IML Term project paper

Julia Palorinne, Pyry Silomaa, Sara Sippola

05.12.2022

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.

Preprocessing of the data

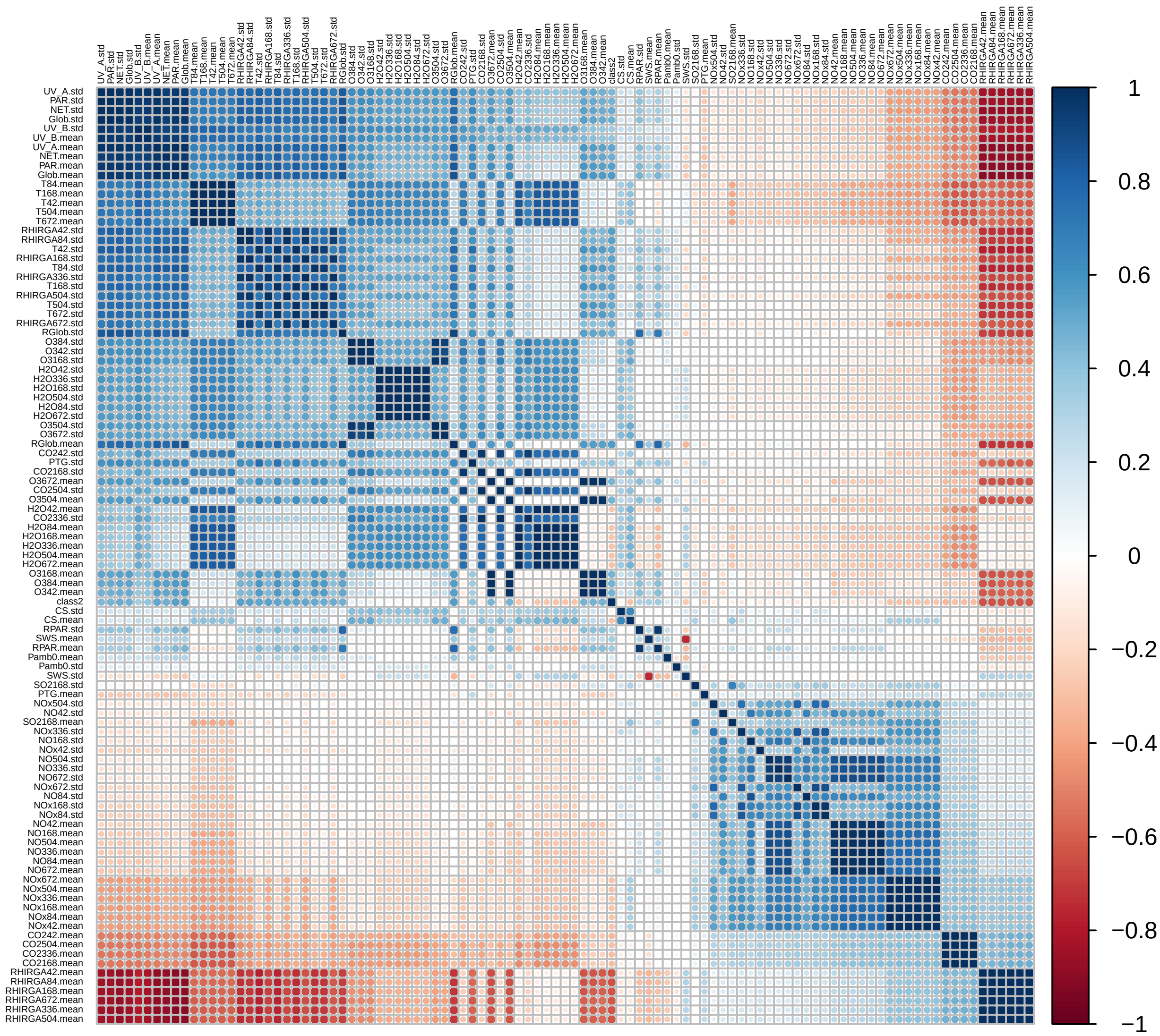
Initial data analysis

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 related to radiation (see figure below). We take a more detailed look into this in the section *Correlation between 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?)



