Seminarska naloga iz strojnega učenja

Umetna Inteligenca

Nejc Vrčon Zupan (63200327)

Luka Šveigl (63200301)

Vsebina

[**1.** **Uvod** 3](#_Toc89417717)

[**2.** **Dodajanje atributov** 3](#_Toc89417718)

[**3.** **Vizualizacija podatkov** 4](#_Toc89417719)

[**4.** **Ocenjevanje atributov** 9](#_Toc89417720)

[**5.** **Modeliranje in evalvacija modelov** 9](#_Toc89417721)

[**5.1** **Klasifikacijsko modeliranje** 9](#_Toc89417722)

[**5.2** **Regresijsko modeliranje** 10](#_Toc89417723)

[**5.3** **Primerjava uspešnosti modelov na podatkih ene regije** 10](#_Toc89417724)

[**5.3.1** **Odločitveno drevo** 10](#_Toc89417725)

[**5.3.2** **Linearna regresija** 10](#_Toc89417726)

[**5.4** **Kombiniranje modelov strojnega učenja** 11](#_Toc89417727)

[**5.5** **Kronologija podatkov** 11](#_Toc89417728)

[**6.** **Zaključek** 12](#_Toc89417729)

# **Uvod**

Za seminarsko nalogo iz strojnega učenja je bilo potrebno iz podanih množic (učne in testne) oblikovati modele, ki napovedujejo porabo električne energije in namembnost stavbe.

Začela sva z branjem in čiščenjem podatkovnih množic. Spreminjala sva tipe določenih stolpcev v podatkovnih okvirjih in jih uredila po številki stavbe in datumu merjenja porabe, naraščajoče.

# **Dodajanje atributov**

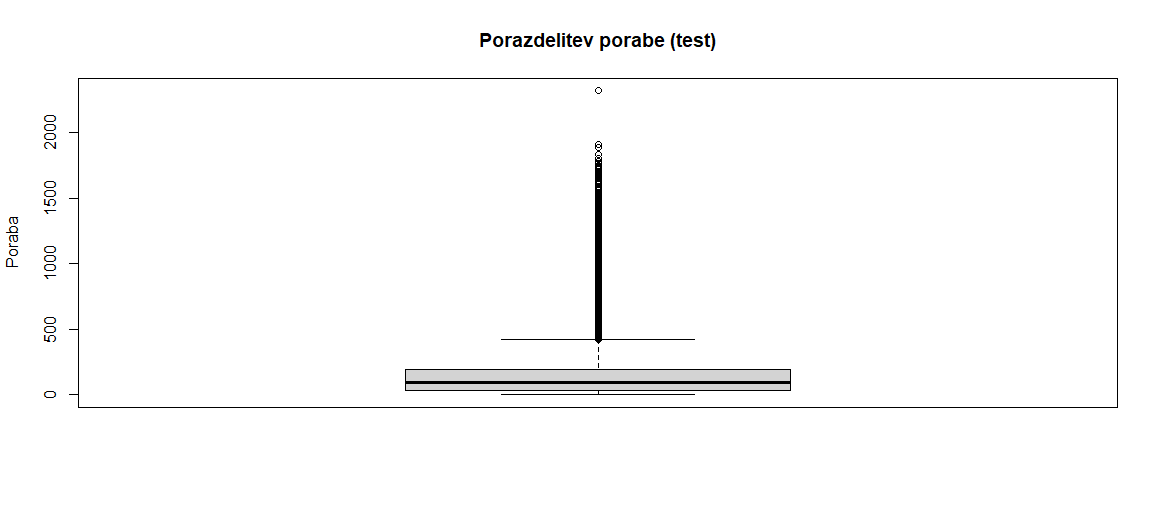
Pri dodajanju atributov sva pazljivo premislila o njihovi smiselnosti oz. odvisnosti od porabe energije. Cilj dodajanja atributov je bil izoblikovati čim bolj natančen model. Leto sva razbila na četrtletja / letne čase, teden pa na delovne dni in vikend. Prav tako sva dodala atribut povprečne porabe električne energije prejšnjega tedna. Atribute sva dodala v oba podatkovna okvirja (učni in testni). Ker so algoritmi precenili atribute stavba in datum sva se odločila, da jih bova odstranila iz obeh podatkovnih okvirjev.

