



Univerza v Ljubljani
Fakulteta *za računalništvo*
in informatiko

Umetna Inteligenca

Seminarska naloga

Avtorja:
Dragan Spasovski
Domen Žukovec

Datum: 6. 12. 2020

Pregled podatkov

Najprej sva uvozila in pregledala osnovne podatke v datoteki, ki sva jo dobila pri nalogi.

```
podatki <- read.table("dataSem1.txt",sep=";",header=T, stringsAsFactors = T)
summary(podatki)
```

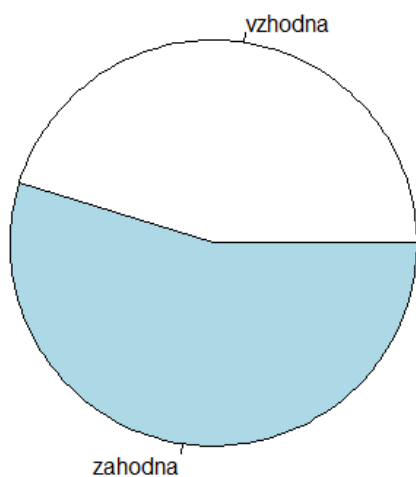
Dobila sva naslednje rezultate.

datum	ura	regija	stavba	namembnost	povrsina	
2016-01-01:	772	Min. : 5.00	vzhodna: 93555	Min. : 1.00	izobrazevalna :105293	Min. : 329.3
2016-01-02:	772	1st Qu.: 5.00	zahodna:113230	1st Qu.: 39.00	javno_storitvena : 27219	1st Qu.: 3445.1
2016-01-03:	772	Median :11.00		Median : 77.00	kulturno_razvedrilna: 29293	Median : 6619.1
2016-01-04:	772	Mean :13.98		Mean : 86.16	poslovna : 26228	Mean :10575.4
2016-01-06:	772	3rd Qu.:23.00		3rd Qu.:133.00	stanovanjska : 18752	3rd Qu.:12733.8
2016-01-07:	772	Max. :23.00		Max. :193.00		Max. :79000.4
(other) :202153						
leto_izgradnje	temp_zraka	temp_rosisca	oblacnost	padavine	pritisk	smer_vetra
Min. :1900	Min. : -10.00	Min. : -22.800	Min. : 0.000	Min. : -1.0000	Min. : 991.9	Min. : 0.0
1st Qu.:1949	1st Qu.: 10.00	1st Qu.: -2.800	1st Qu.: 0.000	1st Qu.: 0.0000	1st Qu.:1009.8	1st Qu.: 90.0
Median :1968	Median : 19.40	Median : 2.800	Median : 4.000	Median : 0.0000	Median :1014.1	Median :170.0
Mean :1968	Mean : 18.91	Mean : 3.877	Mean : 3.397	Mean : 0.2634	Mean :1015.1	Mean :170.3
3rd Qu.:1993	3rd Qu.: 27.80	3rd Qu.: 10.600	3rd Qu.: 6.000	3rd Qu.: 0.0000	3rd Qu.:1019.9	3rd Qu.:270.0
Max. :2017	Max. : 46.10	Max. : 25.000	Max. : 9.000	Max. :56.0000	Max. :1040.9	Max. :360.0
hitrost_vetra	poraba	norm_poraba				
Min. : 0.000	Min. : 0.00	NIZKA :48335				
1st Qu.: 2.100	1st Qu.: 38.22	SREDNJA :76435				
Median : 3.100	Median : 92.70	VISOKA :38660				
Mean : 3.432	Mean : 173.57	ZELONIZKA :15945				
3rd Qu.: 4.600	3rd Qu.: 185.61	ZELOVISOKA:27410				
Max. :14.900	Max. :3095.44					

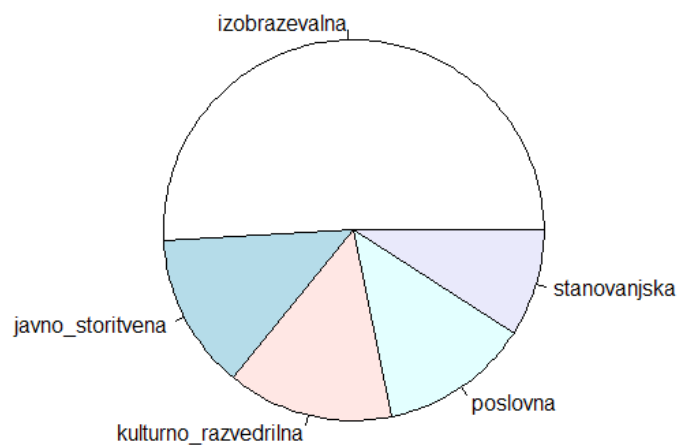
Iz ukaza summary je razvidno, da imamo na voljo 16 atributov, katerih je 12 zveznih in 4 diskretnih. Lotila sva se vizualizacije atributov, ki so se nama zdeli pomembni. Probala sva ugotoviti, kje se nahaja največja povezava med porabo in določenim atributom

Vizualizacija

Najprej naju je zanimalo ali je število meritev med regijami enakomerno razdeljeno.

