

Report for the Course Modelling in Computational Science, HT23

Project 3: Biome classification

Theo Koppenhöfer
(with Anna and Carmen, Group 4)

Lund
October 29, 2023

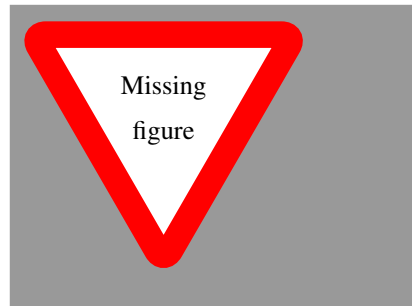


Figure 1: Number of data points with 'desert' and 'arid shrub' in selected countries.

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

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 of regions was Egypt and China. 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 the training and Libya for the testing. The variable we trained for was Biome_obs.

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 output 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 output 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`.

For the implementation we made extensive 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 results did not improve). The regression model was implemented using `RandomForestRegressor`.

To analyse 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

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

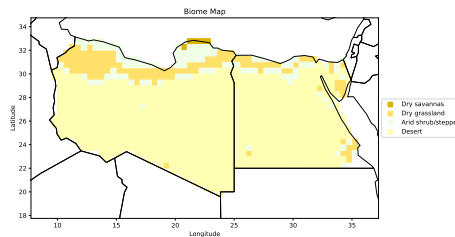


Figure 2: Biome map for Egypt and Libya.

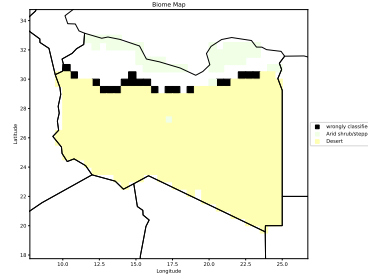


Figure 3: Map of biomes which our model classified wrongly.

In figures 4 and 5 we see a plot of the temperature, precipitation and radiation in the egyptian shrub and desert landscapes. In both climates the temperature, precipitation and radiation peak around the summer. The summer temperatures are quite similar whereas the winter temperatures differ. The variation in temperature seems to be far greater for the shrub. For the desert the winter precipitation and standard deviation is slightly higher than for the arid shrub. Similarly the radiation levels in both biomes behave almost identically though it varies more in the shrub.

more precise

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.

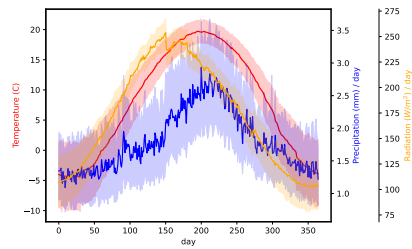


Figure 4: Mean temperature, precipitation and radiation in egyptian shrubs.

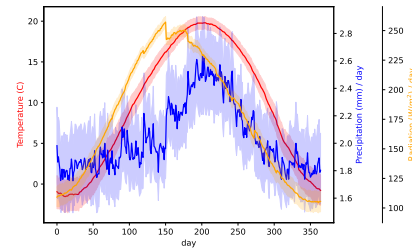


Figure 5: Mean temperature, precipitation and radiation in egyptian deserts.

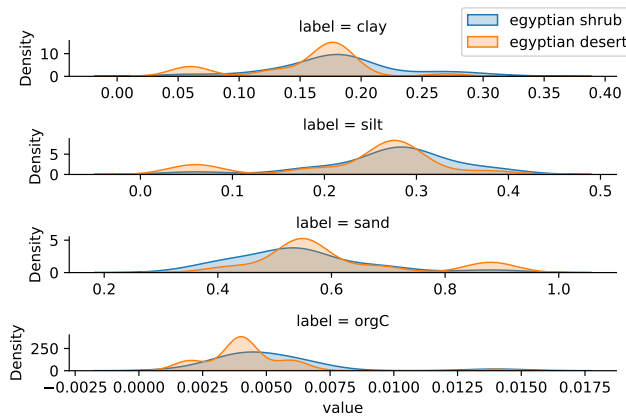


Figure 6: Distribution of the various soil features for the egyptian desert and shrub.

When we tested our trained model on Libya we got the results depicted in the 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.

Now we analyse the importance of features for the binary classification. In figure 7 we can see from the MDI (random forest feature importance) that the precipitation levels play an important role and the soil plays an insignificant role. For the permutation importance on the test set clay and some precipitation data play an insignificant and the temperature and its variation in the summer months play a large role. In the permutation performance for the test set the radiation in the summer and spring and the precipitation in the winter and fall stick out.

We also created a dendrogram 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.

