Building meaningful machine learning models for disease prediction

Dr Shirin Glander

Dep. of Genetic Epidemiology
Institute of Human Genetics
University of Münster

shirin.glander@wwu.de

https://shiring.github.io
https://github.com/ShirinG

Friday, 31st March 2017

About me

since 2015	Bioinformatics Postdoc Next Generation Sequencing autoinflammatory diseases & innate immunity
2011 - 2015	PhD in Biology Is the immune system of plants required to adapt to flowering time change?
2005 - 2011	BSc and MSc of Science in Biology evolutionary genetics, immune memory in Drosophila

Table of contents

Building meaningful machine learning models for disease prediction

1. What makes a model meaningful?
2. Machine Learning (ML) in disease modeling
3. A quick recap of ML basics
4. How to build ML models in R
5. Evaluating model performance

What makes a model meaningful?

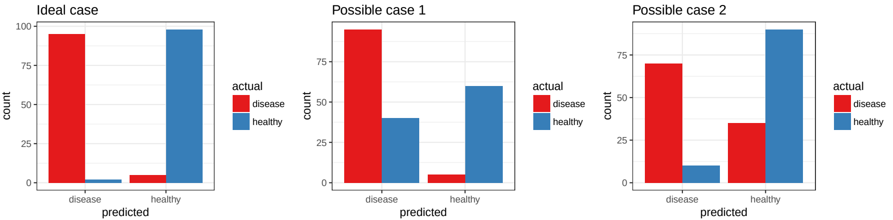
What makes a model meaningful?

- creating ML models is relatively easy
- creating good or meaningful models is hard

Meaningful models

- are generalizable
- answer the question(s) posed...
- ... with sufficient accuracy to be trustworthy

Accuracy depends on the problem!



Machine Learning (ML) in disease modeling

ML in disease modeling

- tools that can interpret "big medical data"
- and provide fast, accurate and actionable information
- for precision or personalized medicine

Examples:

- computer-aided diagnosis of breast cancer from mammograms[1]
- identifying gene defects with facial recognition software[2]
- identifying signatures of Brain Cancer from MRSI[3]
- ... and many more ...

Image source: Wikimedia Commons

[1] Doi 2007.

[2] Levenson 2014.

[3] Sadja 2006.

Can we trust a model?

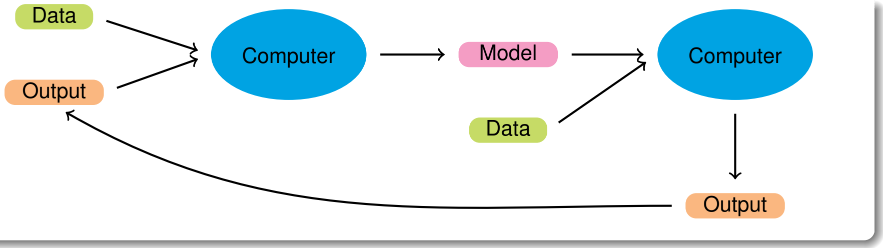
- most ML algorithms model high-degree interactions between variables
- we often don’t know WHY ML models make decisions
- inherent problem with ML models: they are hard (or impossible) to interpret
- therefore, it is crucial that our models are meaningful

Image source: Pixabay

A quick recap of ML basics

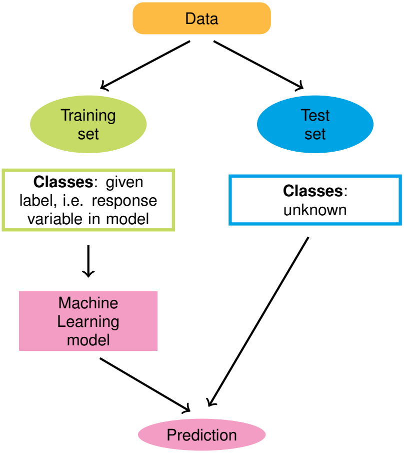
Machine learning

- artificial intelligence (AI)
- data-driven
- algorithms learn by being trained on observed data...
- ... and predict unknown data
- ML concepts are not new, but the increase in computational capacity has made them more accessible



Supervised vs Unsupervised algorithms

Supervised



Data

Training set

Test set

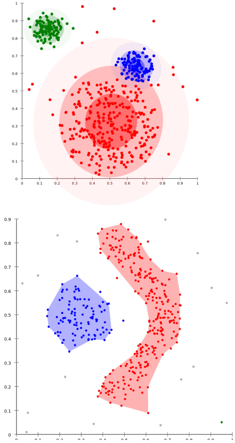
Classes: given label, i.e. response variable in model

