

Umetna inteligenca 2021-2022

Seminarska naloga 1

Janez Sedeljšak, 63200261  
Marko Vrečer, 63200329

Vsebina

[1 Priprava podatkov za konstrukcijo modelov 4](#_Toc89620338)

[1.1 Analiza in izbor novih atributov 4](#_Toc89620339)

[1.2 Priprava atributov v Javi 4](#_Toc89620340)

[1.3 Fakturiranje podatkov 4](#_Toc89620341)

[1.4 Odpravljanje napak v podatkih 4](#_Toc89620342)

[2 Vizualizacija podatkov 5](#_Toc89620343)

[2.1 Grafi za analizo korelacije med atributi 5](#_Toc89620344)

[2.1.1 Korelacija med že obstoječimi atributi 5](#_Toc89620345)

[2.1.2 Prikaz porazdelitev podatkov 7](#_Toc89620346)

[2.1.3 Porazdelitev podatkov z novimi atributi 9](#_Toc89620347)

[2.1.4 Korelacija med porabo in novimi atributi 10](#_Toc89620348)

[2.1.5 Vizualizacije ocenjevanja atributov 11](#_Toc89620349)

[3 Ocenjevanje in konstrukcija atributov 12](#_Toc89620350)

[3.1 Konstrukcija 12](#_Toc89620351)

[3.2 Klasifikacija 12](#_Toc89620352)

[3.3 Regresija 12](#_Toc89620353)

[4 Modeliranje 13](#_Toc89620354)

[4.1 Klasifikacija 13](#_Toc89620355)

[4.1.1 Odločitveno drevo 13](#_Toc89620356)

[4.1.2 Naivni Bayes 13](#_Toc89620357)

[4.1.3 K-najbližjih sosedov – knn 13](#_Toc89620358)

[4.1.4 Bagging 13](#_Toc89620359)

[4.1.5 Izvedbe z naključnim gozdom 13](#_Toc89620360)

[4.1.6 Boosting 13](#_Toc89620361)

[4.2 Regresija 13](#_Toc89620362)

[4.2.1 Linearni model – lm 13](#_Toc89620363)

[4.2.2 Regresijsko drevo – rpart 13](#_Toc89620364)

[4.2.3 Support-vector machine – svm 13](#_Toc89620365)

[4.2.4 K-najbližjih sosedov – knn 13](#_Toc89620366)

[4.3 Kombinirano učenje 13](#_Toc89620367)

[4.4 Primerjanje uspešnosti glede na podatke iz posameznih regij 13](#_Toc89620368)

[5 Evalvacija modelov 14](#_Toc89620369)

[5.1 Klasifikacija 14](#_Toc89620370)

[5.2 Regresija 14](#_Toc89620371)

# Priprava podatkov za konstrukcijo modelov

## Analiza in izbor novih atributov

Po krajši analizi podatkov in korelaciji sva opazila, da lahko pripraviva nove atribute, ki bodo bistveno izboljšale same napovedi modelov, prvi atribut ki sva ga razcepila je bil »datum«. Pridobila sva nove atribute – »mesec«, »vikend« (ali gre za podatek porabe med vikendom ali med tednom) in »letni čas«. Kot drugi atribut, pa sva uporabila še poraba, in sicer statistiko za nazaj. Za vsako posamezno stavbo sva dodala podatek »včerajšnja poraba« in »tedenska poraba«.

## Priprava atributov v Javi

Pri izvedbi generiranja novih atributov sva sprva uporabljala kar R. Ampak, pri samem izračunu porabe za pretekli teden je sama izvedba v R dosti »nerodna«, zato sva se odločila za izvedbo v Javi, ki izkorišča hitrost podatkovne strukture »HashMap«, ki za vsako stavbo hrani vrednosti o preteklih porabah na določen dan. Tako se sama konstrukcija atributov, zgodi v ločenem programu in generira nove datoteke, katere potem uporabljava za konstrukcijo modelov.

## Fakturiranje podatkov

V naši množici podatkov imamo, kar nekaj atributov, ki imajo določeno »zalogo vrednosti«, v ta namen sva pripravila metodo, ki te podatke fakturira kar ob branju datoteke. Fakturirala sva »namembnost«, »sezona« – letni čas, »regija«, »oblacnost« - [1,10] »vikend« (1/0), »mesec«- [1,12].