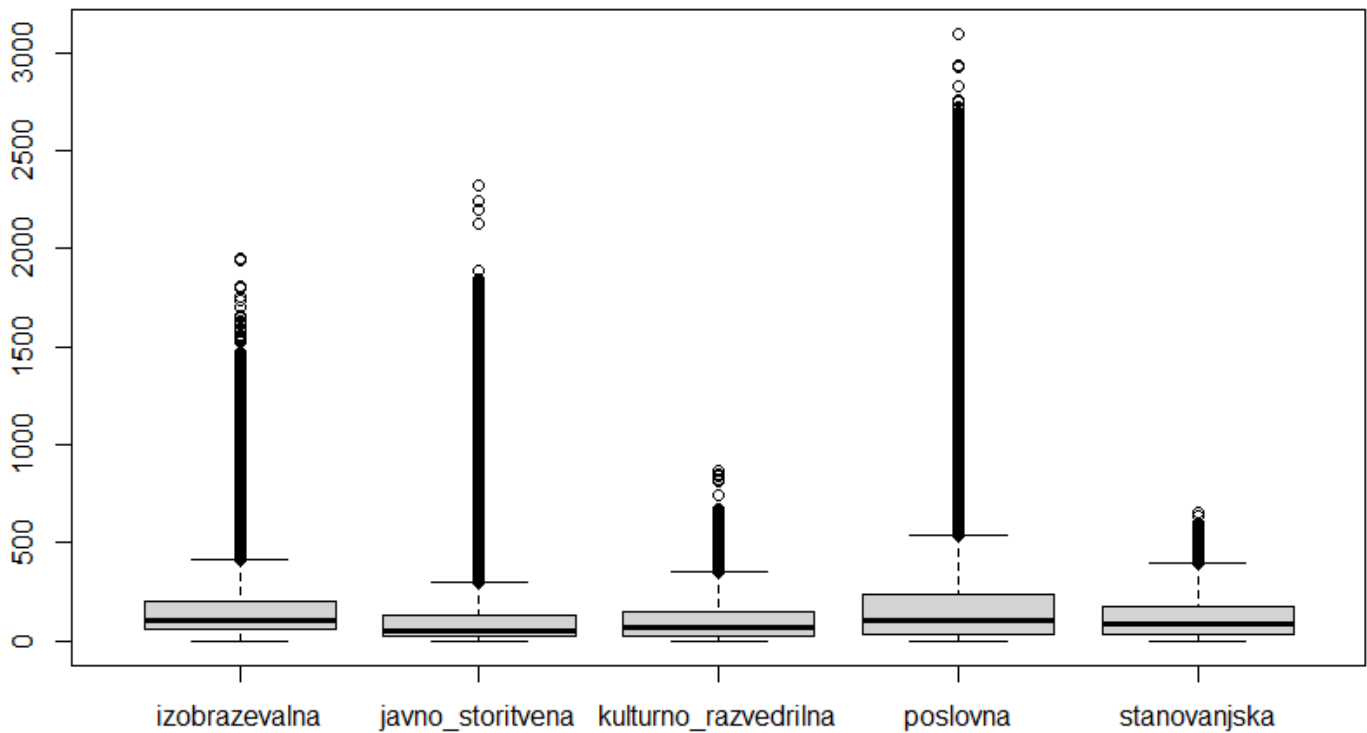


Nato med namembnosti.

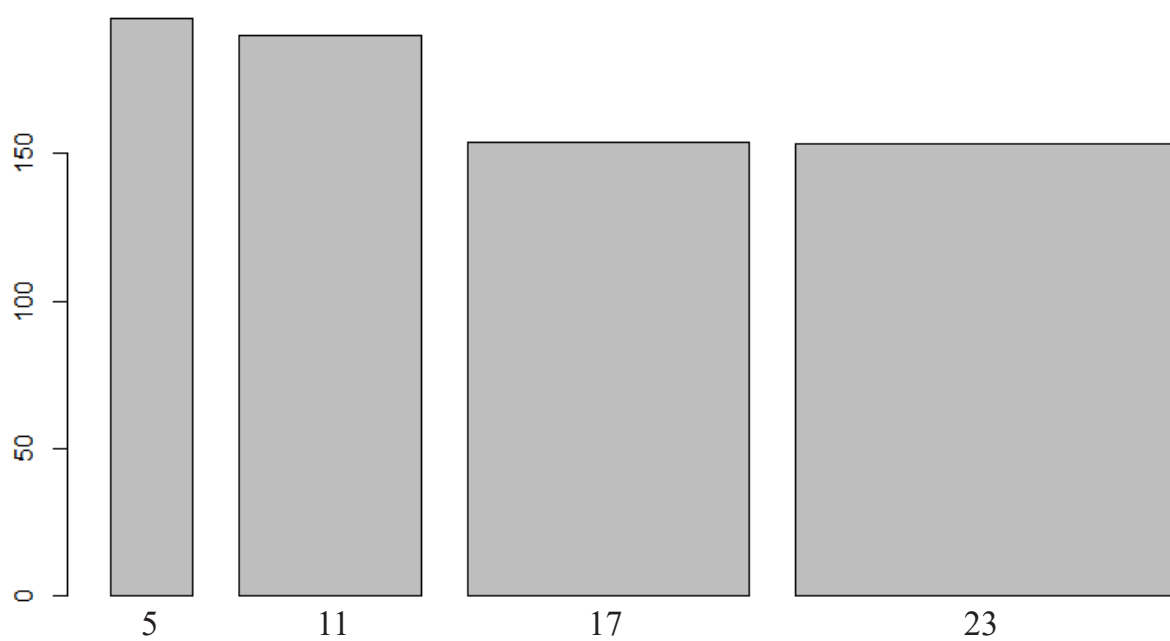


Vizualizacija

Videla sva, da je več kot polovica meritev iz namembnostih stavb. Pogledala sva še če je povezava med porabo in namembnostjo.



Očitno je, da poslovne stavbe imajo največjo porabo.



Stavbe ponavadi porabijo največ dopoldne.

