The task for the MRL lab

The dataset comes from here:

https://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Study&DataMode=AllData&StudyID=ST000007&StudyType=MS&ResultType=1#DataTabs

The data come from a metabolomics study with LC/MS. The study is of fingerprinting type. Some of the important metabolites are identified but some of them are known only with the exact mass.

Study summary from the report:

Bacterial leaf blight (BLB), caused by Xanthomonas oryzae pv. oryzae (Xoo), gives rise to devastating crop losses in rice. Disease resistant rice cultivars are the most economical way to combat the disease. The TP309 cultivar is susceptible to infection by Xoo strain PXO99. A transgenic variety, TP309_Xa21, expresses the pattern recognition receptor Xa21, and is resistant. PXO99?raxST, a strain lacking the raxST gene, is able to overcome Xa21-mediated immunity. We used a single extraction solvent to demonstrate comprehensive metabolomics and transcriptomics profiling under sample limited conditions, and analyze the molecular responses of two rice lines challenged with either PXO99 or PXO99?raxST. LCTOF raw data file filtering resulted in better within group reproducibility of replicate samples for statistical analyses. Accurate mass match compound identification with molecular formula generation (MFG) ranking of 355 masses was achieved with the METLIN database. GCTOF analysis yielded an additional 441 compounds after BinBase database processing, of which 154 were structurally identified by retention index/MS library matching. Multivariate statistics revealed that the susceptible and resistant genotypes possess distinct profiles. Although few mRNA and metabolite differences were detected in PXO99 challenged TP309 compared to mock, many differential changes occurred in the Xa21-mediated response to PXO99 and PXO99?raxST. Acetophenone, xanthophylls, fatty acids, alkaloids, glutathione, carbohydrate and lipid biosynthetic pathways were affected. Significant transcriptional induction of several pathogenesis related genes in Xa21 challenged strains, as well as differential changes to GAD, PAL, ICL1 and Glutathione-S-transferase transcripts indicated limited correlation with metabolite changes under single time point global profiling conditions.

In this lab, you have a complicated data pre-processing task. The data and metadata are given in two different files and the data file does not contain information about the classes of the classification parameter `Xoo infection`. Also, both files contain sample id but their names are different. Additionally, pay attention to what are the features and what are the group related parameters (rows vs columns). NB! We always want to have features as columns and group variables/samples as rows.

Aim of the lab: find if the metabolites profile can be used to track back if TP309_Xa21 was infected with PXO99 or RaxST.

Compare logistic regression, KNN, LDA, decision trees and random forests for this task.

Read package `caret` to see which package can be used in collaboration of cross-validation. NB! One of the models does have variable selection model available and two do not need variable selection.

Explain the results.

Submit your R code as "*name.R*" file. Kepe your code commented. All comments are to be written into the code. Use

```
#comment
```

for this.