Seminarska naloga 1

“UNIVERZA V LJUBLJANI”
Fakulteta za računalništvo in informatiko

Ardit Nela
Bisera Nikoloska

Predmet: Umetna inteligenca

November, 2020

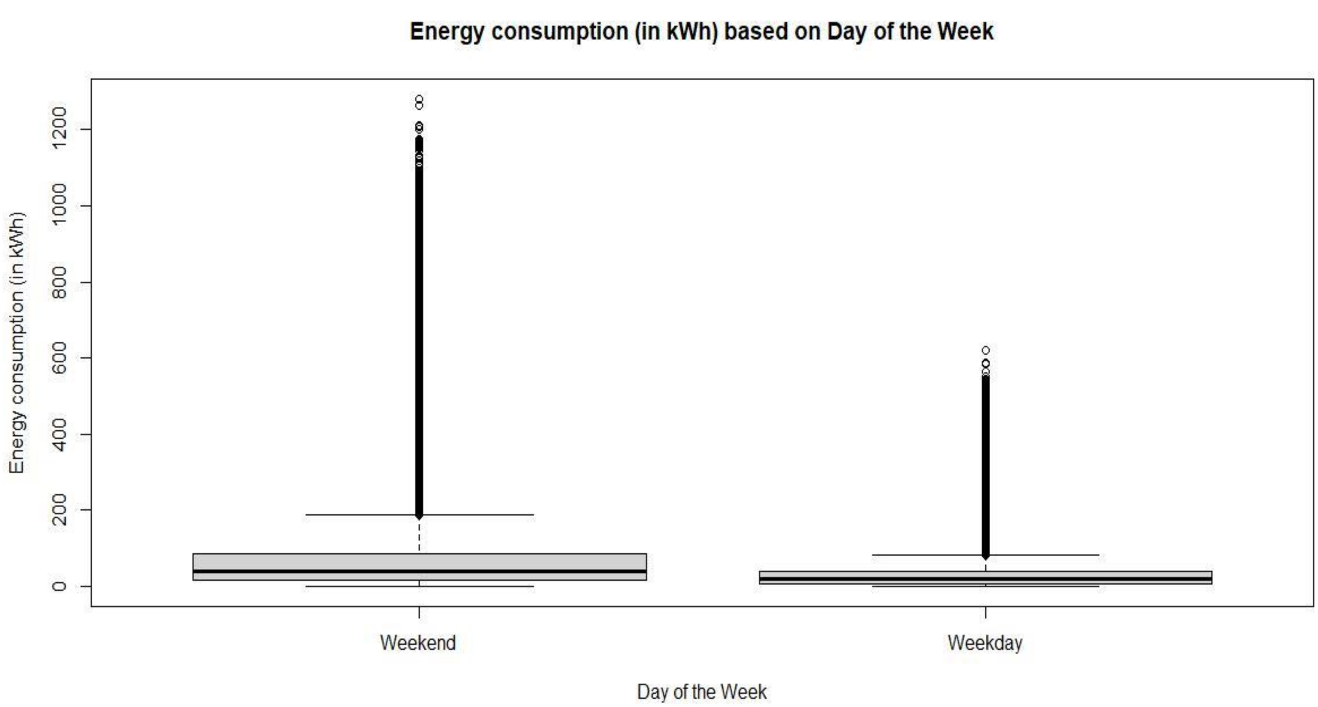
Contents

1. Attribute selection and visualization 3
2. Classification models 11
3. Regression models 13
4. Combining the models 16
5. Comparing the performance of models 19
6. Used literature: 21

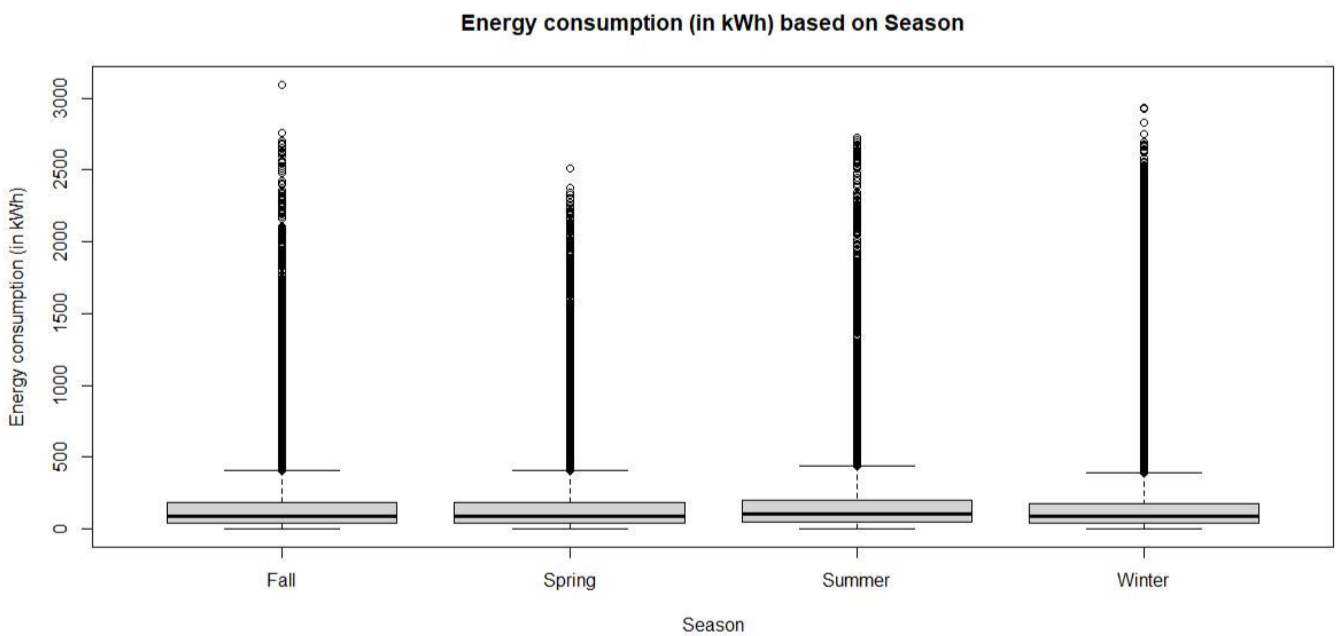
1. Attribute selection and visualization

