Report for the Course Modelling in Computational Science, HT23

Project 3: Biome classification

Theo Koppenhöfer
(with Anna and Carmen, Group 4)

Lund

October 31, 2023

Introduction

The following report is part of the third project of the course Modelling in Computational Science, BERN01, taken at Lund university. In this project we will use machine learning to classify biomes based on climate and soil data. We will test the performance of our machine learning model in binary and multiclass classification. We will also compare our model with LPG GUESS output and modify our model to predict continuous outputs of LPG GUESS. For this we will discuss the choice of regions and biomes, the setup of our model, give some interesting results, discuss these and finally give a conclusion. The code to the project was implemented in a `jupyter notebook`. The project report and code can be found online under [1].

Methods

For the binary classification model we chose the biomes ‘arid shrub’ and ‘desert’. For the choice of regions for training and testing we choose two countries which contained sufficient samples of both biomes. In figure 1 we plot the amount of shrub and desert data points in countries with at least 30 samples of each biome. Our initial choice of regions was Egypt (EGY) and China (CHN). It turned out however that when we took out the LPG GUESS output from the training data, our model could not handle the classification well. The reason is that deserts in these countries have very different climates. Thus we chose Egypt for training and Libya (LBY) for testing. The variable we trained for was the observed biome `Biome_obs` which was determined with the help of satellite data.

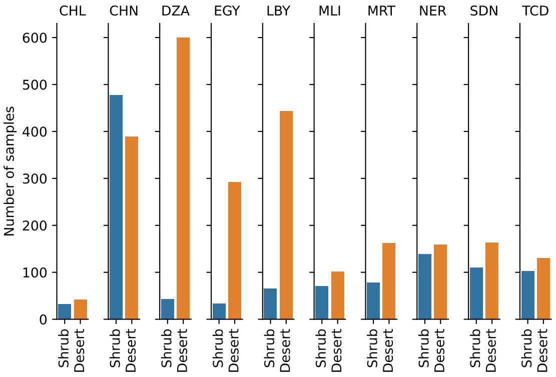

Figure 1: Amount of shrub and desert landscape in selected countries.

For the multiclass classification we initially chose Africa and China but that too turned out to be a poor choice. Thus we switched to the regions to Russia for training and Canada for testing. We trained and tested for both `Biome_obs` and the LPG GUESS biome classification `Biome_cmax`.

For the regression model we chose Canada to train and Russia to test the model. The reason for this switch of roles lies in the performance of the training. The LPG GUESS outputs we trained the model for were `NPP` (net primary productivity) and `VegC` (vegetation carbon pool).

If not otherwise stated we used as training data all climate data in the file `’data_index_2.csv’` together the soil data `clay`, `silt`, `sand` and `orgC`. The climate data includes the standard deviation, mean and median over the four seasons for radiation, precipitation and minimum, maximum and mean temperature measurements over the years 1961-1990.

For the implementation we made extensive use of the `sklearn` library. We implemented the classification model with `RandomForestClassifier` and analysed the permutation importance with `permutation_importance`. The hyperparameter tuning was initially implemented with `GridSearchCV` and after some bad initial results we switched to `HalvingGridSearchCV` (the results did not improve). We varied the parameters `max_depth`, `n_estimators` `min_samples_leaf` each between 4 reasonably chosen values around their default. The regression model was implemented using `RandomForestRegressor`. The clustering was performed according to [6]. To analyse the importance of features we also implemented a crude routine which runs the model whilst dropping some features and then plots the results for each run.

Results

In this section we will first discuss our results for the binary classification, then for the multiclass classification and finally the regression problem.

Binary classification

In figure 2 one can see the geographical distribution of the biomes in Egypt and Libya. As expected the desert is inland whereas the dry shrub is closer to the sea.

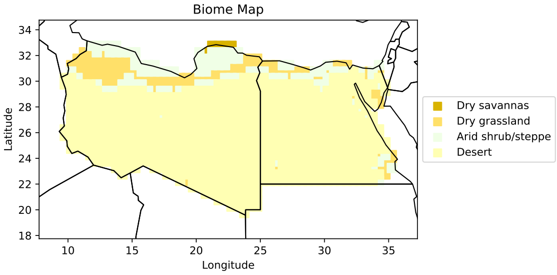

Figure 2: Biome map for Egypt and Libya.

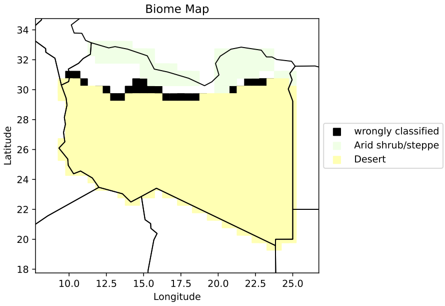

Figure 3: Map of biomes which our model classified wrongly.

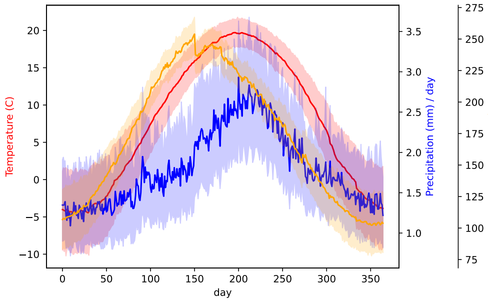

Figure 4: Mean temperature, precipitation and radiation in Egyptian shrubs.

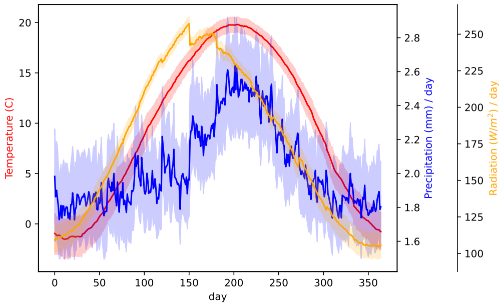

Figure 5: Mean temperature, precipitation and radiation in Egyptian deserts.

