Umetna inteligenca 2021-2022

Seminarska naloga 1

Jan Vovk, Aljaž Hribar

30 November – 5. December 2021

# Vizualizacija učne množice

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

A picture containing chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Ugotovila sva da je med temperaturo zraka in temperaturo rosišča pozitivna korelacija, vendar ta ni povsem linearna. Med pritiskom in številom padavin ni bilo opazne korelacije. Prav tako ni bilo opazne korelacije med hitrostjo in smerjo vetra. Med porabo in površino je bila opazna pozitivna korelacija, vendar tudi ta ni bila linearna. Porazdelitev podatkov po stavbah ni enakomerna, vendar je zanimivo da obstaja neka enakomernost po segmentih, stavbe do cca 80 so enakomerno porazdeljene in imajo največjo frekvenco, nato je en kratek segment stavb z malo manjšo frekvenco, nato je bil zopet daljši segment stavb od cca 90 do 190, kjer je bila frekvenca še manjša, in nato zadnjih 10 stavb je pa imelo najmanjšo frekvenco, z zelo opazno razliko (cca 4-krat manjša kot frekvenca prejšnjega segmenta stavb. Presenetljivo je bila povprečna poraba pozimi manjša kot poleti, prav tako so bile pozimi tudi večje maksimalne vrednosti.

# Ocenjevanje in konstrukcija atributov

Sprva sva naredila data frame za testne in učne podatke.

ob pregledu učne množice z ukazom summray() sva opazimo da lahko je smer vetra podana koz zvezni podatek ampak bi nam bila bolj koristna kot diskretni zato sva ga faktorizirala na 9 stopenj; 8 smeri in brezvetrovje.

table(ucna$smer\_vetra) ##

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## brezveterje | jug | jugo\_vzhod | jugo\_zahod | sever | severo\_vzhod |
| ## 2482 | 4210 | 3325 | 1044 | 1740 | 2670 |
| ## severo\_zahod | vzhod | zahod |  |  |  |
| ## 2831 | 3799 | 2024 |  |  |  |

iz atributa “datum” sva naredila atributa “season”, ki nam pove letni čas meritve in atribut “vikend” ki nam pove ali je na ta dan bil vikend ali delovni teden, saj je atribut “datum” v trenutni obliki precej neuporaben

table(ucna$season)

##

## Fall Spring Summer Winter ## 6741 4235 5140 8009

table(ucna$vikend)

##

## FALSE TRUE ## 17184 6941

prav tako sva iz atributa “poraba” izvlečemo atributa “dosedanja\_povpreča” in “dosedanja\_skupna” ki nam povesta kolikšna je povprečna in skupna poraba stavbe do vključno trenutnega datuma meritve

summary(ucna$dosedanja\_povprecna)

## Min. 1st Qu. Median Mean 3rd Qu. Max. ## 3.616 60.733 118.774 228.265 204.453 2196.688

summary(ucna$dosedanja\_skupna)

## Min. 1st Qu. Median Mean 3rd Qu. Max. ## 3.7 4531.1 12638.2 28874.0 30728.0 424149.1

sedaj lahko izločimo atribut stavba saj je za klasifikacijo odvečen atribut, ki bi samo kvaril modele

ucna$stavba<-NULL testna$stavba<-NULL

# Klasifikacijski problem

z attrEval() funkcijo ocenimo atribute,

library(CORElearn)

sort(attrEval(namembnost~., ucna,"Relief"),decreasing =TRUE)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## | dosedanja\_povprecna | povrsina | poraba | dosedanja\_skupna |
| ## | 0.061888110 | 0.035969465 | 0.035938575 | 0.035583211 |
| ## | leto\_izgradnje | regija | padavine | season |
| ## | 0.016055510 | 0.000000000 | -0.000613520 | -0.002901554 |
| ## | vikend | temp\_zraka | temp\_rosisca | pritisk |
| ## | -0.005015544 | -0.007857875 | -0.011347804 | -0.013097497 |
| ## | oblacnost | smer\_vetra | hitrost\_vetra | datum |
| ## | -0.013471503 | -0.015046632 | -0.015718982 | -0.057616580 |