Before constructing the models, we decided to create new attributes that we thought could prove useful in improving their overall quality. We experimented with a number of attributes, which can be seen in the following images.

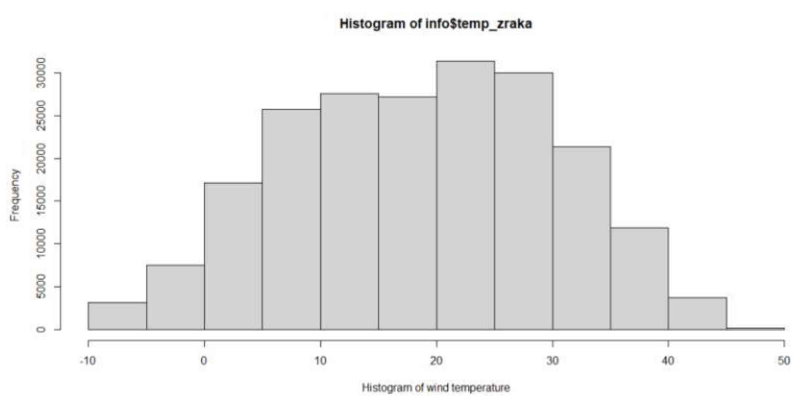
First off, we noticed a correlation between energy usage in sampled buildings and whether or not the day of the sample was a weekend or weekday. After normalizing the data (as there are more weekdays compared to weekends), we decided that the difference was big enough, so we made it into an attribute.



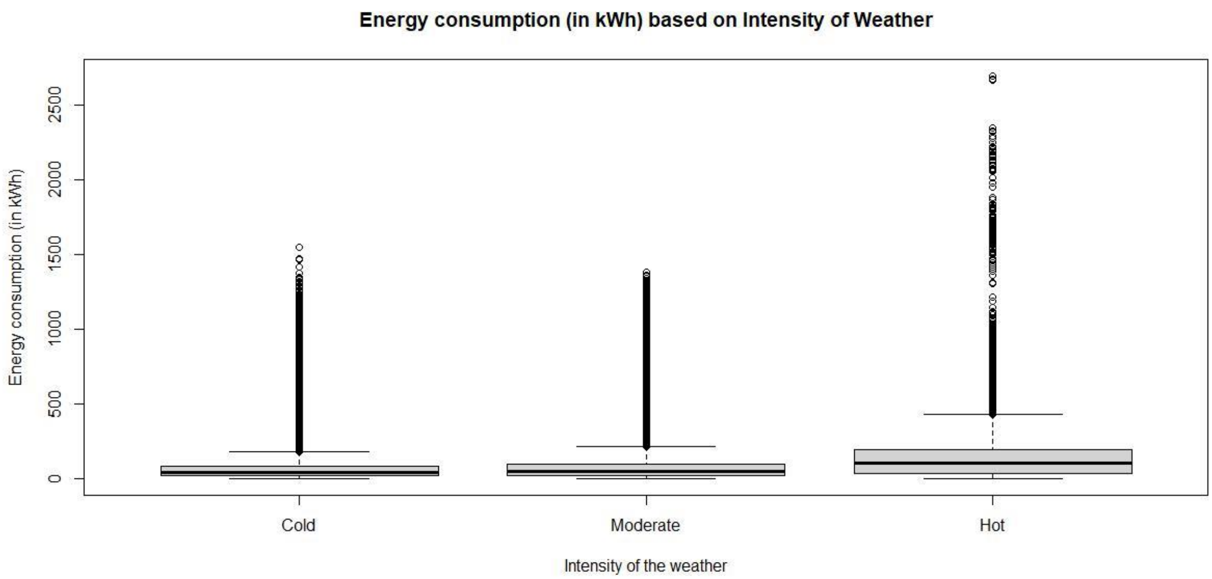
We also explored the idea of different seasons showing different amounts of energy usage (as during certain seasons there are different weather conditions and people tend to spend more time at home during the colder winter months). After grouping the data by seasons, this is what we got:



We can't really see too big of a difference but decided to keep the attribute for now (as we would be able to remove it later on, when building the models if necessary).

