Umetna Inteligenca

Seminarska naloga 1

December, 2020

Authors: Kiril Tofiloski, 63190385 and Ivan Nikolov, 63190378

1. Visualization of data

In order to add new attributes to the data, we need to find some interesting correlations, distributions, recurring samples etc. This is done by plotting and analyzing the graphs that might give us useful information.

The first graph(*image 1*) shows the changes of electricity usage(poraba) in kWh of the building(stavba) with id 1 measured at 11am throughout the whole year. By looking at the graph it is visible that the energy usage is higher at the summer months and lower in spring, autumn and winter.

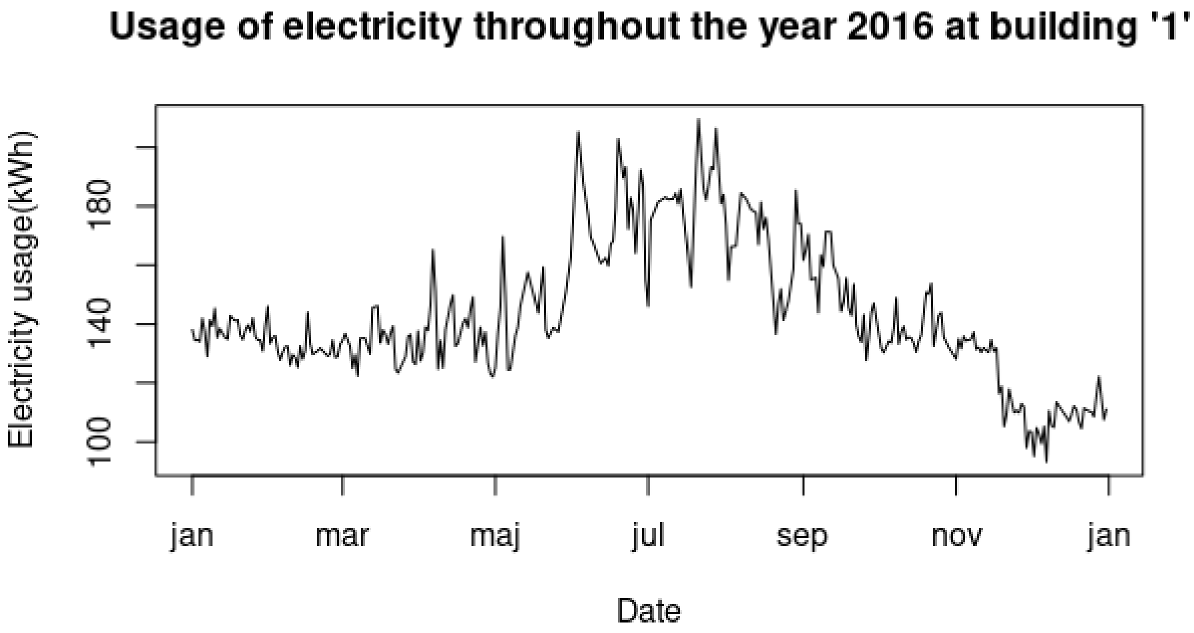


Image 1

This might be caused by the use of air conditioning if the temperatures are high, so we plot the correlation between the air temperature(temp_zraka) and energy use(poraba)*(Image 2)*.

