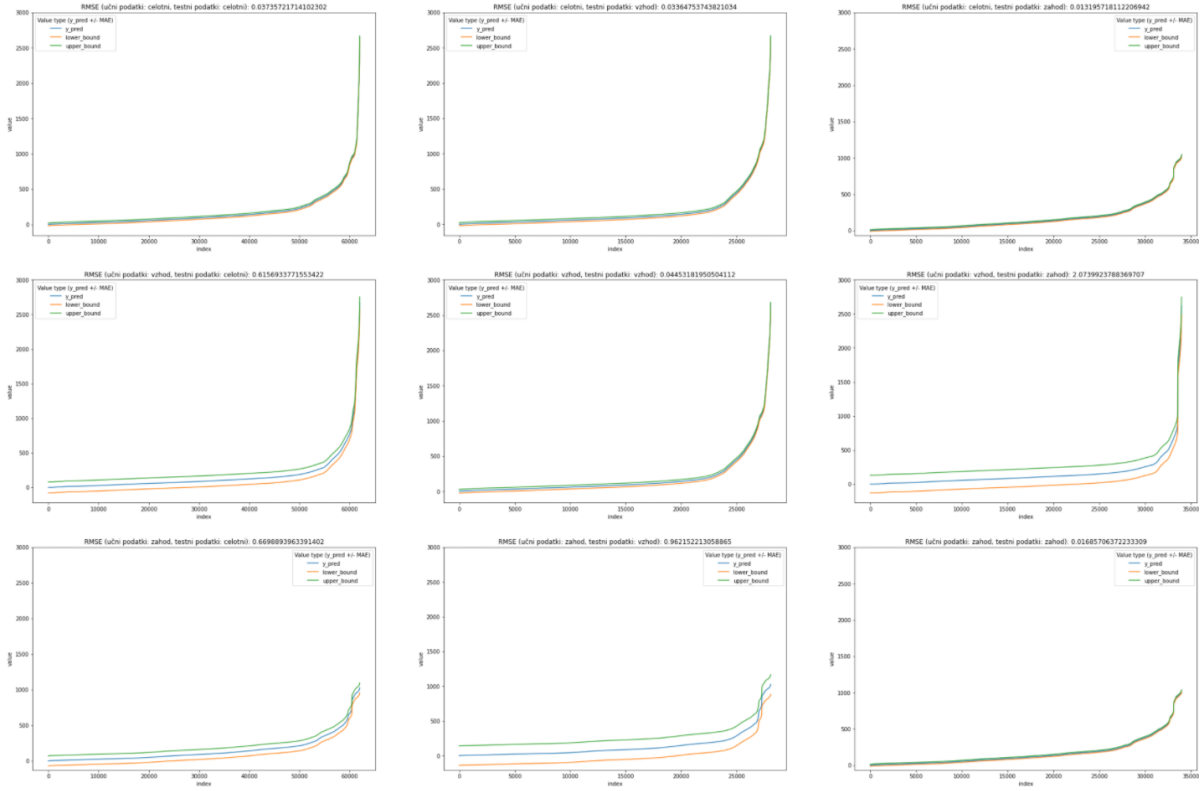


4.2.3.2. Ovrednotenje modela na mesečnih podatkih ({januar} proti {februar}, {januar, februar} proti {marec}, ...)

```
Število mesecev v učni množici: 1, RMSE: 0.05944988957569023
Število mesecev v učni množici: 2, RMSE: 0.07157149643414498
Število mesecev v učni množici: 3, RMSE: 0.057713932429109646
Število mesecev v učni množici: 4, RMSE: 0.06728539716908996
Število mesecev v učni množici: 5, RMSE: 0.07771464213283569
Število mesecev v učni množici: 6, RMSE: 0.06026289037323355
Število mesecev v učni množici: 7, RMSE: 0.0856493060318153
Število mesecev v učni množici: 8, RMSE: 0.07092901241897932
Število mesecev v učni množici: 9, RMSE: 0.10445569872207636
Število mesecev v učni množici: 10, RMSE: 0.09045127955445796
Število mesecev v učni množici: 11, RMSE: 0.10310552009743798
Povprečna RMSE: 0.07714446044898828
```



4.2.3.3. Primerjava modelov naučenih na podatkih posameznih regiji



4.2.4. Regresija z glasovanjem in uteženim glasovanjem

Model je sestavljen iz regresijskega drevesa, regresijskega naključnega gozda in KNN regresije. Uteži za uteženo glasovanje na podlagi povprečja napovedanih verjetnosti sem izbral z večkratnim poizkušanjem in prišel do uteži (0.8, 0.1, 0.1), ki dajejo malenkost slabši rezultat kot če uporabimo samo regresijski naključni gozd.