## Odpravljanje napak v podatkih

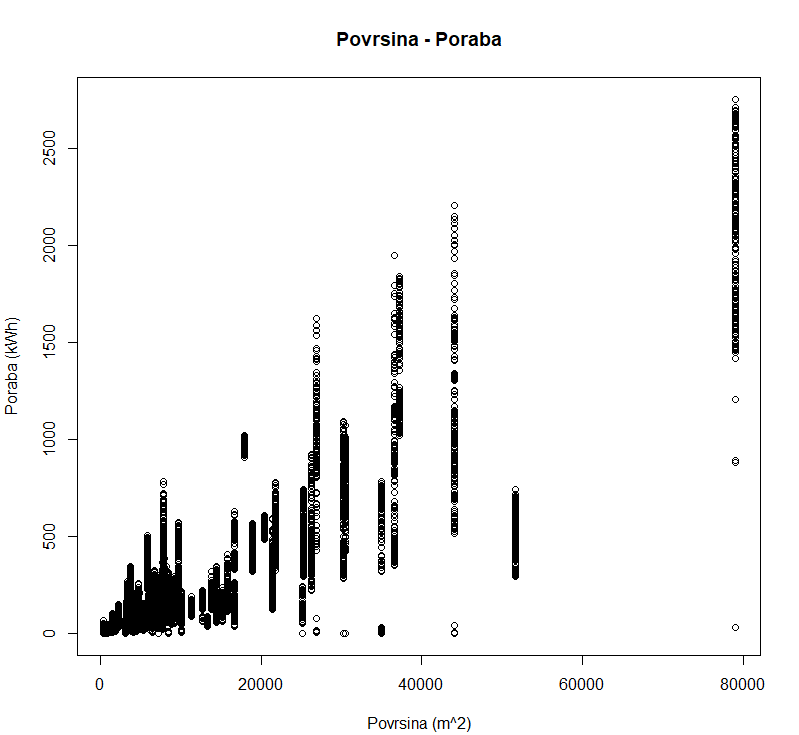
Pri konstrukciji atributov, sva opazila nekaj manjših napak, najprej sva morala podatke, ki smo jih iz Jave dobili, kot »-1« (zapisi, ki zaradi manjkajočih podatkov za nazaj nimajo izračunane zgodovine porabe). Poleg tega, pa ima vrednost padavin, kadar gre za »pršenje« vrednost -1, kar ni v skladu s podatki, saj »0« pomeni »popolnoma jasno«, zato sva vrednost »-1« preslikala v vrednost »1«.

# Vizualizacija podatkov

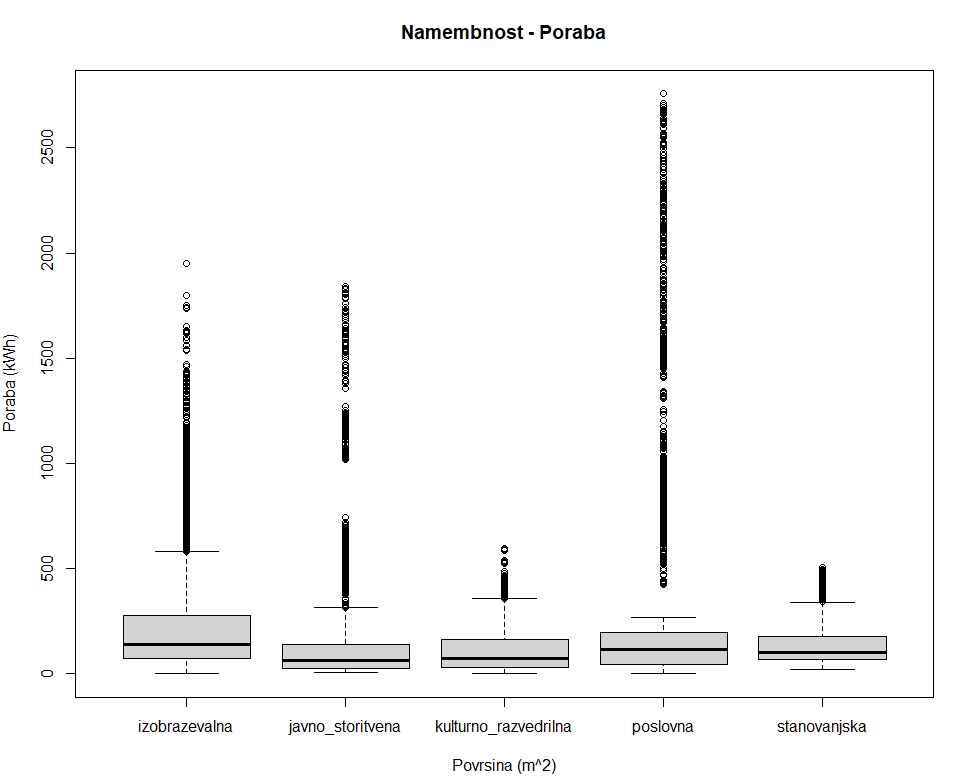
## Grafi za analizo korelacije med atributi

