Report for the Course Modelling in Computational Science, HT23

Project 3: Biome classification

Theo Koppenhöfer

(with Anna and Carmen, Group 4)

Lund

October 28, 2023

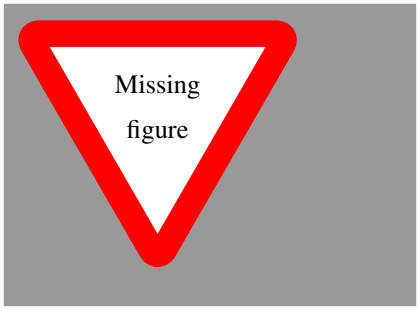

Figure 1: Number of data points with ‘desert’ and ‘arid shrub’ in selected countries.

Introduction

The following report is part of the second 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 classification and in distinguishing multiple biomes for different regions. We will also compare our model with LPG_guess output and modify our model to predict continuous variables 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].

‘net primary productivity’ (*NPP*) and ‘vegetation carbon pool’ (*VegC*)

Methods

To test our first binary classification model we chose the biomes ‘arid shrub’ and ‘desert’. For the choice of regions we had to choose two countries which contained sufficient amount of both regions. A plot of regions with sufficient amounts of both biomes can be seen in figure Our initial choice was Egypt and China. It turned out however that when we took out soil data our model could not handle the classification well since the deserts in both countries have very different climates. Thus we decided for Egypt for the training and Libya for the testing.

For the classification of multiple biomes we initially chose Africa and China but this quickly turned out to be a poor choice as both regions have very different climate data. Thus we switched to the regions to Russia for training and Canada for testing.

For the regression model chose Canada to train and Russia to test the model. The reason for this switch of roles lies in the performance of the training.

If not otherwise stated we used as training parameters all climate data in the file `'data_index_2.csv'` together the soil parameters 'clay', 'silt', 'sand' and 'orgC'.

For the implementation we made heavy use of the `sklearn` library. We implemented the classification model with `RandomForestClassifier`. We analysed the permutation importance with the function `permutation_importance`. The hyperparameter tuning was initially implemented with `GridSearchCV` and after some bad initial results we switched to `HalvingGridSearchCV`. The regression model was implemented using `RandomForestRegressor`.

For analysing the importance of features we also implemented a 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

Insert the maps

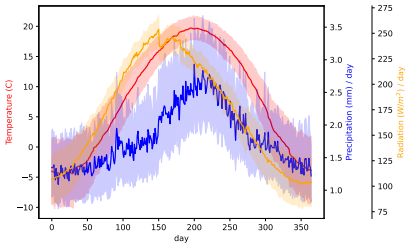

Figure 2: Mean temperature, precipitation and radiation in egyptian shrubs.

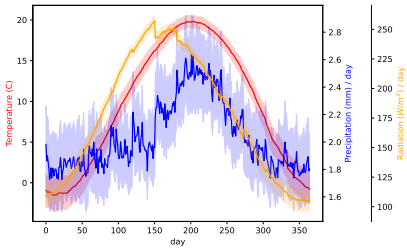

Figure 3: Mean temperature, precipitation and radiation in egyptian deserts.

We start by giving some statistics on the desert and arid shrub landscape in egypt and libya. In figures 3 and 3 we see a plot of the temperature, precipitation and radiation in the egyptian shrub and desert landscapes. One can see in both climates that the temperature, precipitation and radiation peak around the summer. We hypothesis that in machine learning model the features that differ between these two plots will be more important and all others are of lesser importance so we procede in highlighting the differences and similarities. Regarding temperature it is notable that the summer temperatures are quite similar whereas the winter temperatures differ. The variation in temperature seems to be also far greater for egyptian shrubs. Though the standard deviation in the figure is between the respective biomes in egypt, it is reasonable to assume that the standard variation in the corresponding regions is correlated the parameters `tmp_SummerStd` and `tmp_WinterStd` where the temperature plateaus.

more precise

It is apparent from the plots that the precipitation level is a little higher in the desert in winter though its variance seems roughly identical in the winter and summer months. Similarly we see that the radiation levels in both biomes behave almost identically though in the shrubs it varies more. Although not shown here the climate plots for libya are quite similar.