Evalvacija atributov

Za evalvacijo atributov sva uporabila "MDL", "InfGain", "ReliefEqualK". Po ukazu attrEval sva dobila boljšo idejo kateri atributi so pomembni in katere bi mogla vpoštevati pri najinih testih

```
sort(attrEval(norm_poraba ~ .,podatki,"MDL"), decreasing=T)
  poraba leto_izgradnje namembnost površina ura stavba temp_rosisca regija
1. 352364e-01 5.747618e-02 5.437178e-02 2.394871e-02 2.343338e-02 1.687218e-02 1.103000e-02 5.840108e-03
  datum temp_zraka pritisk smer_vetra oblacnost hitrost_vetra padavine
3. 173134e-03 2.558398e-03 1.668027e-03 1.289557e-03 4.428373e-04 1.135268e-04 5.421448e-05
```

```
sort(attrEval(norm_poraba ~ .,podatki,"InfGain"), decreasing=T)
  poraba leto_izgradnje namembnost datum površina ura stavba temp_rosisca
0. 1353764256 0.0576120974 0.0548670491 0.0289169413 0.0240757138 0.0235681488 0.0170020276 0.0111544847
  regija temp_zraka pritisk smer_vetra oblacnost hitrost_vetra padavine
0. 0059742970 0.0026924803 0.0017944931 0.0014230198 0.0005750612 0.0001852129 0.0001569170
```

```
> sort(attrEval(norm_poraba ~ .,podatki,"ReliefEqualK"), decreasing=T)
  namembnost leto_izgradnje površina poraba stavba regija padavine ura
0. 1756139219 0.1755928525 0.1551206946 0.1208954587 0.1208553772 0.0006403402 -0.0004163083 -0.0020136406
  temp_zraka pritisk hitrost_vetra temp_rosisca oblacnost smer_vetra datum
-0. 0163807487 -0.0192241938 -0.0232825075 -0.0250257193 -0.0288196767 -0.0343903390 -0.1044915923
```

attrEval pokaže, da razen porabe, so leto izgradnje, namembnost, regija, površina in ura pomembni atributi. Ker je sedanja oblika atributa datum dokaj neuporabna, sva hotela pri dodajanju atributov iz njega izluščiti čim več uporabnih podatkov.

Dodajanje atributov

Za dodatne attribute sva se najprej lotila izluščevanja atributov iz datuma. Naredila sva dodatni atribut Mesec, Dan, Weekend in LetniCas.

```
##### ATRIBUT DAN

podatki$Dan <- as.factor(weekdays(as.Date(podatki$datum)))

##### ATRIBUT VIKEND

sel <- podatki$Dan %in% c("Saturday","Sunday")

podatki$weekend <- FALSE
podatki$weekend[sel] <- TRUE

podatki$weekend <- as.factor(podatki$weekend)

##### Letni Casi

podatki$LetniCas <- "nedoloceno"
sel <- podatki$Mesec %in% c(1,2, 3)
podatki$LetniCas[sel] <- "zima"
sel <- podatki$Mesec %in% c(4,5, 6)
podatki$LetniCas[sel] <- "pomlad"
sel <- podatki$Mesec %in% c(7,8, 9)
podatki$LetniCas[sel] <- "poletje"
sel <- podatki$Mesec %in% c(10,11, 12)
podatki$LetniCas[sel] <- "jesen"
podatki$LetniCas <- as.factor(podatki$LetniCas)

#####
```

Nato sva dodala še atributa avgPoraba in avgNorm_Poraba.

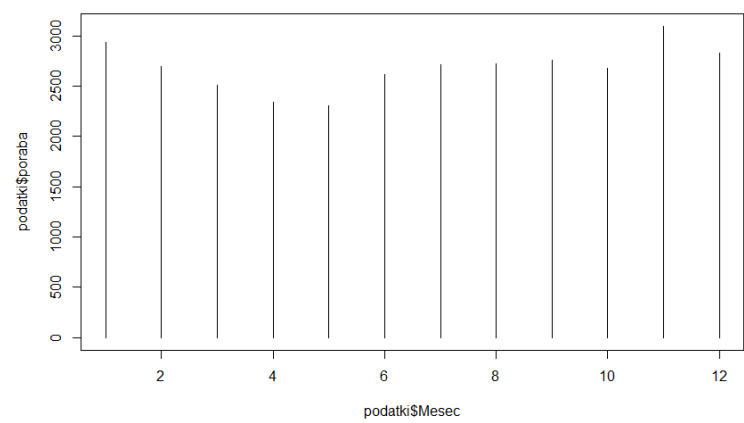
```
##### POVPREČNA PORABA IN POVPREČNA NORM PORABA

for(x in 1:max(podatki$stavba))
{
  sel <- podatki$stavba == x

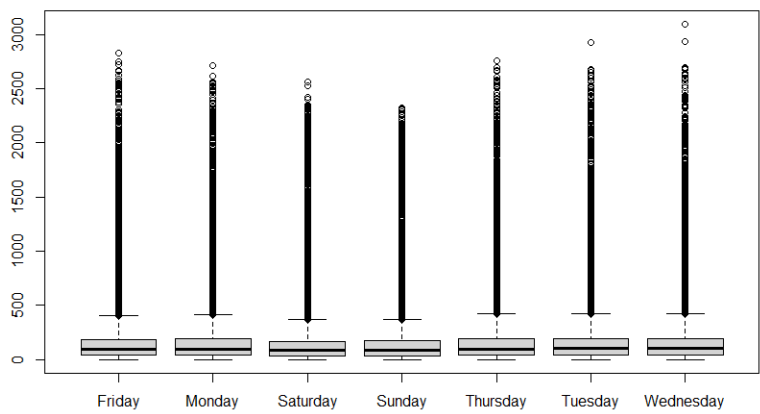
  for(i in 1:nrow(podatki[sel,]))
  {
    if(i <= 20)
    {
      podatki$avgNorm_Poraba[sel][i] <- names(sort(table(podatki$norm_poraba[sel][1:i]), decreasing = T))[1]
      podatki$avgPoraba[sel][i] <- mean(podatki$poraba[sel][1:i])
    }
    else
    {
      podatki$avgNorm_Poraba[sel][i] <- names(sort(table(podatki$norm_poraba[sel][i-20:i]), decreasing = T))[1]
      podatki$avgPoraba[sel][i] <- mean(podatki$poraba[sel][i-20:i])
    }
  }
}
```

