Fakulteta za računalništvo in informatiko

1.seminarska raziskovalna naloga za predmet

Umetna inteligenca

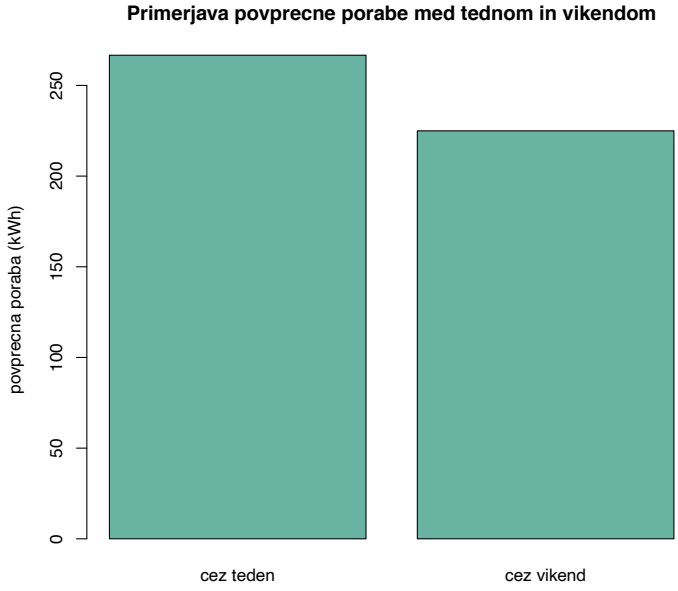
Gašper Rataj in Hana Čurk

1. Vizualizacija

V 1. točki seminarske raziskovalne naloge, sva želela prikazati nekaj razmerja med atributi in porazdelitve podatkov. Programska koda je shranjena v datoteki »vizualizacija.R«.

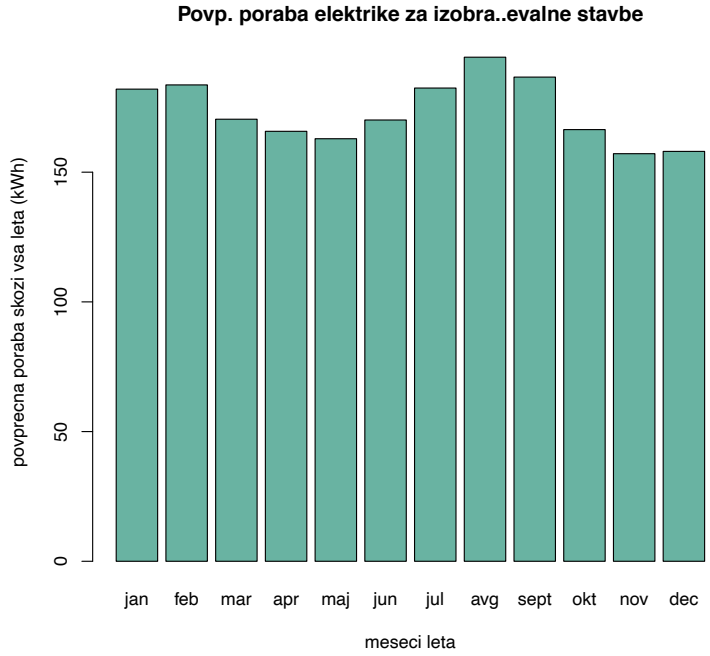
1.1 Povprečje porabe v poslovnih objektiv za delavne dni in med vikendom

S to meritvijo sva želela preveriti hipotezo, da je v poslovnih stavbah poraba električne energije večja čez teden kot čez vikend.
Neodvisno sva prikazala povprečje vseh meritev med vikendom in vseh meritev za delovne dni. Kot pričakovano, je med delovnimi dnevi poraba v poslovnih prostorih večja, a ne opazimo ogromne razlike.



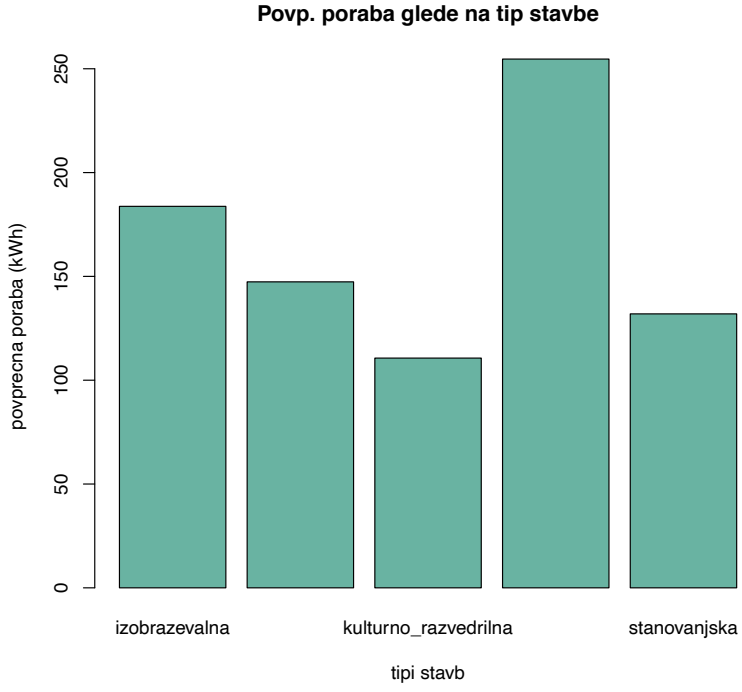
1.2 Poraba elektrike skozi leto v izobraževalnih stavbah

S to vizualizacijo sva želela ugotoviti, ali je poleti poraba elektrike v izobraževalnih stavbah kaj manjša kot skozi šolsko leto. Končni podatki so čuda pokazali nasprotno.