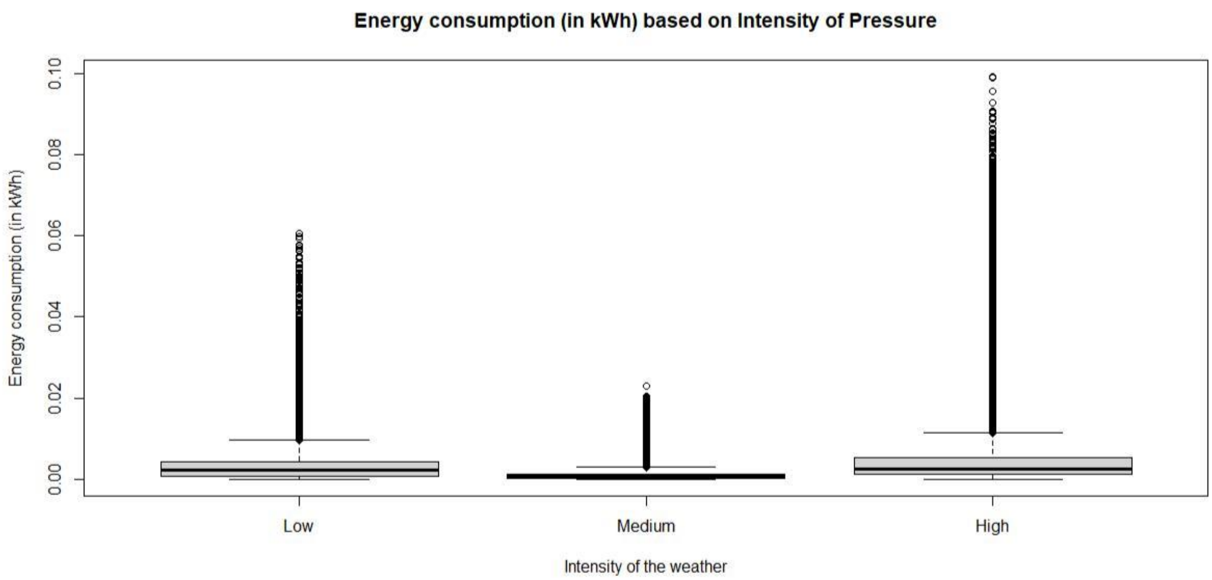


Since the data with seasons doesn't look to promising, we decided to group the data by the intensity of the weather (temp_zraka <= 15 being "Cold", 30 <= temp_zraka being "Hot" and everything in between being "Moderate").



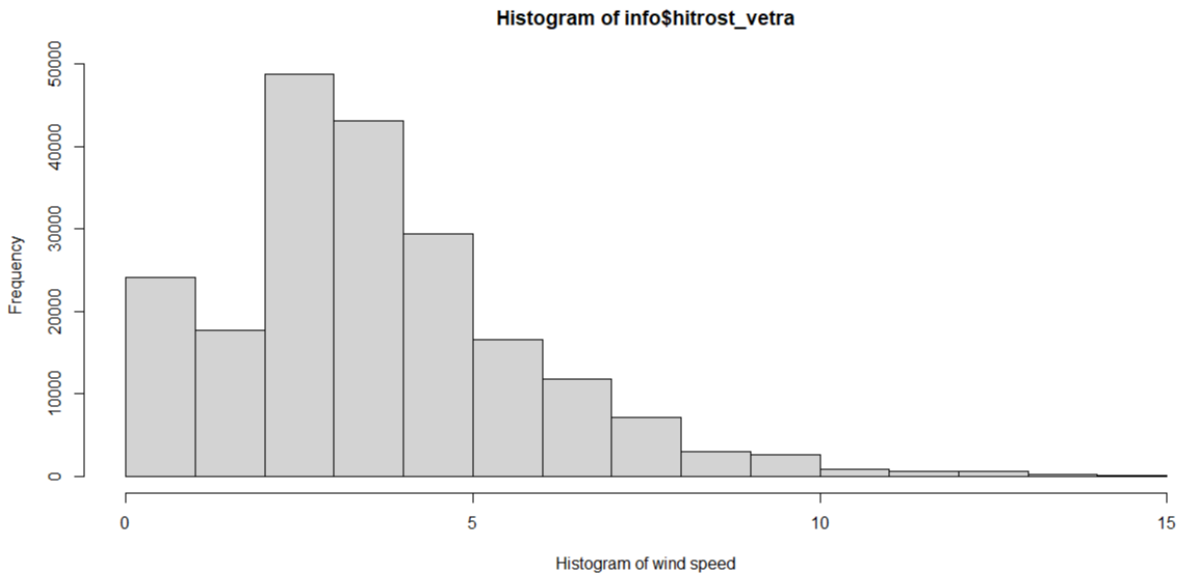
It becomes clear from the graph that people tend to use more energy when the weather is hotter, in comparison to more moderate or colder weather.