Visualizacija dodatnih atributov

Atribut Mesec

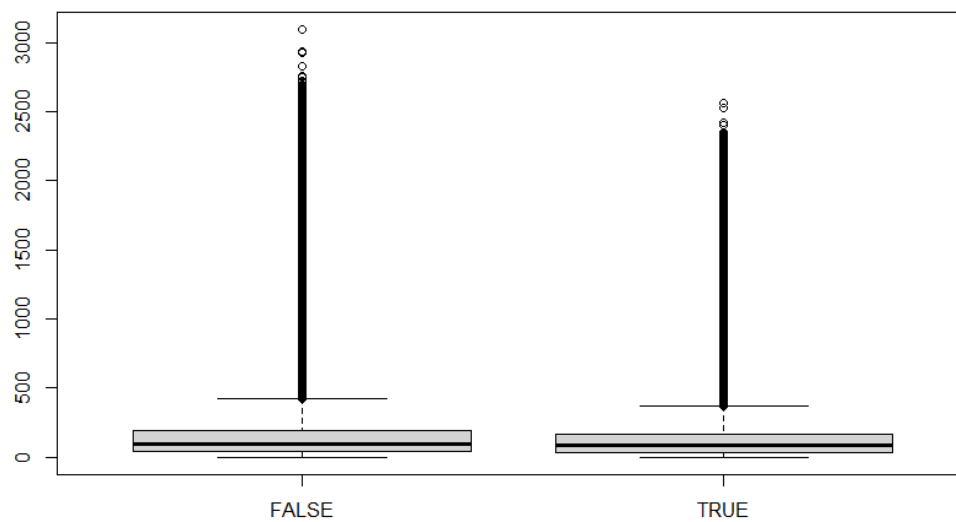


Atribut Dan

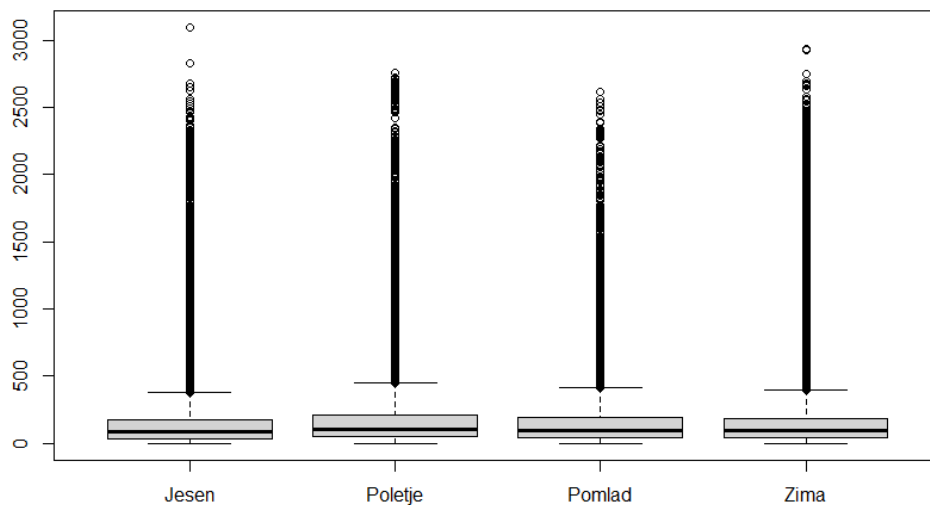


Visualizacija dodatnih atributov

Atribut Weekend

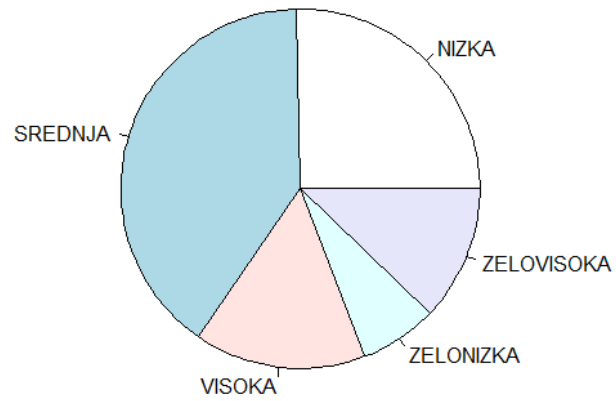


Atribut Letni Čas



Visualizacija dodatnih atributov

Atribut avgNorm_Poraba



Klasifikacijski del

Evalvacija osnovnih podatkov

Za učne modele sem se odločil za Tree, Random Forest in Naive Bayes. Testiral sem jih na nespremenjenih podatkih z atributi leto izgradnje, namembnost, površina, ura, stavba in regija ter dobil takšne rezultate

CA

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
1	0.7893212	0.7533558	0.7397112	0.6750438	0.6022546	0.6014609	0.5933530	0.6456428	0.6880014	0.6773276	0.6861088
2	0.7904249	0.7530244	0.7387619	0.6748687	0.6017975	0.6029218	0.5926472	0.6466745	0.6875189	0.6770434	0.6860203
3	0.4282561	0.4310335	0.4078243	0.4105079	0.3884530	0.3818115	0.3737970	0.3795309	0.4001930	0.4084814	0.4173120

Brier score

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
1	0.3186583	0.3605960	0.3729109	0.4433928	0.5596178	0.5498953	0.5398664	0.4639194	0.4380036	0.4358538	0.4324368
2	0.2985114	0.3512912	0.3606522	0.4430288	0.5516562	0.5358248	0.5280302	0.4476206	0.4255349	0.4285935	0.4232783
3	0.6872003	0.6872845	0.6934367	0.7079058	0.7191025	0.7262549	0.7363058	0.7244522	0.7113207	0.7102620	0.7078510

Iz dobljenih rezultatov je razvidno, da Tree (1) in Random forest (2) veliko bolje kot Naive Baye (3). Kar je zanimivo je tudi to, da pri polovici leta pri mesecu juliju, CA začne spet rasti in brier score padati. Zdi se mi, da se to zgodi zaradi tega, ker v prvih sedmih mesecih se nahaja večino meritev.