We also plotted the distribution of the various soil features in figure 4. One can see that the soil features do not differ much between the egyptian desert and shrub. Thus they will probably not play much of a role.

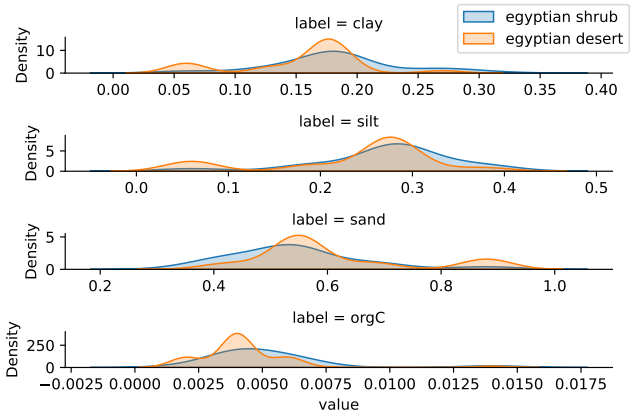

Figure 4: Distribution of the various soil features for the egyptian desert and shrub.

When we tested our trained model on libya we got the results depicted in the cocnfusion table 1. The accuracy of our model was approximately 0.96 which is reasonably good. However since there is inherently a large imbalance in our dataset the balanced accuracy is a better measure here. This is with approximately 0.85 only slightly more modest.

Truth Predicted	16	17
16	46	0
17	20	443

Table 1: Confusion table.

Now we analyse the importance of features for the binary classification. In figure 5 we can see from the mdi_importances that indeed as hypothesised previously the precipitation levels play an important role and the soil plays an insignificant role. Since there is a lot of colinearity between many parameters the permutation importance is a better measure in this case. Here again at least 'clay' plays an insignificant role. As predicted the temperature and its variation in the summer months play a large role. The third plot shows how good a predictor the various features are at predicting the

look this up

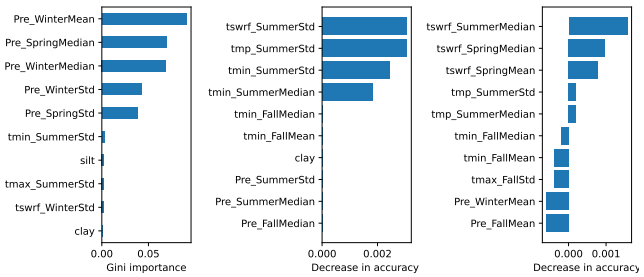

Figure 5: The five most and least important features.

biome in libya. Here we see that the radiation in the libyan climate the desert and shrub landscapes seem to differ significantly from the egyptian.

insert titles for the plots

We also created a dendogramm 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.

A little more interpretation here

The results for the error rate be seen in figure 6. Here the abbreviations 'pre', 'tmp|tmin|tmax' and 'tswrf' stand for the parameters representing precipitation, temperatures and radiation respectively. Since we have a large imbalance in the sample sizes of the biomes we look at the balanced error. For one, one sees that most modifications have very little impact. When we drop all the climate data and only train our model on the soil data it performs very badly as we previously hypothesised. Surprisingly dropping the spring, temperature or radiation data significantly improves the performance of our model. This indicates that these parameters differ significantly for deserts and shrubs in egypt and libya.

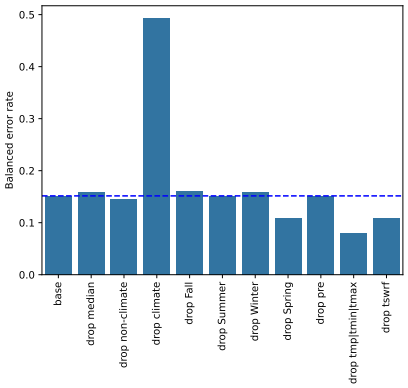

Figure 6: Balanced error rates for various experiments.