sort(attrEval(namembnost~., ucna,"ReliefFequalK"),decreasing =TRUE)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## | leto\_izgradnje | povrsina | regija | dosedanja\_povprecna |
| ## | 0.242451468 | 0.166456481 | 0.139201471 | 0.136221147 |
| ## | poraba | dosedanja\_skupna | datum | temp\_zraka |
| ## | 0.115882015 | 0.105300619 | 0.102695830 | 0.082560008 |
| ## | smer\_vetra | pritisk | temp\_rosisca | hitrost\_vetra |
| ## | 0.067548928 | 0.062092321 | 0.047547866 | 0.040365334 |
| ## | oblacnost | season | vikend | padavine |
| ## | 0.038386700 | 0.027151592 | 0.007127530 | 0.004231328 |
| sort(attrEval(namembnost~., ucna,"ReliefFexpRank"),decreasing =TRUE) | | | | |
| ## | leto\_izgradnje | povrsina | regija | datum |
| ## | 0.260707133 | 0.166830040 | 0.164866503 | 0.119054407 |
| ## | dosedanja\_povprecna | poraba | temp\_zraka | dosedanja\_skupna |
| ## | 0.115823728 | 0.095175641 | 0.094856759 | 0.085446344 |
| ## | smer\_vetra | pritisk | temp\_rosisca | oblacnost |
| ## | 0.079984195 | 0.074110398 | 0.056642911 | 0.048480491 |
| ## | hitrost\_vetra | season | vikend | padavine |
| ## | 0.046881204 | 0.033750883 | 0.008384743 | 0.004979457 |
|  | | | | |

iz avaluacije atributov sva opazila da imajo atributi ki opisujejo vremenske razmere in datum meritev le teh

zelo majhno povezavo z namembnostjo stavbe zato sva jih izločila

## gradnja modelov

### odločitveno drevo

najprej sva zgradila model z vsemi atributi

rpart.plot(dt)



izobrazevalna

.55 .12 .14 .13 .06

100%

*yes* **povrsina >= 3495** *no*

kulturno\_razvedrilna

.07 .38 .40 .14 .00

18%

**leto\_izgradnje < 1981**

izobrazevalna

.66 .07 .08 .12 .08

82%

**leto\_izgradnje < 1972** izobrazevalna

.46 .08 .17 .22 .07

37%

**leto\_izgradnje >= 1991**

kulturno\_razvedrilna

.00 .10 .75 .15 .00

8%

**dosedanja\_povprecna >= 8.7**

izobrazevalna

.81 .06 .00 .05 .08

45%

**povrsina < 29e+3** javno\_storitvena

.00 .67 .00 .33 .00

3%

**povrsina >= 34e+3**

poslovna

.13 .07 .39 .42 .00

13%

**leto\_izgradnje < 1981**

javno\_storitvena

.14 .64 .09 .14 .00

9%

**regija = vzhodna** izobrazevalna

.50 .00 .00 .50 .00

3%

**povrsina < 1914**

izobrazevalna

.86 .02 .00 .03 .09

43%

**regija = vzhodna** izobrazevalna

.76 .00 .00 .06 .18

21%

**leto\_izgradnje >= 1964**

izobrazevalna

.64 .09 .05 .11 .11

24%

**povrsina >= 16e+3** izobrazevalna

.40 .15 .09 .18 .18

14%

**leto\_izgradnje >= 2014**

javno\_storitvena

.00 .87 .13 .00 .00

7%

poslovna **dosedanja\_p**k**o**u**v**lt**p**u**r**r**e**n**c**o**n**\_**a**ra**<**zv**7**e**7**drilna