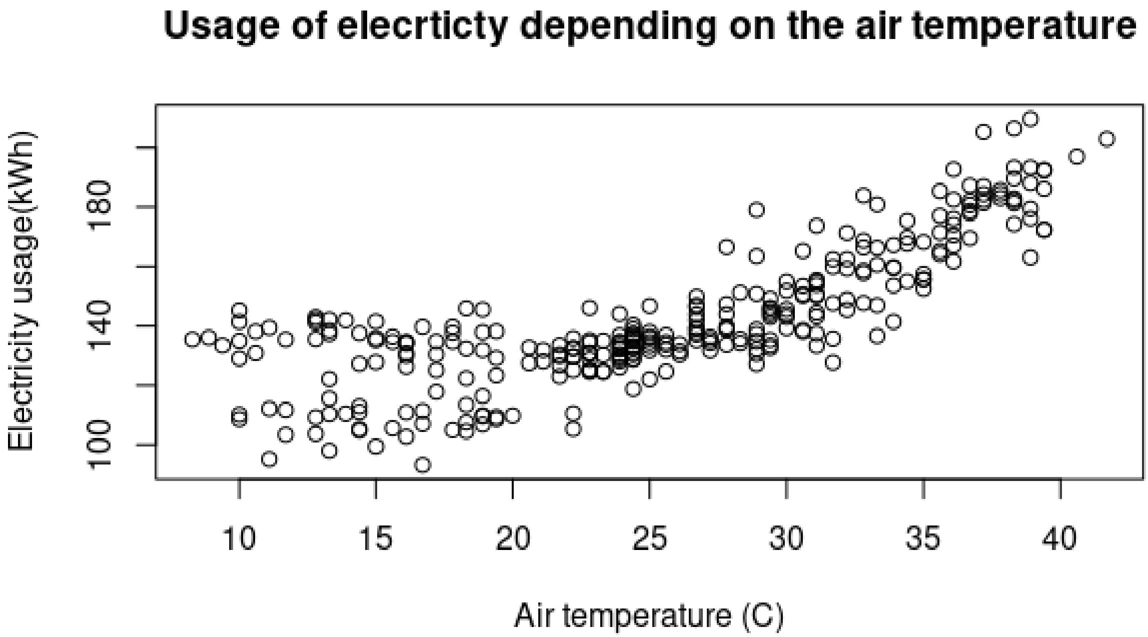


Image 2

From the plot(*Image 2*) it is visible that there is a positive linear correlation between these two variables, the Pearson coefficient of correlation is *0.7956* and confirms that. We continue looking for more patterns based on the season.
Next, we check the ratio of normalized electricity usage(norm_poraba) based on the time of the year.
We examine how the ratio of *low* electricity usage(nizka, zelo_nizka)(*Image 3) and high* el.usage*(*zelo_visoka, visoka)(*Image 4)* changes.

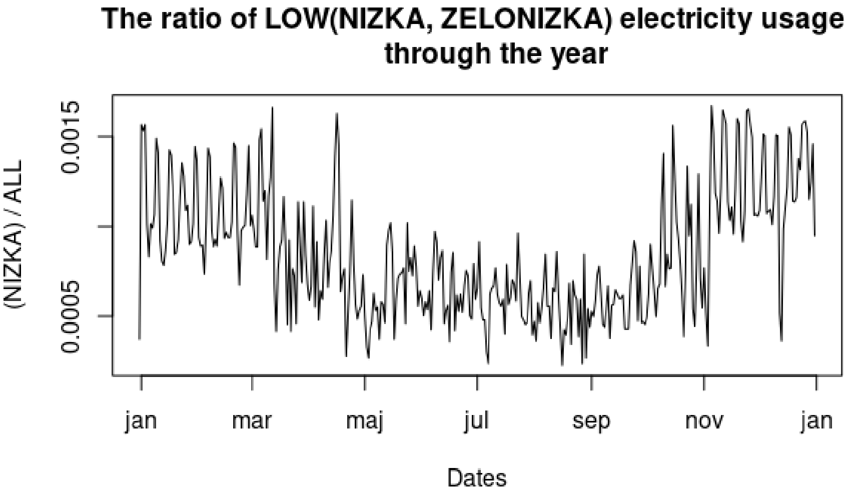


Image 3

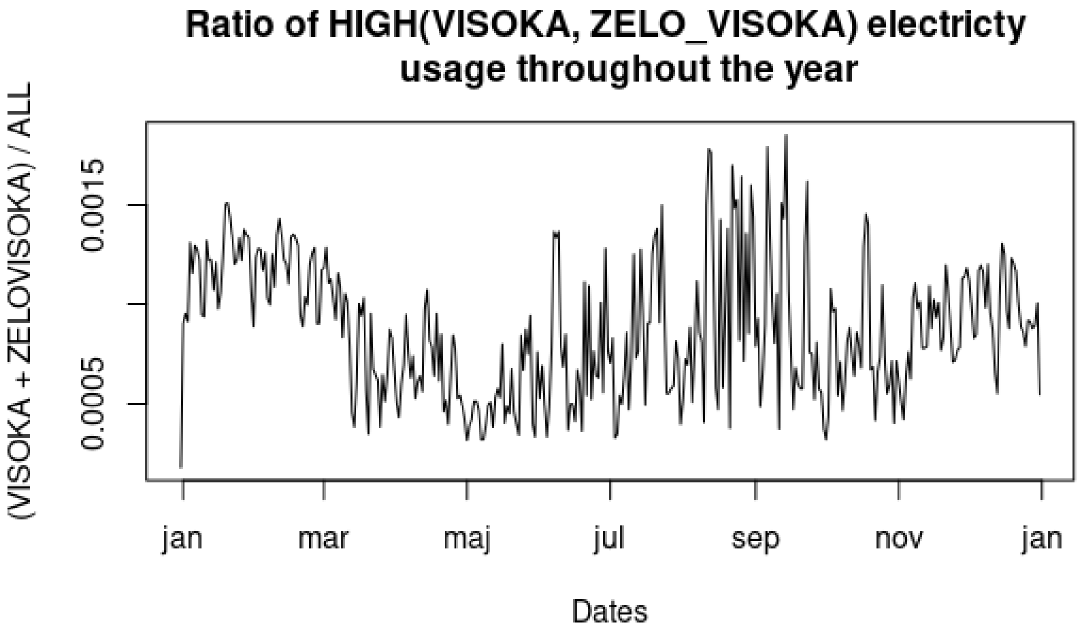
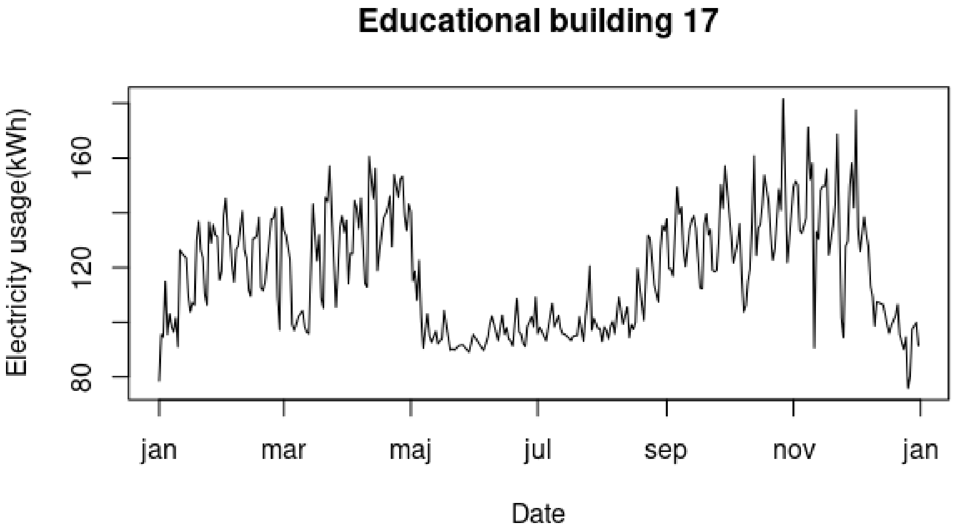
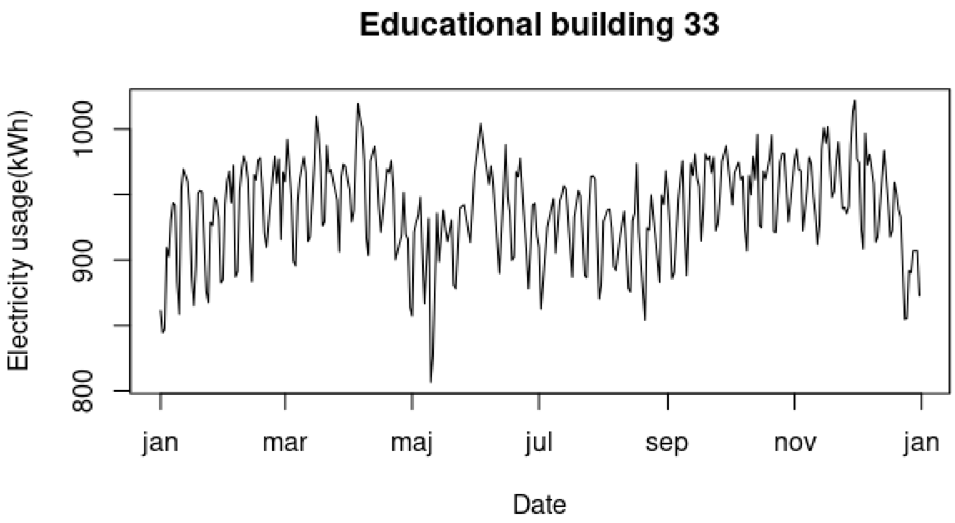


