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

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

- Welcome everybody!
- Thank you very much for the invitation and the opportunity to present this talk
- I will go over my work on using machine learning models in disease prediction
- I want to specifically give a hands-on demonstration of how you can build meaningful models yourselves using R
- I will demonstrate how to evaluate model performance and
- how to optimize models to address different disease-related questions

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

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

- Before I start, I want to quickly introduce myself:
- I am a bioinformatics postdoc
- working with next generation sequencing data,
- like RNA-seq for transcriptomics,
- whole genome sequencing for variant analysis,
- ATAC- or Chip-seq for chromatin and epigenetic information.
- My own research focuses on questions relating to autoinflammatory diseases
- and innate immune mechanisms

- I earned my PhD in biology from the University of Münster in 2015
- working with RNA-seq data to determine how plant defense has co-evolved with and potentially shaped different life-history strategies

- Before that, during my BSc and MSc I worked on questions about evolutionary genetics and immune memory in Drosophila

Table of contents

Building meaningful machine learning models for disease prediction

- Now, let's start with the topic about which you're here:
- Machine learning is a powerful approach for developing sophisticated, automatic, and objective models for the analysis of complex biomedical data

- I titled my talk: "Building meaningful machine learning models for disease prediction"
- with the emphasis on **meaningful**
- So, first, I want to introduce to you what it is I mean exactly with **meaningful models**
- Then, I will give you a few examples of how machine learning is currently being used in disease modeling and clinical data science
- Before I delve into the nitty-gritty of modeling, I will quickly recap the most important concepts of ML
- And finally, I will show how to build ML models in R
- and how to evaluate the performance of such models

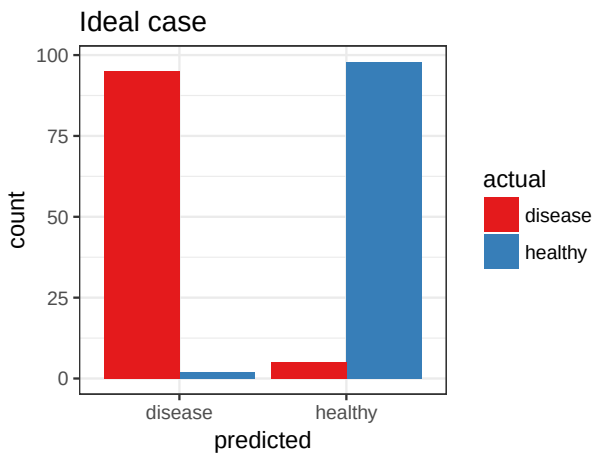
What makes a model meaningful?

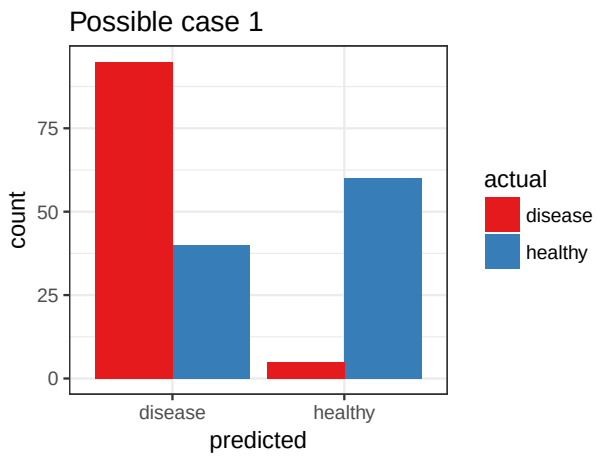
What makes a model meaningful?

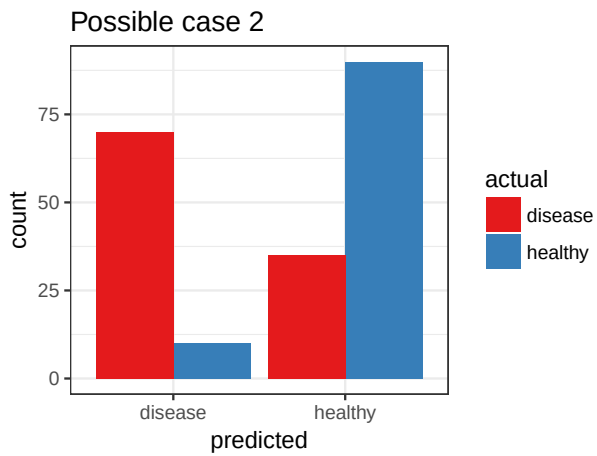
Meaningful models

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

Accuracy depends on the problem!





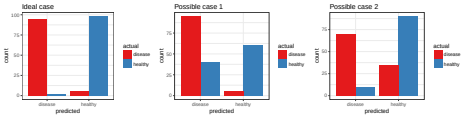


What makes a model meaningful?

Meaningful models

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

Accuracy depends on the problem!



- I want to begin with an introduction to the main question of my talk:
- what makes a model *good* or *meaningful*?

1. It answers a **specific** question or addresses a specific problem,
   e.g. does a mammogram image show a healthy breast or is there a tumor?
   Or does an ECG show a normal heart rhythm or arrhythmia?
2. And it produces a correct outcome (e.g. a diagnosis) often enough that we trust it!

- If we build models, we therefore need to evaluate its accuracy
- before we can decide whether it is trustworthy enough to implement in real-life,
- like in a hospital where it could e.g. assist doctors in making decisions on treatment

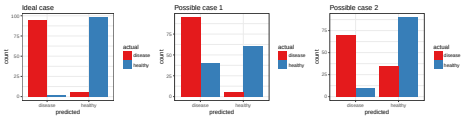
- But what exactly is *high accuracy* can not be defined with a one-size-fits-all approach:
- it depends on the problem we want to model.