### Korelacija med že obstoječimi atributi

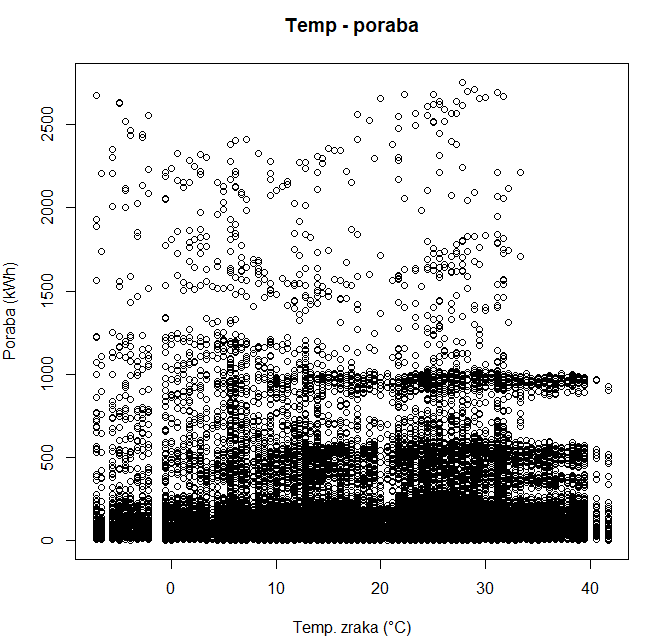
Spodnji graf prikazuje korelacijo med Površino n Porabo. Pričakovano se poraba veča skupaj z površino imamo nekaj manjših odstopanj, ampak v primeru maksimalnih porab, za neko površino, pa imamo zelo lepo naraščanje.



Naslednji graf predstavlja porabe glede na namembnost neke stavbe. Iz grafa lahko vidimo, da imajo tiste najvišje porabe »poslovne« stavbe. »Kulturno razvedrilne« in »Stanovanje« stavbe, pa imajo nekako bistveno manjšo porabo. Iz grafa tako vidimo, da bo površina dober atribut za napoved »namembnosti« stavbe.