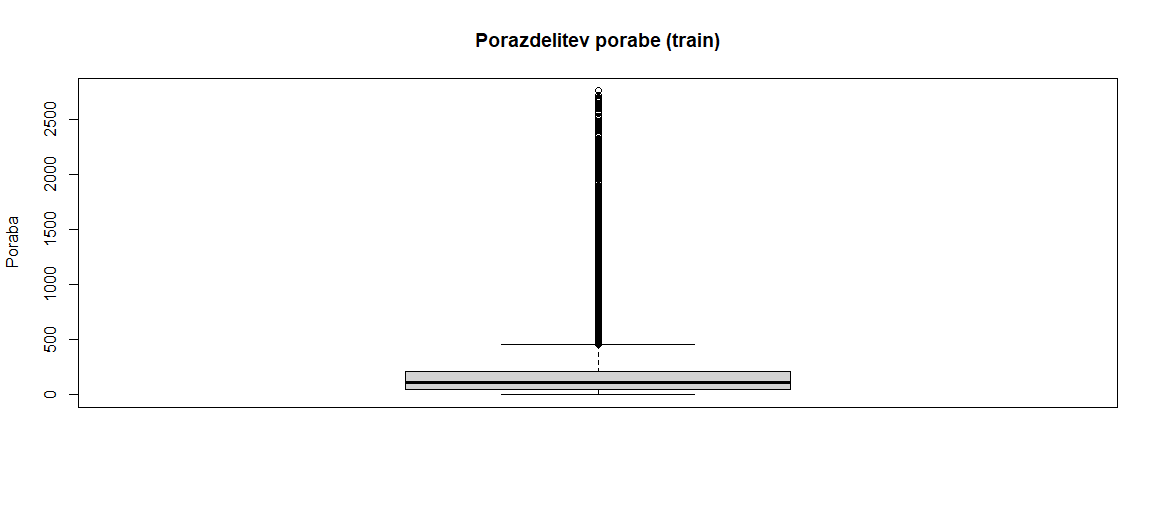
V podatkovnih okvirijih sva imela naslednje atribute:

* regija
* namembnost
* povrsina
* leto\_izgradnje
* temp\_zraka
* temp\_rosisca
* oblacnost
* padavine
* pritisk
* smer\_vetra
* hitrost\_vetra
* poraba
* *letni\_cas*
* *vikend*
* *poraba\_prteden*

# **Vizualizacija podatkov**

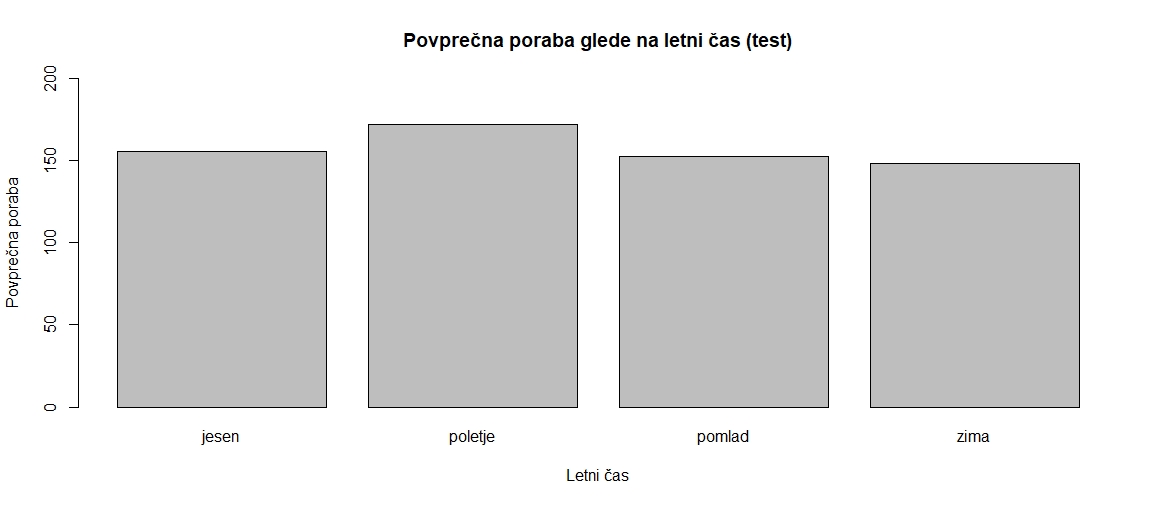
Prva grafa prikazujeta porazdelitev atributa porabe. Večina vrednosti je strnjenih med 0-250, vendar imamo tudi velika odstopanja. Opazimo lahko, da poraba ni normalno porazdeljena.