In figures 4 and 5 we see the time evolution of the mean temperature, precipitation and radiation in the Egyptian shrub and desert landscapes over a year. In both climates the temperature, precipitation and radiation peak around the summer. At closer inspection one can spot subtle differences between the diagrams for the different variables. We note however that the mean radiation levels look almost identical. Although not shown here the climate plots for Libya are quite similar.

We also plotted the distribution of the various soil features in figure 6. One can see that the soil features do not differ much between the Egyptian desert and shrub.

When we tested our trained model on Libya we got the results depicted in the confusion table 1. The accuracy of our model was approximately 0.96 which is reasonably good. The balanced accuracy is with approximately 0.85 slightly more modest. The reason for this lies in the bad recall rate for the arid shrub biome. In the map 3 one can see that all misclassified regions lie at the boundary of the desert and shrub biomes.

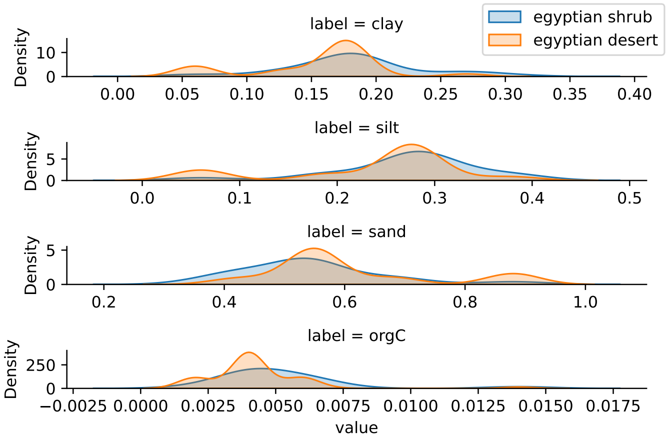

Figure 6: Distribution of the various soil features for the Egyptian desert and shrub.

Now we analyse the importance of features for the binary classification. In figure 7 the MDI (random forest feature importance) shows that the precipitation levels played an important role in training and the soil played an insignificant role. Note also that the decrease in accuracy score for all features in the permutation importance is insignificant.

Truth Predicted	arid shrub	desert
arid shrub	46	0
desert	20	443

Table 1: Confusion table.

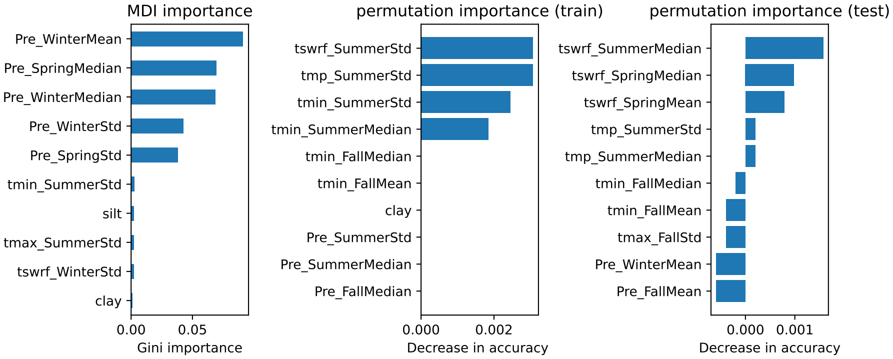

Figure 7: The five most and least important features.

We also created a dendogram which for space reasons has to be admired in the appendix. Unsurprisingly we found that the medians and means were strongly correlated. One could also see that the soil data was relatively independent from the rest of the data. Notably the standard deviation of the temperatures in the summer and fall were also quite independent. The results from the following clustering are then shown in figure 8. Here one can notice that the soil features are relatively unimportant as is the cluster including the mean radiation in the summer.

The results of our routine dropping various features can be seen in figure 9. On the y-axis we depict the balanced error rate rather than the classical error rate but the qualitative behaviour is identical. The abbreviations ‘pre’, ‘tmp|tmin|tmax’ and ‘tswrf’ correspond to data representing precipitation, temperatures and radiation respectively. Firstly one sees that most modifications have little impact. When we drop all the

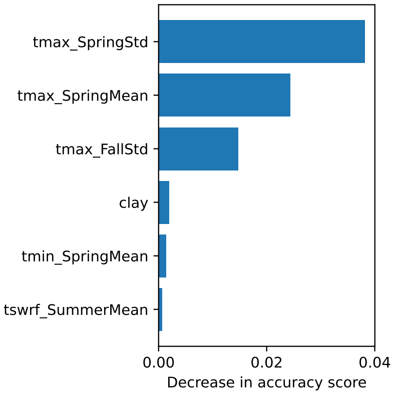

Figure 8: Feature importance after clustering.

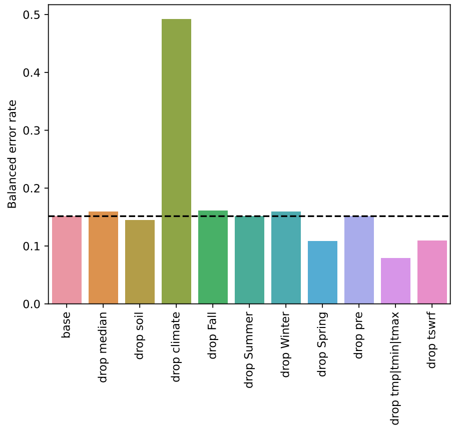

Figure 9: Results for experiment series.

climate data and only train our model on the soil data it performs very badly. Surprisingly, dropping the spring, temperature or radiation data significantly improves the performance of our model.

Multiclass classification

We start by giving an overview of the biomes which were classified. Figure 10 shows the geographical distribution of biomes in Russia and Canada. One sees that the order in which the biomes appear in Canada as one travels northwards is quite similar to the order in which the biomes appear in Russia as one travels northeast. Although not shown here the climate diagrams for Russia and Canada are quite similar.