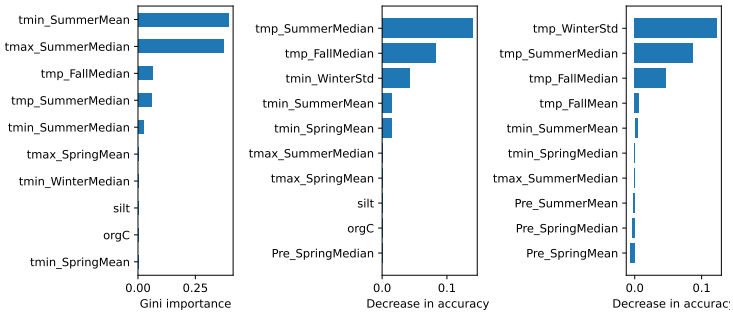

Figure 7: The five most and least important features.

Multiclass classification

For the multiclass classification the error rates were significantly higher than for the binary classification.

We start by interpreting the importance of the various features.

	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 2: Class report

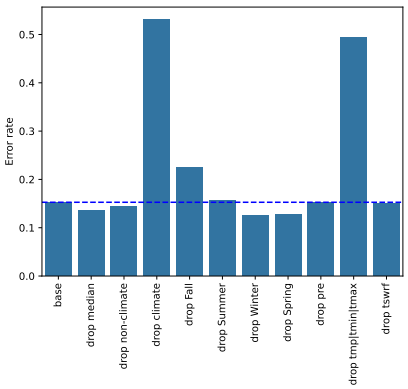

Figure 8: Error rates for various experiments.

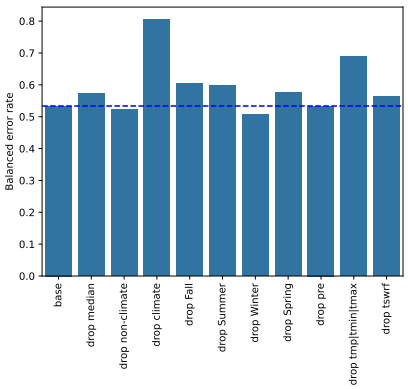

Figure 9: Balanced error rates for various experiments.

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 10. We see that VegC has a lot of values close to 0 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.

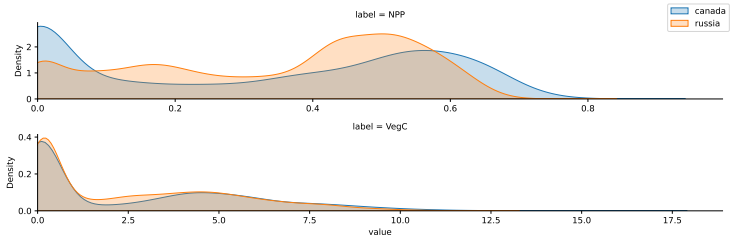

Figure 10: Distribution of NPP and VegC in Canada and Russia.

We first trained the model for NPP and tested it on russia. The results of this test can be seen in the scatterplot 13. It can also be seen in the distribution plot of the residues in figure 14.

After this we trained the model for VegC. The results of this can analogously be seen in figures 13 and 14. We see that the residual for VegC has a more pronounced spike at the origin. Comparing with 14 we see after taking into account that VegC is about an order of magnitude larger than NPP that there also seem to be more outliers in the prediction.