The results of our routine dropping various features can be seen in figure 2. On the y-axis we depict the balanced error rate but the qualitative behaviour of the error rate is identical. The abbreviations 'pre', 'tmp|tmin|tmax' and 'tswrf' correspond to data

insert titles
for the plots

A little more
interpretation
here

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

Table 1: Confusion table.

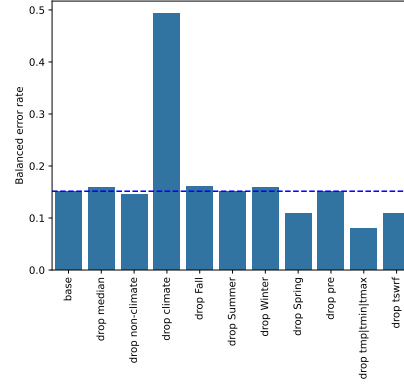


Table 2: Results for experiment series.

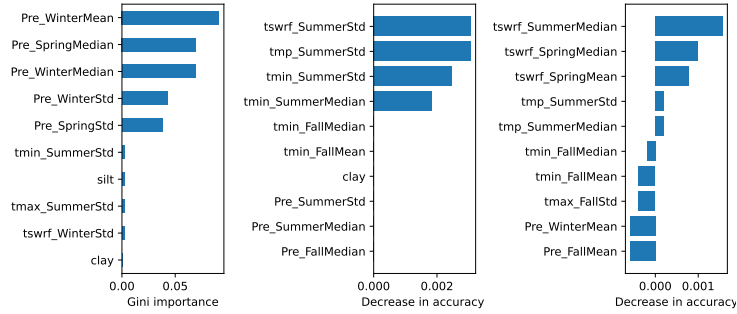


Figure 7: The five lowest and highest values for the permutation importance.

representing precipitation, temperatures and radiation respectively. For one, one sees that most modifications have little impact. When we drop all the 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 8 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. The results for the multiclass classification are shown in table 3. The accuracy of 0.85 is worse than for the binary classification but not to much worse. On the other hand the balanced accuracy of about 0.47 is significantly worse.

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

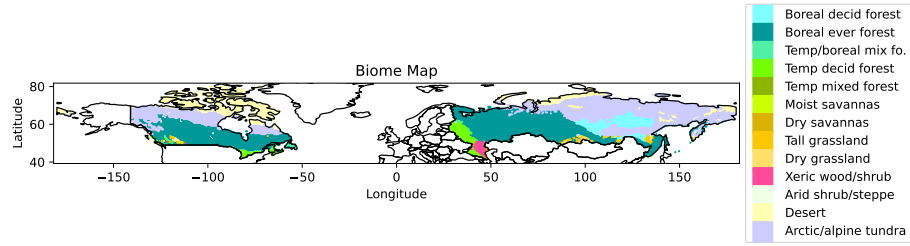


Figure 8: Overview of the biomes in Egypt and libya.

	precision	recall
macro average	0.57	0.47
weighted average	0.85	0.85

Table 3: Average precision and recall for the multiclass classification.

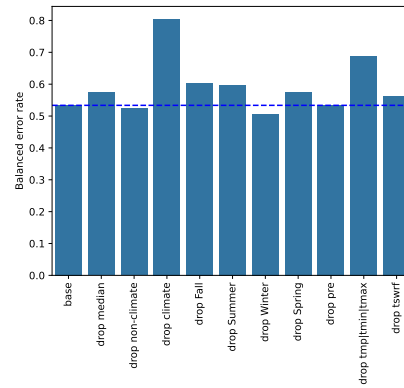


Table 4: Results for experiment series.

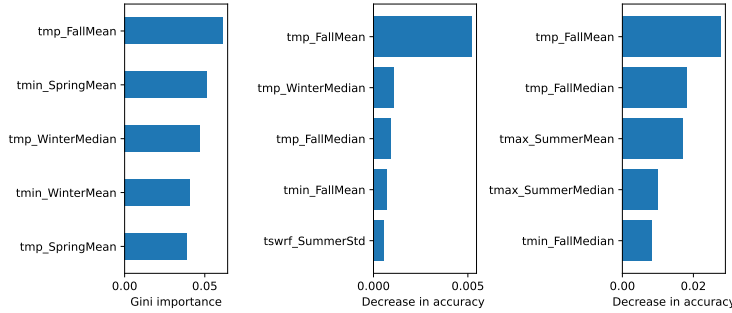


Figure 9: The five most and least important features.

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 compared the output of LPG GUESS with t

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 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.