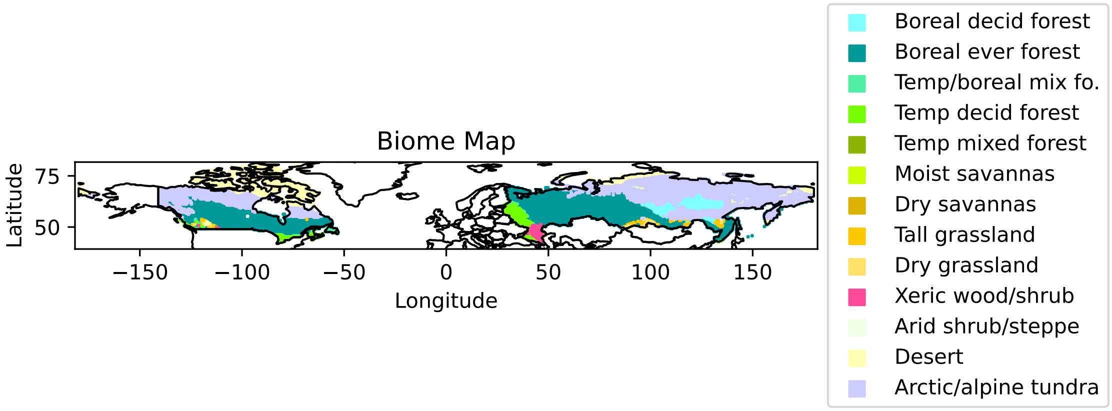

Figure 10: Overview of the biomes `Biome_obs` in Russia and Canada.

The results for the multiclass classification are shown in table 2. The accuracy of 0.85 is worse than for the binary classification but not by much. The balanced accuracy of about 0.47 is on the other hand significantly worse. Once again a map of the misclassified regions can be seen in 11 and we note that these regions tend to lie between two biomes.

	precision	recall
macro average	0.57	0.47
weighted average	0.85	0.85

Table 2: Average precision and recall for the multiclass classification.

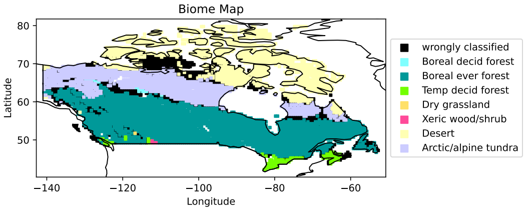

Figure 11: Misclassified regions for `Biome_obs`.

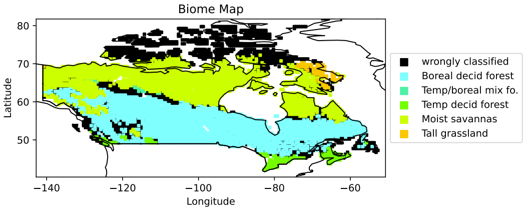

Figure 12: Misclassified regions for `Biome_Cmax`.

The results regarding MDI and permutation importance are shown in figure 13. The dendogram once again was banished to the appendix. The results of our own set of experiments dropping selected features are shown in figures 15. Here dropping the climate and the temperature has the biggest negative impact. Dropping the winter and the soil on the other hand have a mild positive impact.

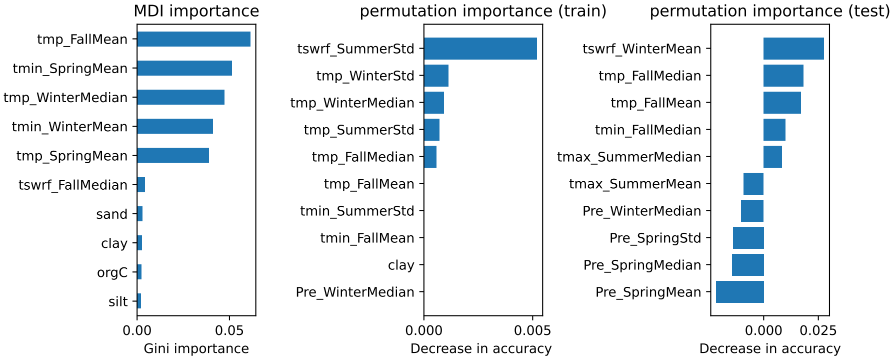

Figure 13: The five most and least important features.

Regarding hyper parameter tuning our results were quite mixed. When we initially used Africa and China as training and test regions the hyperparameter tuning could in certain circumstances improve our model. In most cases it tended to overfit if the ranges for the hyperparameters was not chosen carefully. The overfitting was visible in a very high accuracy on the train and a very low accuracy on the test set. For Russia and Canada the hyperparameter tuning did not improve the performance our model. Here too it tended to overfit.

We also trained and ran the model to predict the LPG guess output `Biome_Cmax`. Due to the strict page limit the author will have to spare the reader the rather uninteresting details but we note that the accuracy and the weighted accuracy were with 0.72 and 0.57 high. We also give in figure 12 a map showing the results. One should note here the large difference between the biome types to `Biome_Cmax` in figure 11. More interestingly we tested a model trained on `Biome_obs` with `Biome_Cmax` data. Here the accuracy plummeted to 0.14 and the balanced accuracy to 0.31.

When we tested a model trained on `Biome_Cmax` on `Biome_obs` the results were similarly abysmal and will be skipped here.

Regression

We begin by plotting the distribution of the parameters `NPP` and `VegC` in both Canada and Russia. The results can be seen in figure 16. We see that `VegC` has a lot of very small values whereas `NPP` is more spread out in both domains. We also note that the values for `VegC` are about an order of magnitude larger than those of NPP.