What makes a model meaningful?

Meaningful models

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

Accuracy depends on the problem!



Let me illustrate what I mean with the following examples:

- Ideal case: Of course, we all want to achieve ideal modeling results where overall prediction accuracy is very high. With a model like that, we can be very confident that a healthy person is indeed healthy and a sick person is not.
- But in reality, we often achieve prediction accuracies that are much less nice.
- Scenarios 1 and 2: Le's consider two possible scenarios:
- in scenario 1, we can be very confident that a person who got classified as "healthyïs indeed healthy, while a person who has been diagnosed as diseased might as well be healthy based on these prediction accuracies
- in case 2, it is the other way around.
- We now need to make a decision which scenario is better and in which direction we want to optimize our model: do we rather want to refer a few healthy people for further checking because the model predicted them as diseased? Or do we rather want to be as certain as possible that a predicted disease is actually true and accept that we might miss a few disease cases?

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 &
- early diagnosis of osteoporosis from chest radiographs[1]
- identifying signatures of Brain Cancer from MRSI[2]
- ... and many more ...

---

[1] K. Doi (2007). "Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential." In: *Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society* 31.4-5, S. 198–2011.

[2] P. Sadja (2006). "Machine Learning for Detection and Diagnosis of Disease". In: *Annual Review of Biomedical Engineering* 8.1. PMID: 16834566, S. 537–565.

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 &
- early diagnosis of osteoporosis from chest radiographs[1]
- identifying signatures of Brain Cancer from MRSI[2]
- ... and many more ...

[1] K. Doi (2007). "Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential." In: *Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society* 31.4-5, S. 198–2011.

[2] P. Sadja (2006). "Machine Learning for Detection and Diagnosis of Disease". In: *Annual Review of Biomedical Engineering* 8.1. PMID: 16834566, S. 537–565.

- There is not a place in medicine where AI doesn't have potential applications:
- more data is being collected that needs to be interpretated
- increasingly data driven diagnosis, e.g. in radiology
- image recognition of e.g. tumors or pneumonia in medical images
- similar to training ML models to recognize images of cats (classification in images)
- ML allows incorporation of data from heterogenous inputs:
- clinical, genomics, drugs, electronic health records, etc.
- = personalised medicine

- A key aspect of precision medicine is the development of informatics tools
- that can analyze and interpret 'big data' sets
- in an automated and adaptive fashion
- while providing accurate and actionable clinical information
- ML based models can improve detection, diagnosis, and therapeutic monitoring of disease

- You can find things with ML that you wouldn't be able to find otherwise

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 &
- early diagnosis of osteoporosis from chest radiographs[1]
- identifying signatures of Brain Cancer from MRSI[2]
- ... and many more ...

[1] K. Doi (2007). "Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential." In: *Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society* 31.4-5, S. 198–2011.

[2] P. Sadja (2006). "Machine Learning for Detection and Diagnosis of Disease". In: *Annual Review of Biomedical Engineering* 8.1. PMID: 16834566, S. 537–565.

- Many models today perform better than humans!
- identifying breast cancer, lung cancer, osteoporosis, brain tumors, etc. from medical images
- predicting response of different cancer types to different treatments
- Computer-aided diagnosis (CAD) has become one of the major research subjects in medical imaging and diagnostic radiology
- With CAD, radiologists use the computer output as a 'second opinion' and make the final decisions
- In vivo magnetic resonance spectroscopy imaging (MRSI) allows noninvasive characterization
- and quantification of molecular markers of potentially high clinical utility
- for improving detection, identification, and treatment for a variety of diseases, most notably brain cancers

- A patient with a rare difficult condition can be matched to similar cases from the past
- faster diagnosis, better treatment

- Doctors are still important!
- computer does the tasks that we humans are not good at, like interpreting complex images
- we have more data than humans can manage to make sense of (e.g. genomics data)
- the clinicians will be freed up to think about best treatment options and talk more with the patients
- Good models are built on strong knowledge of the question and the biology behind it
- features need to be evaluated in context

A quick recap of ML basics

Machine learning

- artificial intelligence (AI)
- 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 vs Unsupervised algorithms

1. Unsupervised matrix decomposition methods, such as nonnegative matrix factorization, which impose general, although physically meaningful, constraints, are able to recover biomarkers of disease and toxicity, generating a natural basis for data visualization and pattern classification.

2. Supervised discriminative models that explicitly address the bias-variance trade-off, such as the support vector machine, have shown great promise for disease diagnosis in computational biology, where data types are disparate and high dimensional.

3. Generative models based on Bayesian networks offer a general approach for biomedical image and signal analysis in that they enable one to directly model the uncertainty and variability inherent to biomedical data as well as provide a framework for an array of analysis, including classification, segmentation, and compression.

Classification vs Regression

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

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

Machine learning uses so called features (i.e. variables or attributes) to generate predictive models. Using a suitable combination of features is essential for obtaining high precision and accuracy. Because too many (unspecific) features pose the problem of overfitting the model, we generally want to restrict the features in our models to those, that are most relevant for the response variable we want to predict. Using as few features as possible will also reduce the complexity of our models, which means it needs less time and computer power to run and is easier to understand.

extraction of salient structure in the data that is more informative than the raw data itself (the feature extraction

problem)

Take home messages:

- ...