.00 .00 .23 .77 .00 .00 .50 .50 .00 .00

kulturno\_razvedrilna

.00 .12 .88 .00 .00

7%

**leto\_izgradnje >= 2010**

6%

**leto\_izgradnje >= 1988**

2%

**povrsina >= 2733**

izobrazevalna

.96 .04 .00 .00 .00

21%

**povrsina >=**izo**4**b**2**r**9**a**6**zevalna

.67 .33 .00 .00 .00

3%

**povrsina < 4135**

izobrazevalna

.56 .00 .00 .11 .33

11%

**dosedanja\_povprecna >= 164**

stanovanjska

.11 .26 .16 .16 .32

8%

**povrsina < 12e+3**

kulturno\_razvedrilna

.22 .11 .50 .17 .00

8%

**regija = vz**k**h**u**o**lt**d**u**n**rn**a**o\_razvedrilna

.00 .00 .75 .25 .00

5%

**leto\_izgradnje >= 1974**

stanovanjska

.41 .00 .00 .15 .45

9%

**povrsina < 4561**

izobrazevalna

.79 .00 .00 .21 .00

6%

**leto\_izgradnje**iz**>**o**=**br**2**a**0**ze**1**v**5**alna

.50 .00 .00 .50 .00

3%

**povrsina >= 10e+3**

izobrazevalna

.67 .33 .00 .00 .00

3%

**povrsina >= 5275**

kulturno\_razvedrilna

.00 .00 .60 .40 .00

2%

**povrsina < 46e+3**

stanovanjska

.16 .00 .00 .21 .63

6%

**leto\_izgradnje < 1941**

stanovanjska

.13 .31 .19 .00 .37

7%

**leto\_izgradnje >**s**=**ta**2**n**0**o**1**v**2**anjska

.16 .38 .00 .00 .46

5%

**povrsina >= 9050**

stanovanjska

.20 .00 .00 .00 .80

5%

**dosedanja\_povprecna >= 128**

stanovanjska

.00 .46 .00 .00 .54

5%

**povrsina < 4820**

stanovanjska

.00 .25 .00 .00 .75

3%

**povrsina >= 7153**

stanovanjska

.43 .00 .00 .00 .57

2%

**povrsina < 9409**

izobrazevalna izobrazevalna izobrazevalna izobrazevalna stanovanjska

poslovna

izobrazevalna poslovna

izobrazevalna javno\_storitvena poslovna javno\_storitvena

poslovna

poslovna javno\_storitvena kulturno\_razvedrilna

poslovna kulturno\_razvedrilna

1.00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 1.00 .00 1.00 .00 .00 .00 .00 .00 .00 .00 1.00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 1.00 .00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 1.00 .00 .00

2% 10% 3% 1% 3% 1% 3% 1% 1% 1% 1% 1% 1% 1% 5% 1% 1% 6%

izobrazevalna javno\_storitvena izobrazevalna poslovna stanovanjska javno\_storitvena izobrazevalna izobrazevalna kulturno\_razvedrilna javno\_storitvena stanovanjska izobrazevalna kulturno\_razvedrilna kulturno\_razvedrilna poslovna javno\_storitvena izobrazevalna javno\_storitvena poslovna

1.00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 1.00 .00 .00 .00 1.00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 .00 1.00 1.00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 .00 1.00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 1.00 .00

19% 1% 3% 1% 1% 2% 10% 1% 1% 1% 3% 2% 4% 1% 3% 1% 1% 1% 1%

CA(observed,predicted) [1] 0.5093227

brier.score(obsMat, predMat)

[1] 0.9812315

nato pa še poskusila maksimizirati točnost z fukcijo wrapper(), ki je vrnila da so najboljšo točnost imeli atributi povrsina in leto\_izgradnje z pričakovano napako 0.007543904

rpart.plot(dt)



izobrazevalna

.55 .12 .14 .13 .06

100%

*yes* **povrsina >= 3495** *no*

kulturno\_razvedrilna

.07 .38 .40 .14 .00

18%

**leto\_izgradnje < 1981**

izobrazevalna

.66 .07 .08 .12 .08

82%

**leto\_izgradnje < 1972** izobrazevalna

.46 .08 .17 .22 .07

37%

**leto\_izgradnje >= 1991**

kulturno\_razvedrilna

.00 .10 .75 .15 .00

8%

**dosedanja\_povprecna >= 8.7**

izobrazevalna

.81 .06 .00 .05 .08

45%

**povrsina < 29e+3** javno\_storitvena

.00 .67 .00 .33 .00

3%

**povrsina >= 34e+3**

poslovna

.13 .07 .39 .42 .00

13%

**leto\_izgradnje < 1981**

javno\_storitvena

.14 .64 .09 .14 .00

9%

**regija = vzhodna** izobrazevalna

.50 .00 .00 .50 .00

3%

**povrsina < 1914**

izobrazevalna

.86 .02 .00 .03 .09

43%

**regija = vzhodna** izobrazevalna

.76 .00 .00 .06 .18

21%

**leto\_izgradnje >= 1964**

izobrazevalna

.64 .09 .05 .11 .11

24%

**povrsina >= 16e+3** izobrazevalna

.40 .15 .09 .18 .18

14%

**leto\_izgradnje >= 2014**

javno\_storitvena

.00 .87 .13 .00 .00

7%

poslovna **dosedanja\_p**k**o**u**v**lt**p**u**r**r**e**n**c**o**n**\_**a**ra**<**zv**7**e**7**drilna

