IML Term project paper

Julia Palorinne, Pyry Silomaa, Sara Sippola

03.12.2022

Introduction

In this project we trained a classifier on a data set of atmospheric measurements. The task is to predict whether new particle formation (NPF) happens or not on a given day based on the atmospheric data.

Some title that covers data analysis and classifier testing & choosing

Initial data analysis

The training data consists of several variables measured on 464 non-consecutive days. The variables are daily means and standard deviations of measurements such as carbon dioxide concentration, solar radiation and air temperature. Some of the variables are of the same phenomenon measured at different heights.

Many classifiers are affected by correlation and colinearity between variables. As expected, we found that many of the variables describing the same phenomenon are correlated, as are variables such as humidity and temperature. We take a more detailed look into this in the section *Steps we took to select good features and model parameters*.

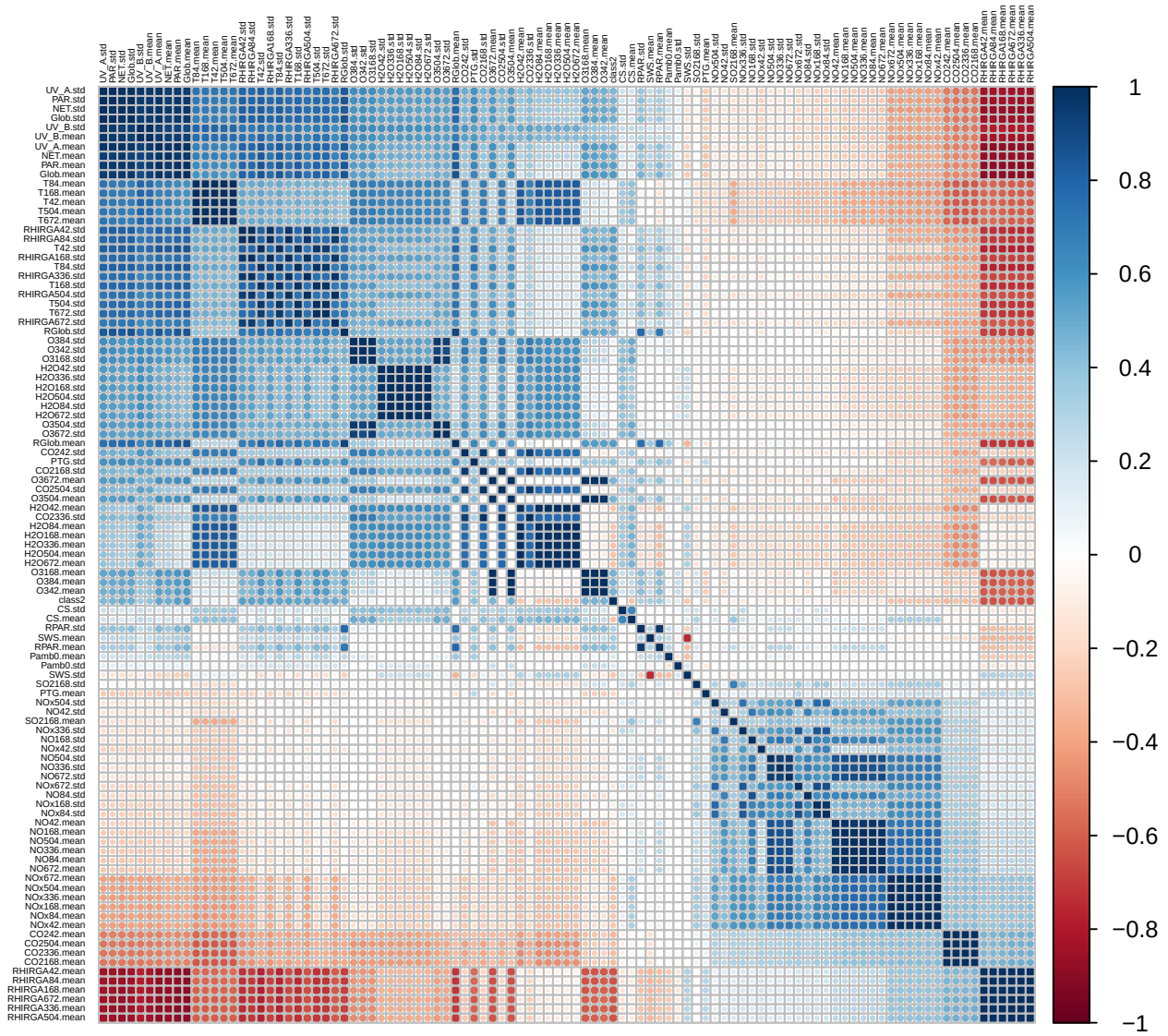
We also familiarized ourselves with the data through the smear.avaa.scs.fi webpage that offers further visualizations details about the data and the measurement site, and with the variable names and details available at wiki.helsinki.fi.