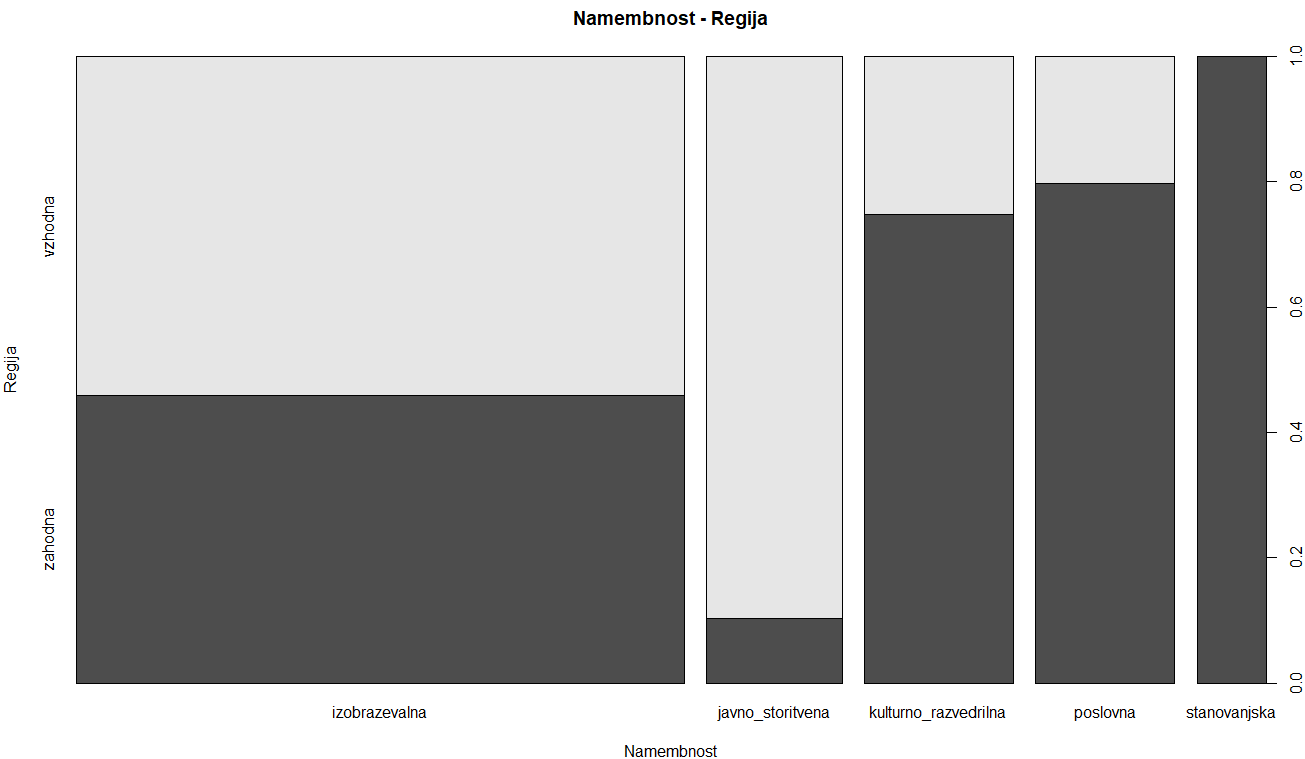


Po prvem pregledu učnih podatkov sva pričakovala, da bodo podatki za »temperaturo zraka« močno vplivali na porabo elektrike, saj ob nižjih temperaturah porabimo nekaj več elektrike (npr. za ogrevanje). Ta hipoteza, pa je bila kaj hitro zavrnjena, saj lahko vidimo, da med porabo in »temperaturo zraka« nimamo dobrega nivoja korelacije.