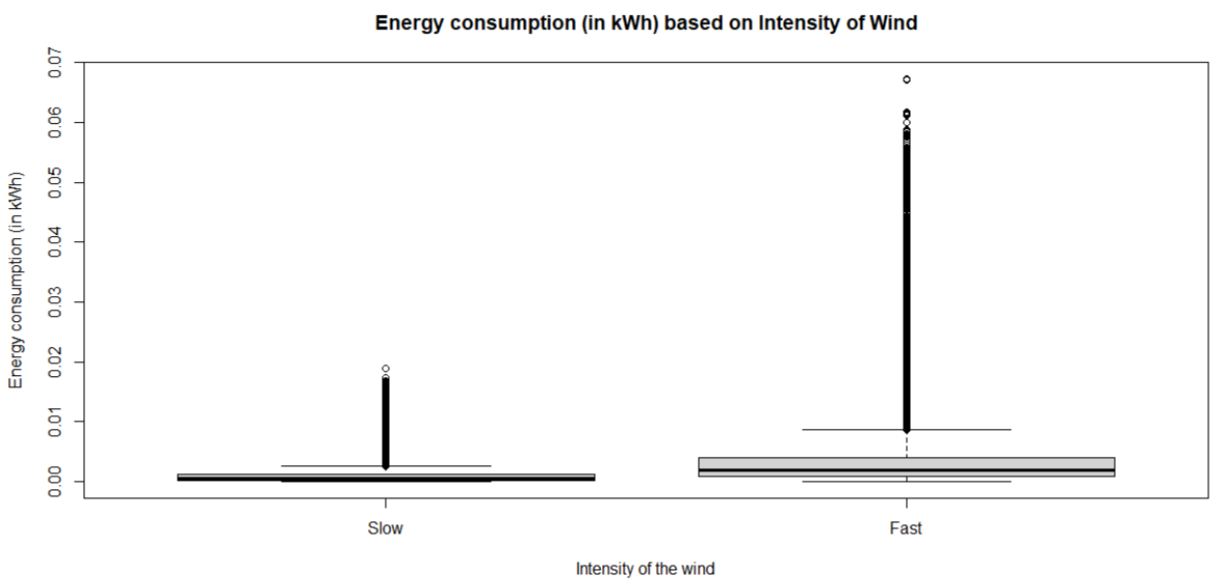
Since the intensity of pressure is directly linked to weather (Low means rainier, High means warmer/clearer, Medium means present conditions persist), we grouped the data using that criterium and got the following (note the data is scaled to make up for any inequalities in the number of items in a given factor):



This makes sense because, as previously mentioned, people tend to waste more when the weather is hotter (and higher wind pressure is linked to warmer weather conditions).

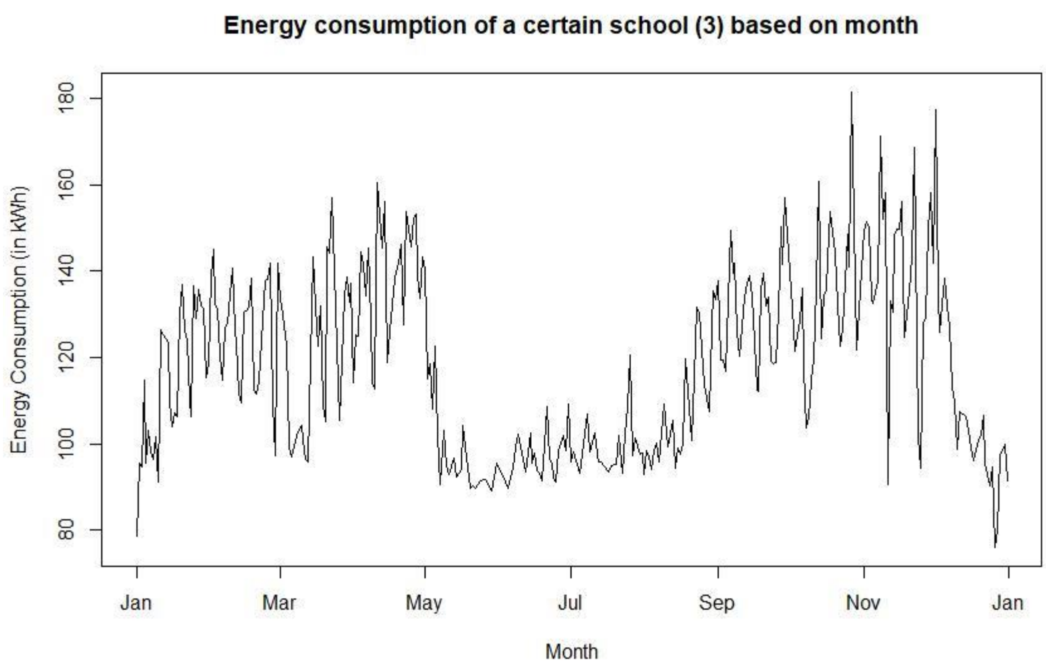
Afterwards, we decided to see if there's any correlation between usage of energy and wind speed.

