Identifying new particle formation event and nonevent days using machine learning

Jimi Hytönen, Aino Ovaska, Eino Roine

20.12.2020

Introduction

This report is part of Introduction to Machine Learning course. As the term project we created a classifier that uses atmospheric measurements to predict new particle formation event classes.

New particle formation (NPF) is a process where in suitable conditions the vapours in the atmosphere can condense and form secondary aerosol particles. The days when NPF occurs are called event days and can be classified into three different types: Ia, Ib and II. The days when NPF does not occur are called nonevent days. (Dal Maso et al., 2005)

For this project we used atmospheric measurements from SMEAR II station in Hyytiälä, Finland. The training data contained 430 daily means for 50 different atmospheric variables, their standard deviations and additionally the event class for each day (Ia, Ib, II or nonevent). We were also given test data containing atmospheric variables for another 965 days.

The main goal of this project was to develop a binary classifier that can classify the test data into event and nonevent days based on the atmospheric measurements. Additionally, we created a multiclass classifier that is able to estimate the event type for event days. The end product was the predicted event type for each day, its probability and binary accuracy.

Data

The variables in the given data sets and their explanations are listed in table 1. For each atmo-

Table 1: Variables in the dataset and their explenations

Variable name	**Description**	**Measurement heights (m)**
id	Index	
date	Date	
class4	Event types (Ia, Ib, II, nonevent)	
partlybad	Quality flag	
T	Temperature	4.2, 8.4, 16.8, 33.6, 50.4, 67.2
WS	Wind speed	4.2, 8.4, 16.8, 33.6, 50.4, 67.2
WD	Wind direction	17, 34, 50
RH	Relative humidity	4.2, 8.4, 16.8, 33.6, 50.4, 67.2
Pamb0	Ambient pressure	0
PTG	Potential temperature gradient	
SWS	Surface wetness sensor	18
UV-A	UV-A	18
UV-B	UV-B	18
Glob	Global radiation	18
RGlob	Reflected global radiation	70
PAR	Photosynthetically active radiation	18
RPAR	Reflected PAR radiation	70
NET	Net radiation	70
O3	Concentration	4.2, 8.4, 16.8, 33.6, 50.4, 67.2
SO2	Concentration	4.2, 8.4, 16.8, 33.6, 50.4, 67.2
NOx	Concentration	4.2, 8.4, 16.8, 33.6, 50.4, 67.2
NO	Concentration	4.2, 8.4, 16.8, 33.6, 50.4, 67.2
H2O	Concentration	4.2, 8.4, 16.8, 33.6, 50.4, 67.2
CO2	Concentration	4.2, 8.4, 16.8, 33.6, 50.4, 67.2
CO	Concentration	4.2, 8.4, 16.8, 33.6, 50.4, 67.2
CS	Condensation sink	

spheric variables both daily means and standard deviations were given. We were mostly interested in the "class4" event types and the different atmospheric variables which contained information of meteorological and radiation conditions, consentrations of trace gases, and condensation sink. Quite many of the variables were measured at multiple heights.

Developing the classifier

Figure 1: Correlation matrix of all the measured variables and their standard deviations where the color desrcibes the absolute value of correlation and below 0.5 the variables are interpreted not to correlate.

We approached the classification in two separate parts: selecting the optimal variables for the classification and choosing the best model. Therefore, we decided to use three different methods for feature selection and dimensionality reduction to reduce the number of variables and then try these methods on several models with 10-fold cross validation. We then chose the combination that gave best accuracy.

The three variable selection methods we chose were: variable selection by hand, principal component analysis (PCA) and recursive feature elimination (RFE) with cross-validation.

Feature selection by hand

Our first variable selection method was doing the selection by hand. After the initial look of the data, we dropped "id", "date" and "partlybad" columns and plotted correlation matrix of the variables (figure 1).

Based on these absolute correlations we noticed that measurements at different heights correlate strongly with each other. We decided to use only one measurement height for these variables. We chose 16.8 m as it is the first level above treetops but still within surface layer (Hyytiälä Database). Additionally, we decided to remove standard deviations completely as they should not affect NPF as much based on our initial trial. This method leaves us with 19 variables.