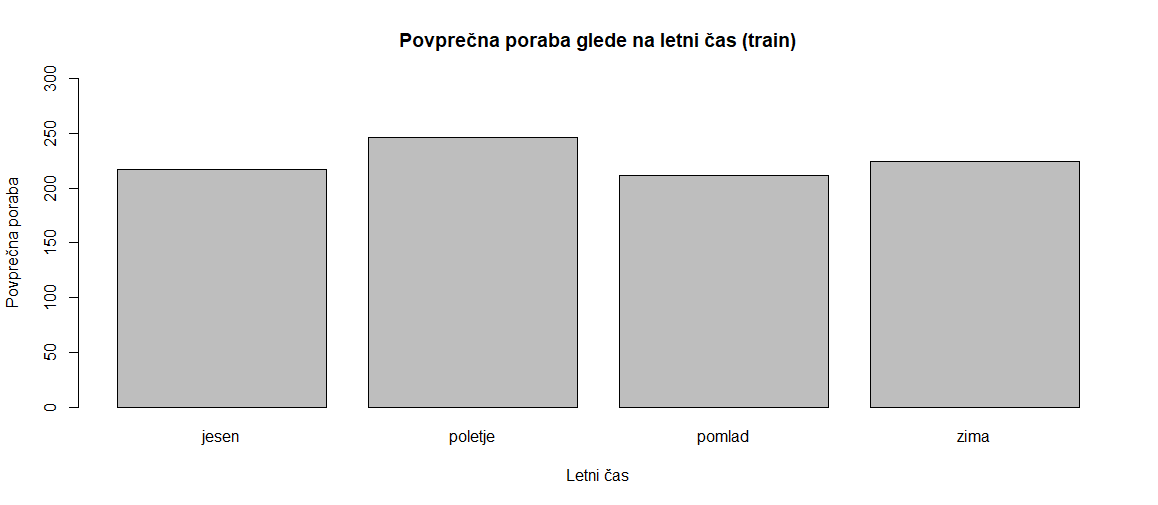


Slika 1: Porazdelitev porabe (test)

Slika 2: Porazdelitev porabe (train)

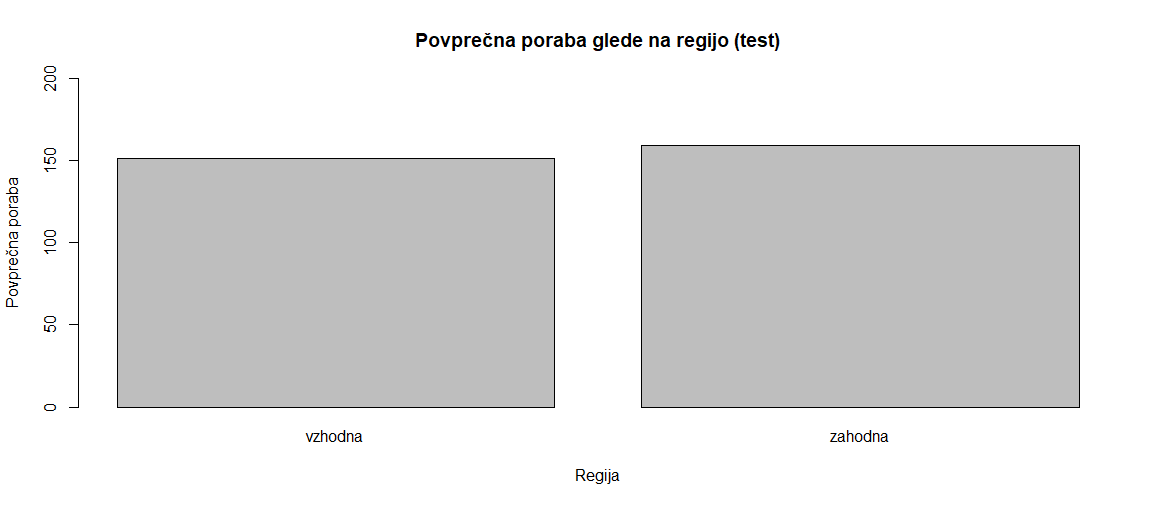
Naslednja dva grafa prikazujeta povprečno porabo glede na letni čas. Opazimo, da je poraba električne enrgije največja v poletnem letnem času. Sklepava, da se poraba poveča zaradi potrebe hlajenja prostorov.