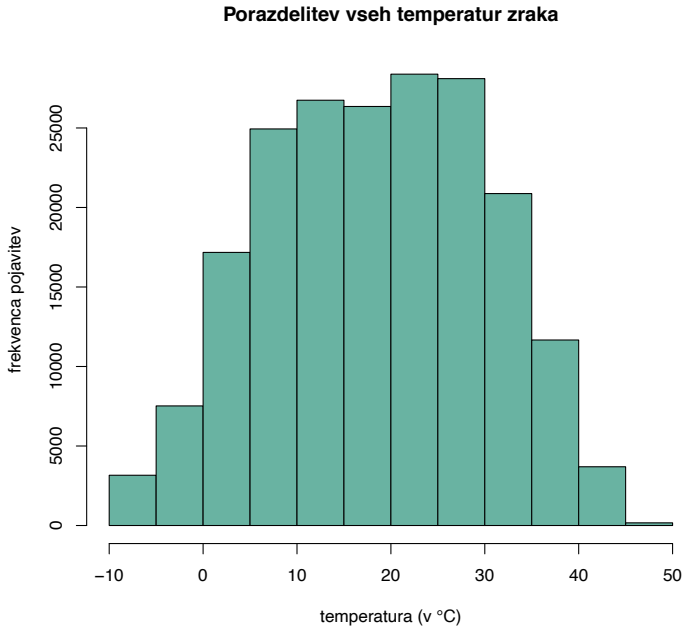
1.3 Primerjalna povprečna poraba vsakega tipa izgradb

Želela sva ugotoviti, kateri tipi stavb porabijo največ električne energije. Ugotovila sva, da prednjačijo stavbe za poslovno uporabo, najmanj pa porabijo tiste, ki so namenjene kulturi in razvedritvi.



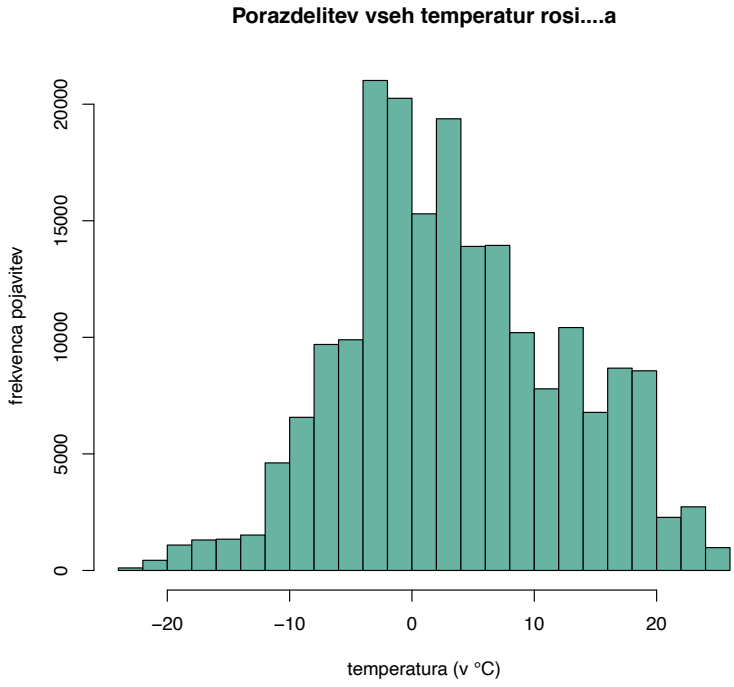
1.4 Porazdelitev temperatur zraka

S tem sva poiskala porazdelitev podatkov temperatur zraka. S tehniko ostrega pogleda bi lahko sklepali, da so podatki normalno porazdeljeni, ne pa nujno tudi simetrični.



1.5 Porazdelitev temperatur rosišča

Poleg temperatur zraka sva želela izmeriti tudi vse temperature rosišč. Tudi tu bi oster pogled narekoval, da so podatki normalno porazdeljeni.