PCA

For the second variable selection method we used dimensionality reduction. When using all 100 variables and plotting the variance explained by each feature, we get figure 2, from which we choose 30 principal components, which seems to explain almost all the variation.

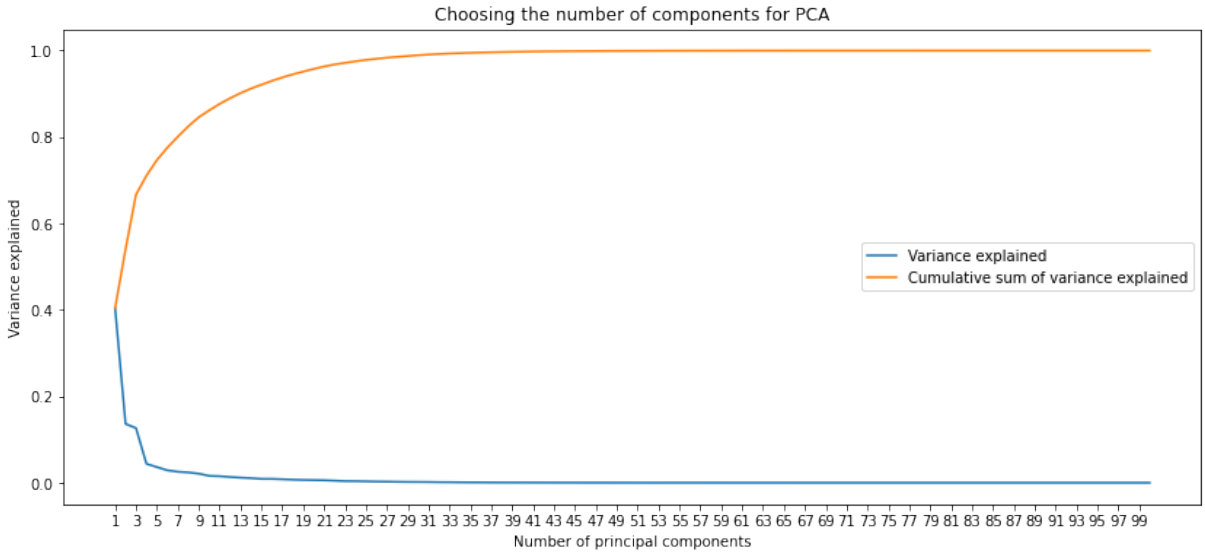

Figure 2: From PCA the variance explained by each principal component against number of principal components using all 100 variables.

We also tried combining hand selection and PCA. In that case we found that 11 principal components of the 19 left by hand selection would explain almost all the variation (figure 3).