Evalvacija dodanih atributov

```
> sort(attrEval(norm_poraba ~ .,podatki,"MDL"), decreasing=T)
```

avgNorm_Poraba	poraba	avgPoraba	leto_izgradnje	namembnost	povrsina	ura	stavba
6.885177e-01	1.352364e-01	7.330354e-02	5.747618e-02	5.437178e-02	2.394871e-02	2.343338e-02	1.687218e-02
temp_rosisca	LetniCas	Dan	weekend	regija	Mesec	datum	temp_zraka
1.103000e-02	8.662091e-03	6.808768e-03	6.468332e-03	5.840108e-03	3.384957e-03	3.173134e-03	2.558398e-03
pritisk	smer_vetra	oblacnost	hitrost_vetra	padavine			
1.668027e-03	1.289557e-03	4.428373e-04	1.135268e-04	5.421448e-05			

```
> sort(attrEval(norm_poraba ~ .,podatki,"InfGain"), decreasing=T)
```

avgNorm_Poraba	poraba	avgPoraba	leto_izgradnje	namembnost	datum	povrsina	ura
0.6890894714	0.1353764256	0.0734389357	0.0576120974	0.0548670491	0.0289169413	0.0240757138	0.0235681488
stavba	temp_rosisca	LetniCas	Dan	weekend	regija	Mesec	temp_zraka
0.0170020276	0.0111544847	0.0090440412	0.0075396906	0.0065996882	0.0059742970	0.0035164356	0.0026924803
pritisk	smer_vetra	oblacnost	hitrost_vetra	padavine			
0.0017944931	0.0014230198	0.0005750612	0.0001852129	0.0001569170			













```
> sort(attrEval(norm_poraba ~ .,podatki,"ReliefEqualK"), decreasing=T)
```

avgNorm_Poraba	namembnost	povrsina	leto_izgradnje	avgPoraba	poraba	stavba	regija
0.4494030936	0.1513502374	0.1491698996	0.1323071678	0.1181163221	0.1077320324	0.0889913090	0.0011984241
LetniCas	padavine	ura	weekend	Mesec	temp_zraka	pritisk	temp_rosisca
0.0001255983	-0.0004808915	-0.0006771472	-0.0017788415	-0.0088820377	-0.0124750060	-0.0153484189	-0.0203818491
hitrost_vetra	oblacnost	smer_vetra	Dan	datum			
-0.0241739122	-0.0290797329	-0.0304230997	-0.0372093770	-0.0660670420			

Dodatna atributa avgPoraba in avgNorm_Poraba se izkažejo kot dokaj pomembna. Atributi Mesec, Dan, Weekend in Letni cas, pa varirajo s svojo pomembnostjo ampak so ševedno uporabni.

Evalvacija podatkov z dodatnimi atributi

Za mojo drugo evalvacijo, sem najboljše rezultate dobil z atributi Leto izgradnje, namembnost, površina, stavba, avgNorm_Poraba, Dan, Mesec in Weekend.

	 V1 	V2 	V3 	V4 	V5 	V6 	V7 	V8 	V9 	V10 	V11 
1	0.8088208	0.7907529	0.8120551	0.7276708	0.7117069	0.7460190	0.7533042	0.7859550	0.6623635	0.7550344	0.7341240
2	0.8223418	0.7945092	0.8157163	0.7288091	0.7304441	0.7545654	0.7616451	0.7840979	0.6799783	0.7459370	0.7338585
3	0.6923749	0.6682318	0.6685877	0.6109457	0.5333993	0.5345508	0.5173874	0.5769998	0.6081317	0.6265340	0.6470328

Evalvacija podatkov zahodnih/vzhodnih stavb

Vzhodne stavbe

CA

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
1	0.6925926	0.6350017	0.6502528	0.5379046	0.5026557	0.6295215	0.5783826	0.5678832	0.4628088	0.6259502	0.6630525
2	0.6950081	0.6308210	0.6423891	0.5396082	0.5065186	0.5971692	0.5926905	0.5710115	0.4615385	0.6256007	0.6629009
3	0.6840580	0.6015561	0.5864070	0.5132027	0.4232255	0.3776679	0.3849145	0.4231491	0.4752294	0.5731761	0.6217302

Brier Score

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
1	0.4113014	0.5047279	0.4753824	0.6266471	0.6580915	0.5228765	0.5744047	0.5559804	0.7811457	0.4875861	0.4557556
2	0.4035650	0.5014322	0.4846677	0.6443943	0.6741094	0.5491242	0.5779813	0.5625542	0.7616725	0.5035714	0.4521249
3	0.4770702	0.5710983	0.5823876	0.6825610	0.8002825	0.8098593	0.8179680	0.7046285	0.6912229	0.6090310	0.5141804