How to build ML models in R

Session setup

- RStudio
- Breast Cancer Wisconsin Dataset
- caret
- h2o

Code will be available on my website and on Github

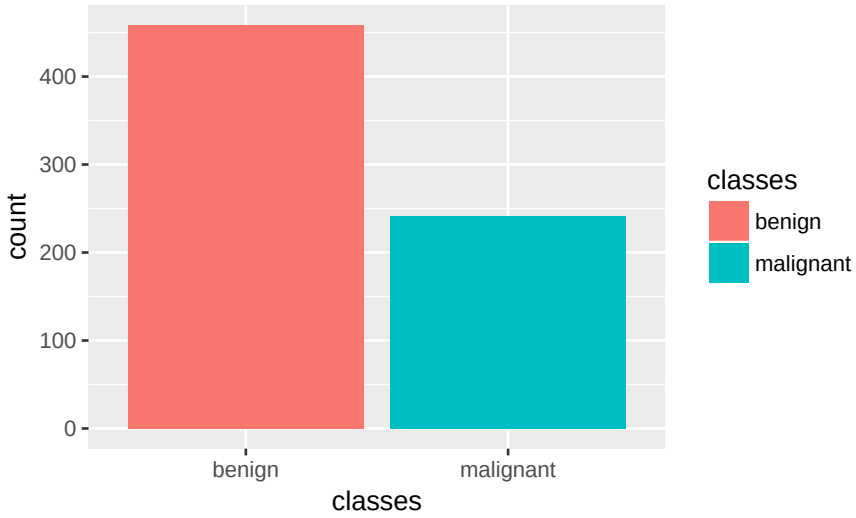
Get to know your data

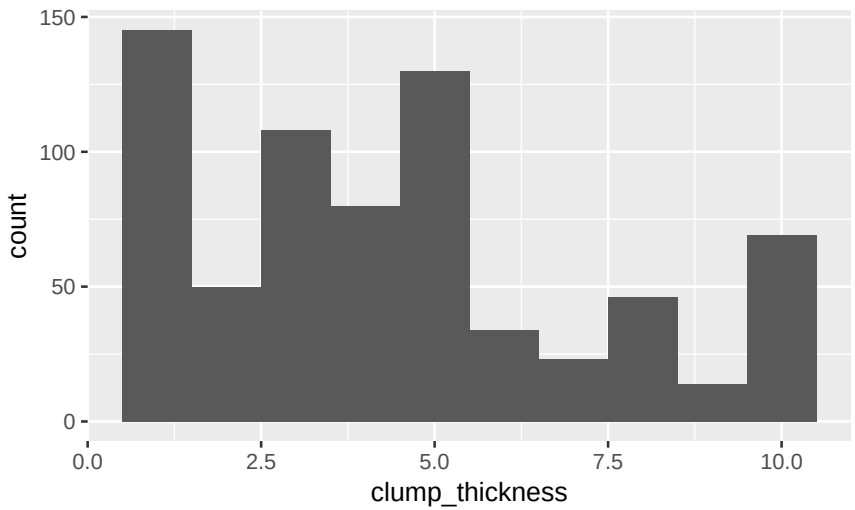
Missing data

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

Response variable

- Is it balanced?





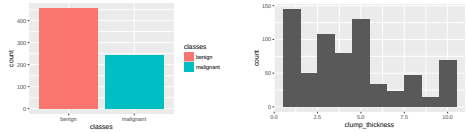
Get to know your data

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)?

Response variable

- Is it balanced?



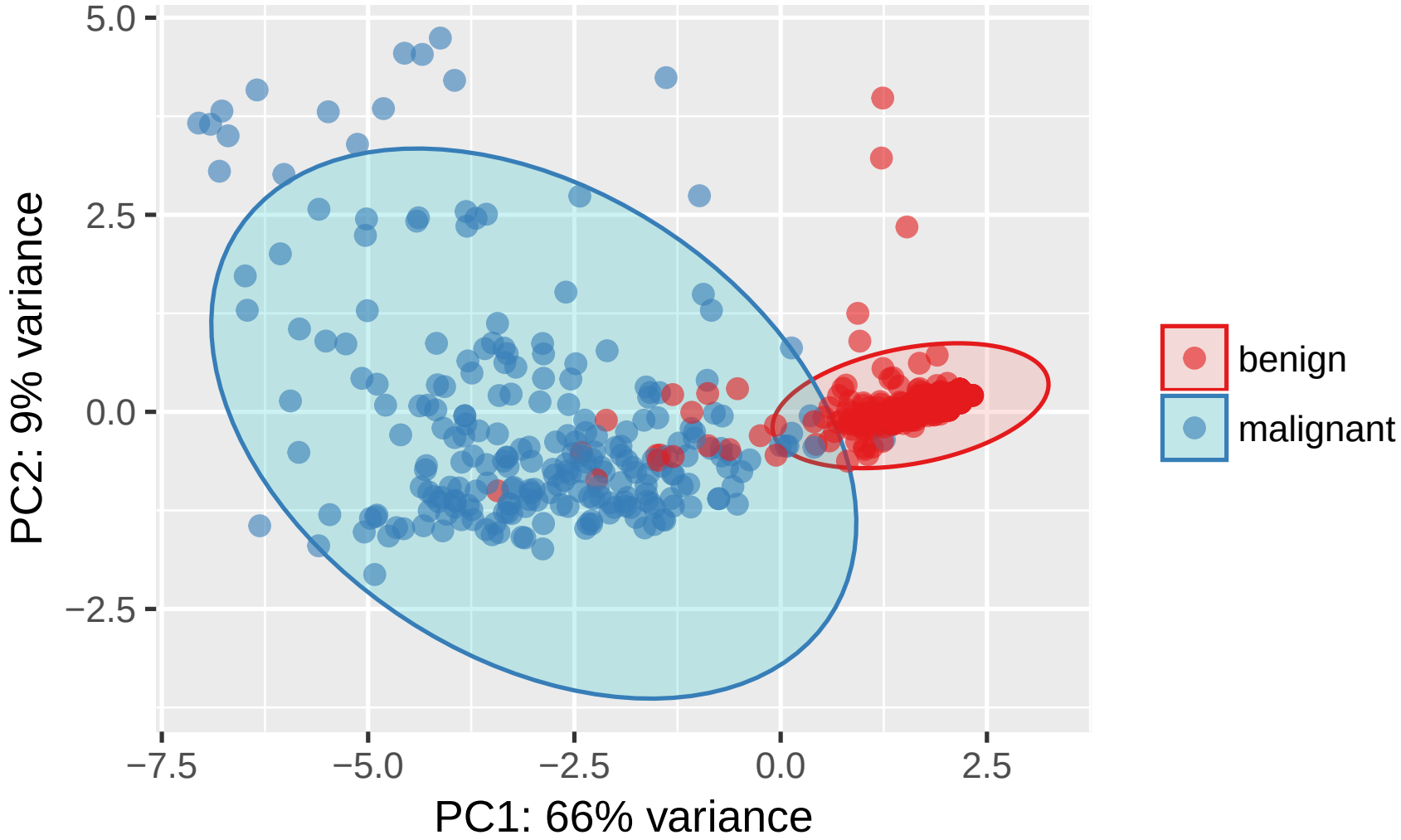
If we have lots of data and few missing values, we can afford to loose these data points.
If we don't have that much data though, our model will probably loose significant power if we remove the samples.
In that case, we would rather introduce a certain uncertainty by imputing missing values.