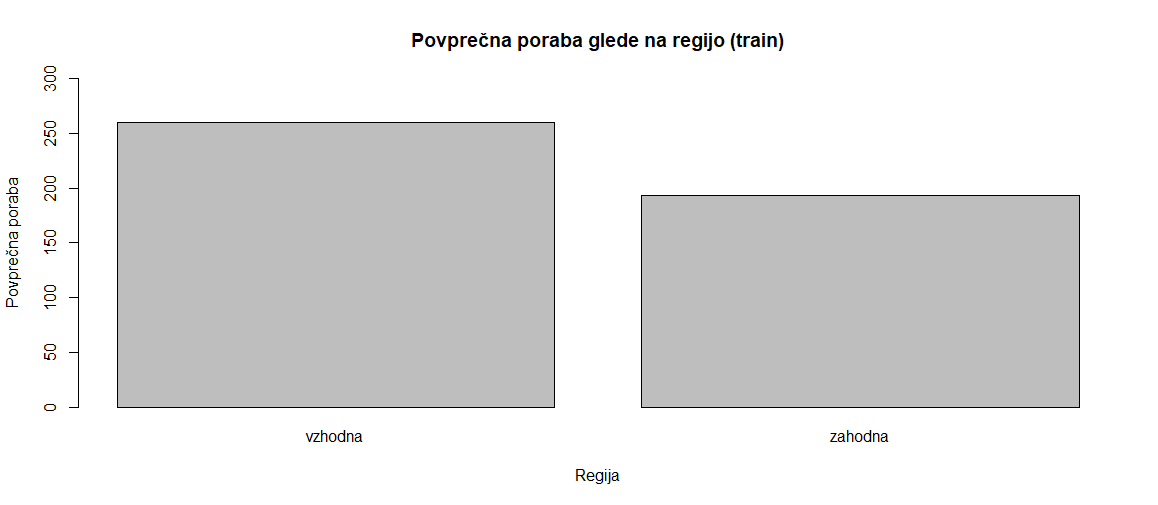
Slika 3: Povprečna poraba glede na letni čas (test)



Slika 4: Povprečna poraba glede na letni čas (train)

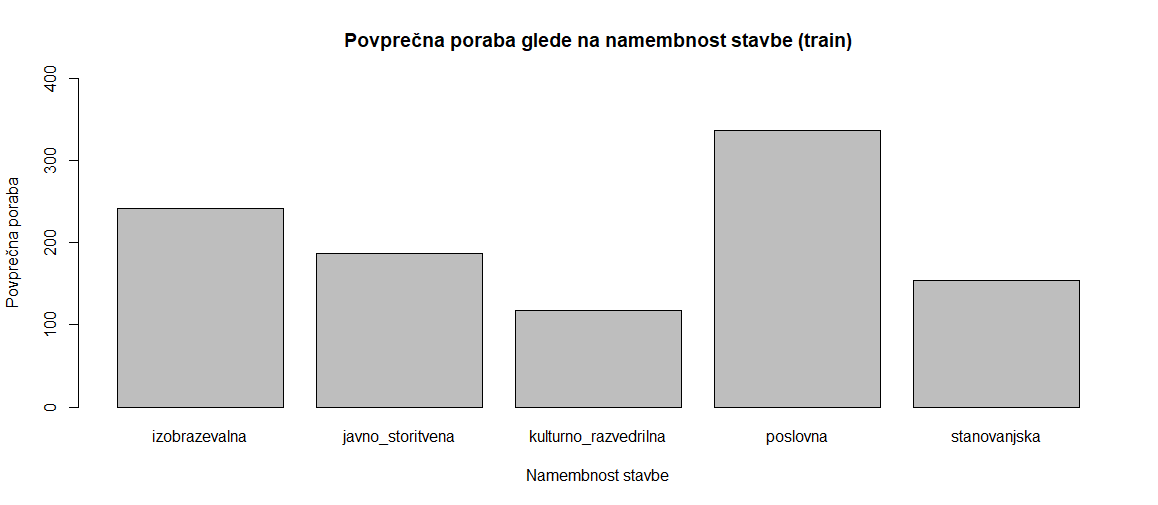
Naslednja dva grafa prikazujeta povprečno porabo glede na regijo. Opazimo, da se regiji z največjo povprečno porabo obrneta v test in train podatkovnih okvirjih.