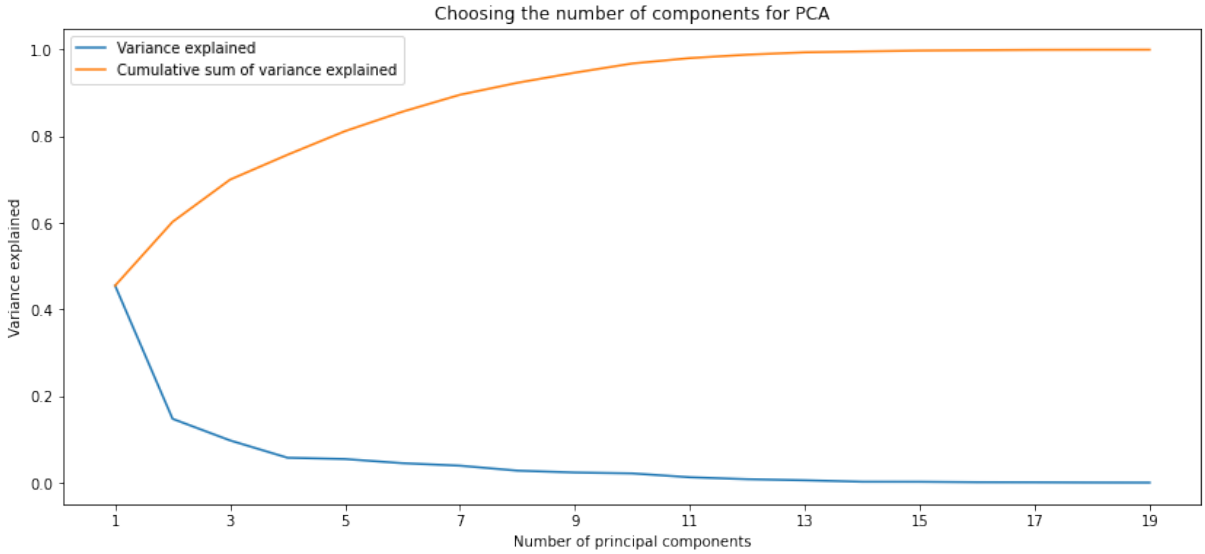

Figure 3: From PCA the variance explained by each principal component against number of principal components using 19 hand selected variables.

RFE with cross-validation

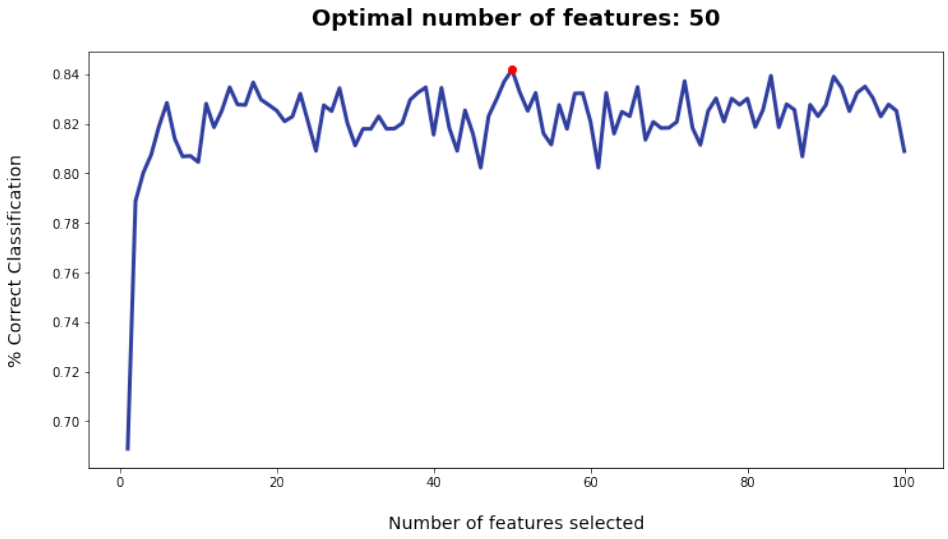

Figure 4: Optimal number of features selected with RFE based on their effect on accuracy using all 100 variables.

The final variable selection method was recursive feature elimination with cross-validation. We used the sklearn ready RFECV function with 10-fold cross validation, which is similar to backward selection. In this method inside 10-fold cross validation loop it first trains the estimator with all features, chooses the least important of them and removes it, and repeats this until leaving only wanted features. The estimator we used was DecisionTreeClassifier().

With this method the optimal number of features varies with each run but can go up to 90 variables. In figure 4 we show one example run with 50 features as optimal number. However, it should be noted that around 10 features would already give us a good estimation. This same observation can be made from figure 5 where we show what the chosen features are. Only the 24 first features have significant importance.