.00 .00 .23 .77 .00 .00 .50 .50 .00 .00

kulturno\_razvedrilna

.00 .12 .88 .00 .00

7%

**leto\_izgradnje >= 2010**

6%

**leto\_izgradnje >= 1988**

2%

**povrsina >= 2733**

izobrazevalna

.96 .04 .00 .00 .00

21%

**povrsina >=**izo**4**b**2**r**9**a**6**zevalna

.67 .33 .00 .00 .00

3%

**povrsina < 4135**

izobrazevalna

.56 .00 .00 .11 .33

11%

**dosedanja\_povprecna >= 164**

stanovanjska

.11 .26 .16 .16 .32

8%

**povrsina < 12e+3**

kulturno\_razvedrilna

.22 .11 .50 .17 .00

8%

**regija = vz**k**h**u**o**lt**d**u**n**rn**a**o\_razvedrilna

.00 .00 .75 .25 .00

5%

**leto\_izgradnje >= 1974**

stanovanjska

.41 .00 .00 .15 .45

9%

**povrsina < 4561**

izobrazevalna

.79 .00 .00 .21 .00

6%

**leto\_izgradnje**iz**>**o**=**br**2**a**0**ze**1**v**5**alna

.50 .00 .00 .50 .00

3%

**povrsina >= 10e+3**

izobrazevalna

.67 .33 .00 .00 .00

3%

**povrsina >= 5275**

kulturno\_razvedrilna

.00 .00 .60 .40 .00

2%

**povrsina < 46e+3**

stanovanjska

.16 .00 .00 .21 .63

6%

**leto\_izgradnje < 1941**

stanovanjska

.13 .31 .19 .00 .37

7%

**leto\_izgradnje >**s**=**ta**2**n**0**o**1**v**2**anjska

.16 .38 .00 .00 .46

5%

**povrsina >= 9050**

stanovanjska

.20 .00 .00 .00 .80

5%

**dosedanja\_povprecna >= 128**

stanovanjska

.00 .46 .00 .00 .54

5%

**povrsina < 4820**

stanovanjska

.00 .25 .00 .00 .75

3%

**povrsina >= 7153**

stanovanjska

.43 .00 .00 .00 .57

2%

**povrsina < 9409**

izobrazevalna izobrazevalna izobrazevalna izobrazevalna stanovanjska poslovna izobrazevalna poslovna izobrazevalna javno\_storitvena poslovna javno\_storitvena poslovna poslovna javno\_storitvena kulturno\_razvedrilna poslovna kulturno\_razvedrilna 1.00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 1.00 .00 1.00 .00 .00 .00 .00 .00 .00 .00 1.00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 1.00 .00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 1.00 .00 .00

2% 10% 3% 1% 3% 1% 3% 1% 1% 1% 1% 1% 1% 1% 5% 1% 1% 6%

izobrazevalna javno\_storitvena izobrazevalna poslovna stanovanjska javno\_storitvena izobrazevalna izobrazevalna kulturno\_razvedrilna javno\_storitvena stanovanjska izobrazevalna kulturno\_razvedrilna kulturno\_razvedrilna poslovna javno\_storitvena izobrazevalna javno\_storitvena poslovna

1.00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 1.00 .00 .00 .00 1.00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 .00 1.00 1.00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 .00 1.00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 1.00 .00

19% 1% 3% 1% 1% 2% 10% 1% 1% 1% 3% 2% 4% 1% 3% 1% 1% 1% 1%

CA(observed,predicted) [1] 0.5005853

brier.score(obsMat, predMat)