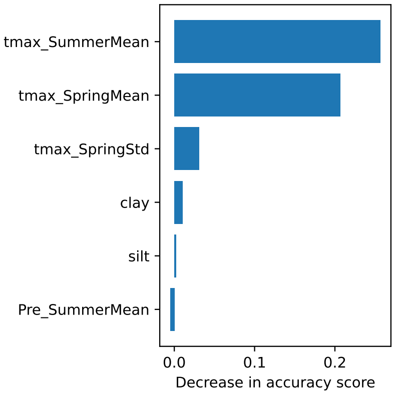

Figure 14: Feature importance after clustering.

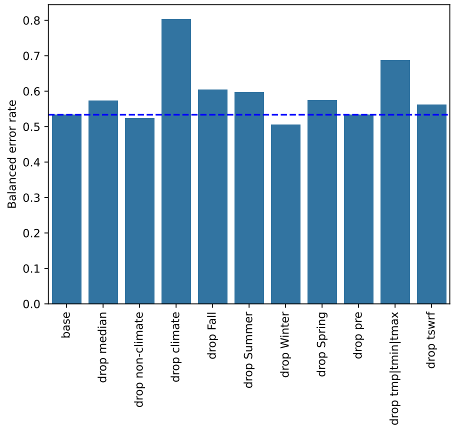

Figure 15: Results for experiment series.

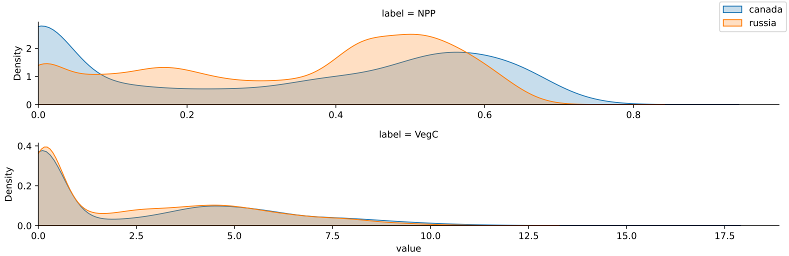

Figure 16: Distribution of NPP and VegC in Canada and Russia.

We first trained the model for NPP and tested it on Russia. The results can be seen in the scatterplot 17. It can also be seen in the distribution of the residues $\hat{Y}_{\text{test}} - Y_{\text{test}}$ in figure 18. We also plotted a map of the residues seen in figure 19. From this one sees that the model works best in the north and east of Russia.

After this we trained and tested the model for VegC. The plots of the residuals and the predicted versus true values are very similar to NPP which is why they will will not be discussed here and the keen reader is referred to the appendix.

We will proceed in taking a closer look at the model predicting NPP. The feature importance can be seen in figures 20 and after clustering in 21. Note that the temperature in the spring, summer and fall the most important factors for for the training data. For the test radiation is the most important factor. For a dendogram of this experiment we refer to the appendix.

The results of our own set of experiments is shown in figure 22. Here we show the mean square error on the y-axis. We created the same chart also for the R^2 error, the mean absolute error and the maximal error but the qualitative behaviour for these different error metrics was the same. Once again we remark on the outliers. Unsurprisingly the experiment dropping all the climate data performs terribly. Dropping the temperatures decreases accuracy significantly. Similarly, though not as pronounced, dropping the fall data decreases accuracy. When dropping the data for the summer on the other hand the error decreased.

Discussion

During the project it became apparent that the machine learning model gives features that differ between the categories a higher importance than features that are largely identical. Here we note that our interpretation

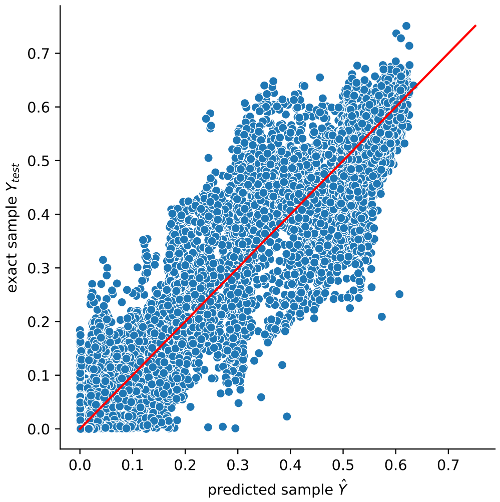

Figure 17: Predicted versus true values for the parameter NPP.

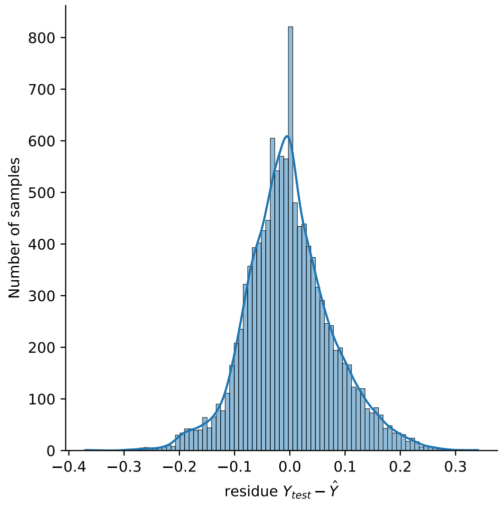

Figure 18: Distribution of the residues for the parameter NPP.

was hampered by the strong correlation between features. This correlation made itself noticeable in the low decreases in accuracy in the permutation importance plots 7 and 13. Nonetheless figure 7 showed that the soil cluster was relatively unimportant. On the other hand 6 showed that the soil features in the training set did not differ much. The unimportance of soil features was a common theme during the project. Similarly the cluster of the summer radiation was relatively unimportant. Previously we remarked in the description of the climate diagrams 4 and 5 that the radiation was quite similar in both biomes.

From the figure 9 produced by our experiment series we can get an idea of some measures that differ between the Egyptian and Libyan biomes. The accuracy increases when dropping spring, temperature or radiation. This means that at least for these measures there seems to have been some overfitting. We note that according to figure 7 they also have a big impact on the response of the model on the test set.