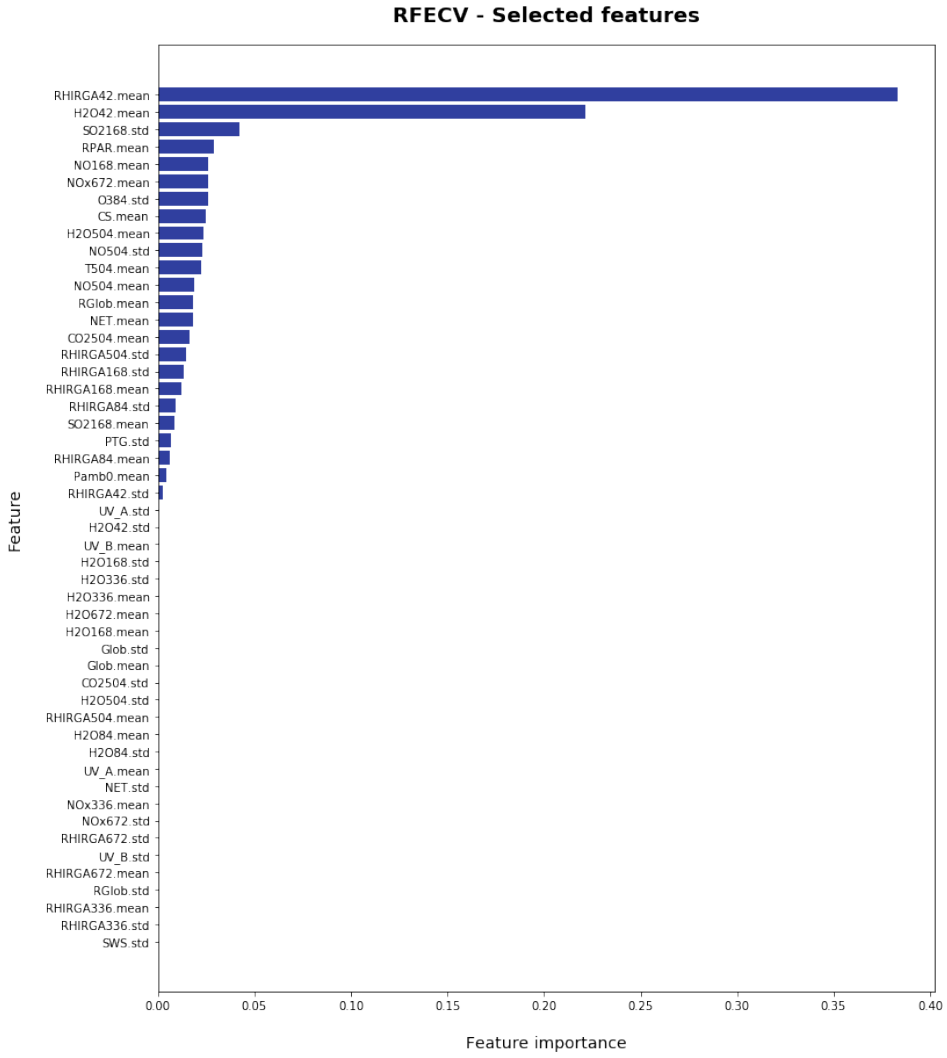

Figure 5: Features selected with RFE and their importance, where importance describes how much that feature affects the cross-validation score

Based on this we also decided to try combining hand selection and RFE. In this case RFE

method found that 14 features out of 19 hand selected features would be optimal (figure 6). The chosen features and their importances can be seen in figure 7.

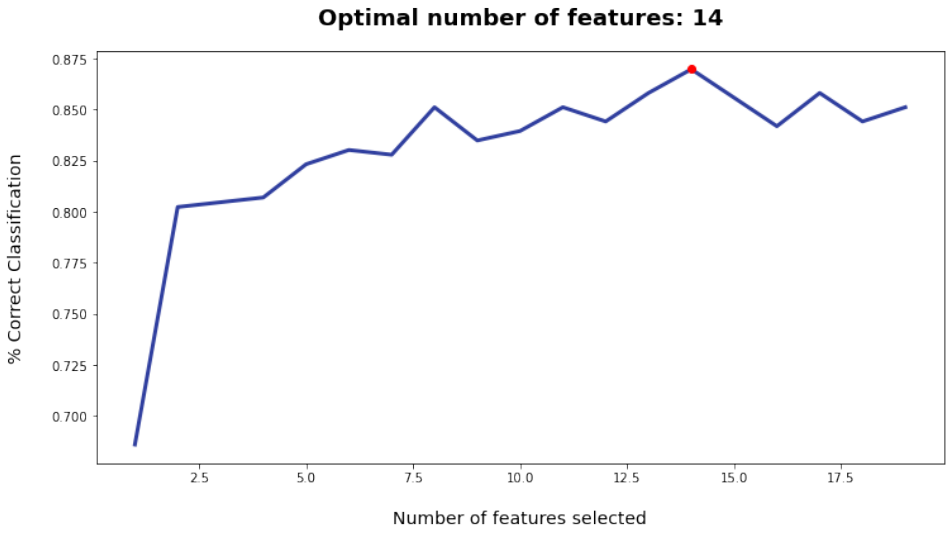

Figure 6: Optimal number of features selected with RFE based on their effect on accuracy using 19 hand selected variables.