We are especially interested in our response variable, that we want to classify or regress on.
Most important to know is the distribution: are the classes balanced?
If we have unbalanced data, this will effect our model later on

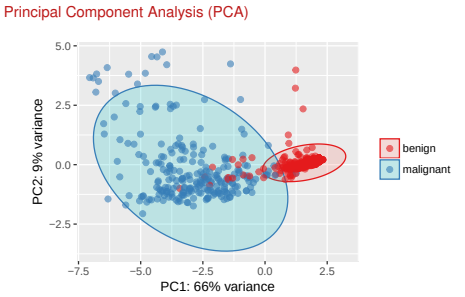
We would have to consider over- or undersampling.

Get to know your data

Principal Component Analysis (PCA)



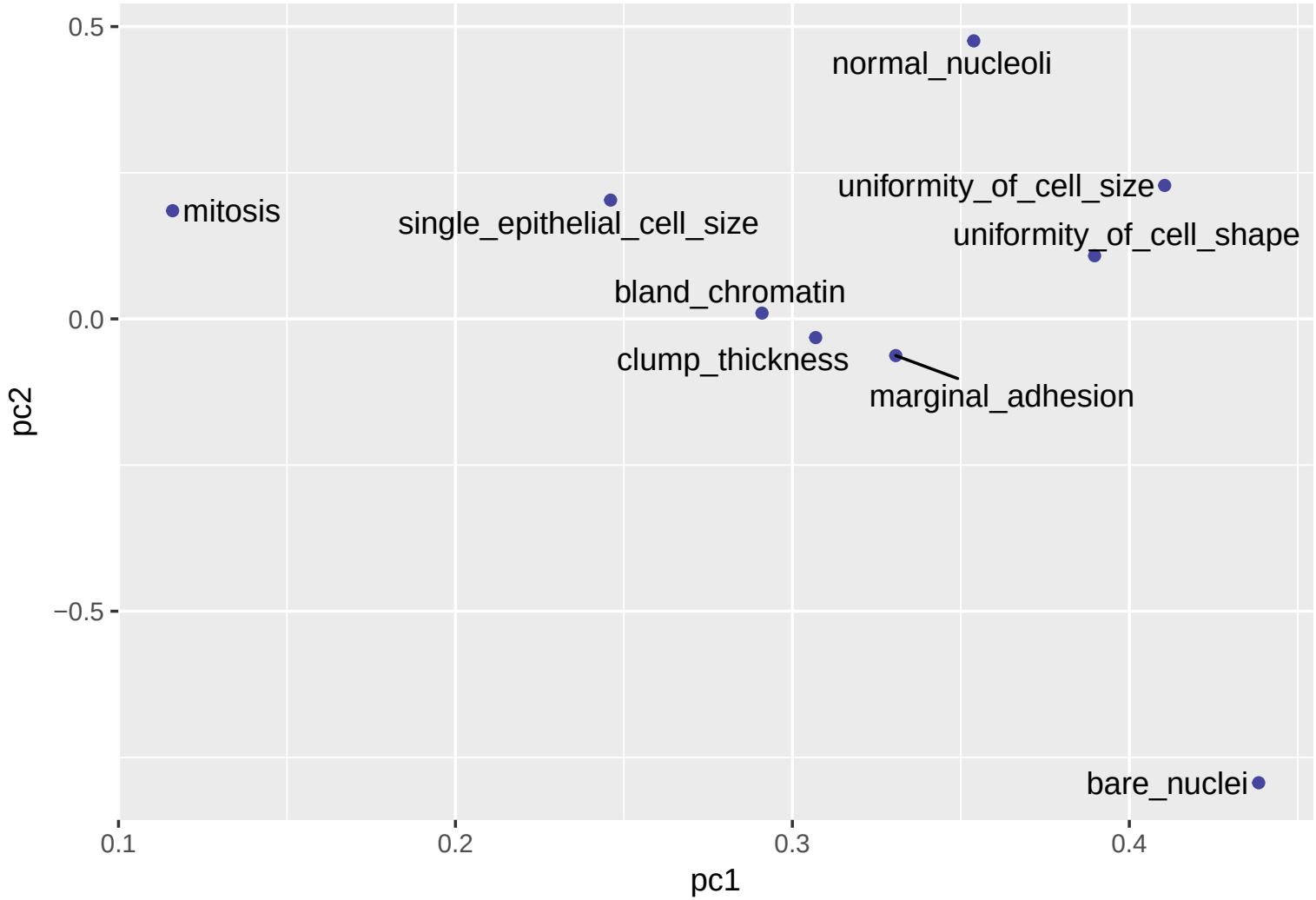
Get to know your data



To get an idea about the dimensionality and variance of the datasets, I am first looking at PCA plots for samples and features. The first two principal components (PCs) show the two components that explain the majority of variation in the data.