Image 3 shows that the ratio of low energy use falls in the summer mouths and is high in winter. By contrast the ratio of high energy use has peaks in the summer and autumn (*Image 4*).

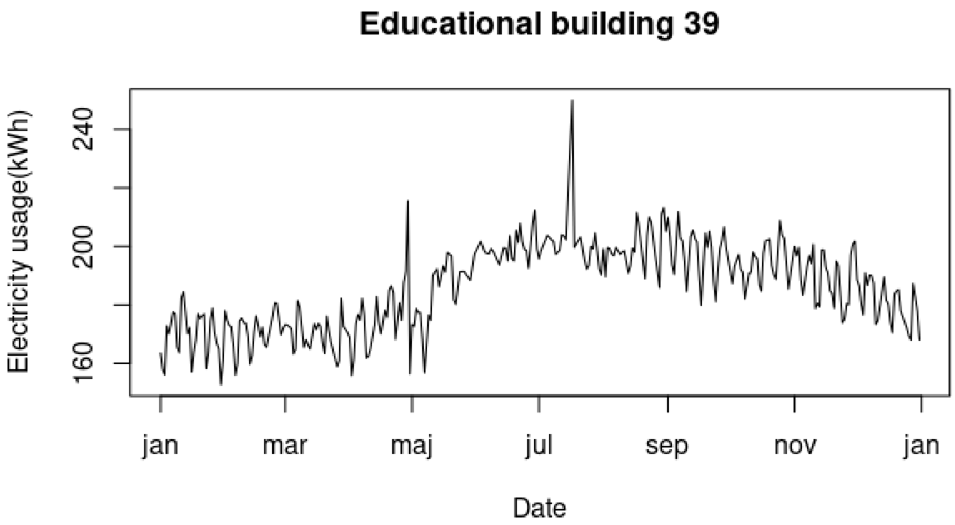
All of this confirms our hypothesis that energy use is dependent on the season, so we add the attribute season(letni_cas) that has four values(zima, pomlad, leto, jesen) that are factors.

We also tried to find if the educational institutions had a summer break that would result in lower electricity. Although some buildings had these patterns, overall no good results were found.