Slika 5: Povprečna poraba glede na regijo (test)

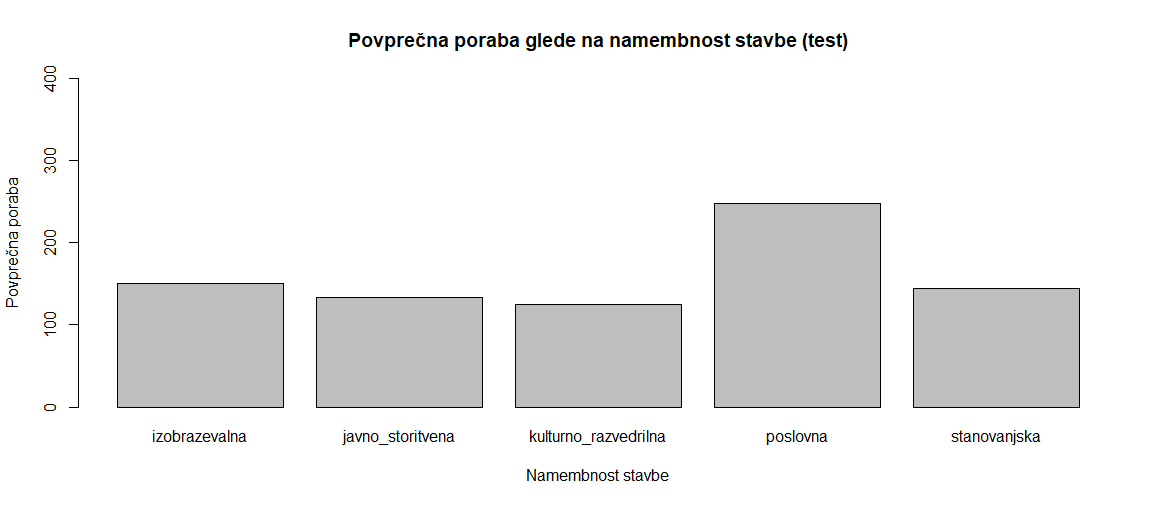


Slika 6: Povprečna poraba glede na regijo (train)

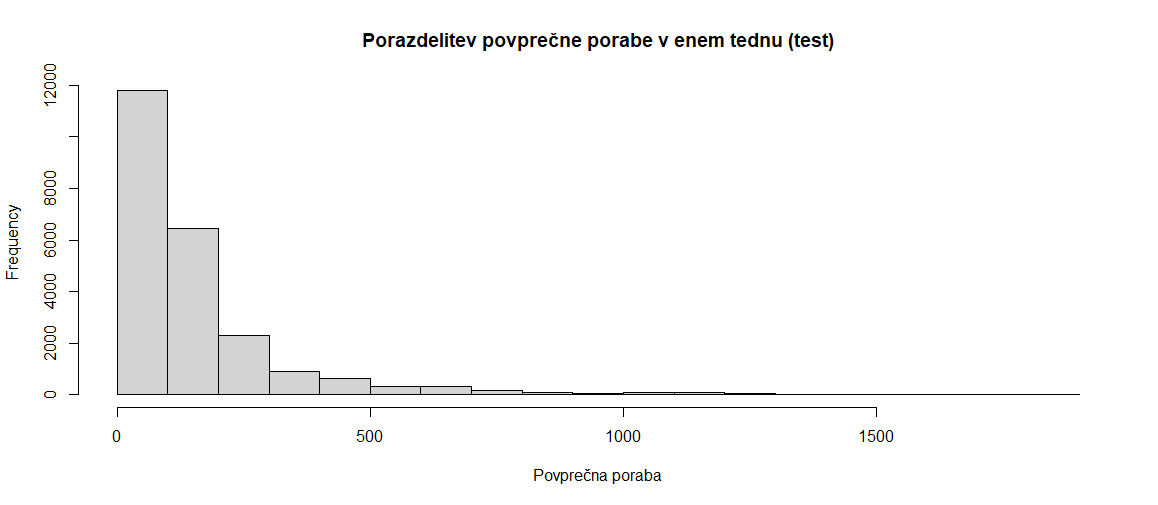
Naslednja dva grafa prikazujeta povprečno porabo glede na namebnost stavbe. Opazimo, da poslovne stavbe v povprečju porabijo največ električne energije. Sklepava, da imajo poslovne stavbe večje število porabnikov električne energije, prav tako pa uporabljajo večje sisteme, ki imajo seveda večjo porabo.