Get to know your data

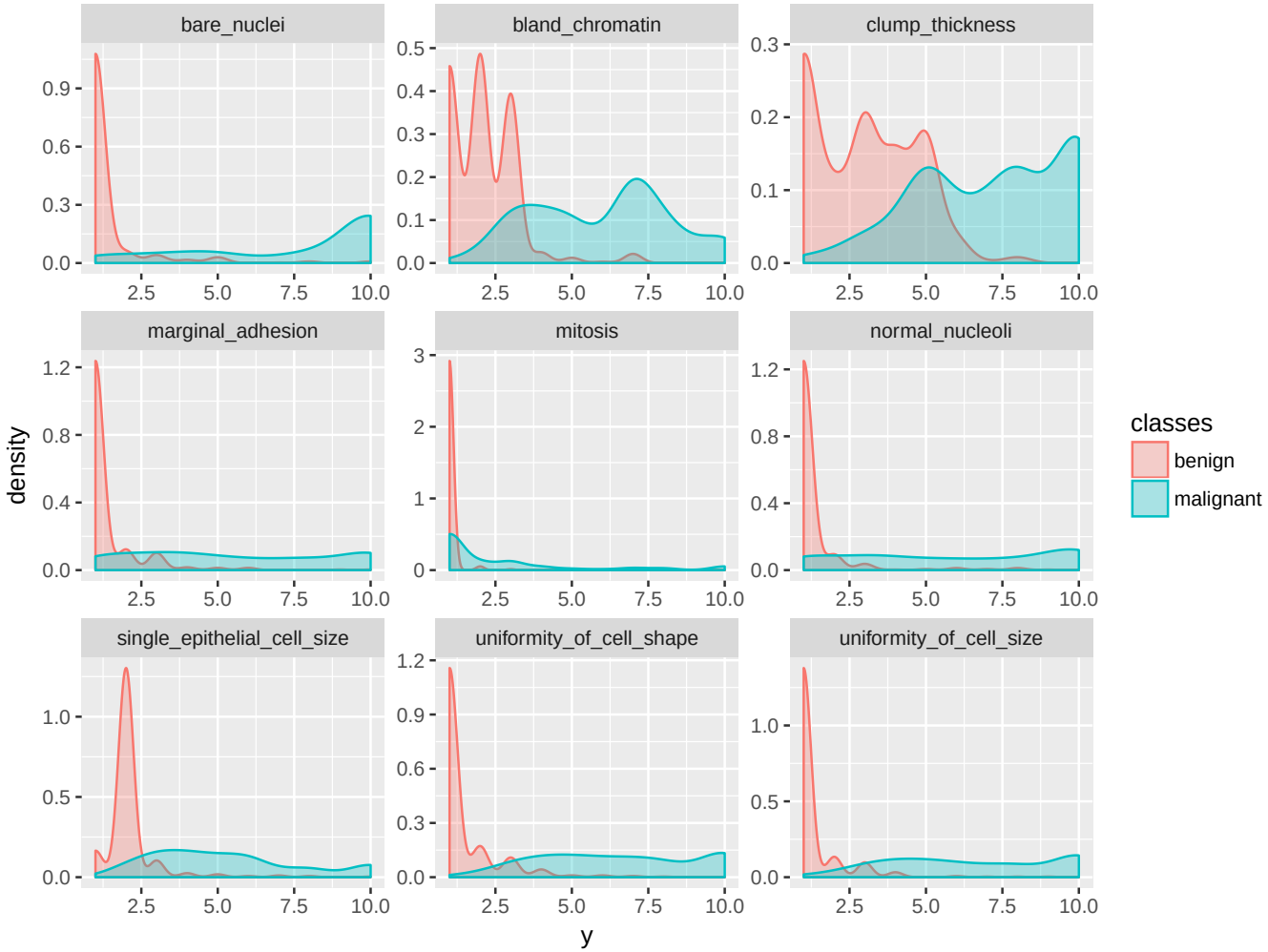
Principal Component Analysis (PCA)