Zahodne stavbe

CA

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
1	0.7360575	0.7532659	0.7520196	0.6839727	0.6855458	0.7177184	0.6924421	0.6817529	0.6916351	0.6916149	0.6844175
2	0.7370227	0.7470501	0.7494685	0.6787919	0.6828753	0.7191254	0.6980122	0.6817529	0.6924779	0.6844720	0.6780400
3	0.7120335	0.7268226	0.7141794	0.6342593	0.5822855	0.5870765	0.6119484	0.6437808	0.6749895	0.6655280	0.6633716

Brier Score

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
1	0.3905083	0.3681202	0.3593693	0.4711154	0.4472747	0.3881012	0.4423013	0.4291377	0.4176937	0.4282447	0.4411753
2	0.3583972	0.3560369	0.3467891	0.4460416	0.4238115	0.3796421	0.4274783	0.4125929	0.3945748	0.4152207	0.4366969
3	0.4431642	0.4359079	0.4508791	0.5563181	0.6065140	0.5971641	0.5588738	0.5204902	0.4908841	0.5072600	0.5168380

Klasifikacijski del

Nespremenjeni podatki

Tudi pri regresiji sem na začetku začel z originalnimi podatki. Za primerjavo sem izbral linearno regresijo, regresijsko drevo in k-najbližjih sosedov, saj sem jih lahko izvajal relativno hitro.

Spisal sem for loop v katerem se modeli izvajajo 11x, rešitve pa se shranjujejo v tabelo. Prvi učna podmnožica je mesec januar, testna pa februar. Nato je naslednja učna januar in februar, testna pa marec. Tako se nadaljuje, dokler ni zadnja učna podmnožica od januarja do novembra in testa december.

Prve teste sem delal z vsemi podatki, ki so mi bili na voljo:

leto_izgradnje+namembnost+povrsina+ura+stavba+temp_rosisca+regija+temp_zraka+oblacnost+padavine+pritisk+smer_vetra+hitrost_vetra

Trivialni model

Za začetek sem začel z trivialnim modelom, samo da sem dobil občutek kakšne rezultate pričakovati oz. da bi takoj videl če bi bili le te preslabi.

```
> mae(observed, predTrivial)
[1] 145.0652
> mse(observed, predTrivial)
[1] 64160.66
```

Linearna regresija, regresijsko drevo in k-najbližjih sosedov

To so trije modeli, ki sem jih poganjal z for loopom. Mae in rmae sem shranjeval v tabelo da jih lahko zdaj vidimo skupaj. Prvi stolpec je linearna regresija, drugi regresijsko drevo in zadnji k-najbližjih sosedov

```
> tab_mae
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]      [,9]     [,10]     [,11]
[1,] 106.3456 104.77817 102.23755 103.88258 105.20274 105.91051 108.34910 105.78856 106.65566 103.32197 103.42713
[2,]  63.9826  69.95946  67.82947  64.70818  69.57297  72.83190  72.93018  71.72087  69.32495  69.81592  67.56951
[3,]  43.4972  42.60427  41.30718  41.34068  45.76915  44.58625  44.59282  41.50737  44.21923  48.24359  48.99922

> tab_rmae
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]      [,9]     [,10]     [,11]
[1,] 0.6487141 0.6720471 0.6995575 0.7354769 0.7095021 0.7039621 0.6902188 0.6981495 0.7109934 0.6882595 0.6894663
[2,] 0.3902976 0.4487199 0.4641212 0.4581266 0.4692099 0.4840964 0.4645888 0.4733204 0.4621375 0.4650653 0.4504321
[3,] 0.2653355 0.2732637 0.2826432 0.2926873 0.3086736 0.2963542 0.2840707 0.2739271 0.2947765 0.3213654 0.3266388
```

Opazimo lahko da se rezultati z večanjem učne množice slabšajo.

Naključni gozd

Ker je model potreboval preveč časa da se izvede sem se odločil, da ga bom zaganjal samo na treh učnih množicah. Učna januar in testna februar, učne od januarja do maja in testa junij, učne od januarja do novembra in testna december. Spet vidimo da nam najmanjša učna množica da najboljši rezultat.

J-N → D

```
> mae(test5$poraba, predicted)
[1] 42.76214
> rmae(test5$poraba, predicted, mean(train5$poraba))
[1] 0.2784876
```

J-M → J

```
> mae(test11$poraba, predicted)
[1] 39.95511
> rmae(test11$poraba, predicted, mean(train11$poraba))
[1] 0.2754286
```

J → F

```
> mae(test1$poraba, predicted)
[1] 28.19448
> rmae(test1$poraba, predicted, mean(train1$poraba))
[1] 0.1707756
```


Ocenjevanje podatkov

nato sem ocenil podatke in zmanjšal množico podatkov iz katerih se modeli učijo na malenkost manjšo:

povrsina + leto_izgradnje + stavba + namembnost + ura + temp_rosisca + regija

```
> sort(attrEval(poraba ~ ., train, "MSEofMean"), decreasing = TRUE)
povrsina leto_izgradnje stavba namembnost ura datum temp_rosisca regija temp_zraka Mesec pritisk smer_vetra oblacnost hitrost_vetra padavine
-38245.86 -58075.50 -62060.33 -62736.87 -63431.26 -63602.93 -63700.95 -63775.22 -63777.39 -63796.93 -63816.00 -63835.13 -63845.59 -63846.24 -63848.84
```