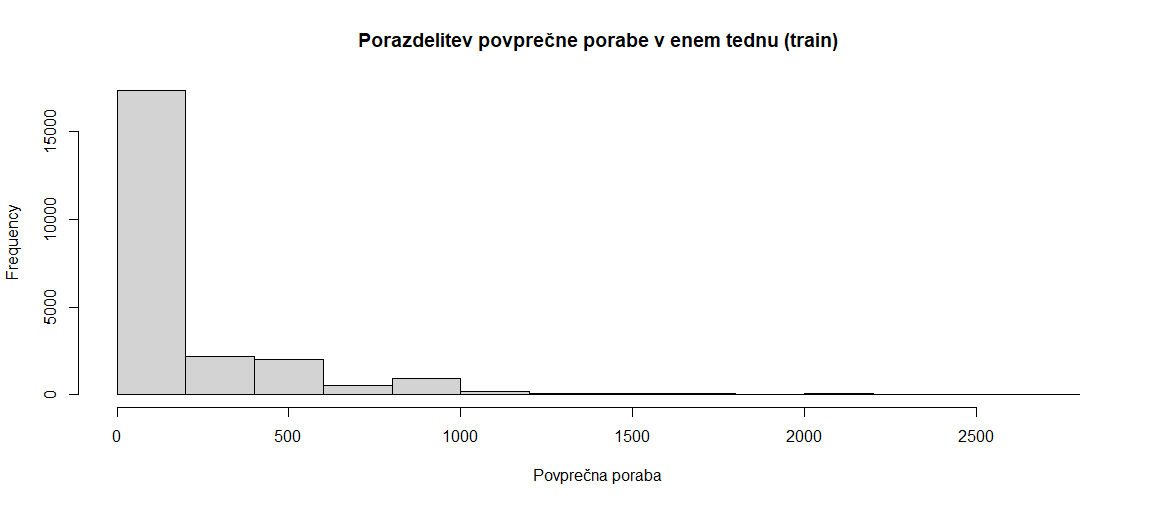


Slika 7: Povprečna poraba glede na namembnost stavbe (train)



Slika 8: Povprečna poraba glede na namembnost stavbe (test)

Naslednja dva grafa prikazujeta porazdelitev povprečne porabe v enem tednu. Opaziva, da poraba ni normalno porazdeljena.

 Slika 9: Porazdelitev povprečne porabe v enem tednu (test)

Slika 10: Porazdelitev povprečne porabe v enem tednu (train)

# **Ocenjevanje atributov**

Atribute sva pri klasifikaciji ocenila z klasifikacijskimi kratkovidnimi metodami (informacijski prispevek, gini, princip najkrajšega opisa), klasifikacijskimi nekratkovidnimi metodami (relief, relifF), pri regresiji pa z metodami MSE, ReliefFwithMSE in reliefF.