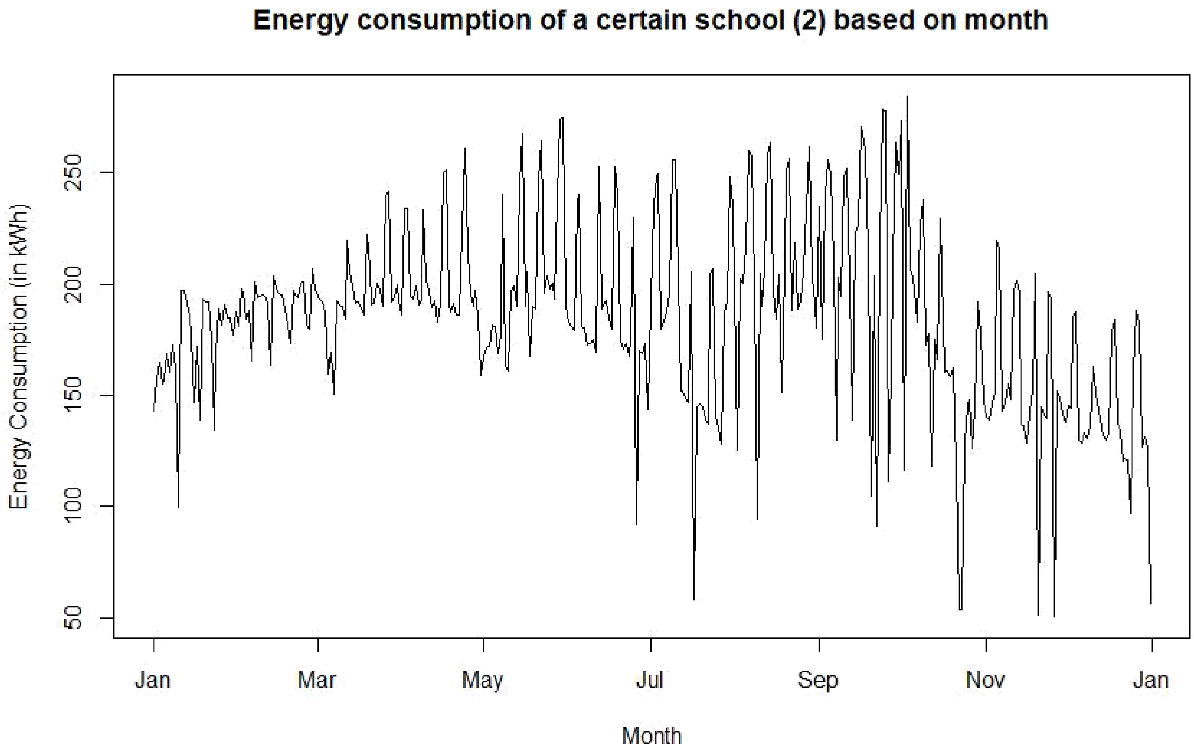




We decided to keep this attribute for the time being.

We tried to see if the month had anything to do with energy consumption but after drawing two different graphs, we decided to discard the attribute.





Lastly, we decided to write a function that would take information for energy consumption based on the previous day (sum, minimum, maximum and average/mean). After running the function, we noticed that certain rows did not contain all the required data and had NaN values instead. For this reason, we decided to omit these rows from our main data frame (lowering the number of rows to 198976). (NOTE: we removed certain attributes at this point).

At this point, we believe we've created enough attributes and it's time to score them. We scored our regression and classification attributes separately.

Classification:

```
> sort(attrEval(norm_poraba ~ ., infoClass, "InfGain"), decreasing = TRUE)
leto_izgradnje    namembnost      povrsina           ura        stavba   temp_rosisca     dayOfWeek        regija        season
  0.0586363693  0.0544900570  0.0253157846  0.0249516916  0.0161259701  0.0091863170  0.0062827951  0.0057038659  0.0042187559
         datum       weather    temp_zraka      pressure       pritisk    smer_vetra     oblacnost hitrost_vetra     windSpeed
  0.0032187835  0.0030693445  0.0024243737  0.0020500712  0.0017623428  0.0016059070  0.0004688430  0.0002005743  0.0002005743
      padavine
  0.0001793959
> sort(attrEval(norm_poraba ~ ., infoClass, "GainRatio"), decreasing = TRUE)
      povrsina        stavba leto_izgradnje  temp_rosisca    namembnost      padavine       pritisk           ura    temp_zraka
  0.2653805425  0.2648819839  0.1860525041  0.0396611009  0.0275568000  0.0140114668  0.0125152493  0.0124999122  0.0123437467
 hitrost_vetra     dayOfWeek        regija         datum        season       weather    smer_vetra      pressure     oblacnost
  0.0079773992  0.0072313174  0.0057759827  0.0056569921  0.0021458721  0.0020204870  0.0016560587  0.0015928185  0.0008543212
     windSpeed
  0.0002696576
> sort(attrEval(norm_poraba ~ ., infoClass, "MDL"), decreasing = TRUE)
leto_izgradnje    namembnost      povrsina           ura        stavba   temp_rosisca     dayOfWeek        regija        season
  5.849650e-02  5.397727e-02  2.518432e-02  2.455426e-02  1.599209e-02  9.058594e-03  6.146790e-03  5.565060e-03  3.823975e-03
         datum       weather    temp_zraka      pressure       pritisk    smer_vetra     oblacnost      padavine hitrost_vetra
  3.082971e-03  2.801990e-03  2.301628e-03  1.789195e-03  1.631862e-03  1.467808e-03  3.414033e-04  7.267827e-05  6.766415e-05
     windSpeed
  6.766415e-05
> sort(attrEval(norm_poraba ~ ., infoClass, "ReliefFequalK"), decreasing = TRUE)
leto_izgradnje    namembnost      povrsina        stavba           ura        regija      padavine     windSpeed       weather
  1.607764e-01  1.603881e-01  1.222748e-01  1.090811e-01  8.167942e-04  7.415437e-04 -3.113834e-05 -2.186895e-03 -2.921708e-03
        season      pressure     dayOfWeek    temp_zraka         datum       pritisk hitrost_vetra  temp_rosisca     oblacnost
 -3.265452e-03 -4.435441e-03 -7.356714e-03 -9.995945e-03 -1.010176e-02 -1.295746e-02 -2.177554e-02 -2.691052e-02 -3.355161e-02
    smer_vetra
 -3.426099e-02
```