Get to know your data

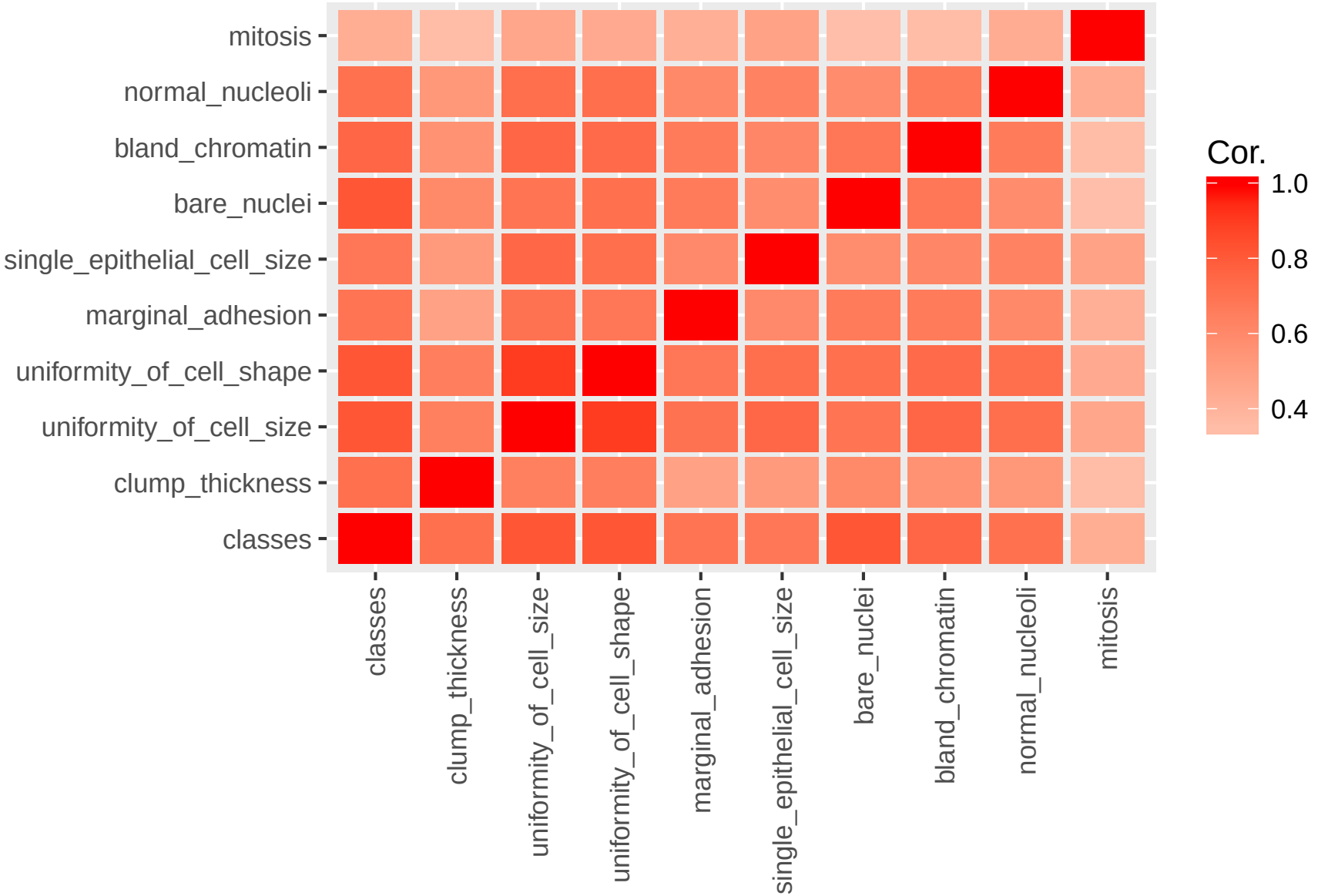
Features

- factors or numeric
- pre-processing



Get to know your data

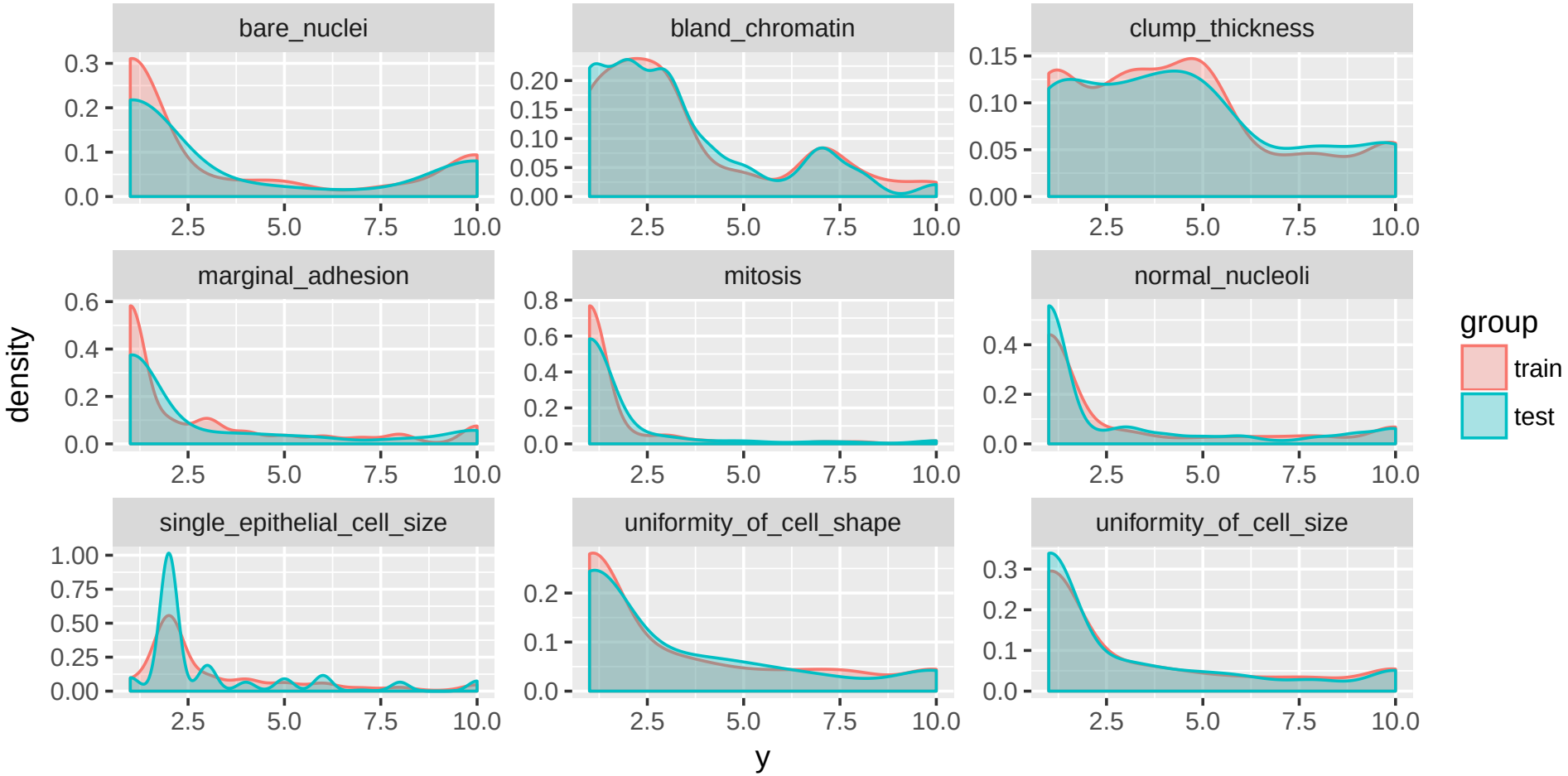
Correlation



Training, validation and test data

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

Density distribution

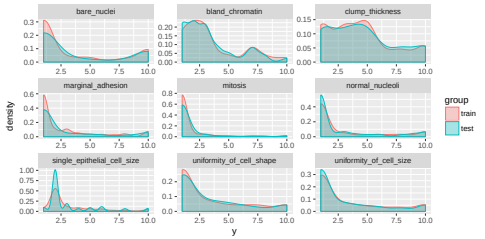