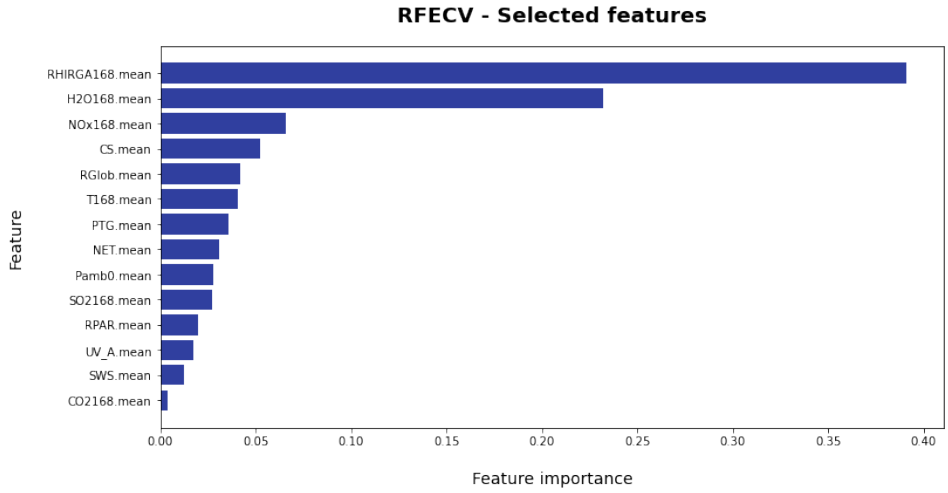

Figure 7: Features selected with RFE and their importance, where importance describes how much that feature affects the cross-validation score using 19 hand selected variables.

Choosing the model

We did the model selection by 10-fold cross validation comparing 10 different models: Logistic Regression (LR), Support Vector Classifier (SVC), Random Forest Classifier (RF), Extra Trees Classifier

(ET), Decision Tree Classifier (DT), K Nearest Neighbour (KNN), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and finally a dummy model (Dummy), which classifies everything as as the most frequent class, in this case nonevent. We run the validation six times using each time a different variable selection method and the resulting accuracies can be seen in figure 8. For figures 8 a-c we used one of the selection methods described above. In figures 8 d-e we combined hand selection to PCA and RFE. Finally figure 8f shows what the results would be without any feature selection.