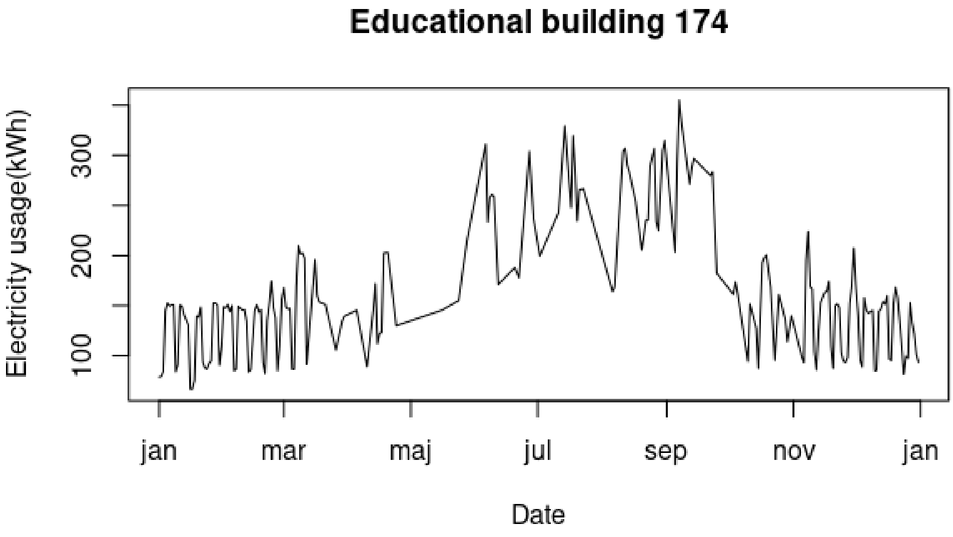


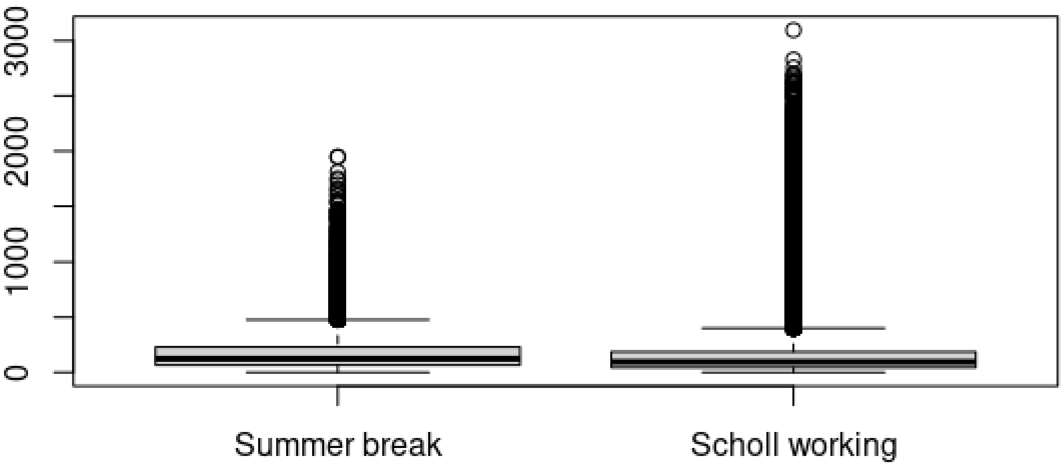


Educational building 33



Educational building 39





The activities of people differ depending on the day of the week, some institutions do not work on the weekends and people may spend more time at home. We check if the energy use is determined whether it is weekend or not. The plot(*Image 5)* shows us that in general more electricity is used during the weekend.

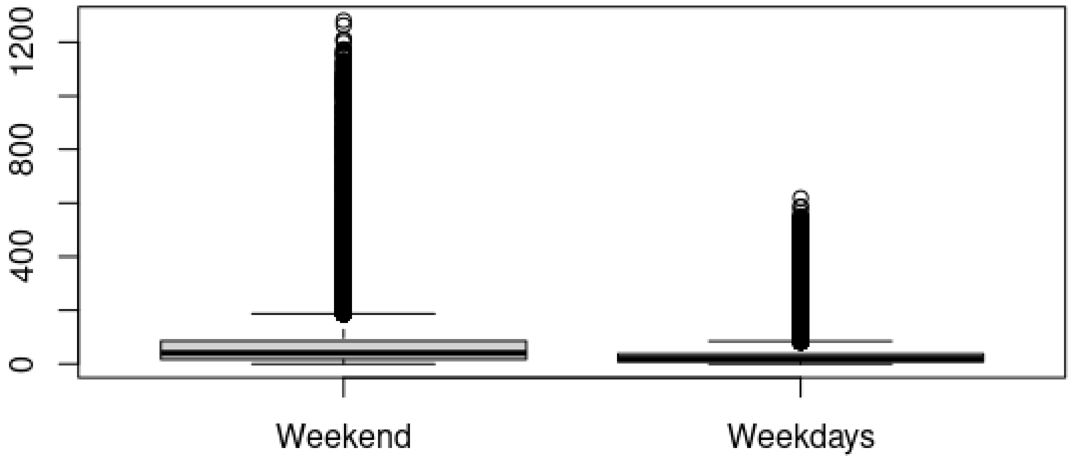
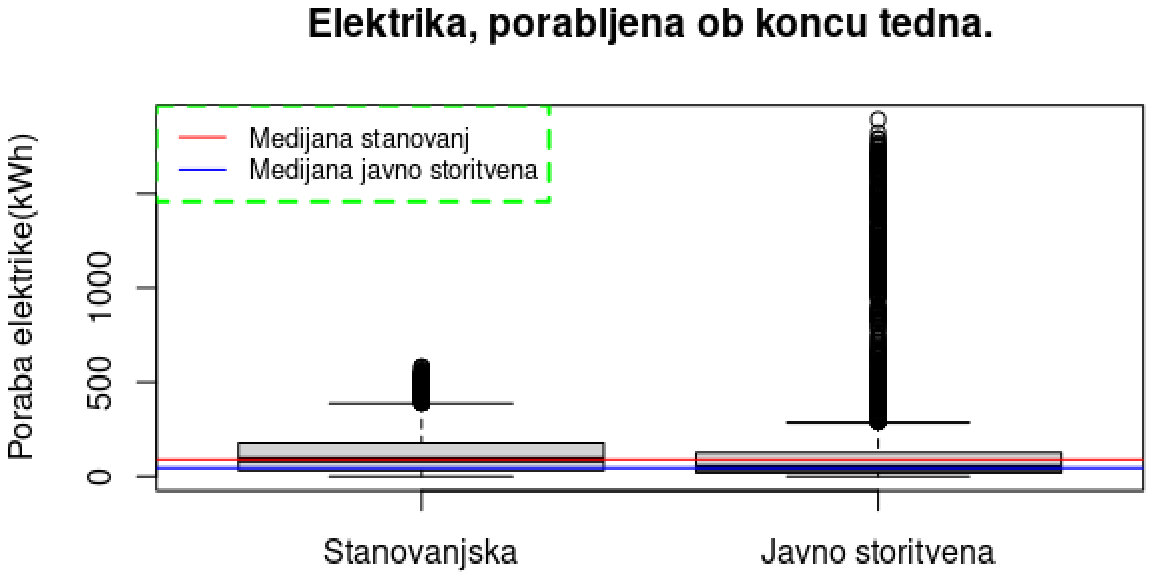