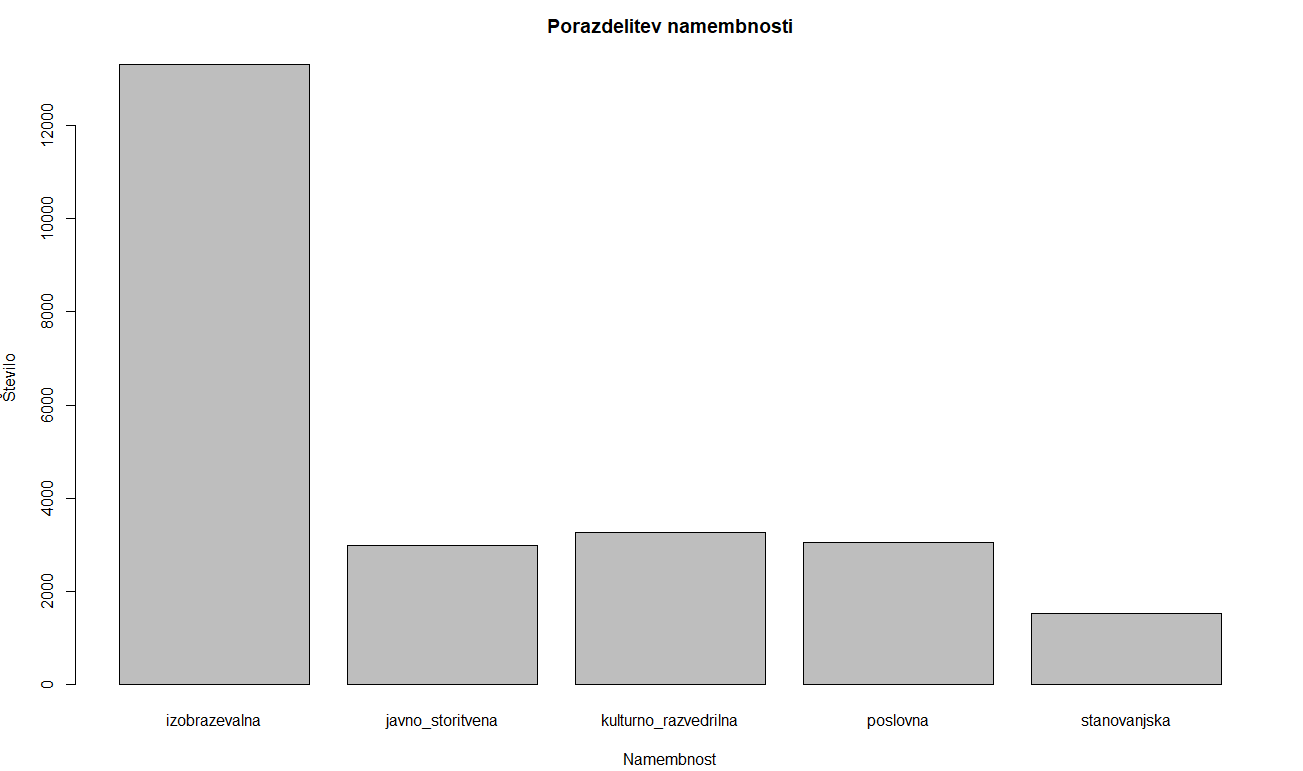


### Prikaz porazdelitev podatkov

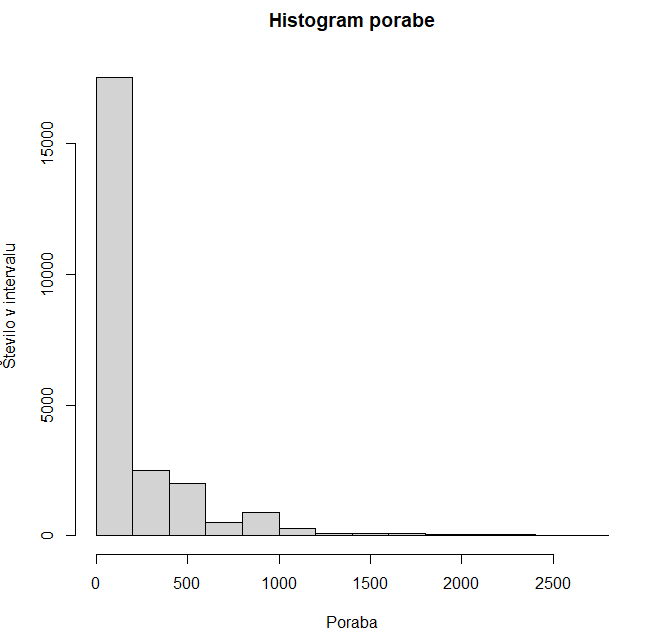
Prvi graf prikazuje porazdelitev glede na regijo in namembnost, razberemo lahko v katerih regijah prevladujejo določene stavbe. Vidimo, da imamo v učnih podatkih »stanovanjske« stavbe le v zahodni regiji, kar je lahko malo zavajajoč podatek, saj lahko, da v učni množici podatki ne sledijo enaki porazdelitvi.



Spodnji graf, pa prikazuje splošno porazdelitev »namembnosti« stavb, razberemo lahko, da je zelo velik del »izobraževalnih« stavb, kar je pozneje pri učenju privedlo, v »zelo dober«, trivialen model, podatki v testni množici sledijo podobni porazdelitvi, in tako dobimo oceno natančnosti (0.47).

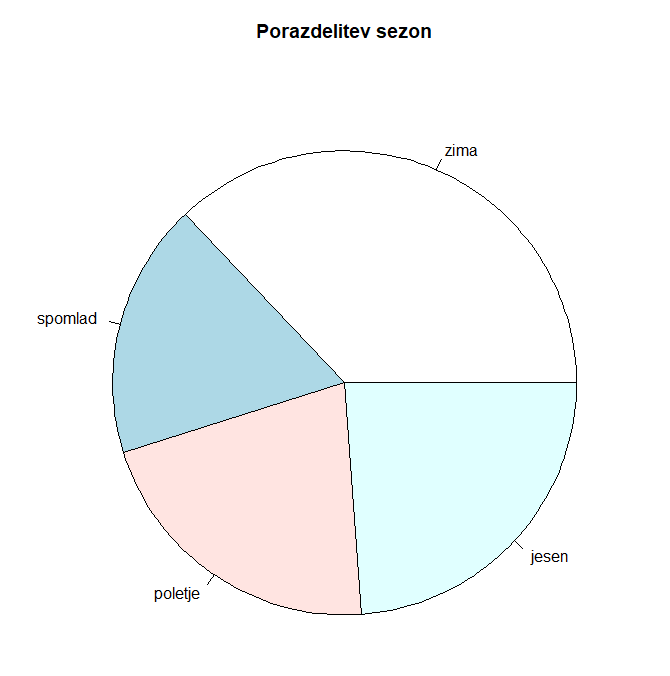


Spodnji histogram prikazuje razporeditev porabe v učnih podatkih vidimo lahko, da prevladujejo zapisi z relativno nizko porabo, torej skupino od 0-200 kWh. Po tej skupini število v skupini močno upade, ta trend se tudi nadaljuje in opazimo, da so podatki z porabo višjo od 100 kWh zelo redki.

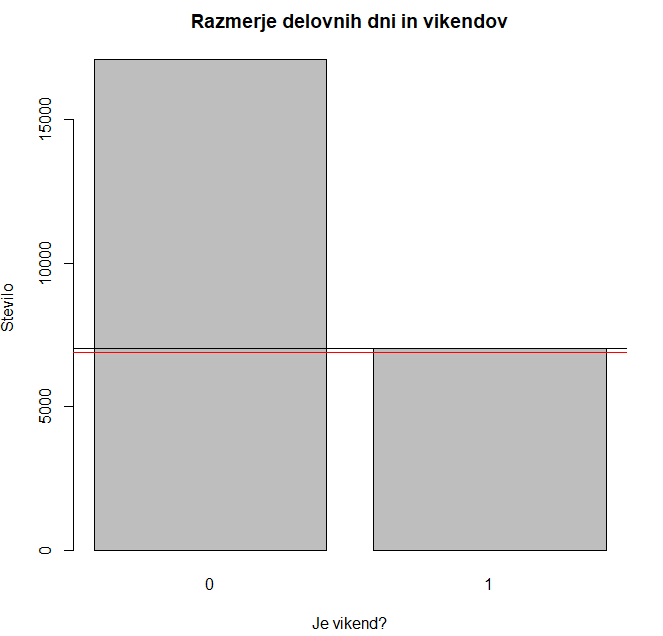


### Porazdelitev podatkov z novimi atributi

Pri porazdelitvi podatkov glede na letni čas opazimo, da je večji delež podatkov iz »zimskega« letnega časa.