Najboljši atributi klasifikacijskih kratkovidnih metod so bili:

* **Informacijski prispevek**: povrsina, regija, leto\_izgradnje, temp\_zraka, poraba\_prteden
* **Gini**: povrsina, leto\_izgradnje, regija, poraba\_prteden, poraba
* **MD**L: povrsina, regija, leto\_izgradnje, temp\_zraka, poraba\_prteden

Najboljši atributi klasifikacijskih nekratkovidnih metod so bili:

* **Relief**: poraba\_prteden, poraba, povrsina, leto\_izgradnje
* **ReliefF**: leto\_izgradnje, povrsina, poraba\_prteden, regija, poraba

Najboljši atributi regresijskih metod so bili:

* **MSE**: poraba\_prteden, povrsina
* **ReliefF**: poraba\_prteden, povrsina, leto\_izgradnje
* **ReliefFwithMSE**: poraba\_prteden, povrsina, leto\_izgradnje

# **Modeliranje in evalvacija modelov**

## **Klasifikacijsko modeliranje**

Preizkusila sva naslednje modele:

* **Naivni bayesov model**
* **Model k-najbližjih sosedov (5, 10, 20)**
* **Odločitveno drevo**

Pri vsakem modelu sva uporabila različne podmnožice atributov:

* **Vsi atributi**
* **Gini**
* **Informacijski prispevek**
* **MDL**
* **Relief**
* **ReliefF**

Najboljši trije modeli so bili:

- **Odločitveno drevo z podmnožico ReliefF:** 0,5587303

- **Naivni bayesov model z podmnožico Relief**: 0,5385538

- **Model k-najbližjih sosedov(20) z podmnožico MDL**: 0,5285727

## **Regresijsko modeliranje**

Preizkusila sva naslednje modele:

* **Linearni**
* **Model k-najbližjih sosedov (5, 10, 20)**
* **Regresijsko drevo**

Pri vsakem modelu sva uporabila različne podmnožice atributov:

* **Vsi atributi**
* **MSE**
* **ReliefFwithMSE**
* **ReliefF**

Najboljši trije modeli po Rmae so bili:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mae | Rmae | Mse | Rmse |
| **Model k-najbližjih sosedov(20) z podmnožico vseh atributov** | 20.00806583 | 0.1222320 | 2520.26675275 | 0.05441129 |
| **Model k-najbližjih sosedov(20) z podmnožico**  **MSE** | 20.00806583 | 0.1222320 | 2520.26675275 | 0.05441129 |
| **Model k-najbližjih sosedov(20) z podmnožico reliefF** | 20.00870636 | 0.1222359 | 2520.11978825 | 0.05440811 |

## **Primerjava uspešnosti modelov na podatkih ene regije**

### **Odločitveno drevo**

Model odločitvenega drevesa je bil veliko bolj natančen pri napovedovanju podatkov iz vzhodne regije, pri zahodni pa opazimo poslabšanje (v primerjavi z obemi regijami).

Dobljeni rezultati:

* **Zahodna regija:** 0,3831382
* **Vzhodna regija:** 0,7036904
* **Obe regije:** 0,4966555

### **Linearna regresija**

Model linearne regresije je bolj natančen, če uporabimo učne podatke ene regije (po Rmae).

Dobljeni rezultati:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mae | Rmae | Mse | Rmse |
| Linearna regresija zahodna regija | 13.24988190 | 0.09557767 | 604.20811081 | 0.02208052 |
| Linearna regresija vzhodna regija | 31.86758990 | 0.12195228 | 4495.65785619 | 0.06647863 |
| Linearna regresija obe regije | 21.45662657 | 0.13439253 | 2348.56817708 | 0.05210445 |

## **Kombiniranje modelov strojnega učenja**

Uporabila sva metode glasovanja, uteženega glasovanja in bagginga. Kombinirala sva modele odločitevenega drevesa, naivnega bayesa in Model k-najbližjih sosedov(5) nad vsemi atributi. Primerjala sva klasifikacijo natančnost in pridobila naslednje rezultate:

* **Glasovanje**: 0,5200908
* **Uteženo glasovanje:** 0,5336703
* **Bagging**: 0,4582762
* **Odločitveno drevo:** 0,4282042
* **Naivni Bayes**: 0,4205363
* **Model k-najbljižjih sosedov(5):** 0,4210504

Ugotovila sva, da so bili vsi kombinirani modeli bolj natančni, najbolj natančen je bil model uteženega glasovanja.

## **Kronologija podatkov**

Združila sva podatkovna okvirja učne in testne množice, nato sva pridobljeni podatkovni okvir razdelila na 12 podmnožic (meseci).