Image 5



Also, the residential buildings(stanovanjska) have higher energy use during the weekends than the public buildings(javno storitvena) which is expected. In residential buildings the usage is noticeably higher in the weekends than the weekdays.

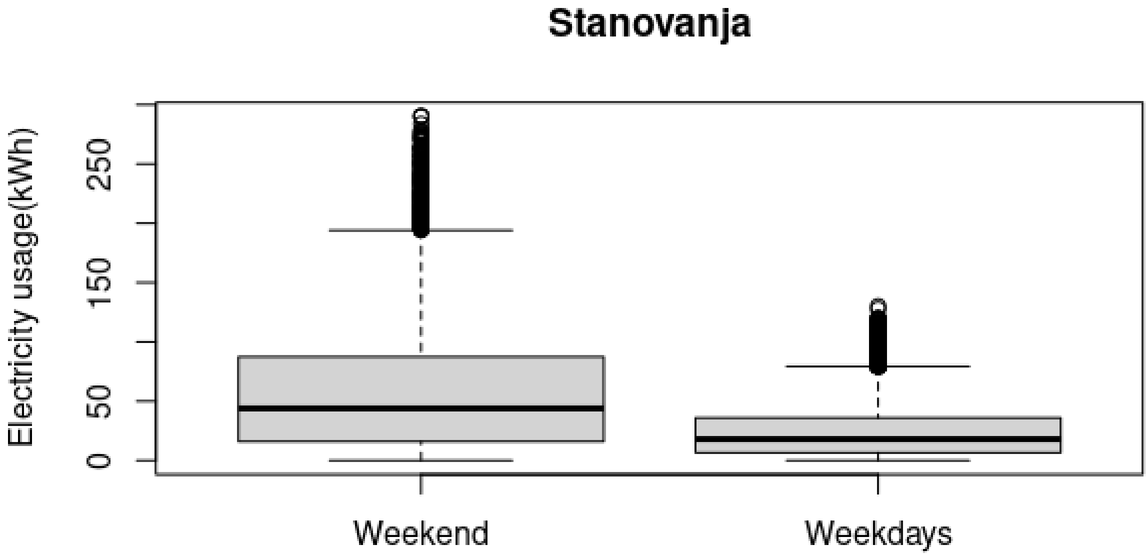


Image 7

This hints that adding a new logical attribute for weekend would be useful for the learning models.

2. Construction and evaluation of attributes

List of attributes we added:

1. vikend (logical vector, true if the day is Saturday or Sunday).
2. letni_cas(vector of factors: pomlad, leto, jesen, zima)
3. Minimum, average, maximum of the air temperature from the previous day for the same building
4. Minimum, average, maximum of the electricity used from the previous day for the same building

When calculating the statistics from the previous day, for the first day of the measurements we used the same results from that hour. For the other days if there were not any measurements from the previous day we went back up to three days. At the end there were 1467 rows with NAN values. We discarded those, because the number was relatively small and will only lower the accuracy of the models.

To evaluate our attributes, we used a variety of measures to sort them by. Listed below are the 4 highest evaluated attributed for classification and regression models based on each method.
Classification:

- **Information Gain**: min_por_prejsni_dan, pov_por_prejsni_dan, max_por_prejsni_dan and leto_izgradnje.
- **Gini Index**: min_por_prejsni_dan, pov_por_prejsni_dan, max_por_prejsni_dan and leto_izgradnje.
- **Gain Ratio**: max_por_prejsni_dan, pov_por_prejsni_dan, min_por_prejsni_dan and povrsina.
- **MDL**: min_por_prejsni_dan, pov_por_prejsni_dan, max_por_prejsni_dan and leto_izgradnje.
- **ReliefF**: namembnost, leto_izgradnje, povrsina and stavba.