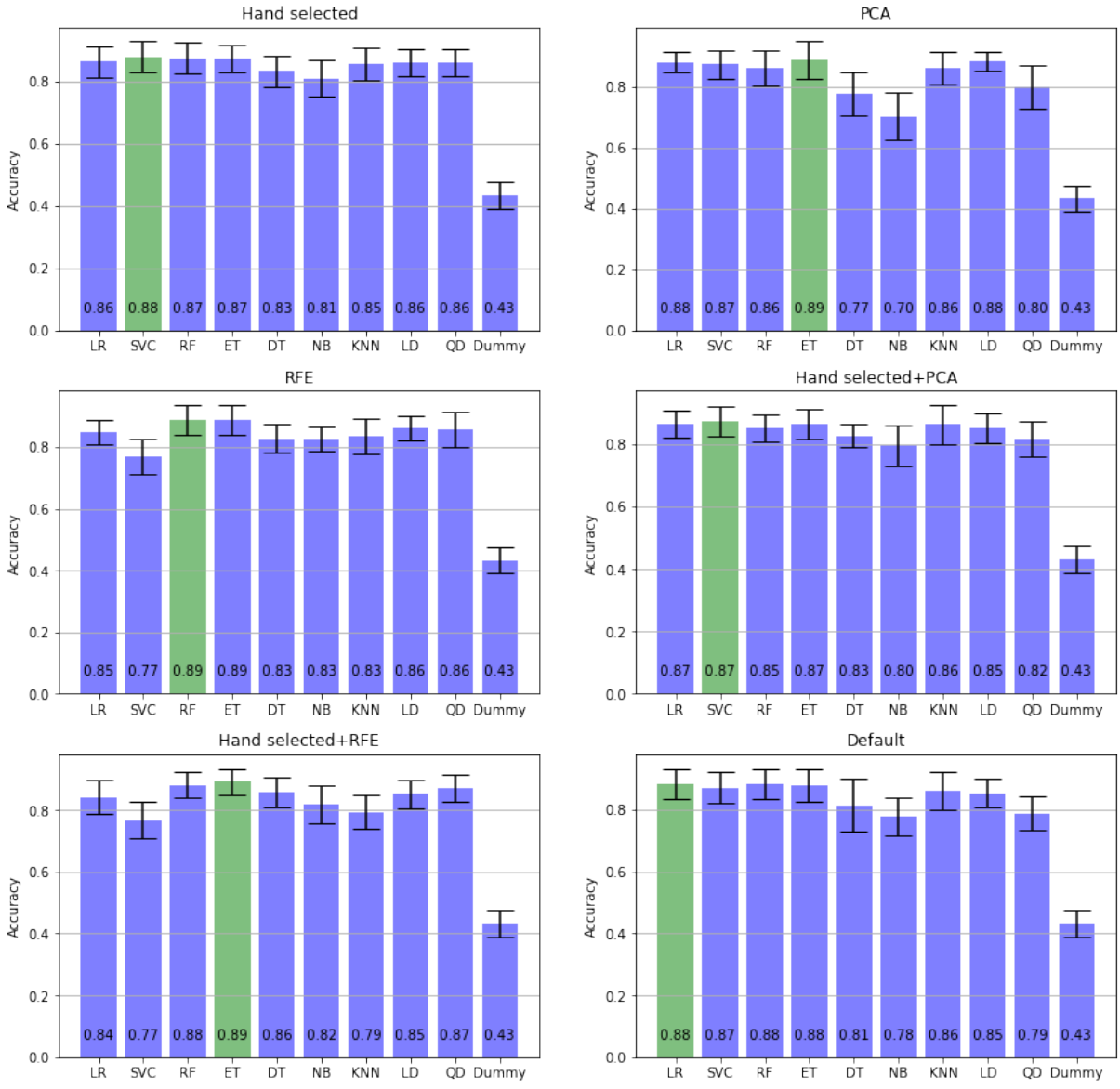

Figure 8: Results of the cross validation for each model with different variable selection methods. Error bar is one standard deviation. Green shows the best accuracy for each method.

From these results we noticed that we get the best accuracy using hand selection and RFE as the

feature selection method and Extra Trees Classifier as the model. However, several other methods have only marginally lower accuracies.

The results above describe the binary classifier. For multiclass classifier we used RFE as feature selection method. From the models Random Forest would give multiclass accuracy 0.695, Support Vector Classifier 0.688 and Linear Regression third best 0.670, but since all of these are within standard deviation, we choose to use LR for multiclasss classification.

Classification

Finally, using the chosen binary and multiclass classifiers, we predicted the event types for the test set. We also calculated the probability for each day being an event day by summing together the Ia, Ib and II probabilities outputted by the classifier.

Results

For the binary classifier our test accuracy is 0.88 and for multiclass classifier 0.66.

In the case of binary classifier our chosen features, from most important to least important, were relative humidity, H_2O concentration, NO_X concentration, reflected global radiation, temperature, potential temperature gradient, net radiation, atmospheric pressure, SO_2 concentration, reflected photosynthetically active radiation, UV-A radiation, surface wetness sensor and CO2 concentration.

Conclusions

In this project we created a binary classifier that can predict if a day is event or nonevent day based on atmospheric variables. Additionally, we created a multiclass classifier that is able to predict the event type occurring during that day.