Classes: unknown

Machine Learning model

Prediction

Unsupervised

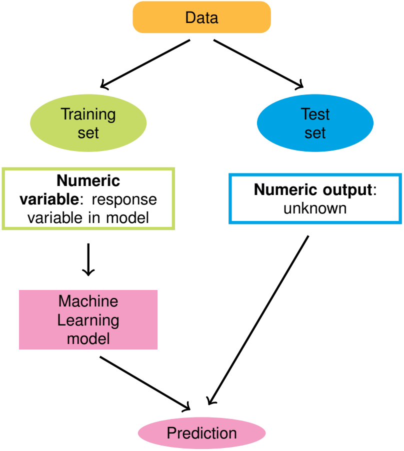


Wikipedia

Classification vs Regression

Regression

e.g. weight loss



Data

Training set

Test set

Numeric variable: response variable in model

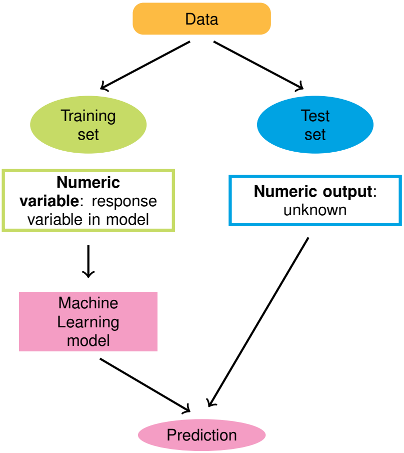
Numeric output: unknown

Machine Learning model

Prediction

Classification vs Regression

Regression
e.g. weight loss



Data

Training set

Test set

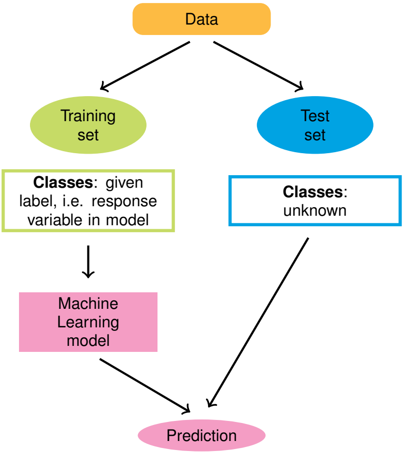
Numeric variable: response variable in model

Numeric output: unknown

Machine Learning model

Prediction

Classification
e.g. healthy vs disease



Data

Training set

Test set

Classes: given label, i.e. response variable in model

Classes: unknown

Machine Learning model

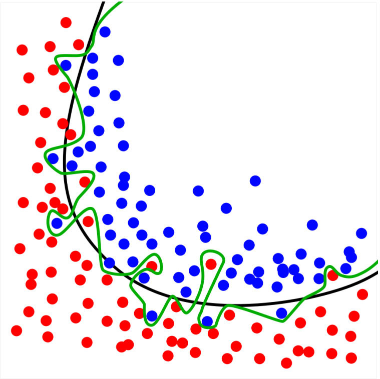
Prediction

Features

- Features are the variables used for model training.
- Using the right features is crucial.

- More is not necessarily better (overfitting)!

- feature selection
- feature extraction/ engineering



Wikipedia

Training, (cross-) validation and test data

Hyper Parameter Tuning

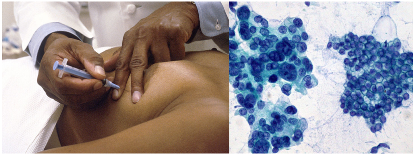
Take home messages:

- ...

How to build ML models in R

Session setup

- Breast Cancer Wisconsin Dataset[4]



- caret[5]
- h2o[6]

Code will be available on my website and on Github

[4] W. H. Wolberg and O. L. Mangasarian (1990). "Multisurface method of pattern separation for medical diagnosis applied to breast cytology." In: *Proceedings of the National Academy of Sciences* 87.23, pp. 9193–9196.

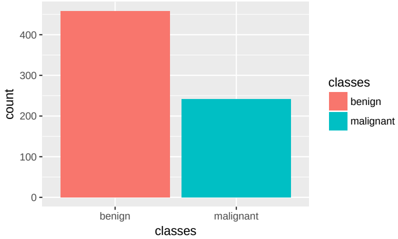
[5] M. Kuhn et al. (2016). *caret: Classification and Regression Training*. R package version 6.0-71.