4.2.4.1. Regresija z glasovanjem in uteženim glasovanjem (atributi izbrani z metriko f-regression)

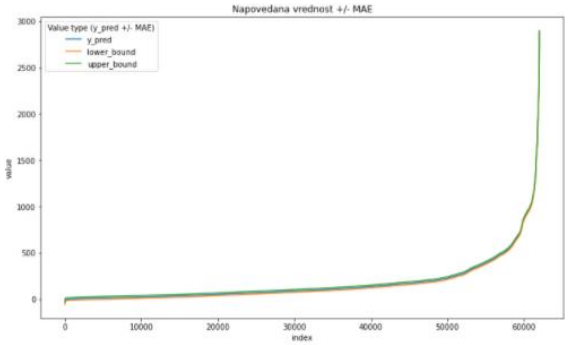
```
RMSE pri glasovanju na podlagi povprečja verjetnosti (soft voting): 0.018759602374150343
RMSE pri glasovanju na podlagi uteženega povprečja verjetnosti (weighted soft voting): 0.01594005447256969
```

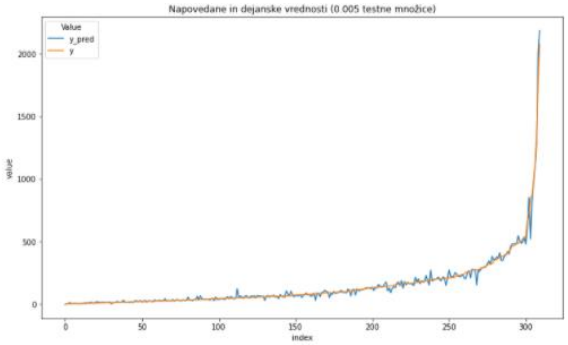
4.2.5. Regresija z globoko nevronsko mrežo

4.2.5.1. Regresija z globoko nevronsko mrežo (atributi izbrani z metriko f-regression)

Mreža je sestavljena iz štirih skritih nivojev velikosti(512, 256, 256, 256), do najboljšega rezultata pa sem prišel ko sem učenje izvajal 100 »epoch-ov«.

```
RMSE na testni množici: 0.02185764041000098
Root Mean Squared Error na testni množici: 37.586491914799375
Test Mean Absolute Error na testni množici: 15.116510818633975
Izbrani atributi (# atributov: 23):
['ura', 'stavba', 'povrsina', 'leto_izgradnje', 'temp_rosisca', 'pritisk', 'mesec', 'teden', 'dan', 'vikend', 'temp_rosisca_MA_
7', 'temp_rosisca_MA_14', 'temp_rosisca_MA_21', 'temp_rosisca_MA_28', 'hitrost_vetra_MA_21', 'hitrost_vetra_MA_28', 'regija_vzh
odna', 'regija_zahodna', 'namembnost_izobrazevalna', 'namembnost_javno_storitvena', 'namembnost_kulturno_razvedrilna', 'namembn
ost_poslovna', 'namembnost_stanovanjska']
```





5. Zaključek

Na vseh grafih, ki predstavljajo uspešnost (naj bo to RMSE ali natančnost) modela pri validaciji po mesecih lahko opazimo, da uspešnost začne padati, ko se model začne učiti tudi na spomladanskih mesecih, se čez poletje nekoliko izboljša, ter potem spet pade ko se model začne učiti tudi na jesenskih mesecih. Sklepam, da je eden od razlogov za to obnašanje sprememba okolijskih spremenljivk (temperatura zraka, temperatura rosišča, …), saj na primer ko se model uči na podatkih od januarja do marca, napoveduje pa za april, ki je načeloma toplejši, z več dežja, itd. od prejšnjih mesecev, se uspešnost skoraj pri vseh modelih enako poslabša, seveda pa je vsak model različno občutljiv.