Regression:

- **MSE of Mean**: pov_por_prejsni_dan, max_por_prejsni_dan, min_por_prejsni_dan and povrsina.
- **ReliefF**: max_por_prejsni_dan , pov_por_prejsni_dan, min_por_prejsni_dan and povrsina.

A problem with most of these methods is that they are short-sighted, which is why we will use the highest evaluated ReliefF attributes for training our models later.

3. Modeling

We used 4 regression models to predict the poraba attribute:

- Linear Regression
- Regression tree
- K – nearest neighbors
- SVM

For the regression tree, we compared the pruned and unpruned tree model. For K-nearest neighbors, we chose the value k = 5.
For our regression models, we used a modified data frame where we removed the datum attribute, because it does not provide valuable information and only slows down the models, and the norm_poraba attribute. From this, we took a random sample for the training and test sets.
To assess the learning of our regression models we used:

- Mean Absolute Error – MAE
- Mean Squared Error – MSE
- Relative Mean Absolute Error – RMAE

- Relative Mean Squared Error – RMSE

Model	MAE	RMAE	MSE	RMSE
Linear regression	30.45863	0.201877	4252.676	0.06699659
Regression tree(unpruned)	42.22185	0.2798425	6157.869	0.09701095
Regression tree(pruned)	18.07195	0.1197792	2236.659	0.03523628
KNN	32.84924	0.2177217	3349.831	0.05277318
SVM	20.66351	0.1369558	2445.672	0.03852907

We can see from these results, that the pruned tree model is the most accurate, closely followed by SVM.

We used 3 classification models to predict the norm_poraba attribute:

- Decision tree
- Naïve Bayes Classifier
- Random Forest

For the decision tree, we again compared the pruned and unpruned tree model. For our classification models, we used a modified data frame where we again removed the datum attribute, and the poraba attribute. From this, we took a random sample for the training and test sets.
To assess the learning of our classification models we used:

- Classification Accuracy – CA
- Brier score

Model	CA	Brier score
Decision tree(unpruned)	0.6025716	0.5515094
Decision tree(pruned)	0.8327489	0.2625211
Naive Bayes	0.3791805	0.8361006
Random Forest	0.8672641	0.1995532

From these results we can see that Random forest has the most accuracy, followed by the pruned Decision tree model. We can see that Naïve Bayes is a weak model, as it does not consider the dependencies between attributes.

Models taught with a subset of attributes

In addition to training each model on the entire data frame, we also trained each model on the highest evaluated Relieff attributes.

For our classification models, these attributes were: namembnost, leto_izgradnje, povrsina and stavba.
For our regression models, the attributes released by the Relieff algorithm were: max_por_prejsni_dan , pov_por_prejsni_dan, min_por_prejsni_dan, and povrsina.
As the top 3 were all derived from the previous day's poraba attribute, we found that it is better to only use one of these and the 2 following highest evaluated attributes.
For our regression models, we thus used: max_por_prejsni_dan , povrsina, ura and pritisk.

Listed below are the learning assessments for models trained with these attributes:

Model	CA	Brier score
Decision tree(unpruned)	0.4679038	0.6710721
Decision tree(pruned)	0.6436782	0.473547
Naïve Bayes	0.3779629	0.7329174
Random Forest	0.6442301	0.6580415

Model	MAE	RMAE	MSE	RMSE
Linear regression	37.12275	0.2460462	6008.596	0.09465932
Regression tree(unpruned)	46.2569	0.3065865	7299.444	0.1149953
Regression tree(pruned)	42.22185	0.2798425	2236.659	0.1197917
KNN	27.85457	0.1846175	4260.888	0.06712595
SVM	30.21697	0.2002753	4377.999	0.06897092

We can see from these results that using these attributes for learning in every model leads to lower accuracy.

Combination of models(bagging, voting, boosting)

We decided to combine the three trivial models(decision tree, naive Bayes and k-nearest neighbors) and make the voting and weighted voting models.

Classification:

Model	CA
Decision tree(pruned)	0.83
Naïve Bayes	0.43
K-nearest neighbors	0.51
Voting	0.64
Weighted voting	0.78
Bagging	0.86
Boosting	0.64

Regression:

Model	RMSE
Bagging	0.1494

The bagging model gave us the best results outperforming the decision tree, while voting and weighted voting had lower results because naive Bayes and k-nearest neighbors had relatively bad accuracies. Also boosting did not preform that well.
The regression bagging had good accuracy.

Comparison between models taught with buildings from just one region

We divided our modified data frames into new ones based on the regija attribute.
We took samples from each for our training and test sets.
The accuracies for each model from these regions are listed below.

