Report for the Course Modelling in Computational Science, HT23

Project 3: Biome classification

Theo Koppenhöfer

(with Anna and Carmen, Group 4)

Lund

October 27, 2023

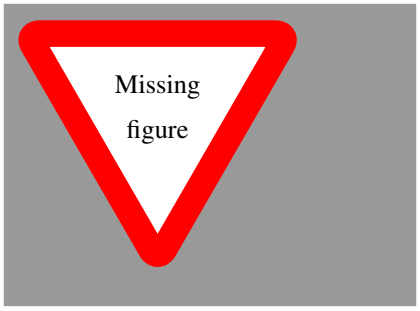

Figure 1: Number of data points with ‘desert’ and ‘arid shrub’ in selected countries.

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 `python`. 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.

In the regression part we used `pythons` `randomForestRegressor` to predict the continuous parameters

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 use as training parameters all the parameters of the file `data_index_2.csv` excluding

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

We start by giving some statistics on the desert and arid shrub landscape in egypt and libya. In figure 3 and figure 3 we see a plot of the precipitation and daily mean temperature in the egyptian shrub and desert landscapes. One can see in both climates that the most rain falls in the summer and that the mean daily temperature also peaks in the summer. It is apparent from the plots that the precipitation level is a little lower in the desert. One can also see that the standard deviation of the temperature is higher for the shrubs and that the temperature is slightly lower for the shrub biome in the winter months. From this we can expect that the mean temperature in the winter will play a greater role in the classification than the mean temperature in the summer. We can also assume that the variance of the temperature will play a greater role. Although not shown

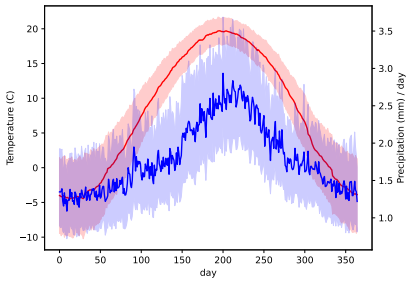

Figure 2: Average Temperature (red) and Precipitation (blue) in egyptian shrubs.

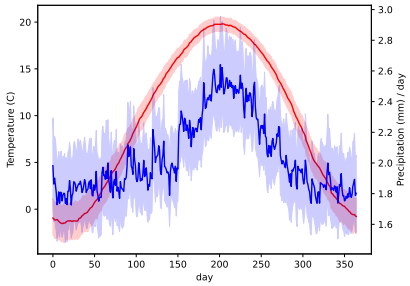

Figure 3: Average Temperature (red) and Precipitation (blue) in egyptian deserts.

here the climate plots for libya are quite similar.

some more statistics

Truth Predicted	16	17
16	46	0
17	20	443

Table 1: Confusion matrix.

We then analysed the importance of features for the binary classification. In a crude analysis we simply dropped the meadians, all weather features, the different seasons, the different climate data categories and the non-climate features. We then collected the error rates and the balanced error rates. The results for the error rate be seen in figure 4. 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 number of biomes present in the training and test data it makes more sense to look at the balanced error rate in figure 5. For one, one sees that the

This belongs into methods

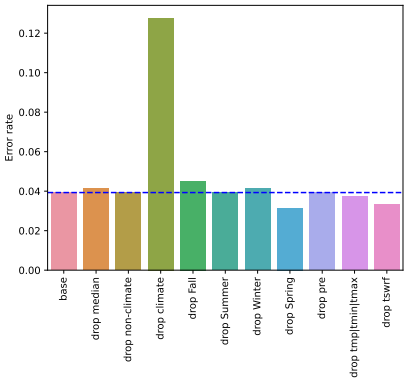

Figure 4: Error rates for various experiments.

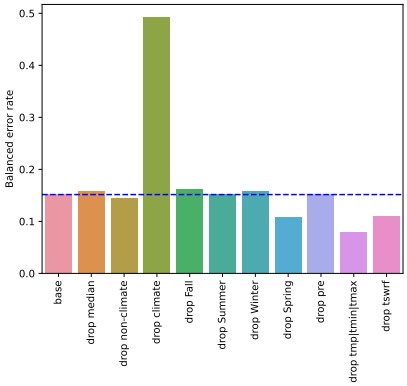

Figure 5: Balanced error rates for various experiments.

Multiclass classification

	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

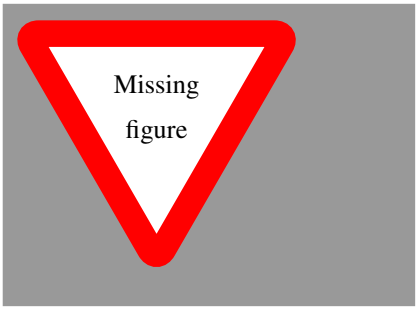

Figure 6: Error rates for various experiments.

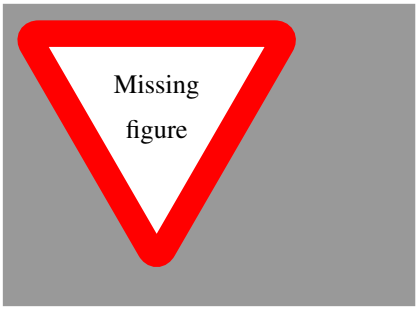

Figure 7: Balanced error rates for various experiments.

Regression

Since the

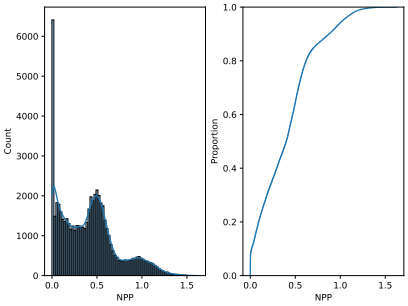

Figure 8: Distribution of the NPP values.