Regression:

```
> sort(attrEval(poraba ~ ., infoReg, "MSEofMean"), decreasing = TRUE)
    avgPoraba     maxPoraba     minPoraba     sumPoraba      povrsina leto_izgradnje        stavba    namembnost           ura
    -25015.28     -25468.73     -26038.64     -27463.56     -37385.79      -56338.99     -60303.48     -60913.62     -61652.84
 temp_rosisca     dayOfWeek         datum    temp_zraka        regija         season       pritisk    smer_vetra      pressure
    -61903.39     -61926.24     -61948.11     -61951.84     -61969.07      -61995.15     -62001.45     -62005.34     -62005.49
      weather     oblacnost      padavine hitrost_vetra     windSpeed
    -62010.08     -62018.30     -62018.34     -62018.81     -62018.89
> sort(attrEval(poraba ~ ., infoReg, "RReliefFexpRank"), decreasing = TRUE)
    maxPoraba     avgPoraba     minPoraba     sumPoraba      povrsina            ura    smer_vetra     oblacnost hitrost_vetra
 0.3077366287  0.2624366327  0.2568964689  0.2381957749  0.2260162557   0.1626677847  0.0917653036  0.0698786453  0.0695170011
      pritisk  temp_rosisca    temp_zraka         datum     dayOfWeek       pressure        season     windSpeed      padavine
 0.0677060911  0.0501097091  0.0479550619  0.0452308484  0.0361791795   0.0227188360  0.0204538233  0.0189486752  0.0069261632
      weather        regija    namembnost        stavba leto_izgradnje
 0.0063507786  0.0001846259 -0.0243624685 -0.0674707697 -0.1200888158
```

We ran multiple tests on the attributes but decided we'd use ReliefF as our main evaluation function.

2. Classification models

We made three different classification models:

- Decision Tree
- Random Forest
- Naïve Bayes

Each model was built twice (once using all our required attributes, and then using the three strongest attributes according to reliefF). We also created a custom evaluation function called "modelEval". These are the scores we got for our models:

Model	Brier Score	Classification Accuracy	"modelEval" score (Brier)
Decision Tree (all attributes)	≈ 0.28	≈ 0.82	≈ 0.30, 0.32, 0.32, 0.42, 0.47, 0.37, 0.40, 0.41 0.43, 0.39, 0.48
Decision Tree (top attributes)	≈ 0.48	≈ 0.64	≈ 0.41, 0.45, 0.46, 0.50, 0.60, 0.59, 0.59, 0.54, 0.51, 0.50, 0.48
Random Forest (all attributes)	≈ 0.25	≈ 0.83	≈ 0.28, 0.28, 0.28, 0.37, 0.41, 0.30, 0.35, 0.35, 0.34, 0.32, 0.43
Random Forest (top attributes)	≈ 0.46	≈ 0.64	≈ 0.40, 0.43, 0.44, 0.48, 0.57, 0.57, 0.57, 0.52, 0.50, 0.48, 0.46

Model	Brier Score	Classification Accuracy	“modelEval” score (Brier)
Naïve Bayes (all attributes)	≈ 0.74	≈ 0.38	≈ 0.79, 0.78, 0.80, 0.79, 0.84, 0.82, 0.78, 0.76, 0.75, 0.74, 0.75
Naïve Bayes (top attributes)	≈ 0.74	≈ 0.37	≈ 0.72, 0.72, 0.72, 0.74, 0.76, 0.78, 0.77, 0.76, 0.74, 0.73, 0.74

3. Regression models

We made four different regression models:

- Linear model
- Regression Tree
- Neural Network
- K-nearest-neighbor

Each model was built twice (once using all our required attributes, and then using the four strongest attributes according to reliefF). Aside from "modelEval", we also wrote a function that's specific to knn, called "modelEvalKNN". We wrote functions for mae, mse, rmae and rmse, but we chose to use rmse and rmae.