Models taught with “vzhodna” region data:

Model	CA	Brier score
Decision tree(unpruned)	0.5968799	0.5499633
Decision tree(pruned)	0.8096427	0.301578
Naive Bayes	0.3199768	0.9380646
Random Forest	0.8463098	0.2319066

Model	RMAE	RMSE
Linear regression	0.2453173	0.08104204
Regression tree (unpruned)	0.3118476	0.1207086
Regression tree (pruned)	0.135979	0.04262494
KNN	0.2399431	0.06332326

Models taught with “zahodna” region data:

Model	CA	Brier score
Decision tree(unpruned)	0.6512703	0.4988423
Decision tree(pruned)	0.8682622	0.2092135
Naive Bayes	0.4865907	0.7098367
Random Forest	0.8872207	0.1721131

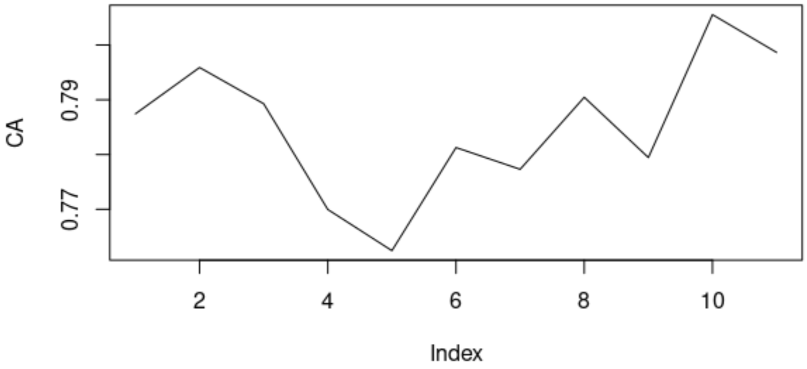
Model	RMAE	RMSE
Linear regression	0.149892	0.02491583
Regression tree (unpruned)	0.2169422	0.04695856
Regression tree (pruned)	0.08804727	0.01260737
KNN	0.1989326	0.03748677

We can see from these results, that models trained with “vzhodna” data have slightly lower accuracy than models trained with both regions, while models trained with “zahodna” data have significantly better accuracy.

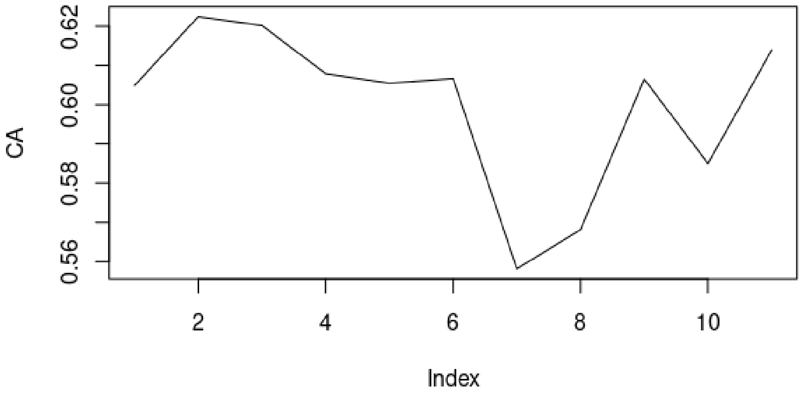
4. Model evaluation

Model evaluation was done on 12 subsets as described in the task. Model accuracy for classification was measured with CA and for regression with RMSE. Accuracies are the following:

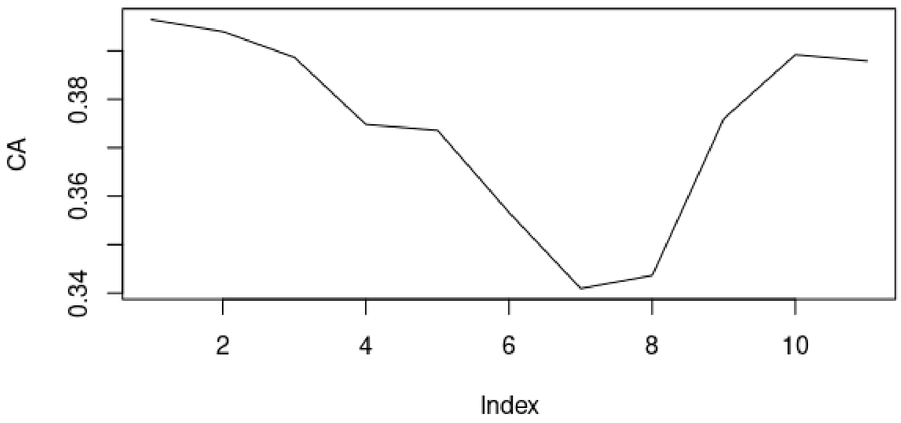
1. Pruned decision tree:
   mean=0.785; standard deviation =0.0127



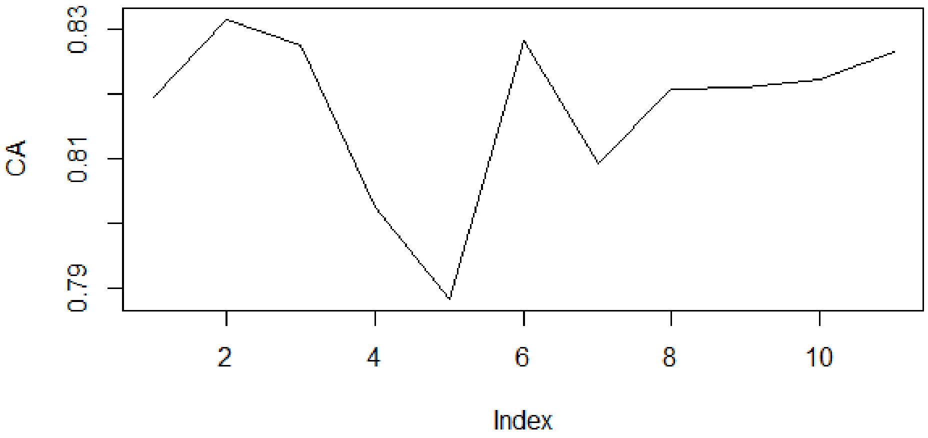
2. Unpruned decision tree:
   mean=0.60; standard deviation = 0.0207