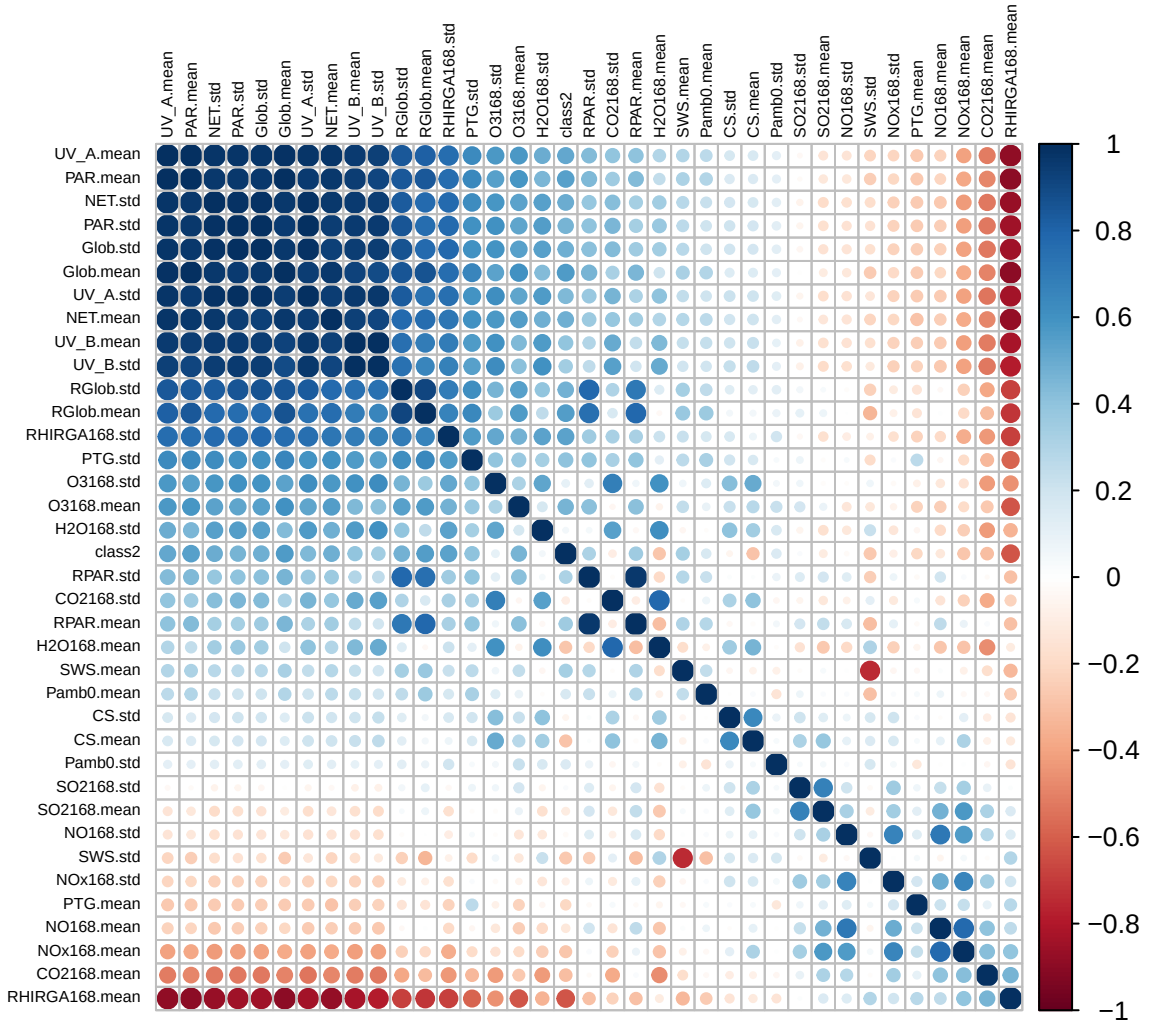
(The data is over non-consecutive days, ok, but are there any significant gaps e.g. a month?)

We found that there are 38 unique columns if we select measurements from only one height. We ended up selecting height 16.8 meter as some measurements were taken only at that height. Plots below describe the correlations in the original data and correlations after omitting all but the selected height.

Selecting columns by correlation

```
## corrplot 0.92 loaded
```


1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1



Description of considered machine learning approaches

Chosen calssifier, pros and cons of this particular classifier for this application

Steps we took to select good features and model parameters

(Should this go before the previous title? Let's see how the writing goes)

Results

Classifier performance (numerical)

Insights, conclusions, discussion etc.