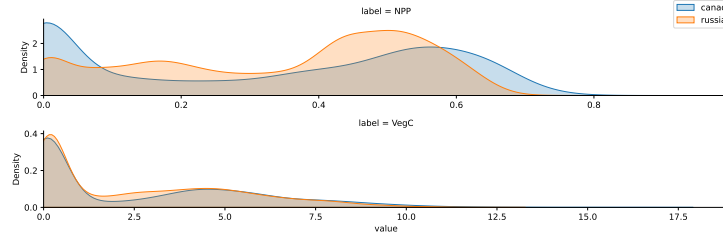


Figure 10: Distribution of NPP and VegC in Canada and Russia.

We first trained the model for NPP and tested it on Russia. The results can be seen in the scatterplot 16. It can also be seen in the distribution of the residues in figure 12.

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 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. From the permutation feature importance plot for the training data in figure 13 we see that the temperature in the summer and fall and the spring precipitation were the most important

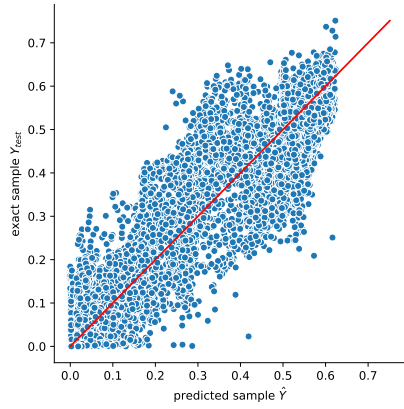


Figure 11: Predicted versus true values for the parameter NPP.

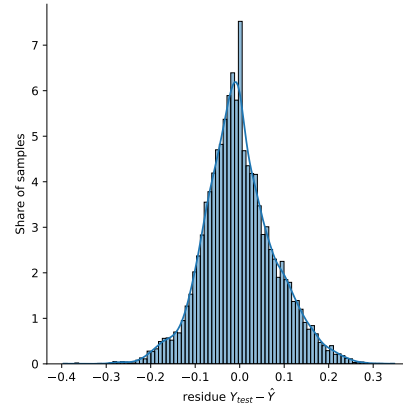


Figure 12: Distribution of the residues for the parameter NPP.

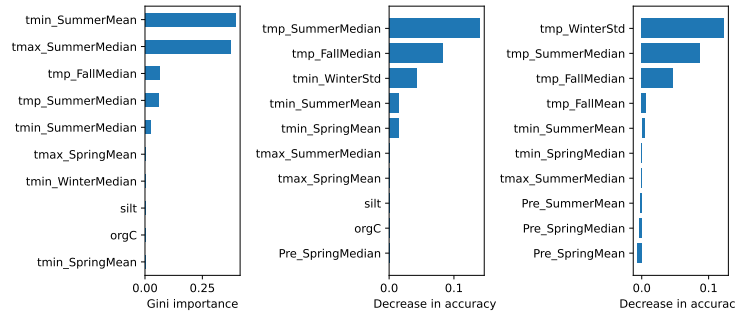


Figure 13: The five most and least important features.

factors for predicting NPP. These factors with the exception of the spring precipitation also show up as the most sensitive factors for the test set.

The results of our own set of experiments is shown in figure 14. 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.

insert titles
for the plots

Discussion

The discussion follows a similar structure to the presentation of the results.

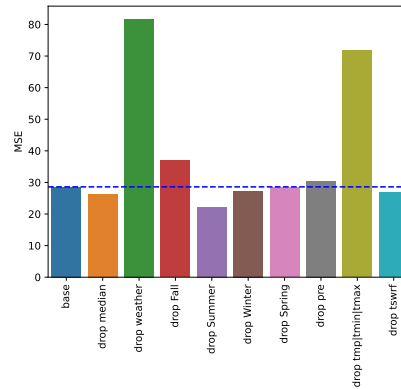


Figure 14: Results for experiment series.

In the binary classification it became apparent that the machine learning model gives features that differ between the categories a higher importance than features that are largely the same. Our results show for instance in figure 6 that the soil features in the training set do not differ much and in figure 7 that they have a negligible impact on our model. On the other hand the winter and spring precipitation and the standard deviation of the summer temperature and radiation of the training set had a big impact on the model. The analogous parameters were remarked to differ in the description of the climate diagrams 4 and 5. Although the standard deviation in the climate diagrams is between the respective biomes in Egypt, it is reasonable to assume that the standard variation in the corresponding months is correlated to the parameters `tmp_SummerStd` and `tmp_WinterStd` where the temperature plateaus.

From the figure 2 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. 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.

In the regression part we see from the permutation importance plot that our model chose the temperature to be the most important feature to determine the NPP. The plot 14 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.

mention the colinearity of parameters as reason to prefer permutation importance. Misclassified regions at the border of biomes

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 that it can only build on preexisting data and does not generalise well to different regions. 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. In this regard the random forest classifier acts a lot like an interpolator and should not be used for extrapolation. 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.