Regarding the performance of the multiclass classification we note that the large difference in the weighted accuracy and the accuracy originates in the large variance of sample sizes for the respective biomes. Additionally the small biomes tend to be classified poorly. Because the classification problem at hand should treat every biome equally the author believes that the balanced accuracy is more appropriate here. We also note that the similarity between the countries almost certainly significantly improved the performance of the model.

Since our model works well for `Biome_obs` and `Biome_Cmax` but performs terribly if we test a model trained on `Biome_obs` with `Biome_Cmax` we can conclude that the `Biome_Cmax` classification is a bad approximation of `Biome_obs`. That the `Biome_Cmax` data is quite different from the `Biome_obs` data is however quite apparent from a glimpse at the maps 11 and 12.

In the regression part we see from the feature importance plot 20 that our model chose the temperature to be the most important feature to determine the `NPP`. The plot 22 showing the result for the experiment series on the other hand showed that the temperature is also a very good predictor. Together this indicates that the model made a good choice when choosing the temperature to be the most important feature. This also intuitively makes sense since temperature is a major factor when it comes to plant growth in the arctic.

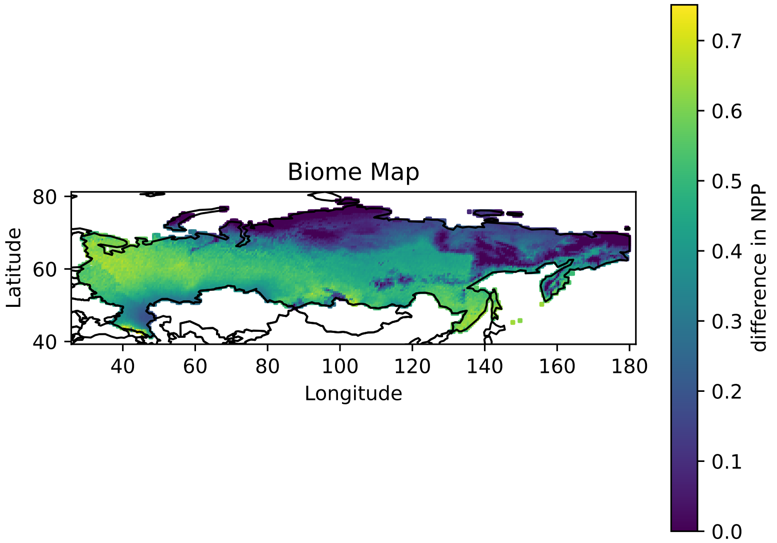

Figure 19: Map of the residue.

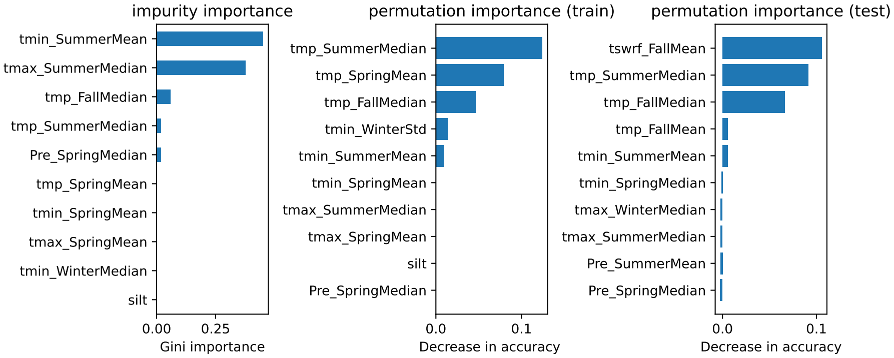

Figure 20: The five most and least important features.

Conclusion

We saw that the multiclass classification only performed well as long as the training and test regions were quite similar. This is a disadvantage of the random forest classifier since it can only build on preexisting data and does not generalise well to new regions. In this regard the random forest classifier acts a lot like an interpolator and should not be used for extrapolation. We also saw that the performance of the model depended a lot more on the quality and quantity of the data that it is fed than the hyperparameters. An advantage over more extensive models like LPG GUESS is that it is quickly set up. It has however the disadvantage that it then is quite cumbersome to understand this model and to gain valuable insights for human minds.

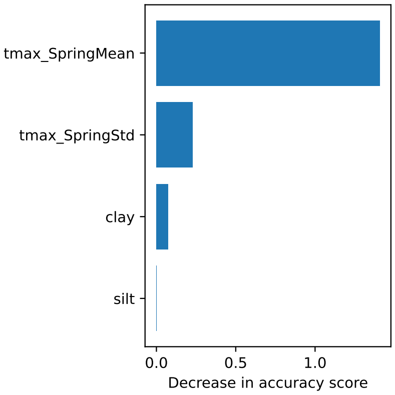

Figure 21: Feature importance after clustering.

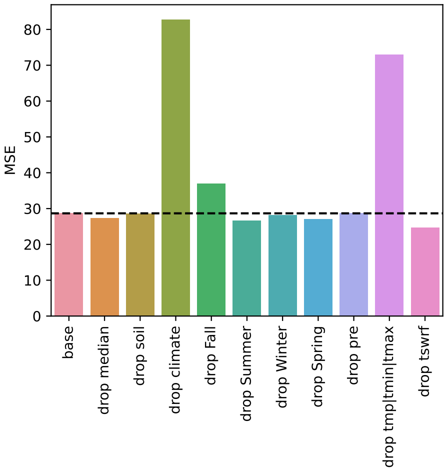

Figure 22: Results for experiment series.

Bibliography

[1] computational-science-HT23, *Github repository to the project.* Online, 2023. [Online]. Available: `https://github.com/TheoKoppenhoefer/computational-science-HT23`.

[2] W. Zhang, *Modelling large scale ecosystems - dynamic global vegetation model*, BERN01, University of Lund, Oct. 2023.

[3] L. Hartman, *Introduction to machine learning, Bern01: Machine learning and big data, lecture 2*, BERN01, University of Lund, Oct. 2023.