Model	rmse	rmae
Neural Network (all)	0.06995541	0.1985665
Neural Network (top)	0.07099577	0.195312

Model	rmse	rmae	modelEval (rmse)
Linear model (all)	≈ 0.064	≈ 0.200	≈ 0.054, 0.055, 0.068, 0.043, 0.061, 0.074, 0.064, 0.070, 0.079, 0.072, 0.065
Linear model (top)	≈ 0.071	≈ 0.196	≈ 0.060, 0.063, 0.078, 0.048, 0.069, 0.080, 0.074, 0.082, 0.090, 0.076 0.068
Regression tree (all)	≈ 0.095	≈ 0.279	≈ 0.089, 0.097, 0.099, 0.073, 0.088, 0.105, 0.100, 0.105, 0.116, 0.112, 0.095
Regression tree (top)	≈ 0.095	≈ 0.279	≈ 0.088, 0.097, 0.099, 0.073, 0.088, 0.105, 0.100, 0.107, 0.114, 0.105, 0.095

Model	rmse	rmae	modelEvalKNN (rmse)
K-Nearest-Neighbor (all)	≈ 0.055	≈ 0.197	≈ 0.057, 0.047, 0.059, 0.041, 0.055, 0.074, 0.046, 0.049, 0.058, 0.089 0.064
K-Nearest-Neighbor (top)	≈ 0.082	≈ 0.211	≈ 0.090, 0.094, 0.094, 0.070, 0.092, 0.106, 0.089, 0.107, 0.108, 0.103, 0.087

4. Combining the models

We used 3 combining models:

- Glasovanje
- Utežno glasovanje
- Bagging

Glasovanje :

We were researching the combination models with Naïve Bayes, Decision tree and K- nearest neighbor. First of all we predicted the test data:

Decision Tree	Naïve Bayes	KNN
SREDNJA	SREDNJA	SREDNJA
SREDNJA	NIZKA	SREDNJA
VISOKA	ZELOVISOKA	VISOKA
ZELONIZKA	NIZKA	ZELONIZKA
SREDNJA	SREDNJA	SREDNJA
ZELONIZKA	NIZKA	ZELONIZKA

In the next step we used the function voting to choose the data that appears the most in one row:

VOTING
SREDNJA
SREDNJA
VISOKA
ZELONIZKA
SREDNJA
ZELONIZKA

The classification accuracy for this model is 0.9939874.

Utežno glasovanje:

Firstly we predict probability for the three models. Then we sum them into one variable. We create a matrix from all the data levels and their probabilities:

NIZKA	SREDNJA	VISOKA	ZELONIZKA	ZELOVISOKA
0.179227180	2.79836000	0.0161424849	2.419467e-03	3.850866e-03
0.567358323	2.36187278	0.0055880911	5.437263e-02	1.080818e-02
0.005069453	0.03720165	2.4698486408	1.351206e-05	4.878667e-01
0.640755678	0.08800530	0.0007887136	2.270419e+00	3.109648e-05
0.139624740	2.82640268	0.0255092717	2.594150e-03	5.869158e-03
0.554812401	0.05792665	0.0006951937	2.386480e+00	8.621839e-05

Furthermore we factorize the index of the maximum value in the column and the levels of our goal variable, in this case "norm_poraba" and we test the accuracy of our model and we get: 0.9940002

The accuracy of the model is assessed e.g. by cross-checking on the train data. Firstly we use 2 functions: mymodel to build our models and mypredict to predict the values of our model. In order to succeed cross checking we used function "errorrest" and we get the values for all three models. Then we sum all the accuracies multiplied with their predicted probability. Again we use the max values of every column and factorize them and we test the accuracy: 0.994068

Bagging:

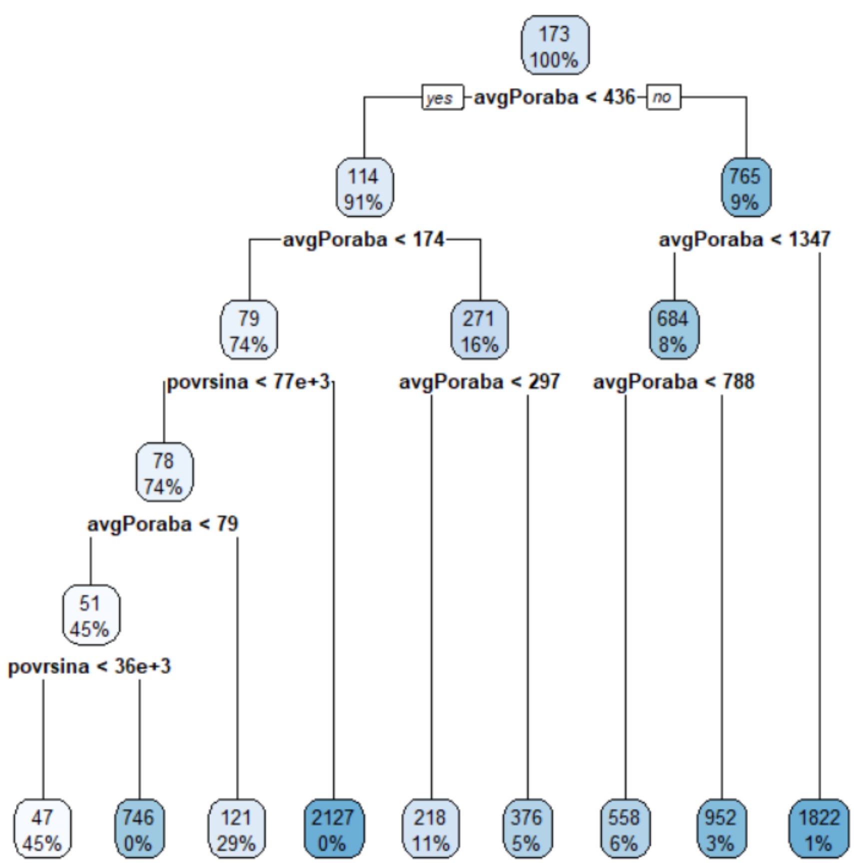
For this model we use decision tree. We create 30 different trees and choose random examples with repeating. Moreover we use another for loop where we build the columns and every column votes for its own class. We factorize the voting values and the levels we have in our data and test the accuracy of our model and we get: 0.9951157