Dodajanje atributov

Želela sva dodati atribute, s katerimi bi bilo učenje modelov čim boljše. Dodala sva tri nove atribute in sicer:

- Letni čas
- Delavnik ali vikend
- Povprečna poraba prejšnega dneva

2. KLASIFIKACIJA

Za nalogo sva morala narediti modele, s katerimi napovedujemo porabo. Za gradnjo modela, sva množico razdelila na učno in testno podmnožico. Za učno množico sva uporabila 70% naključnih podatkov, preostalo pa za testno.
Preizkusila sva naslednje modele:

- Odločitveno drevo
- Naključni gozdovi
- Naivni Bayes

Najprej sva za učno množico uporabila 70% naključnih podatkov, preostale pa za testno. Tako sva testirala na osnovni množici podatkov z vsemi atributi. Potem sva testirala na celotni množici z dodanimi tremi atributi in še na množici najboljše izbranih atributov, ki sva jih izbrala s pomočjo fukncije »wrapper«.

Izbrani atributi: povp_prejsni_dan + povrsina + namembnost + leto_izgradnje + temp_zraka + temp_rosisca

	Vsi atributi		**Najboljši izbrani**	
	CA	BR	CA	BR
ODLOČITVENO DREVO	0.8159778	0.2770505	0.8469598	0.253599
NAKLJUČNI GOZD	0.7741221	0.253599	0.8640628	0.2323974
NAIVNI BAYES	0.4296303	0.7683837	0.4340069	0.7198156

Ugotovila sva, da je Navni Bayes v vsakem primeru najslabši. Najboljši pa naključni gozd testiran na izbranih atributih.

Nato sva na množici z najboljše ocenjenimi atributi naredila evalvacijo modelov, pri kateri sva uporabila k-kratno prečno preverjanje. Na podatkih sva naredila 12 podmnožic in za vsak model imela 11 iteracij.

ODLOČITVENO DREVO

Iteracija	Najboljši izbrani	
	CA	BR
1	0.5678704	0.5686492
2	0.5686492	0.4770872
3	0.5707362	0.4770872
4	0.5891110	0.4495315
5	0.5900549	0.4149412
6	0.5793610	0.3907584
7	0.5775298	0.404451
8	0.5862493	0.3882659
9	0.5974220	0.3866541
10	0.6050524	0.3648421
11	0.5843254	0.3294482

NAKLJUČNI GOZD

Iteracija	Najboljši izbrani	
	CA	BR
1	0.6865078	0.4145615
2	0.6971671	0.4041007
3	0.6998787	0.3993948
4	0.7073995	0.3880128
5	0.7429685	0.3645762
6	0.7579784	0.3452489
7	0.7561502	0.3487112
8	0.7641343	0.3454805
9	0.7692231	0.3361296
10	0.7862569	0.3107149
11	0.8037492	0.3084924

NAIVNI BAYES

Iteracija	Najboljši izbrani	
	CA	BR
1	0.3492242	0.8156802
2	0.3612673	0.7909624
3	0.3875364	0.7708154
4	0.3913766	0.7669845
5	0.3971229	0.7642909
6	0.402956	0.7534526

7	0.4118627	0.7390161
8	0.420867	0.7277477
9	0.4264064	0.7181343
10	0.4284018	0.7162341
11	0.4353815	0.7182263

3. REGRESIJA

Pri regresiji sva ubrala enak način dela, kot pri klasifikaciji. Spodaj so predstavljeni podatki.
Najboljše izbrani podatki: povrsina + leto_izgradnje + namembnost + povp_prejsni_dan

	Vsi atributi		**Najboljši izbrani**	
	RMAE	RMSE	RMAE	RMSE
LINEARNA REGRESIJA	0.1982996	0.06511798	0.1947775	0.0710614
REGRESIJSKO DREVO	0.1184551	0.03267074	0.200669	0.06937854
RANDOM FOREST	0.2538694	0.07937482	0.2322435	0.07528223

REGRESIJSKO DREVO

Iteracija	Najboljši izbrani	
	rmae	rmse
1	0.1779702	0.07044568
2	0.1674597	0.06134495
3	0.1776906	0.06689565
4	0.1752789	0.06597984
5	0.1760204	0.06451683
6	0.1762456	0.06503441
7	0.1723275	0.06211397
8	0.1648365	0.07290199
9	0.1754223	0.08698322
10	0.1677896	0.08187411
11	0.1615627	0.07313208

RANDOM FOREST

Iteracija	Najboljši izbrani	
	rmae	rmse
1	0.2578963	0.08866775
2	0.2710964	0.09552465
3	0.2646489	0.09088415
4	0.2581141	0.09035137
5	0.2567388	0.08966721
6	0.2615143	0.09064329
7	0.2612166	0.08780646
8	0.2588133	0.08759198
9	0.2575392	0.08729522
10	0.2529482	0.08692542
11	0.2523853	0.08632583

LINEARNA REGRESIJA

Iteracija	Najboljši izbrani	
	rmae	rmse
1	0.1907849	0.07099664
2	0.1901512	0.07308739
3	0.1919423	0.07443347
4	0.1938377	0.07442624
5	0.1972753	0.07634786

6	0.1986494	0.07718319
7	0.2004079	0.07682301
8	0.1970974	0.07762016
9	0.1905551	0.0768477
10	0.1840915	0.07264615
11	0.1778737	0.06974729