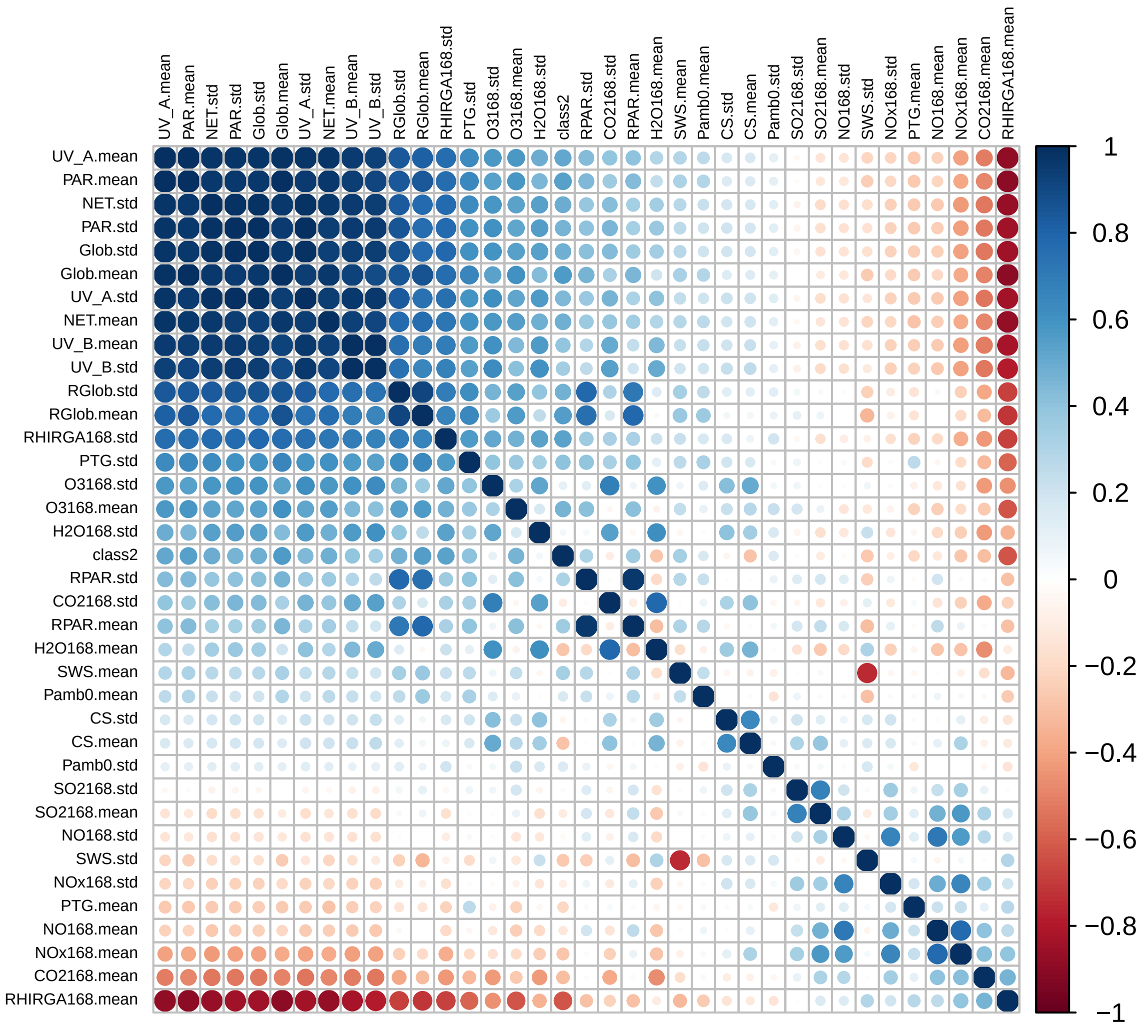
Correlation between parameters

The initial data analysis showed that many of the parameters are correlated. For many classifiers, highly correlated parameters are problematic, and for this reason we set out to drop some of the most highly correlated columns from the data set. We know that many of the parameters are measurements of the same thing at different heights. For example, water vapour concentration (H20) has been measured at heights 4.2m, 8.4m, 16.8m, 33.6m, 50.4m and 67.2m. The correlation coefficients between the daily means at different heights are essentially 1 (see table 1).

Hence we decided to discard most of the measurements of H2O and other variables for which we had multiple correlating measurements, both the means and the standard deviations. We ended up keeping the measurements measured at 16.8 meters as for some parameters, the measurements were taken only at that height. We found that after this, there are 38 unique columns. The figure below describes the correlations after omitting all but the measurements taken at the selected height.

Table 1: Correlation (H20)

	H2O168.mean	H2O336.mean	H2O42.mean	H2O504.mean	H2O672.mean	H2O84.mean
H2O168.mean	1.0000000	0.9998966	0.9997062	0.9997158	0.9994631	0.9998894
H2O336.mean	0.9998966	1.0000000	0.9993506	0.9999302	0.9997498	0.9996330
H2O42.mean	0.9997062	0.9993506	1.0000000	0.9990202	0.9986416	0.9999330
H2O504.mean	0.9997158	0.9999302	0.9990202	1.0000000	0.9998589	0.9993631
H2O672.mean	0.9994631	0.9997498	0.9986416	0.9998589	1.0000000	0.9990316
H2O84.mean	0.9998894	0.9996330	0.9999330	0.9993631	0.9990316	1.0000000



The correlation plot for the remaining parameter shows us that the remaining hihgly correlated parameters are all radiation-related.

Classifier

Description of considered machine learning approaches

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

Results

Classifier performance (numerical)

Insights, conclusions, discussion etc.