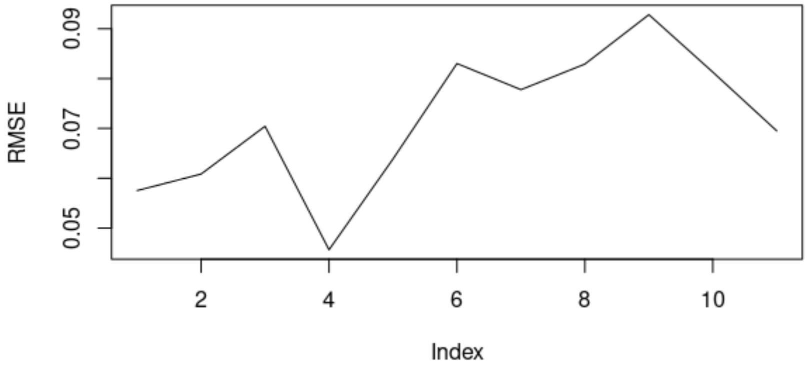
3. Naive Bayes
   mean=0.375; standard deviation=0.0196



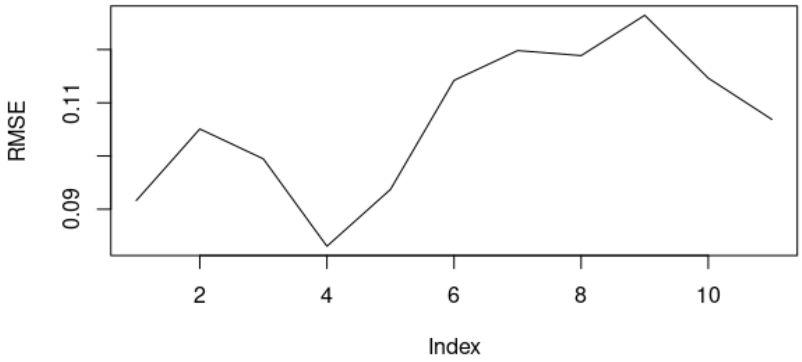
4. Random forest
   mean=0.818, standard deviation=0.0129



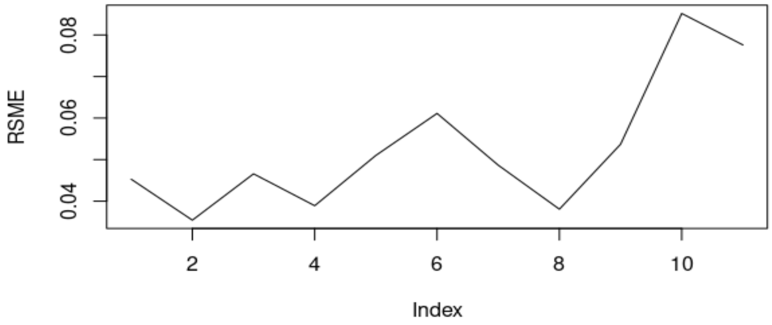
5. Linear regression
   mean=0.0714; standard deviation= 0.0138



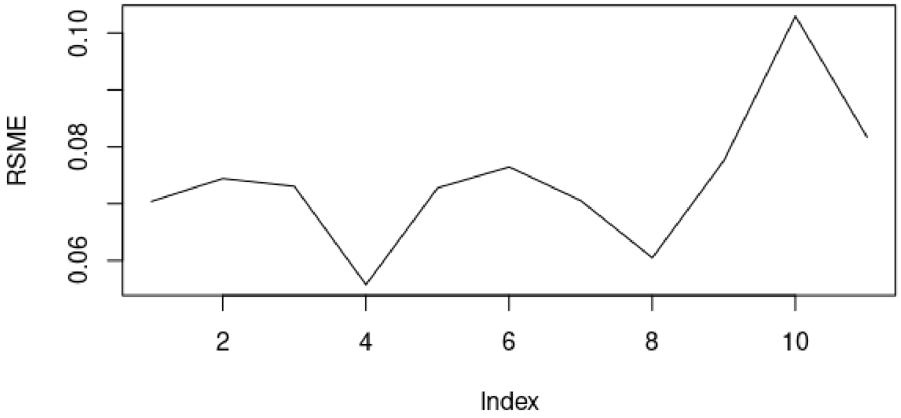
6. Unpruned regression tree
   mean=0.107; standard deviation=0.0136



7. Pruned regression tree
   mean=0.0529; standard deviation=0.016



8. K-nearest neighbors regression
   mean=0.0742; standard deviation=0.012



From the classification models most accurate is random forest with a mean CA of 0.81, while from the regression models most accurate is the pruned regression tree with mean of RSME of 0.0529.
Nearly at every model we can notice a pattern of falling of the accuracy with the change of seasons(data of one season can be different from the others). However once they encountered the new data accuracies get better.