We will proceed in taking a closer look at the model predicting NPP. From the permutation feature importance plot for the training data in figure 15 we see that the temperature in the summer and fall and the spring precipitation were the most important factors for predicting NPP. These factors with the exception of the spring precipitation

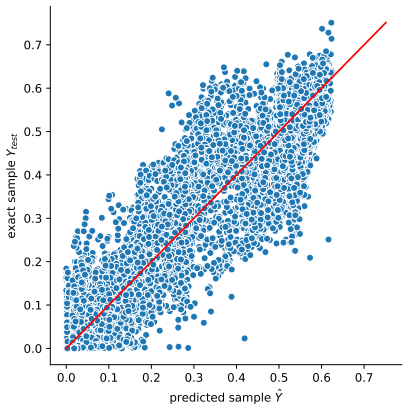

Figure 11: Predicted versus true values for the parameter NPP.

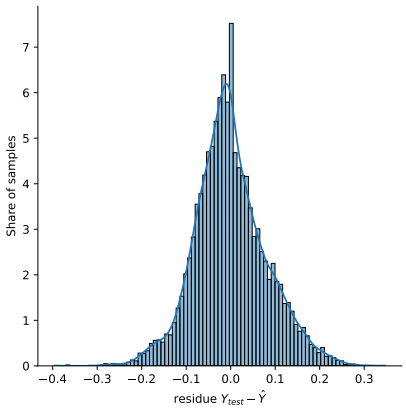

Figure 12: Distribution of the residues for the parameter NPP.

also show up as the most sensitive factors for the test set.

insert titles for the plots

In our own set of experiments the results can be seen in figure 16 where the mean square error is shown. 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. More surprising is the fact that dropping the temperatures decreases accuracy significantly. This indicates that the temperature data indeed was a good choice of predictor for the test set. That it was important we could already see in the previous analysis. Similarly, though not as pronounced the data for the fall also seemed to be a good choice of predictor. When dropping the data for the summer on the other hand the error decreased. This indicates that the choice of parameters regarding the summer was poor and that the relation between the summer data and the NPP differs between Canada and Russia.

Discussion

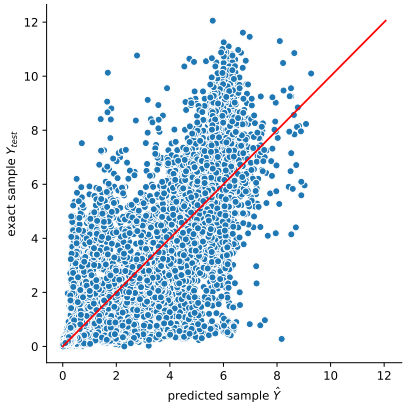

Figure 13: Predicted versus true values for the parameter VegC.

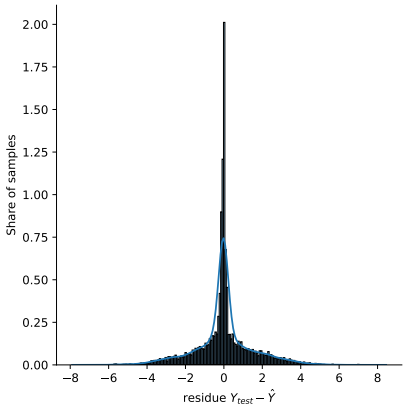

Figure 14: Distribution of the residuals for the parameter VegC.

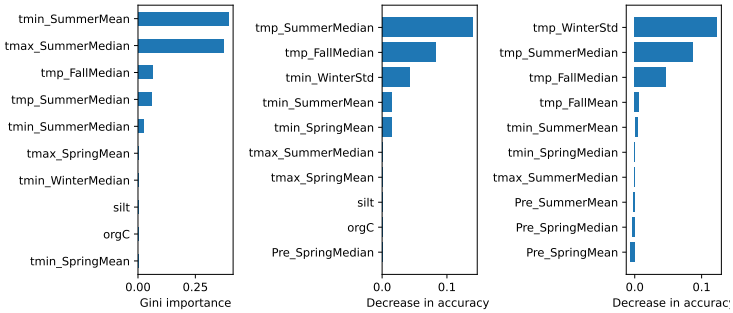

Figure 15: The five most and least important features.

Conclusion

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] *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.