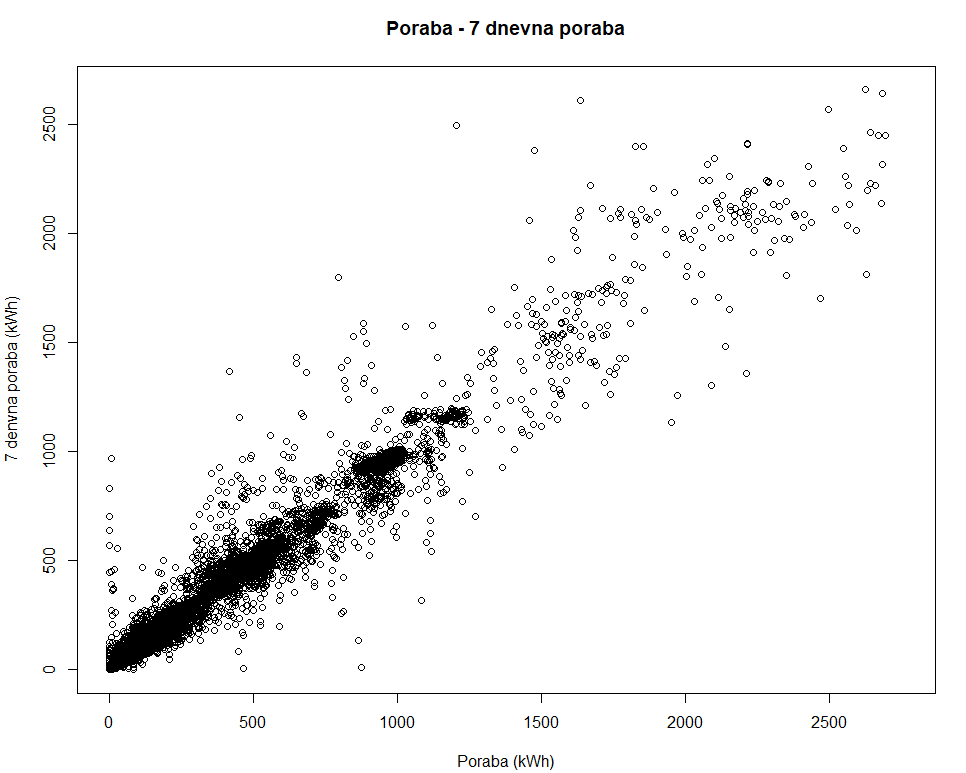


Na spodnjem grafu, pa vidimo porazdelitev »dni med tednu« in »vikendov«, samo razmerje pričakujemo, da bo 2:5, kar prikazuje rdeča črta, desna črta pa prikazuje našo porazdelitev, tako vidimo, da je razmerje podatkov med dnevi dobro uravnoteženo.