[4] ——, *Random forest for regression and classification, Bern01: Machine learning and big data, lecture 3*, BERN01, University of Lund, Oct. 2023.

[5] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An introduction to statistical learning—with applications in Python*, ser. Springer Texts in Statistics. Springer, Cham, [2023] ©2023, pp. xv+607, ISBN: 978-3-031-38746-3; 978-3-031-38747-0. DOI: `10.1007/978-3-031-38747-0`. [Online]. Available: `https://doi.org/10.1007/978-3-031-38747-0`.

[6] Scikit learn, *Permutation importance with multicollinear or correlated features.* Online, Accessed October 2023. [Online]. Available: `https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance_multicollinear.html`.

Appendix

General overview

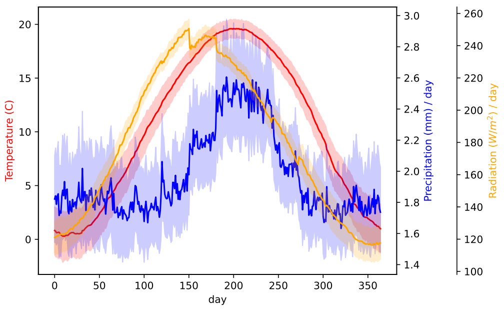

Figure 23: Climate diagram for Libya.

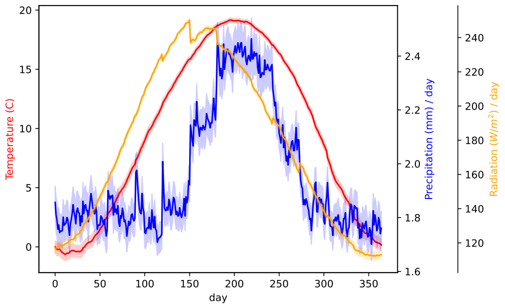

Figure 24: Climate diagram for Canada.

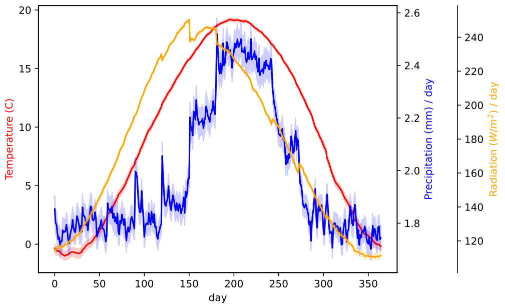

Figure 25: Climate diagram for Russia.

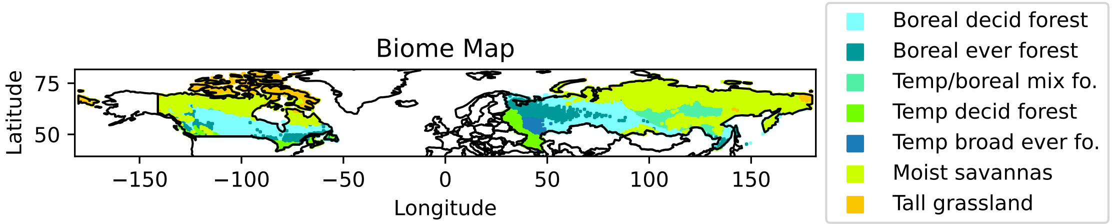

Figure 26: Map of biomes `Cmax` in Russia and Canada.

Binary classification

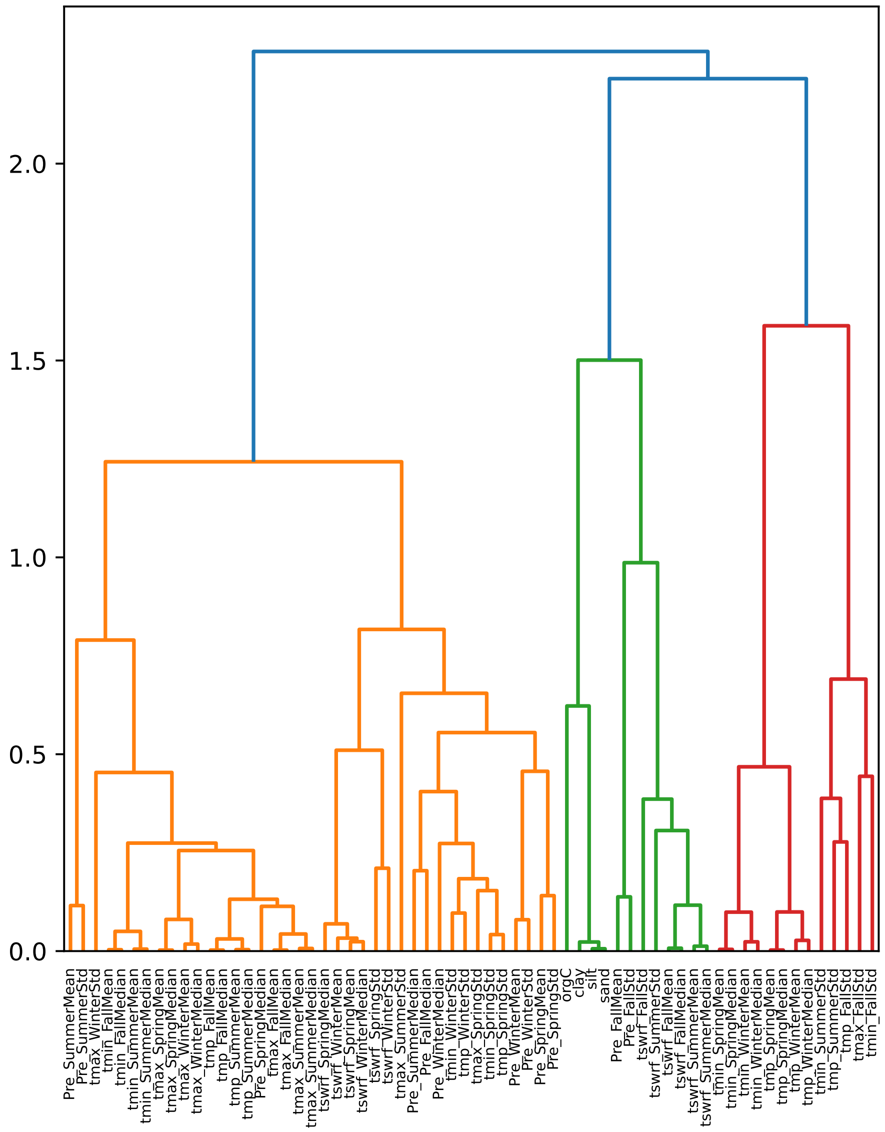

Figure 27: Dendogram to the binary classification.