When creating the classifiers, we compared three different feature selection methods and 10 different models. For binary classifier the best result was achieved when we first removed by hand all standard deviations and the duplicate measurements from different heights using only measurements from height 16.8 m. This reduced the number of atmospheric variables from 100 to 19. After this we did further feature selection with recursive feature elimination. The binary classifier used these 14 features and Extra Tree Classifier to predict event and nonevent days. Our multiclass classifier worked similarly but used only recursive feature elimination in feature selection and Linear Regression as model. However, it should be noted that most of the mean accuracies given by the models did not have statistically significant difference from each other. The only exception was dummy model, which had clearly worse performance than any of our models. Moreover, even without any feature selection (figure 8f) the models performed surprisingly well for training data.

For binary classifier our final accuracy was 0.88 and for multiclass classifier 0.66. With some additional work it might be possible to enhance the performance of the multiclass classifier. However, as the different event types are not necessarily clearly defined and can be difficult to distinguish (Joutsensaari et al., 2018), it is challenging to make excellent multiclass classifier.

One of the benefits of using RFE, instead for example PCA, is that we know what the selected features (figure 5) are and we can estimate their physical significance. Hyvönen et al., 2005, found in their similar study, that condensation sink and relative humidity would be the most important features for determining event and nonevent days from each other as they both restrict nucleation and particle growth, which are both important in new particle formation. We also found these to be important, though condensation sink is a bit lower on the list. In addition, we found temperature, water content and radiation variables to be important features, which agrees with Boy and Kulmala, 2002. In addition, compared to only using RFE, combining RFE and hand selection seems to give physically more reasonable features. When using only RFE we start to see duplicates of same measurements at different heights and the number of features needed to achieve same accuracy is much higher (figure 5), which implies that it is reasonable to remove duplicate variables and standard deviations. However, for multiclass classifier standard deviations might be more important, as conditions that change during day might interrupt the NPF event and cause different type of event to occur (Joutsensaari et al., 2018). Therefore, we did not use hand selection there.

In summary, both our binary and multiclass accuracies are quite good. Furthermore, the features our binary classifier selects seem to agree with other research done in this field.

Group Work

Our group work went well. We had very clear roles and everyone took actively part in the project. Our common goal was to make a good classifier, in which we succeeded. Additionally, we all chose the part of the group work we wanted to work on and set our own goals for that spesific part. Several times during the process we discussed each other's work and gave feedback and ideas what to do next. We drew ideas from exercises and other previous experience and shared it with each other. This group work let us use what we had learnt in this course and search more information about the topics and apply it to a project that simulated real life work project. Our different methods for approaching this project also gave further insight into the topics from this course.

References

Boy, M. and Kulmala, K.: Nucleation events in the continental boundary layer: Influence ofphysical and meteorological parameters, Atmos. Chem. Phys., 2, 1–16, 2002,SRef-ID: 1680-7324/acp/2002-2-1.7579,7594

Dal Maso, M., Kulmala, M., Riipinen, I., Wagner, R., Hussein, T., Aalto, P. P. & Lehtinen, K. E. J.: Formation and growth of fresh atmospheric aerosols: eight years of aerosol size distribution data from SMEAR II, Hyytiälä, Finland. Boreal Env. Res. 10: 323–336, 2005

Hyvönen,S., Junninen, H., Laakso, L., Dal Maso, M., Grönholm, T.,et al.. A look at aerosol formationusing data mining techniques. Atmospheric Chemistry and Physics Discussions, European GeosciencesUnion, 5 (4), pp.7577-7611, 2005, hal-00301731

Hyytiälä Database, 2018, https://wiki.helsinki.fi/pages/viewpage.action?pageId=243959901, visited 19.12.2020.

Joutsensaari, J., Ozon, M., Nieminen, T., Mikkonen, S., Lähivaara, T., Decesari, S., Facchini, M., Laaksonen, A., Lehtinen, K.,. Identification of new particle formation events with deep learning. Atmospheric Chemistry and Physics, 18 (13) , 9597-9615. 2018, 10.5194/acp-18-9597-2018.