[1] 0.9988294

ko sva pognala wraper() za minimizacijo “brier score” sva dobila podobne rezultate saj nama je funkcija vrnila “best model: estimated error = 0.01390878 , selected feature subset = namembnost ~ povrsina + leto\_izgradnje”

rpart.plot(dt)



izobrazevalna

.55 .12 .14 .13 .06

100%

*yes* **povrsina >= 3495** *no*

kulturno\_razvedrilna

.07 .38 .40 .14 .00

18%

**leto\_izgradnje < 1981**

izobrazevalna

.66 .07 .08 .12 .08

82%

**leto\_izgradnje < 1972** izobrazevalna

.46 .08 .17 .22 .07

37%

**leto\_izgradnje >= 1991**

kulturno\_razvedrilna

.00 .10 .75 .15 .00

8%

**leto\_izgradnje >= 2010**

izobrazevalna

.81 .06 .00 .05 .08

45%

**povrsina < 29e+3** javno\_storitvena

.00 .67 .00 .33 .00

3%

**povrsina >= 34e+3**

poslovna

.13 .07 .39 .42 .00

13%

**leto\_izgradnje < 1981**

javno\_storitvena

.14 .64 .09 .14 .00

9%

**povrsina <** j**9**a**0**vn**2**o\_storitvena

.00 .74 .11 .16 .00

8%

**povrsina < 1962**

kulturno\_razvedrilna

.00 .00 .83 .17 .00

8%

**povrsina >= 737**

izobrazevalna

.86 .02 .00 .03 .09

43%

**leto\_izgradnje < 1950** izobrazevalna

.81 .04 .00 .00 .16

24%

**leto\_izgradnje >= 1964**

izobrazevalna

.64 .09 .05 .11 .11

24%

**povrsina >= 16e+3** izobrazevalna

.40 .15 .09 .18 .18

14%

**leto\_izgradnje >= 2014**

poslovna

.00 .00 .23 .77 .00

6%

**leto\_izgradnje >= 1988**

poslovna

.00 .29 .29 .43 .00

3%

**leto\_izgradnje >= 1959**

poslovna

.00 .00 .40 .60 .00

2%

**leto\_izgradnje >= 2000**

izobrazevalna

.93 .00 .00 .07 .00

19%

**povrsina >**iz**=**o**7**b**3**ra**2**z**0**evalna

.83 .00 .00 .17 .00

8%

**povrsina < 6679**

izobrazevalna

.65 .07 .00 .00 .29

13%

**leto\_izgradnje < 1962**

stanovanjska

.11 .26 .16 .16 .32

8%

**povrsina < 12e+3**

kulturno\_razvedrilna

.22 .11 .50 .17 .00

8%

**leto\_izgradnje >= 1974**

izobrazevalna

.79 .00 .00 .21 .00

6%

**leto\_izgradnj**i**e**zo**>**b**=**ra**2**z**0**e**1**v**5**alna

.50 .00 .00 .50 .00

3%

**povrsina >= 10e+3**

kulturno\_razvedrilna

.27 .13 .60 .00 .00

6%

**leto\_izgradnje >=**ku**1**lt**9**u**7**rn**7**o\_razvedrilna

.18 .00 .82 .00 .00

5%

**povrsina < 12e+3**

kulturno\_razvedrilna

.00 .00 .60 .40 .00

2%

**povrsina < 46e+3**

javno\_storitvena

.00 .50 .50 .00 .00

2%

**povrsina >= 2733**

izobrazevalna

.71 .07 .00 .00 .21

12%

**povrsina >= 8**i**2**z**5**o**0**brazevalna

.47 .13 .00 .00 .40

6%

**leto\_izgradnje >= 1958**

stanovanjska

.13 .31 .19 .00 .37

7%

**leto\_izgradnje >**st**=**an**2**o**0**v**1**a**2**njska

.16 .38 .00 .00 .46

5%

**povrsina >= 9050**

izobrazevalna

.50 .50 .00 .00 .00

2%

**povrsina >= 5275**

stanovanjska

.27 .18 .00 .00 .54

5%

**povrsina < 4418**

stanovanjska

.00 .46 .00 .00 .54

5%

**povrsina < 4820**

kulturno\_razvedrilna

.40 .00 .60 .00 .00

2%

**povrsina >= 5941**

stanovanjska

.00 .25 .00 .00 .75

3%

**povrsina >= 7153**

izobrazevalna

.60 .40 .00 .00 .00

2%

**povrsina < 4190**

izobrazevalna izobrazevalna izobrazevalna javno\_storitvena stanovanjska poslovna izobrazevalna poslovna izobrazevalna javno\_storitvena poslovna javno\_storitvena kulturno\_razvedrilna poslovna poslovna izobrazevalna javno\_storitvena poslovna kulturno\_razvedrilna poslovna