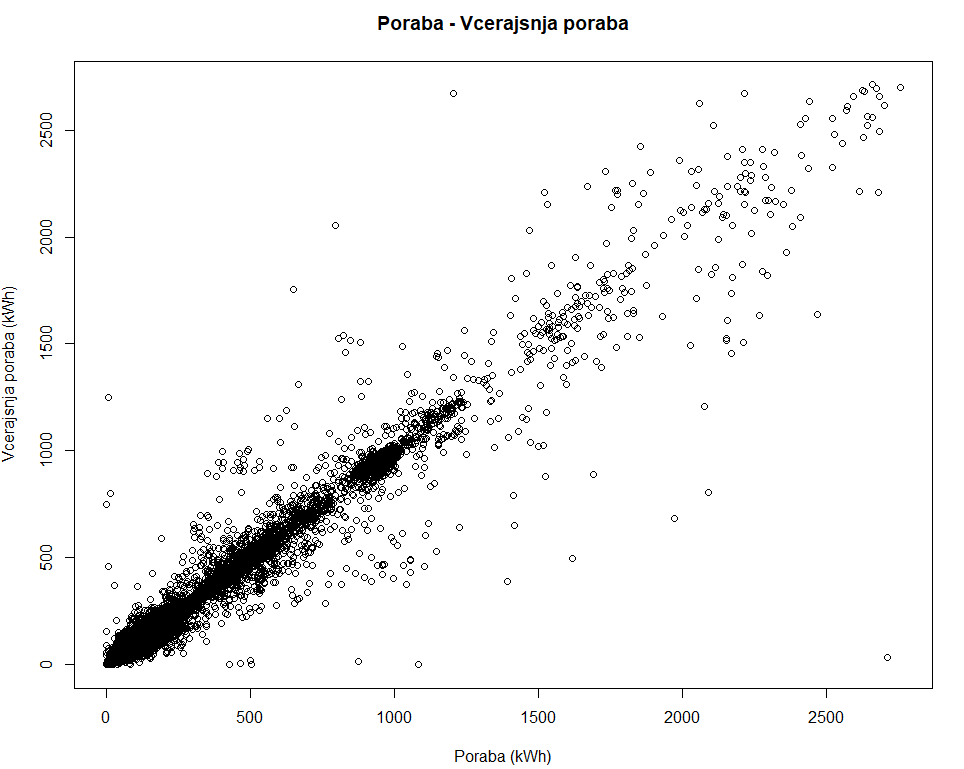


### Korelacija med porabo in novimi atributi

Pri konstrukciji novih atributov, je pri prvih nekaj zapisih računanje pretekle porabe onemogočene, saj teh podatkov nimamo na voljo, zato pri sami konstrukciji grafa te podatke izpustimo (vrednosti »tedenske porabe« za prve zapise imajo vrednost NA). Vidimo pa zelo lepo linearno korelacijo, ki bo napovedi za regresijsko napoved porabe bistveno izboljšala.

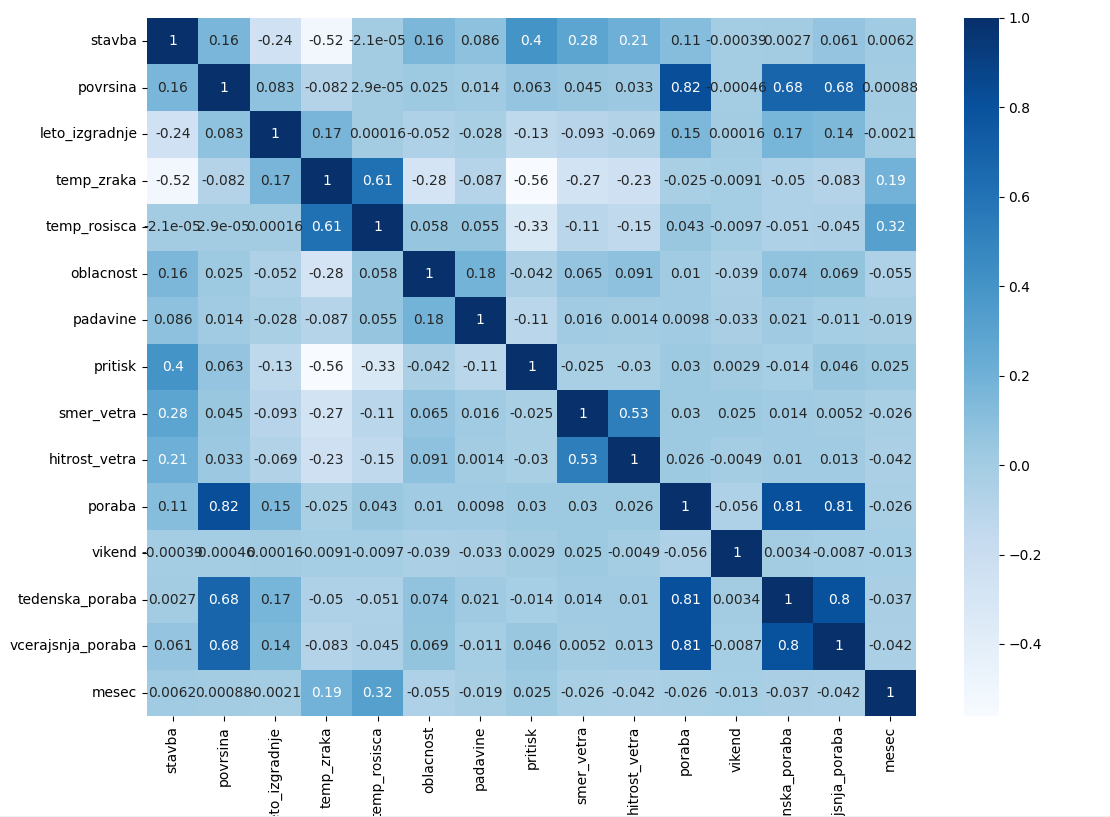


Še boljša je korelacija je med »včerajšnjo« in »današnjo« porabo. To je seveda pričakovano, imamo pa vseeno nekaj manjših »izpadanj podatkov«, ki ne sledijo korelaciji, so pa ti zelo redki. Včerajšnja poraba, bo tako postal ključen atribut, pri napovedi »porabe«.