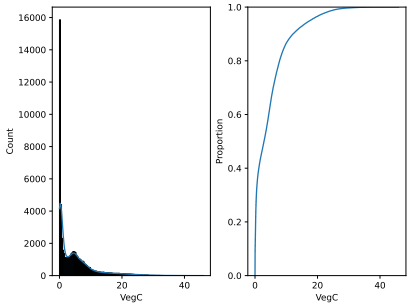

Figure 9: Distribution of the VegC values.

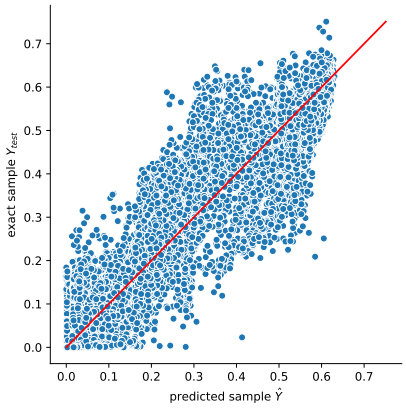

Figure 10: Predicted versus true values for the parameter NPP.

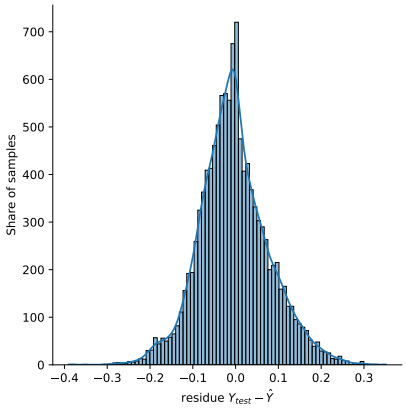

Figure 11: Distribution of the residues for the parameter NPP.

Discussion

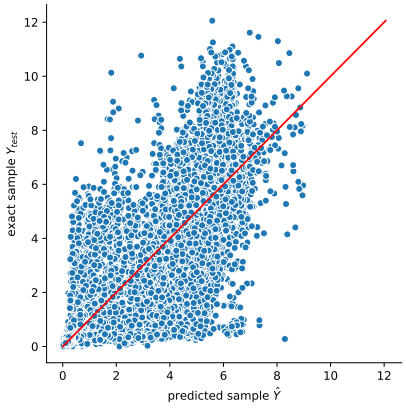

Figure 12: Predicted versus true values for the parameter VegC.

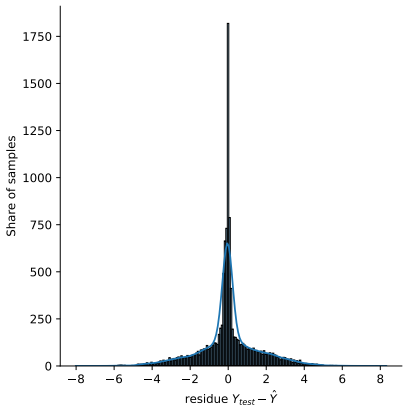

Figure 13: Distribution of the residuals for the parameter VegC.

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.

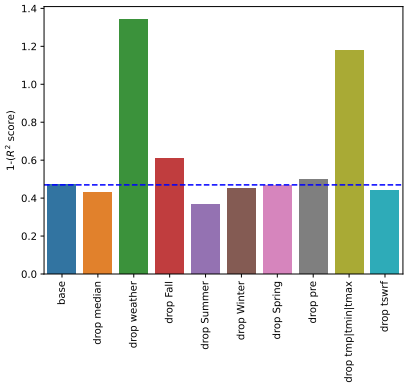

Figure 14: Error rates for various experiments.

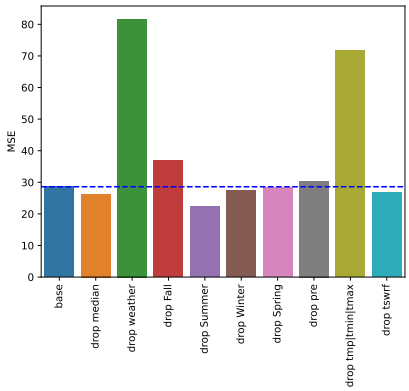

Figure 15: Balanced error rates for various experiments.

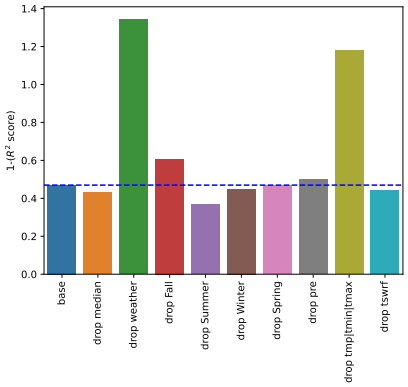

Figure 16: Error rates for various experiments.

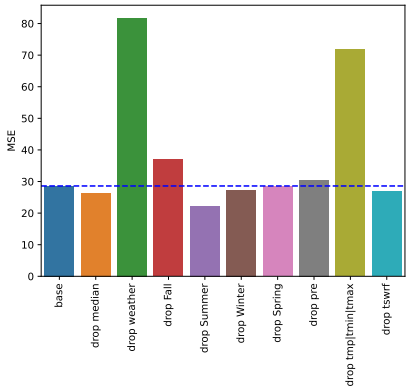

Figure 17: Balanced error rates for various experiments.