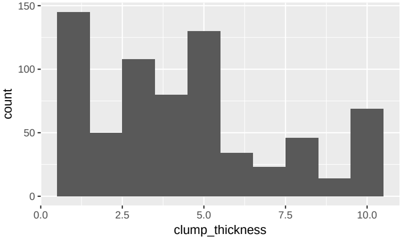
[6] H2O.ai (2017). *h2o: R Interface for H2O*. . R package version 3.10.3.6.

Get to know your data

Response variable

- Is it balanced?



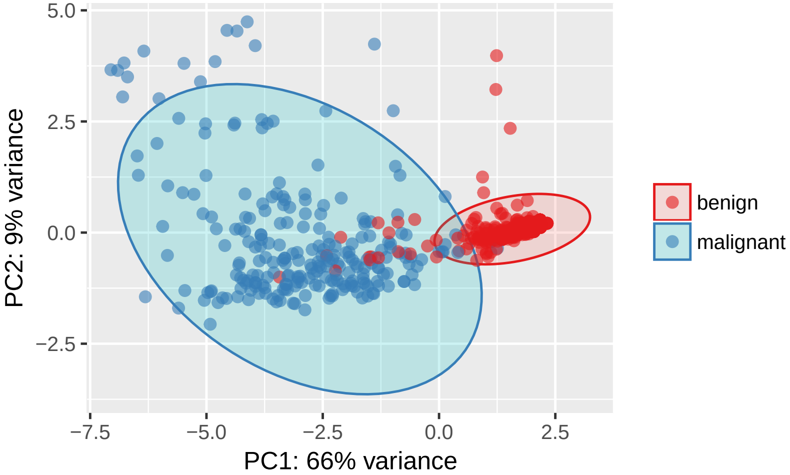


Missing data

- Is there missing data?
- Can we afford to loose data points with missing values?
- Or do we use imputation (and introduce additional uncertainty)?

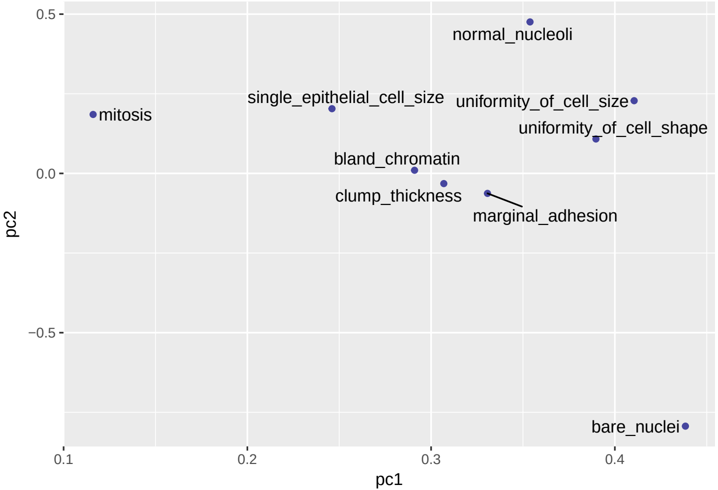
Get to know your data

Principal Component Analysis (PCA)



Get to know your data

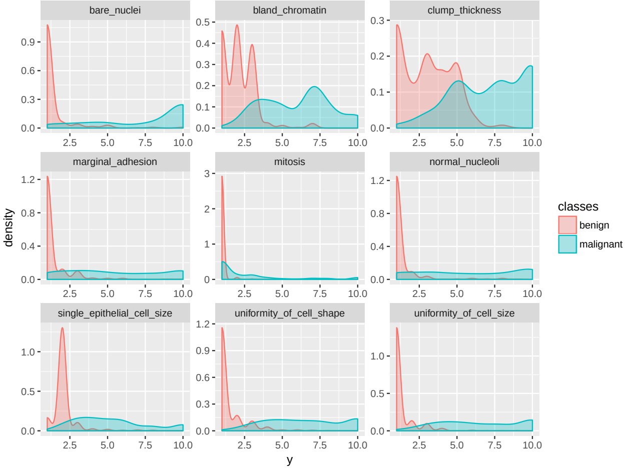
Principal Component Analysis (PCA)