Linearna regresija, regresijsko drevo in k-najbližjih sosedov

Tu je vse teklo po istem principu kot prej samo da se je model zdaj učil iz malenkost manjše množice podatkov.

```
> tab_mae
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]      [,9]     [,10]     [,11]
[1,] 104.86776 105.12056 103.02781 103.82785 104.90389 105.7423 108.29915 105.80268 106.18310 103.14480 103.10430
[2,]  63.98260  69.95946  67.82947  64.70818  69.57297  72.8319  72.93018  71.72087  69.32495  69.81592  67.56951
[3,]  27.45382  28.88264  30.95249  28.36970  36.06203  29.7848  31.27563  30.21396  35.51721  38.83877  38.25681

> tab_rmae
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]      [,9]     [,10]     [,11]
[1,]  0.6396995  0.6742432  0.7049648  0.7350894  0.7074866  0.7028438  0.6899006  0.6982426  0.7078432  0.6870793  0.6873143
[2,]  0.3902976  0.4487199  0.4641212  0.4581266  0.4692099  0.4840964  0.4645888  0.4733204  0.4621375  0.4650653  0.4504321
[3,]  0.1674699  0.1852532  0.2117915  0.2008543  0.2432074  0.1979725  0.1992359  0.1993964  0.2367666  0.2587170  0.2550277
```

Opazimo lahko da so se rezultati za malenkost poboljšali. Modeli pa so se izvajali hitreje. Še zmeraj pa je trend da večja kot je učna množica, slabši so rezultati.

Naključni gozd

Tu je spet vse delovalo po istem principu, samo da smo imeli manjšo množico podatkov.

J – N → D

```
> mae(test5$poraba, predicted)
[1] 41.61328
> rmae(test5$poraba, predicted, mean(train5$poraba))
[1] 0.2710057
```

J – M → J

```
> mae(test11$poraba, predicted)
[1] 40.63487
> rmae(test11$poraba, predicted, mean(train11$poraba))
[1] 0.2801146
```

J → F

```
> mae(test1$poraba, predicted)
[1] 30.05545
> rmae(test1$poraba, predicted, mean(train1$poraba))
[1] 0.1820476
```

Tu pa so se podatki izboljšali samo pri največji učni množici, med tem ko se pa pri drugih poslabšali.

Spreminjanje podatkov (dodajanje atributov)

V množico podatkov sva zdaj dodala že: dan, mesec, vikend, letni čas, povprečno Norm_porabo in pa povprečno porabo.

Prve teste novih atributov sem delal kar na vseh možnih:

ura + regija + stavba + namembnost + površina + leto_izgradnje + temp_zraka + temp_rosisca + oblacnost + padavine + pritisk + smer_vetra + hitrost_vetra + Mesec + Dan + Weekend + LetniCas + avgNorm_Poraba + avgPoraba

Linearna regresija, regresijsko drevo in k-najbližjih sosedov

```
> tab_mae
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]      [,9]     [,10]     [,11]
[1,] 36.28927 38.25926 38.92227 37.97402 49.74434 47.74211 48.79025 39.64985 45.75007 41.57280 43.71685
[2,] 51.31798 51.58184 46.83559 46.02874 49.91463 51.80236 56.70972 50.40316 52.53072 53.63510 53.54108
[3,] 30.16545 30.19932 25.50680 30.30992 35.24279 33.65814 35.25654 30.35181 40.99531 32.98151 37.02727
> tab_rmae
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]      [,9]     [,10]     [,11]
[1,] 0.2213666 0.2453949 0.2663245 0.2688517 0.3354828 0.3173305 0.3108096 0.2616684 0.3049814 0.2769292 0.2914254
[2,] 0.3130427 0.3308459 0.3204712 0.3258782 0.3366313 0.3443180 0.3612593 0.3326346 0.3501829 0.3572799 0.3569157
[3,] 0.1840110 0.1936984 0.1745296 0.2145908 0.2376823 0.2237177 0.2245955 0.2003061 0.2732850 0.2197000 0.2468314
```

Zdaj ko smo dodali nove attribute opazamo, da smo rezultate pri določenih modelih poboljšali za kar več kot 100% najbolj opazno pri linearni regresiji. Sami modeli pa so potrebovali malenkost dalj časa za izvajanje kot prej.

Naključni gozd

J – N → D

```
> mae(test11$poraba, predicted)
[1] 37.20654
> rmae(test11$poraba, predicted, mean(train11$poraba))
[1] 0.2564815
```

J – M → J

```
> mae(test5$poraba, predicted)
[1] 36.85971
> rmae(test5$poraba, predicted, mean(train5$poraba))
[1] 0.2400481
```

J → F

```
> mae(test1$poraba, predicted)
[1] 23.30281
> rmae(test1$poraba, predicted, mean(train1$poraba))
[1] 0.1411465
```

Tu smo dobili najboljši rezultat do zdaj. Zanimivo je, da smo ga dobili ravno pri najmanjši učni množici. Ostala dva pa sta tudi boljša od rezultatov preden smo dodali attribute. Spet opazamo da večja kot je učna množica slabši je rezultat. Res pa je, da sem na rezultate čakal po več 10 minut.

Ocenjevanje atributov z dodatnimi

Spet sem ocenil podatke in zmanjšal množico podatkov iz katerih se modeli učijo na malenkost manjšo:
avgPoraba + površina + avgNorm_Poraba + leto_izgradnje + stavba + namembnost + ura + Dan + Weekend + LetniCas + regija

```
> sort(attrEval(poraba ~ ., train, "MSEofMean"), decreasing = TRUE)
avgPoraba      površina avgNorm_Poraba leto_izgradnje      stavba      namembnost      ura      datum      temp_rosisca      Dan      Weekend      LetniCas      regija      temp_zraka      Mesec
-26823.02      -38245.86      -55668.57      -58075.50      -62060.53      -62736.87      -63431.26      -63602.93      -63700.95      -63748.09      -63748.09      -63762.58      -63775.22      -63777.39      -63796.93
pritisek      smer_vetra      oblacnost      hitrost_vetra      padavine
-63816.00      -63835.13      -63845.59      -63848.24      -63848.84
```