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>.

Appendix

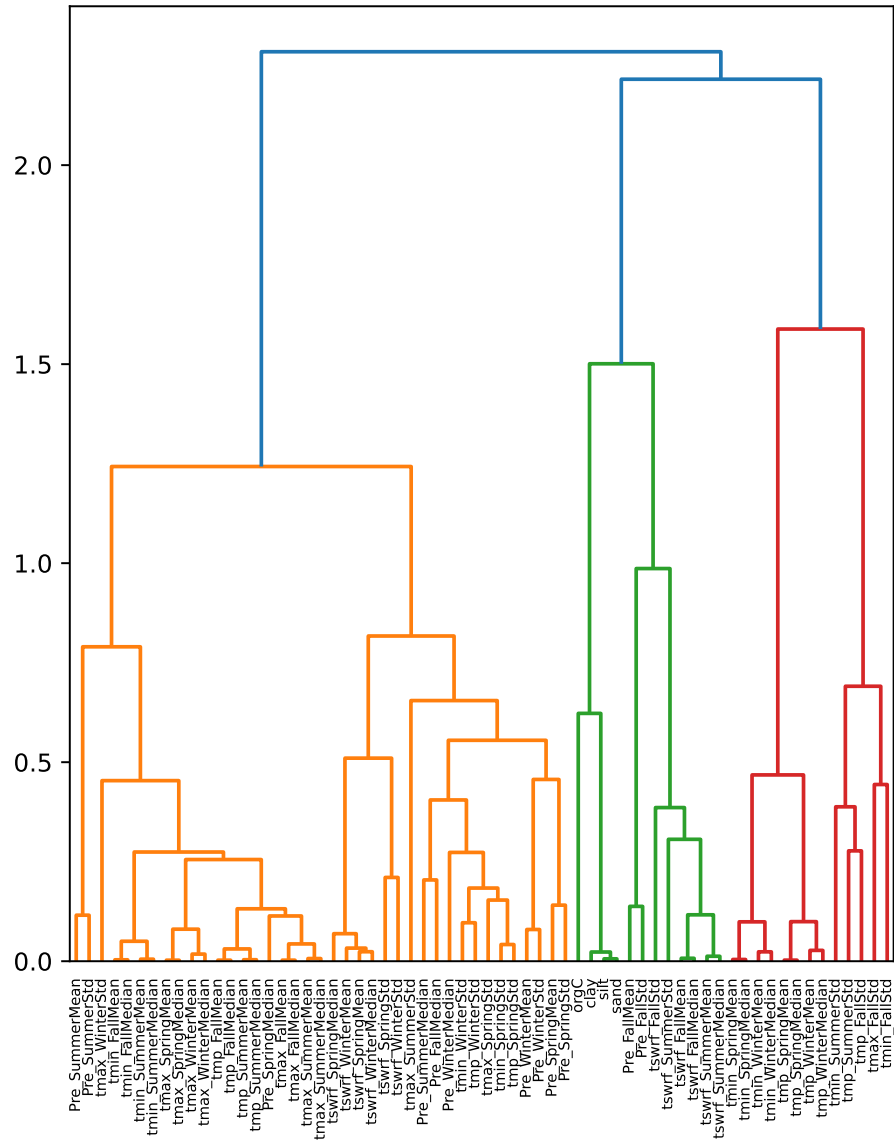


Figure 15: Dendrogram to the binary classification.

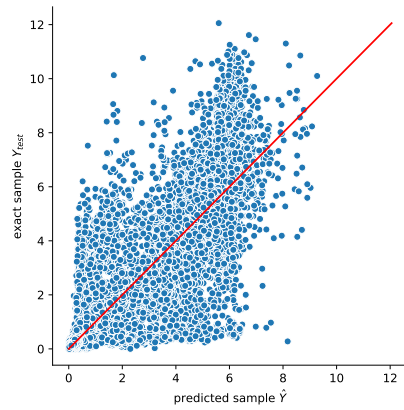


Figure 16: Predicted versus true values for the parameter VegC.

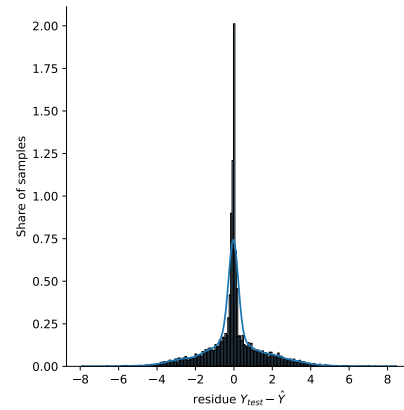


Figure 17: Distribution of the residuals for the parameter VegC.