Get to know your data

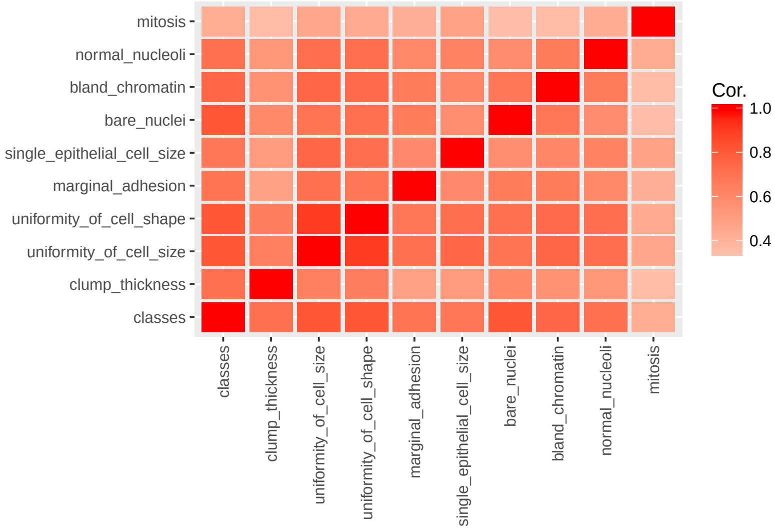
Features

- factors or numeric
- pre-processing



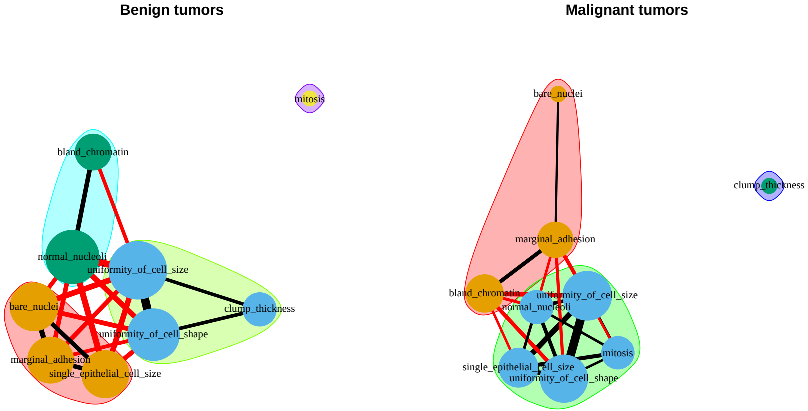
Get to know your data

Correlation



Get to know your data

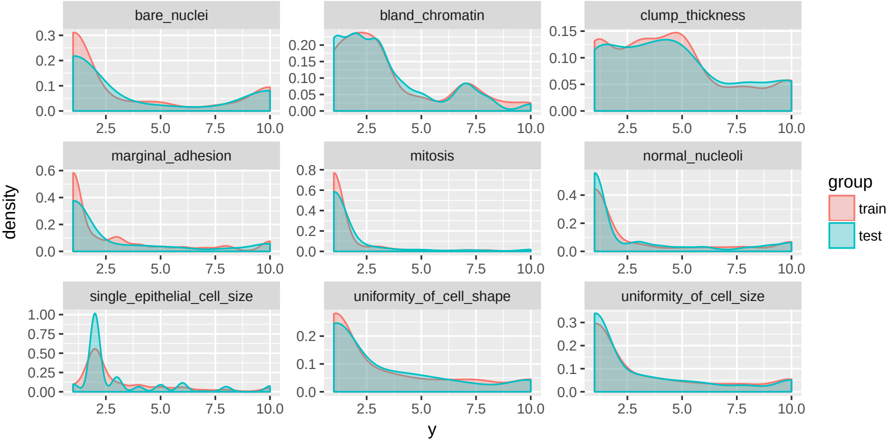
Correlation graphs



Training, validation and test data

We need to split the data into training and test sets - ideally stratified by response class.