Linearna regresija, regresijsko drevo in k-najbližjih sosedov

```
> tab_mae
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]      [,9]     [,10]     [,11]
[1,] 35.80792 38.73470 34.73074 35.48795 42.55186 45.38183 48.92125 39.54995 43.02542 43.85900 42.35736
[2,] 51.31798 51.58184 46.83559 46.02874 49.91463 51.80236 56.70972 50.40316 52.53072 53.63510 53.54108
[3,] 22.17881 27.57239 25.69377 25.35428 30.54080 32.09842 27.95820 24.39678 35.95715 25.28436 31.00975
> tab_rmae
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]      [,9]     [,10]     [,11]
[1,] 0.2184304 0.2484444 0.2376441 0.2512507 0.2869757 0.3016423 0.3116442 0.2610091 0.2868182 0.2921583 0.2823628
[2,] 0.3130427 0.3308459 0.3204712 0.3258782 0.3366313 0.3443180 0.3612593 0.3326346 0.3501829 0.3572799 0.3569157
[3,] 0.1352921 0.1768493 0.1758089 0.1795054 0.2059714 0.2133506 0.1781028 0.1610061 0.2396994 0.1684269 0.2067173
```

Zdaj ko sem modele pognal z manj atributi so se modeli izvajali veliko hitreje, dobili pa smo za malenkost boljše rezultate. Če bi se še malo poigral z atributi s katerimi se model uči, sem prepričan da bi lahko dobili še boljše. Še vedno pa vidimo, da se z večanjem učnih množic rezultati slabšajo.

Naključni gozd

J – N → D

```
> mae(test11$poraba, predicted)
[1] 37.61513
> rmae(test11$poraba, predicted, mean(train11$poraba))
[1] 0.2592981
```

J – M → J

```
> mae(test5$poraba, predicted)
[1] 37.70907
> rmae(test5$poraba, predicted, mean(train5$poraba))
[1] 0.2455796
```

J → F

```
> mae(test1$poraba, predicted)
[1] 24.67511
> rmae(test1$poraba, predicted, mean(train1$poraba))
[1] 0.1494586
```

Spet lahko opazimo, da je najmanjša učna množica najboljša. Čeprav smo zmanjšali število atributov, smo tukaj dobili slabši rezultat.

Svm

Tu sem naredil še svm model. Ker se je izvajal več kot eno uro in pol sem ga izvedel samo enkrat. Tu še nisva imela atributa za letni čas. Učno in testno množico pa sva določala še z 70/30 random. Takrat je bil to zdaleč najboljši rezultat.

```
> svm.model <- svm(poraba ~ Mesec+Dan+Weekend+leto_izgradnje+namembnost+povrsina+ura+stavba+temp_rosisca+regija+avgPoraba+avgNorm_Poraba, train)
> predicted <- predict(svm.model, test)
> mae(test$poraba, predicted)
[1] 38.2851
> rmae(test$poraba, predicted, mean(train$poraba))
[1] 0.2552163
```

Regije

Podatke sem razdelil na dve regiji (zahodna in vzhodna) in nato na vsaki posebej poganjal linearno regresijski model, regresijsko drevo in k-najbližjih sosedov.

Samo vzhodna regija

```
> tab_mae_vz
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]      [,9]     [,10]     [,11]
[1,] 46.30879 55.74451 56.09552 57.26828 65.52903 72.02083 77.26931 61.44350 66.26795 60.07349 53.80032
[2,] 55.28955 61.14978 61.16282 59.25429 65.58713 70.31743 81.02797 70.18348 72.68098 68.03725 59.89970
[3,] 27.64456 40.45264 46.28217 41.59581 47.02799 52.48514 41.25157 34.33406 57.55641 32.50111 36.73900
> tab_rmae_vz
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]      [,9]     [,10]     [,11]
[1,] 0.2402593 0.3003511 0.3233638 0.3228011 0.3646398 0.3914787 0.3999518 0.3273572 0.3716997 0.3503353 0.3185443
[2,] 0.2868532 0.3294746 0.3525744 0.3339956 0.3649631 0.3822197 0.4194069 0.3739218 0.4076707 0.3967782 0.3546579
[3,] 0.1434255 0.2179586 0.2667945 0.2344609 0.2616898 0.2852899 0.2135212 0.1829242 0.3228363 0.1895393 0.2175266
```

Samo zahodna regija

```
> tab_mae_zs
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]      [,9]     [,10]     [,11]
[1,] 22.38197 23.15134 21.55119 28.44299 29.80393 30.64554 27.51748 27.74051 27.08925 27.45462 29.38024
[2,] 28.70605 38.10546 36.25116 41.92675 42.88172 44.38157 40.52041 40.11363 38.29662 38.38794 38.53612
[3,] 13.48316 15.88871 14.61906 21.04441 23.41505 22.27164 17.75045 19.42569 19.55929 15.98697 21.29370
> tab_rmae_zs
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]      [,9]     [,10]     [,11]
[1,] 0.1728973 0.1782598 0.1656789 0.2184816 0.2245552 0.2294212 0.2088687 0.2081341 0.2054671 0.2091688 0.2277881
[2,] 0.2217498 0.2934030 0.2786878 0.3220556 0.3230886 0.3322530 0.3075660 0.3009683 0.2904731 0.2924665 0.2987746
[3,] 0.1041553 0.1223393 0.1123869 0.1616503 0.1764187 0.1667318 0.1347330 0.1457489 0.1483538 0.1218000 0.1650923
```

Ko smo opazovali samo vzhodno regijo so rešitve malenkost slabše, kot če bi opazovali obe regiji skupaj. Če pa se osredotočimo na zahodno regijo so rešitve najboljše, ki smo jih do zdaj videli. Oba pa še veno sledita »pravilu« več kot učnih podatkov, slabši rezultat.