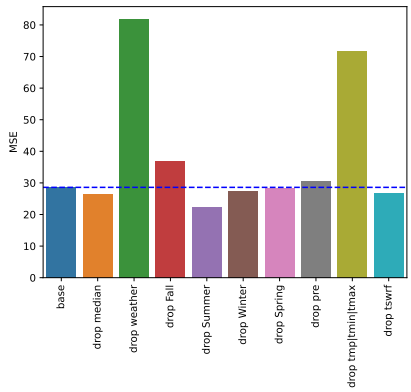

Figure 16: Mean square error for various experiments.

Appendix

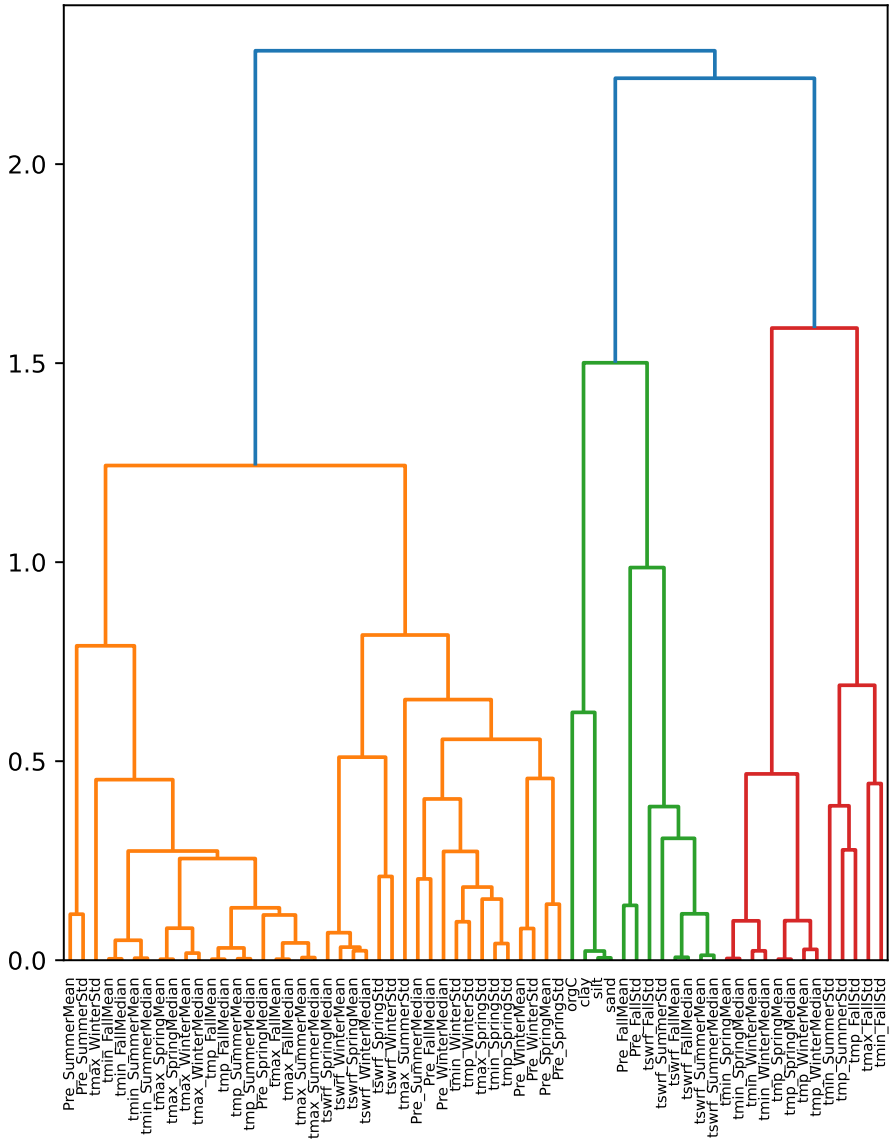

Figure 17: Dendogramm to the binary classification.