Multiclass classification

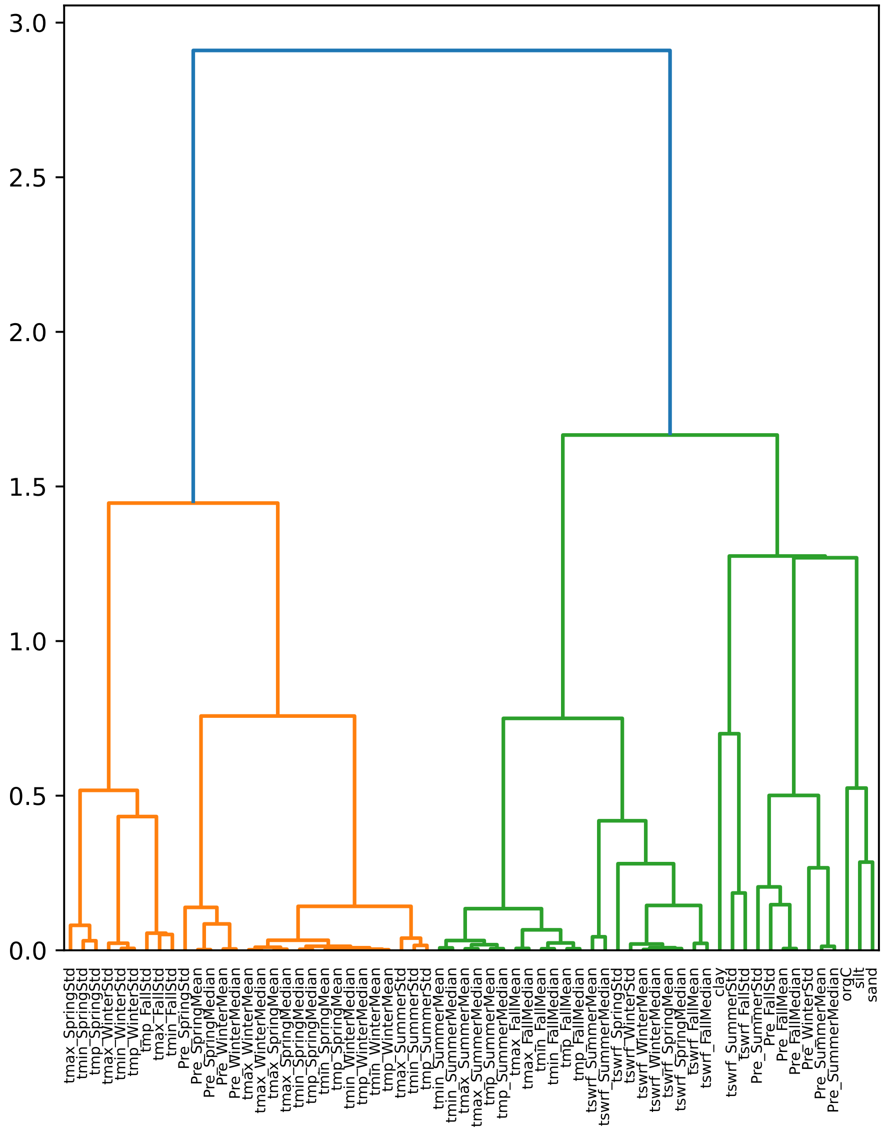

Figure 28: Dendogram to the multiclass classification.

Truth	1	2	3	5	12	13	14	15	17	18
Predicted										
1	19	106	0	0	0	0	0	0	0	9
2	0	2742	61	49	1	17	8	1	2	242
3	0	0	0	2	0	0	0	0	0	0
5	0	0	5	102	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	1	1	0	3	0
15	0	1	0	1	0	0	0	4	0	0
17	0	1	0	0	0	0	0	0	1523	84
18	15	41	0	0	0	0	0	0	342	1114

Table 3: Confusion table for the multiclass classification with `Biome_obs`.

	precision	recall	f1-score	support
1	0.141791	0.558824	0.226190	34.000000
2	0.878002	0.948461	0.911872	2891.000000
3	0.000000	0.000000	nan	66.000000
5	0.953271	0.662338	0.781609	154.000000
12	nan	0.000000	nan	1.000000
13	nan	0.000000	nan	18.000000
14	0.200000	0.111111	0.142857	9.000000
15	0.666667	0.800000	0.727273	5.000000
17	0.947139	0.814439	0.875791	1870.000000
18	0.736772	0.768806	0.752448	1449.000000
accuracy	0.847314	0.847314	0.847314	0.847314
macro avg	0.565455	0.466398	0.631149	6497.000000
weighted avg	0.854244	0.847314	0.857335	6497.000000

Table 4: Classreport for the multiclass classification with `Biome_obs`.