1.00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 1.00 .00 1.00 .00 .00 .00 .00 .00 .00 .00 1.00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 1.00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00

6% 11% 2% 1% 1% 1% 3% 1% 1% 1% 1% 1% 1% 1% 1% 1% 1% 1% 5% 1%

izobrazevalna poslovna izobrazevalna izobrazevalna stanovanjska javno\_storitvena izobrazevalna izobrazevalna kulturno\_razvedrilna javno\_storitvena stanovanjska izobrazevalna izobrazevalna kulturno\_razvedrilna kulturno\_razvedrilna poslovna javno\_storitvena kulturno\_razvedrilna javno\_storitvena kulturno\_razvedrilna 1.00 .00 .00 .00 .00 .00 .00 .00 1.00 .00 1.00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 .00 .00 1.00 .00 1.00 .00 .00 .00 1.00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 .00 1.00 1.00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 1.00 .00 .00

11% 1% 5% 1% 3% 2% 10% 1% 1% 1% 3% 1% 1% 3% 1% 3% 5% 1% 1% 1%

CA(observed,predicted) [1] 0.5005853

brier.score(obsMat, predMat)

[1] 0.9988294

### naivni bayes

tukaj sem uporabil isti postopke kot pri gradnji drevesa in dobil naslednje rezultate: z vsemi atributi

CA(observed, predicted)

[1] 0.4466137

predMat<-predict(nb, testna,type ="prob") brier.score(obsMat, predMat)

[1] 0.7757613

wrapper z minimizacijo napake

CA(observed, predicted) [1] 0.4909699

predMat<-predict(nb, testna,type ="prob") brier.score(obsMat, predMat)

[1] 0.7259014

wrapper z minimizacijo brier

CA(observed, predicted) [1] 0.4909699

predMat<-predict(nb, testna,type ="prob") brier.score(obsMat, predMat)

[1] 0.7259014

### k-najbližjih sosedov

CA(observed, predicted) [1] 0.4701505

predMat<-predict(knn, testna,type ="prob") brier.score(obsMat, predMat)

[1] 1.006113

### naključno gozd

CA(observed, predicted) [1] 0.4965719

predMat<-predict(rf, testna,type ="prob") brier.score(obsMat, predMat)

[1] 0.7265849

### SVM

CA(observed, predicted) [1] 0.6938326

brier.score(obsMat, predMat) [1] 0.5699415

### Umetne nevronske mreže

najprej je bilo potrebno normalizirati zvezne atribute v učni in testni množici nato pa sva dobila te rezultate

CA(observed, predicted) [1] 0.5119147

predMat<-predict(nn, testna\_scaled,type ="raw") brier.score(obsMat, predMat)

[1] 0.8076981

### kombinirani modeli

za kombinirane modele sva se odločila uporabiti modele nevrenoske mreže, naibni bayes in SVM.

najprej sva poskusil z glasovanjem in dobila

CA(observed, predicted)

[1] 0.5402592

nato z vteženim glasovanjem

CA(observed, predicted)

[1] 0.5073997

nazadnje pa še z boostingom

CA(observed, predicted) [1] 0.5048077

Najboljše tri modele glede na klasifikacijsko točnost sva primerjala še glede na njihovo uspešnost pri učenju iz ene regije ali obeh.

Model SVM je pri podatkih iz vzhodne regije imel CA 0.488, nevronske mreže CA 0.128 in kombinacija modelov z glasovanjem 0.519. CA vseh modelov se je zaradi manjšega števila učnih primerov poslabšal prav tako je na njega vplivalo dejstvo da v učnih podatkih ni bilo primerov za stanovanjsko stavbo.

# Regresija

Pripravila sva učne podatke za posamezno regijo, enkrat za vzhodno in enkrat za zahodno.

Testne in učne podatke sva združila v novo spremenljivko in združen data frame razdelila na 12 mesecev.

Za linearno regresijo sva podatke logaritmirala z log(x-2) saj je atribut »padavine« imel minimalno vrednost -1. Logaritmirane podatke sva shranila v nove data frame, nov data frame združenih podatkov sva nato zopet razdelila na 12 mesecev.