Pri klasifikacijskih modelih sva prišla do naslednjih rezultatov:

|  |  |  |  |
| --- | --- | --- | --- |
| Mesec | Odločitveno drevo | Naivni Bayes | KNN (5) |
| 1 | 0.8965650 | 0.5417386 | 0.8401458 |
| 2 | 0.8770858 | 0.5064630 | 0.8470035 |
| 3 | 0.8910066 | 0.4407385 | 0.8228112 |
| 4 | 0.8768015 | 0.3491399 | 0.8335658 |
| 5 | 0.8601089 | 0.4404357 | 0.8631722 |
| 6 | 0.8554386 | 0.4750877 | 0.8449123 |
| 7 | 0.8744681 | 0.5332447 | 0.8268617 |
| 8 | 0.9168901 | 0.5031278 | 0.8745904 |
| 9 | 0.8986702 | 0.5345745 | 0.8505319 |
| 10 | 0.8946228 | 0.5647036 | 0.8652748 |
| 11 | 0.8835320 | 0.5763814 | 0.8618635 |

Kot najbolj natančen model se je izkazalo odločitveno drevo.

Pri regresijskem modeliranju sva prišla do naslednjih rezultatov:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mesec | Model | Mae | Rmae | Mse | Rmse |
| 1 | Linearna regresija | 31.27241597 | 0.17649933 | 3161.47496448 | 0.03699561 |
| 1 | Regresijsko drevo | 21.29896204 | 0.12020985 | 2642.71444635 | 0.03092507 |
| 2 | Linearna regresija | 22.30152945 | 0.13540926 | 2244.49353435 | 0.03089751 |
| 2 | Regresijsko drevo | 22.07171184 | 0.13401387 | 5855.07402861 | 0.08060047 |
| 3 | Linearna regresija | 20.82230011 | 0.14059943 | 1844.33650421 | 0.03425722 |
| 3 | Regresijsko drevo | 18.60185300 | 0.12560620 | 1857.17517905 | 0.03449569 |
| 4 | Linearna regresija | 20.16057971 | 0.14178468 | 1606.88158965 | 0.03343478 |
| 4 | Regresijsko drevo | 18.38597445 | 0.12930429 | 1838.66831745 | 0.03825763 |
| 5 | Linearna regresija | 21.49794477 | 0.12921831 | 1997.73285102 | 0.02701112 |
| 5 | Regresijsko drevo | 22.42745736 | 0.13480536 | 2400.40715597 | 0.03245564 |
| 6 | Linearna regresija | 21.75087419 | 0.12881310 | 3748.35290738 | 0.04878034 |
| 6 | Regresijsko drevo | 24.78260996 | 0.14676765 | 4890.43302802 | 0.06364316 |
| 7 | Linearna regresija | 25.26823454 | 0.13535351 | 3481.01668470 | 0.03689755 |
| 7 | Regresijsko drevo | 24.59801535 | 0.13176337 | 3202.56020519 | 0.03394601 |
| 8 | Linearna regresija | 22.66511510 | 0.12975091 | 2619.09031982 | 0.03319773 |
| 8 | Regresijsko drevo | 21.62831732 | 0.12381556 | 2623.85378862 | 0.03325811 |
| 9 | Linearna regresija | 28.53753488 | 0.17770012 | 4642.64374935 | 0.06714919 |
| 9 | Regresijsko drevo | 27.43312936 | 0.17082310 | 5052.80887028 | 0.07308164 |
| 10 | Linearna regresija | 27.81818899 | 0.19584288 | 2414.14002847 | 0.04809237 |
| 10 | Regresijsko drevo | 25.44448427 | 0.17913176 | 3519.30357900 | 0.07010847 |
| 11 | Linearna regresija | 23.0853486 | 0.1579900 | 2995.4860729 | 0.0491382 |
| 11 | Regresijsko drevo | 23.82578789 | 0.16305734 | 4211.79760124 | 0.06909067 |

Kot najbolj natančen model se je v večini primerov izkazalo Regresijsko drevo. V zadnjem primeru pa je bila najboljša linearna regresija.

# **Zaključek**

Ugotovila sva, da ko sva iz podatkovnih množic odstranila nekaj atributov, sva dobila boljše rezultate pri klasifikacijskih in regresijskih modelih. Prav tako sva ugotovila, da je modeliranje dobrih modelov izredno zahtevno in časovno potratno.