Density distribution



Model examples

Regression with Linear Models

- e.g. Generalized Linear Models
- example with *caret*

Tree-based classification

- Random Forest or Gradient boosting trees
- with *caret*

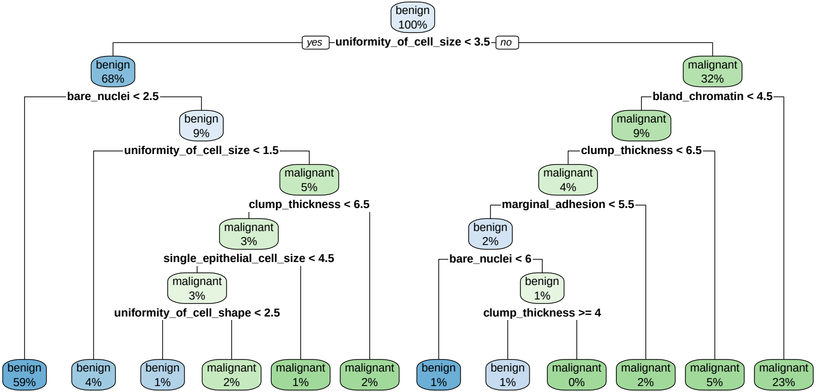
Hyper-parameter tuning

- Grid Search
- with *h2o*

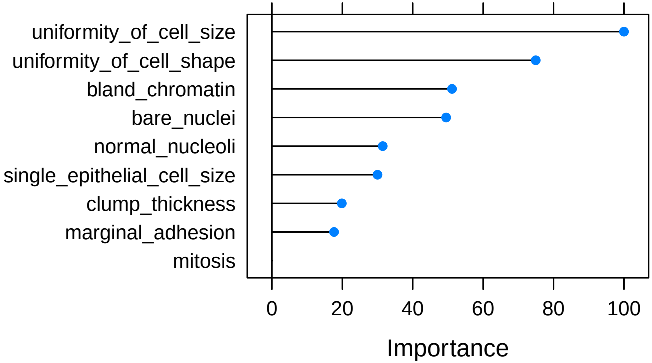
Classification with tree-based models

Decision trees

e.g. Random Forest and gradient boosting trees



Feature importance



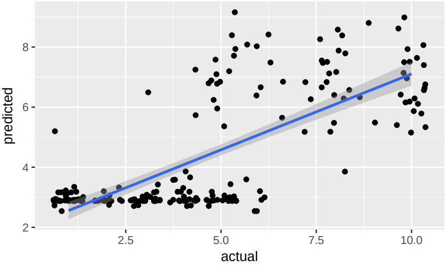
Evaluating model performance

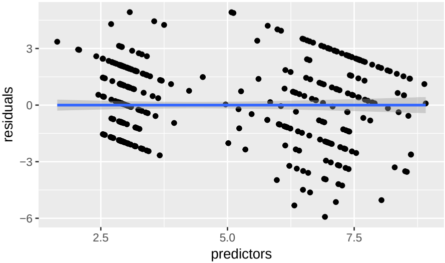
Never use the same data for evaluation that you used for training!