Naredila sva funkciji za izračune relativnih napak (rmae in rmse), nato sva se lotila funkcij za izdelavo in evalvacijo modelov. Pri linearni regresiji sva observed in predicted podatke še potencirala z x^2 + 2, da sva dobila podatke normalne velikosti. Za izbor atributov sva vse atribute ocenila z attrEval (MSEofMean in RreliefFexpRank), in ocenila da so najbolj koristni atributi povrsina, dosedanja\_povprečna uporaba in dosedanja\_skupna\_poraba, saj so imeli najvišje vrednosti. Za random forest sva sprva poskusila uporabiti vse atribute, vendar je evalvacija trajala predolgo in sva zato izbrala iste atribute kot za knn. Da sva lahko uporabila lubridate, randomForest, kknn in CORElearn sva za njih pognala install.packages in library. Za knn in linearno regresijo sva naredila dve funckiji, eno z vsemi atributi in eno z samo izbranimi.

## REZULTATI:

poženemo funkcije brez posebnih pogojev

### linearna regresija

> linregAll(trainLR,testLR)

[1] 0.2252847

[1] 0.0752879

> linreg(trainLR,testLR)

[1] 0.2327933

[1] 0.08356924

>

### knn

> knnAll(train,test)

[1] 0.4593003

[1] 0.1843391

> knn(train,test)

[1] 0.3858386

[1] 0.2289143

>

### random forest

> forest(train,test)

[1] 0.3783901

[1] 0.2428684

>

modeli nauceni le na zahodni regiji

### linearna regresija

> linreg(trainWLR,testLR)

[1] 0.2131239

[1] 0.07756583

>

### knn

> knn(trainW,test)

[1] 0.4609916

[1] 0.413861

>

### random forest

> forest(trainW,test)

[1] 0.426851

[1] 0.2709027

>

modeli nauceni le na vzhodni regiji

### linearna regresija

> linreg(trainELR,testLR)

[1] 0.2698102

[1] 0.1010479

>

### knn

> knn(trainE,test)

[1] 0.4730309

[1] 0.3103466

>

### random forest

> forest(trainE,test)

[1] 0.4242804

[1] 0.3093009

>

izvedemo modele po mesecih

### linearna regresija

> linreg(janLR,febLR)

[1] 0.1364018

[1] 0.02997497

> tr <- rbind(janLR,febLR)

> linreg(tr,marLR)

[1] 0.1598174

[1] 0.04765475

> tr <- rbind(tr,marLR)

> linreg(tr,aprLR)

[1] 0.1774466

[1] 0.05781151

> tr <- rbind(tr,aprLR)

> linreg(tr,mayLR)

[1] 0.1655619

[1] 0.03731816

> tr <- rbind(tr,mayLR)

> linreg(tr,junLR)

[1] 0.2332104

[1] 0.06259062

> tr <- rbind(tr,junLR)

> linreg(tr,julLR)

[1] 0.2540519

[1] 0.09387624

> tr <- rbind(tr,julLR)

> linreg(tr,augLR)

[1] 0.2636536

[1] 0.09123858

> tr <- rbind(tr,augLR)

> linreg(tr,sepLR)

[1] 0.2166214

[1] 0.06394083

> tr <- rbind(tr,sepLR)

> linreg(tr,octLR)

[1] 0.2298427

[1] 0.09858141

> tr <- rbind(tr,octLR)

> linreg(tr,novLR)

[1] 0.2674388

[1] 0.1297452

> tr <- rbind(tr,novLR)

> linreg(tr,decLR)

[1] 0.2463883

[1] 0.09214778

>

### knn

> knn(jan,feb)

[1] 0.2819983

[1] 0.1172519

> tr <- rbind(jan,feb)

> knn(tr,mar)

[1] 0.2295438

[1] 0.08741343

> tr <- rbind(tr,mar)

> knn(tr,apr)

[1] 0.2269071

[1] 0.09232786

> tr <- rbind(tr,apr)

> knn(tr,may)

[1] 0.2304696

[1] 0.07404212

> tr <- rbind(tr,may)

> knn(tr,jun)

[1] 0.2517021

[1] 0.09738879

> tr <- rbind(tr,jun)

> knn(tr,jul)