Training, validation and test data

Training, validation and test data

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

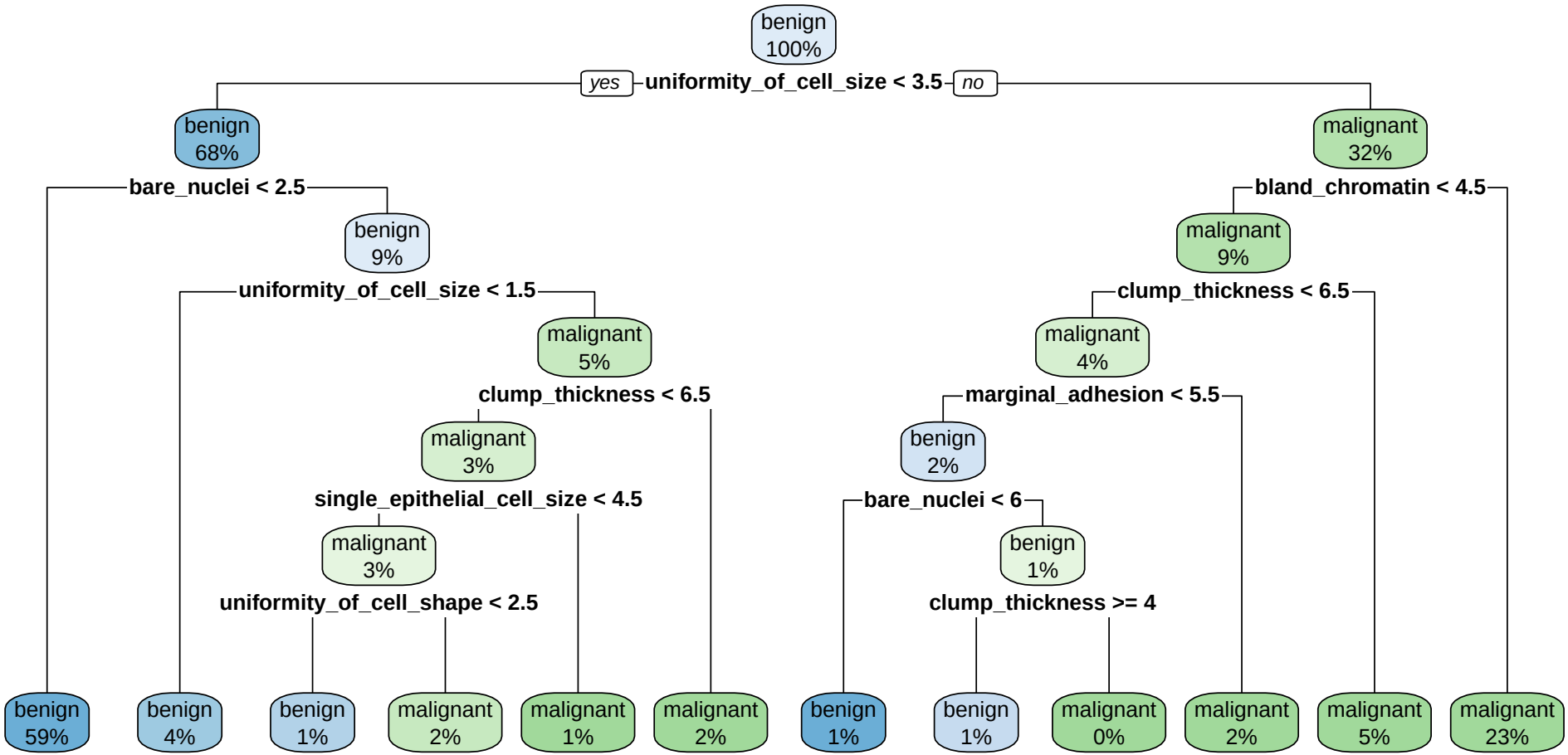
Density distribution



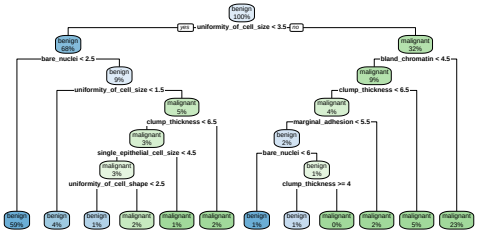
For accurate predictions, density distribution should be similar in training and test data!