Predictions on test data

Regression

- RMSE: 1.97
- R^2: 0.50

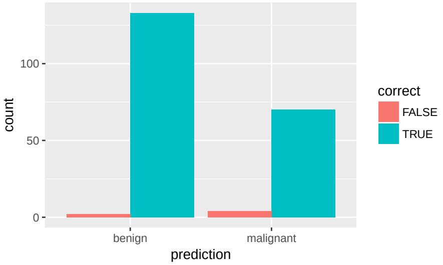


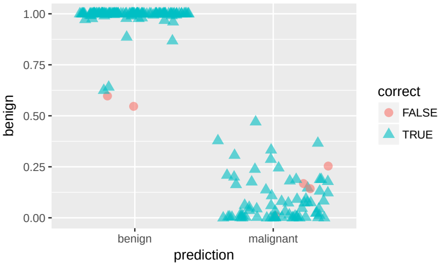


Predictions on test data

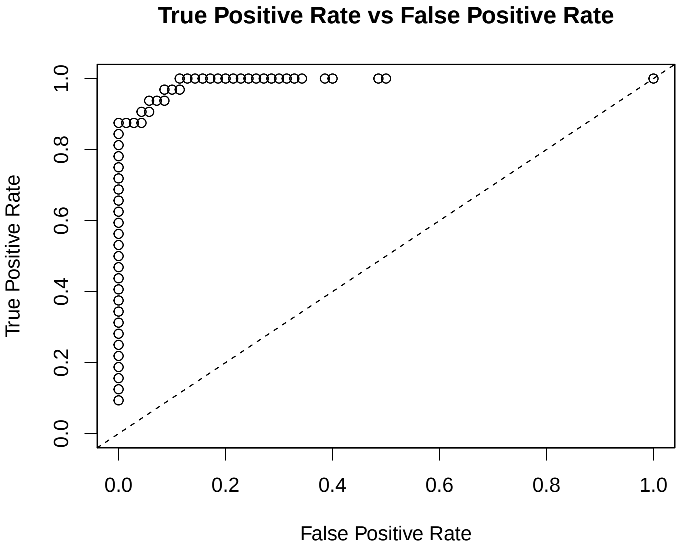
Classification

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction  benign malignant
##   benign       133         2
##   malignant      4        70
##
##                Accuracy : 0.9713
##                  95% CI : (0.9386, 0.9894)
##     No Information Rate : 0.6555
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9369
##  Mcnemar's Test P-Value : 0.6831
##
##             Sensitivity : 0.9708
##             Specificity : 0.9722
##          Pos Pred Value : 0.9852
##          Neg Pred Value : 0.9459
##              Prevalence : 0.6555
##          Detection Rate : 0.6364
##    Detection Prevalence : 0.6459
##       Balanced Accuracy : 0.9715
##
##        'Positive' Class : benign
##
```

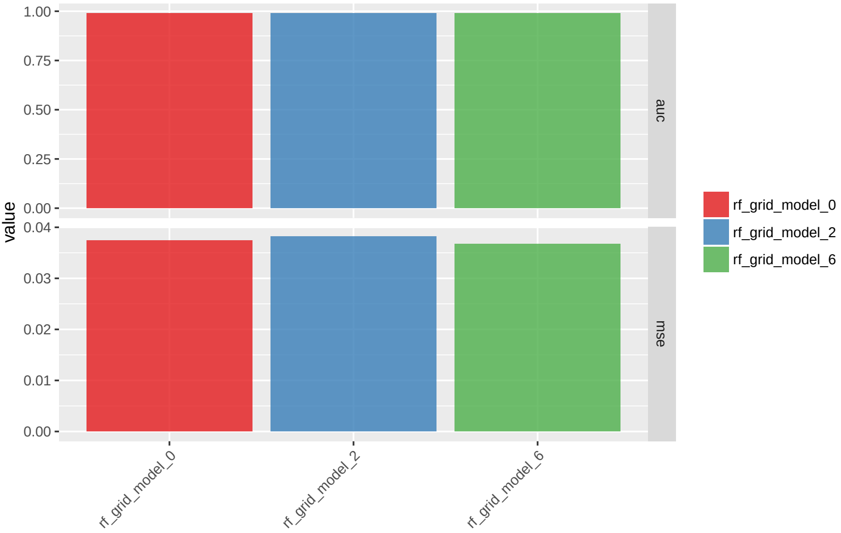




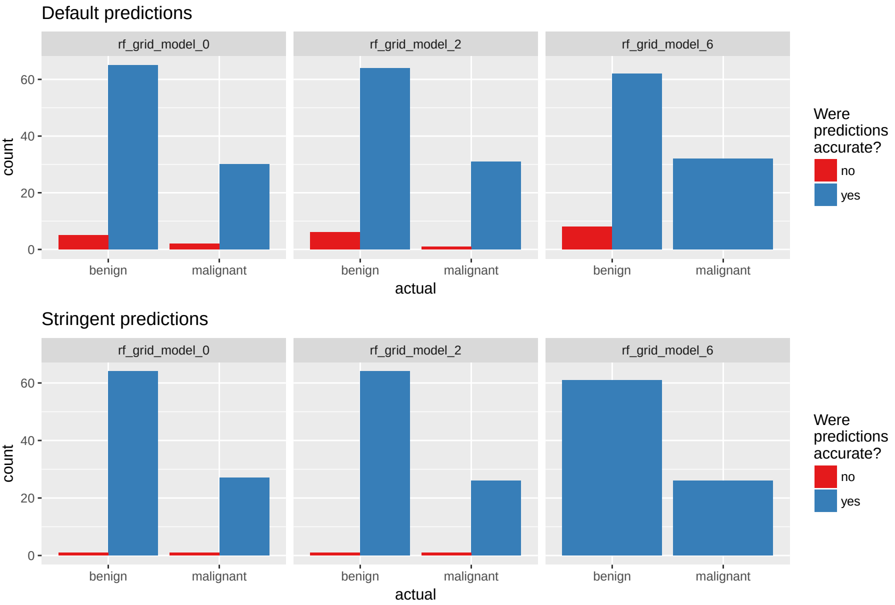
Area Under the Curve (AUC



AUC and mean squared error (MSE)



Predictions on test data



Take home messages:

- ...

Outlook

- ‘big data’ needs to be big!
- for really meaningful models, data needs to be shared
- the more data, the more accurate and generalizable the models will be
- issues: privacy, platform, quality standards
- ML could make health care more cost-effective by reducing the energy required for interpretation

Thank you for your attention!

Questions?

Slides and code will be available on Github:
https://github.com/ShirinG/Webinar_ML_for_disease

Code will also be on my website: https://shiring.github.io

shirin.glander@wwu.de