[1] 0.2181064

[1] 0.1026969

> tr <- rbind(tr,jul)

> knn(tr,aug)

[1] 0.2568558

[1] 0.1074451

> tr <- rbind(tr,aug)

> knn(tr,sep)

[1] 0.1869878

[1] 0.05661387

> tr <- rbind(tr,sep)

> knn(tr,oct)

[1] 0.2946156

[1] 0.209515

> tr <- rbind(tr,oct)

> knn(tr,nov)

[1] 0.2513466

[1] 0.1176842

> tr <- rbind(tr,nov)

> knn(tr,dec)

[1] 0.2532888

[1] 0.1184119

>

### random forest

> forest(jan,feb)

[1] 0.2342854

[1] 0.07663957

> tr <- rbind(jan,feb)

> forest(tr,mar)

[1] 0.2242694

[1] 0.08060193

> tr <- rbind(tr,mar)

> forest(tr,apr)

[1] 0.2269962

[1] 0.08337074

> tr <- rbind(tr,apr)

> forest(tr,may)

[1] 0.2452446

[1] 0.08028867

> tr <- rbind(tr,may)

> forest(tr,jun)

[1] 0.2442828

[1] 0.08367229

> tr <- rbind(tr,jun)

> forest(tr,jul)

[1] 0.2322788

[1] 0.0942318

> tr <- rbind(tr,jul)

> forest(tr,aug)

[1] 0.2385721

[1] 0.08724658

> tr <- rbind(tr,aug)

> forest(tr,sep)

[1] 0.2114393

[1] 0.06534948

> tr <- rbind(tr,sep)

> forest(tr,oct)

[1] 0.3154071

[1] 0.200416

> tr <- rbind(tr,oct)

> forest(tr,nov)

[1] 0.3166819

[1] 0.1730465

> tr <- rbind(tr,nov)

> forest(tr,dec)

[1] 0.3111148

[1] 0.1362544

Pri uporabi navadnih testnih in učnih podatkov so bili rezultati za linearno regresijo najboljši, pri uporabi vseh atributov sta bila rmae in rmse boljša od takrat ko smo izbrali le specifične atribute. Rezultati pri knn so bili tudi dokaj dobri, pri uporabi vseh atributov je bil rmse boljši, pri uporabi le izbranih atributov pa je bil boljši rmae. Rezultati random forest so bili boljši od rezultatov knn, vendar se bi bili še bolj natančni če bi uporabila najbolj optimalne atribute. Te sva poskusila dobiti s pomočjo wrapper.R, vendar se je funkcija izvajala predolgo in sva zato izbrala kar iste atribute za vse tri funkcije.

Podatki naučeni le na eni regiji so bili načeloma slabši od tistih naučenih na obeh regijah, z izjemo rmae linearne regresije, ki je bil manjši pri uporabi podatkov naučenih le na vzhodni regiji. Načeloma so bili rezultati podatkov naučenih na vzhodnji regiji boljši od tistih naučenih na zahodnji regiji.

Pri uporabi podatkov razdeljenih na mesece, se je natančnost podatkov spreminjala, z najbolj natančnimi rezultati po učenju na podatkih iz januarja. Natančnost je nato nihala, kdaj sta bila rmae ali rmse bolj natančna od prejšnjega rezultata, kdaj pa sta bila manj natančna. Rezultati so bili pri učenju modelov na vseh mesecih do novembra in testiranju na podatkih iz decembra precej podobni podatkom naučenih na originalni tesni množici. Podatki random forest so bili bolj najmanj natančni, nato so bili podatki knn, najbolj natančna pa je bila zopet linearna regresija.

## ZAKLJUČKI:

Linearna regresija je bil najhitrejši in presenetljivo tudi najbolj natančen model. Rezultati so bili bolj natančni pri uporabi vseh atributov le v primeru linearne regresije. Random forest ni uporabil najbolj optimalnih atributov, tako da se bi lahko natančnost tu lahko povečala, vendar bi bilo izvajanje opazno daljše. Modeli naučeni na podatkih iz ene regije so bili slabši od tistih naučenih na podatkih obeh regij. Natančnost modelov naučenih po mesecih se je spreminjala od meseca do meseca, ni le naraščala ali le padala. Na splošno so bili rezultati več ali manj dokaj natančni.