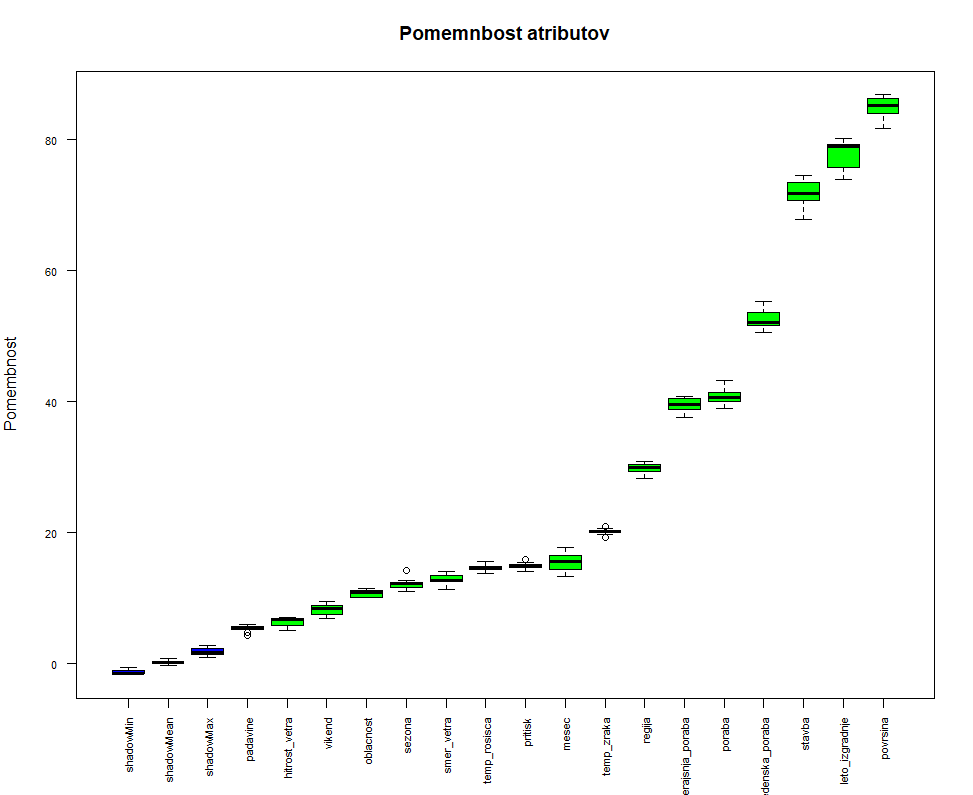
### Vizualizacije ocenjevanja atributov

Za lažje razumevanje, kako se atributi med sabo povezujejo, sva pripravila skripto v »Pythonu«, ki pokaže korelacijo med nekaterimi atributi za regresijo. Opazimo lahko na primer močno korelacijo med »porabo« in »tedensko porabo«, »tedensko porabo« ter »površino«.



Zgornji prikaz sploh dobro prikaže, kateri atributi bodo koristni za regresijski model, ki napoveduje »porabo«. Graf pa ne prikazuje atributa »namembnost« in korelaciji. V ta namen, sva pripravila vizualizacijo s pomočjo knjižnice »Boruta«, kar v R.

Graf napove, da bodo za napovedovanje »namembnosti« najboljši naslednji atributi (napisani od najboljšega proti slabšim) – »poraba«, »leto izgradnje«, »stavba«, »poraba«, »včerajšnja poraba«, »regija«…



# Ocenjevanje in konstrukcija atributov

## Konstrukcija

(pomoje bi tuki opisal mal ono Parse.java)

## Klasifikacija

## Regresija

Za ocenjevanje atributov pri regresiji sva uporabila knjižnico CORElearn in pognala funkcijo attrEval, ki oceni, koliko je posamezen atribut koristen. Za lažje ugotavljanje vrstnega reda sva rezultat funkcije sortirala, prav tako pa sva za lažje kopiranje združila atribute s presledki.

Uporabila sva več različnih ocen: MSEofMean, RReliefFequalK, RReliefFbestK, MSEofModel in RReliefFsqrDistance, ki so bolj ali manj konsistentno napovedovali, da sta najmočnejša atributa za napovedovanje izpeljana atribut vcerajsnja\_poraba in tedenska\_poraba, kasnejši atributi pa so manj konsistentni: poraba, vikend, leto\_izgradnje...

Rezultati preverjanja so precej pričakovani, saj sva že pri vizualizaciji podatkov opazila, da s porabo predlagani atributi precej korelirajo.

# Modeliranje

## Klasifikacija

### Odločitveno drevo

### Naivni Bayes

### K-najbližjih sosedov – knn

### Bagging

### Izvedbe z naključnim gozdom

### Boosting

## Regresija

### Linearni model – lm

### Regresijsko drevo – rpart

### Support-vector machine – svm

### K-najbližjih sosedov – knn

## Kombinirano učenje

(tuki pomoje kr za oba skupi napiševa)

## Primerjanje uspešnosti glede na podatke iz posameznih regij

# Evalvacija modelov

## Klasifikacija

## Regresija