5. Comparing the performance of models

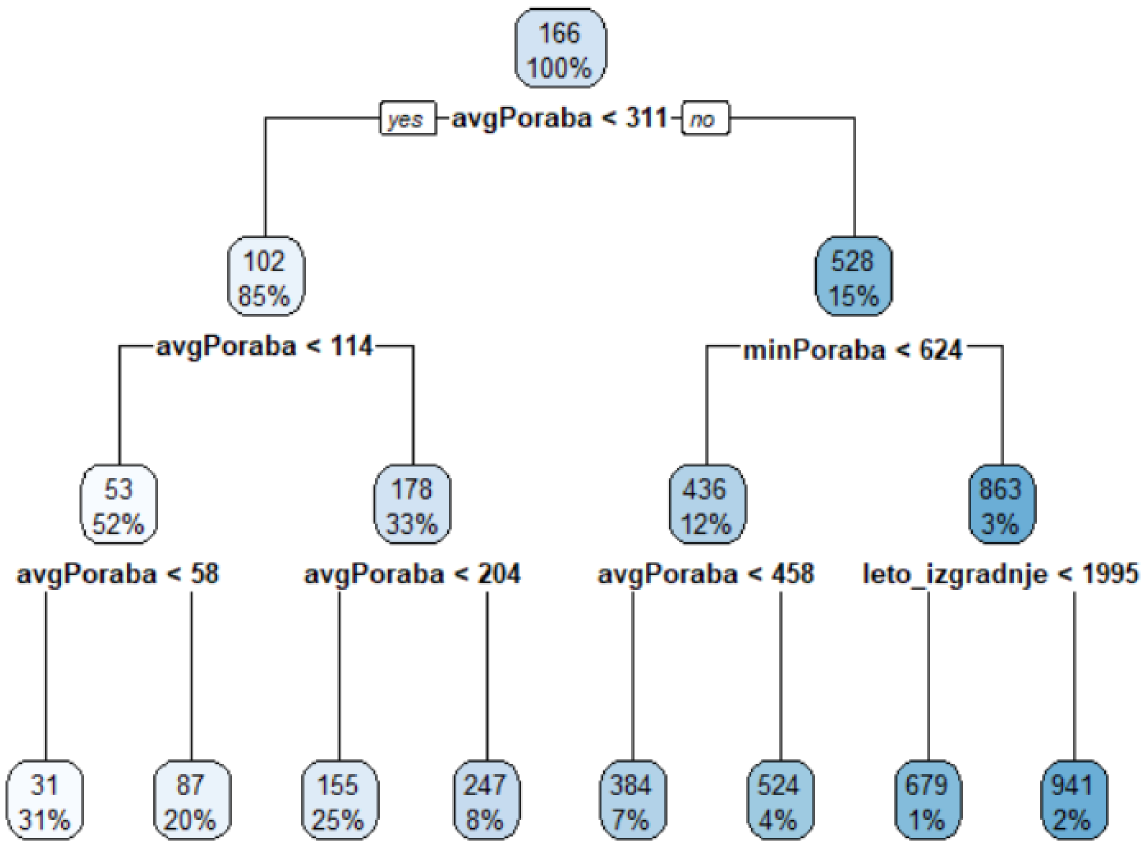
Firstly, we separated the data by regions, we used the data of "zahodna" region and we compared to the whole data. For that purpose we used "random forest" classification model and "rpart" regression model. We used three attributes for the classification model "povrsina", "namembnost" and "temp_zraka".

For the classification model with random forest for whole we get the classification accuracy 0.6841511 and the brier score 0.4269679 and for the data of "zahodna" region we get 0.7055845 and brier score 0.3900271.

The mean absolute error of the linear regression of the whole data is : 49.60903 and the relative average absolute error is 0.3270741. We plot the linear regression as shown below:



The mean absolute error of the linear regression of the “zahodna” region is : 32.22558and the relative average absolute error is 0.2446756. We plot the linear regression as shown below:



6. Used literature:

https://ucilnica.fri.uni-lj.si/course/view.php?id=21

https://www.datasciencecentral.com/profiles/blogs/implemetation-of-17-classification-algorithms-in-r

https://www.r-project.org/