Team Hyvärinen
Linda Hyvärinen
Eero Rantala
Hilma Holappa

Term project initial report

Preprocessing:

- Loading the data
- Adding new column class2: event/nonevent
- Using model answers correlations matrix to choose with features to look with boxplot
- Take a look at the data with box plots for chosen columns/features
- There seems to be few features that have more promise with classifying events/nonevents
    - RHIRGAxxx.mean
    - PAR.mean
    - UV_A.mean
    - NET.mean

Classifiers we considered:

- Regression Tree
- Random Forest
    - We didn't fully understand this method yet, so we left it out for now
- Naive Bayes
    - Not so good
- Dummy Classifier
    - Surprisingly good

Features included in the classifier:

- We chose to try naive Bayes with RHIRGA168.mean, PAR.mean, UV_A.mean, NET.mean. This gave us about 61% correct rate.
- Dummy classifyer with just PAR.mean. We found one value (397) that divided the data between event and nonevent quite well. Predicting event/nonevent based on PAR.mean value being over or under fixed value gave us about 75% correct.

Summary of what we learned so far:

- Naive Bayes seems to perform worse than dummy classifier, which is surprising in a way.
- Random forest seems very promising when looking at the confusion matrix, but we still have some work to do with understanding the logic.
- When the preprocessing is done well, the rest of the work is easier.
- Getting to know the data with eg box plots is helpful when choosing the features to be included in the classifier.
-