Truth	1	2	3	5	6	11	13
Predicted							
1	2424	352	0	32	0	161	1
2	0	0	0	0	0	0	0
3	47	1	19	0	0	19	0
5	18	38	0	168	0	0	0
6	2	3	0	1	0	0	0
11	68	1	12	0	0	2099	1016
13	0	0	0	0	0	0	17

Table 5: Confusion table for the multiclass classification with `Biome_Cmax`.

	precision	recall	f1-score	support
1	0.816162	0.947245	0.876831	2559.000000
2	nan	0.000000	nan	395.000000
3	0.220930	0.612903	0.324786	31.000000
5	0.750000	0.835821	0.790588	201.000000
6	0.000000	nan	nan	0.000000
11	0.656758	0.921018	0.766758	2279.000000
13	1.000000	0.016441	0.032350	1034.000000
accuracy	0.727343	0.727343	0.727343	0.727343
macro avg	0.573975	0.555571	0.558263	6499.000000
weighted avg	0.782587	0.727343	0.687038	6499.000000

Table 6: Classreport for the multiclass classification with `Biome_Cmax`.

Regression

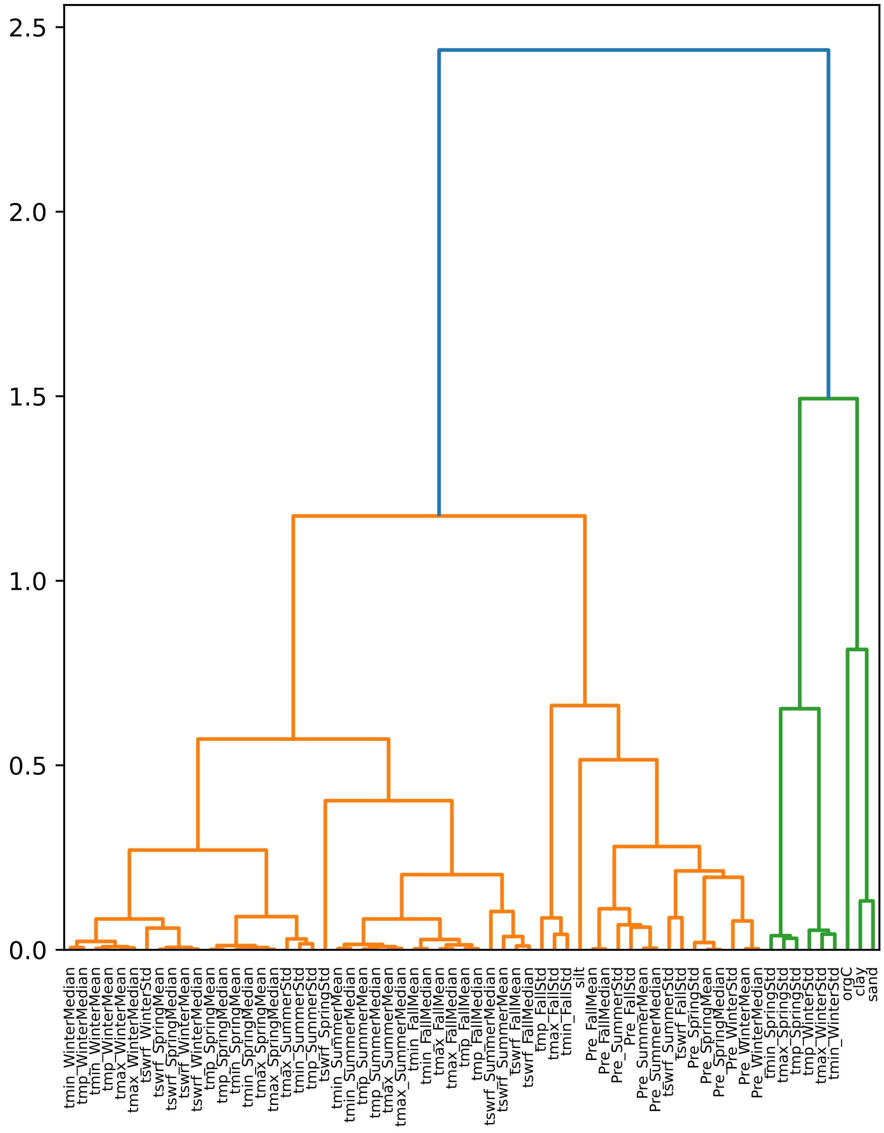

Figure 29: Dendogram to the regression with NPP.

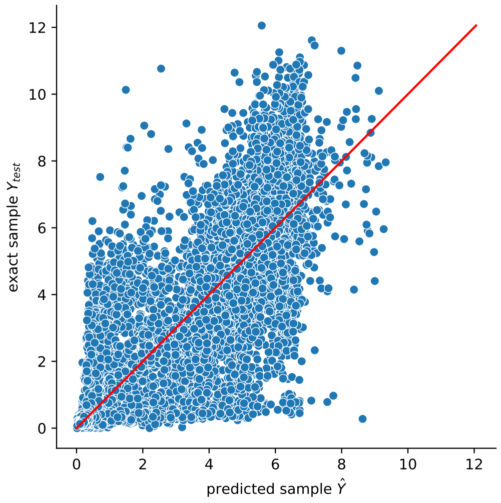

Figure 30: Predicted versus true values for the parameter VegC.

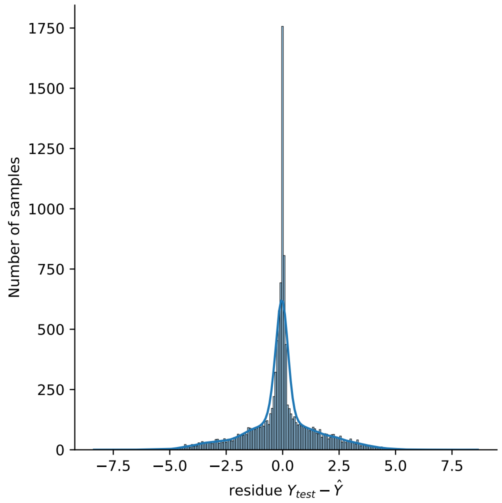

Figure 31: Distribution of the residuals for the parameter VegC.