Classification with tree-based models

Decision trees



Decision trees



To get an idea about how each feature contributes to the prediction of the outcome, I first built a decision tree based on the training data.

Tree-based classification with caret

Random Forest or Gradient boosting

add verbatim code cite caret

Random Forest or Gradient boosting
add verbatim code cite caret

This function sets up a grid of tuning parameters for a number of classification and regression routines, fits each model and calculates a resampling based performance measure.
train can be used to tune models by picking the complexity parameters that are associated with the optimal resampling statistics. For particular model, a grid of parameters (if any) is created and the model is trained on slightly different data for each candidate combination of tuning parameters. Across each data set, the performance of held-out samples is calculated and the mean and standard deviation is summarized for each combination. The combination with the optimal resampling statistic is chosen as the final model and the entire training set is used to fit a final model.

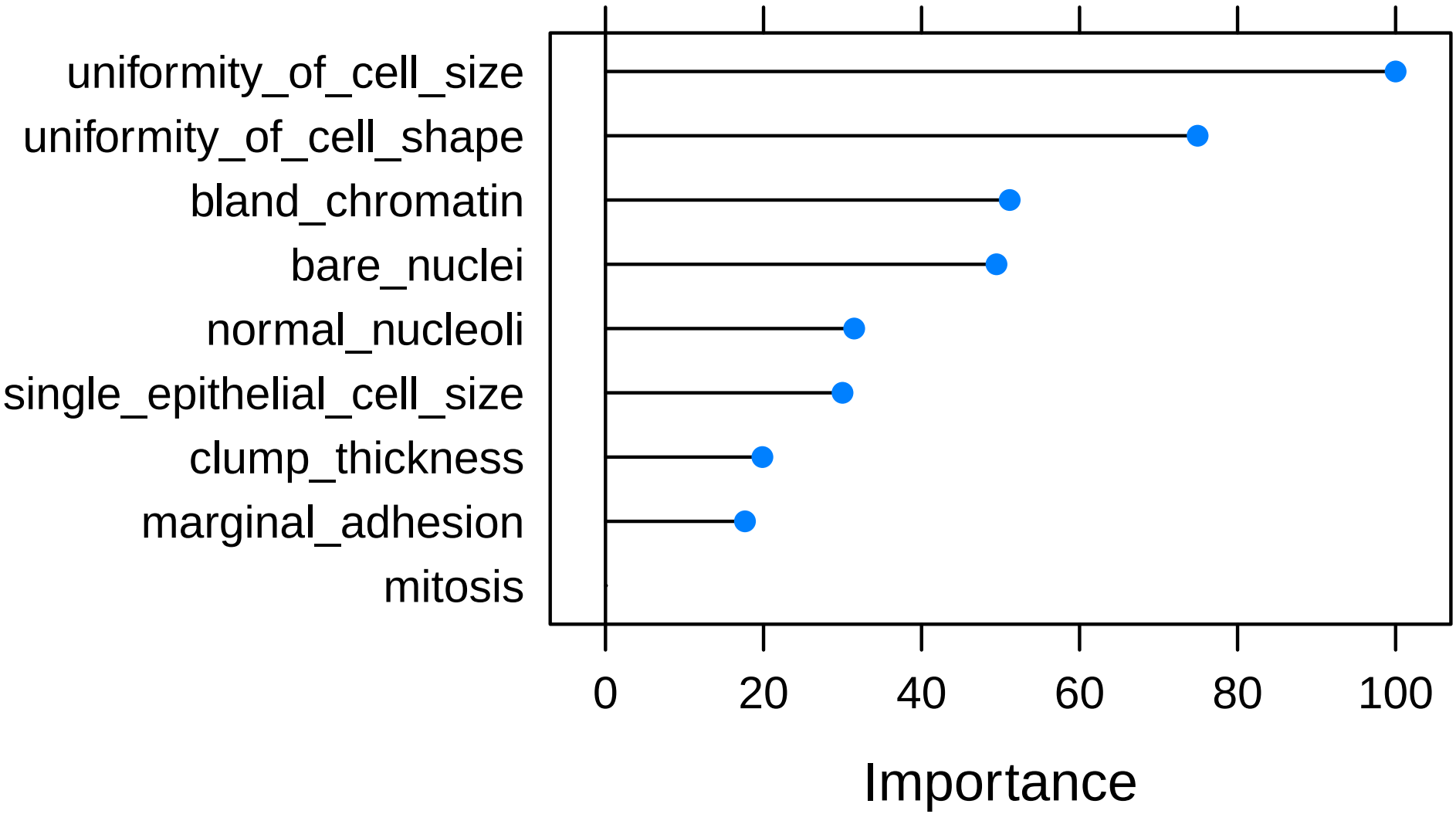
The predictors in x can be most any object as long as the underlying model fit function can deal with the object

class. The function was designed to work with simple matrices and data frame inputs, so some functionality may not

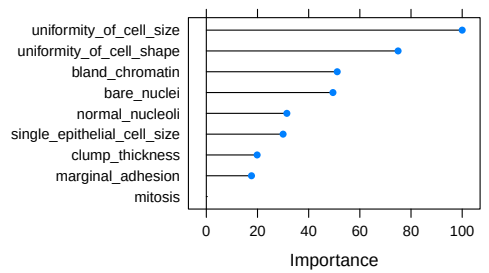
work (e.g. pre-processing). When using string kernels, the vector of character strings should be converted to a

matrix with a single column.

Feature importance



Feature importance



Not all of the features I created will be equally important to the model. The decision tree already gave me an idea of which features might be most important but I also want to estimate feature importance using a Random Forest approach with repeated cross validation.

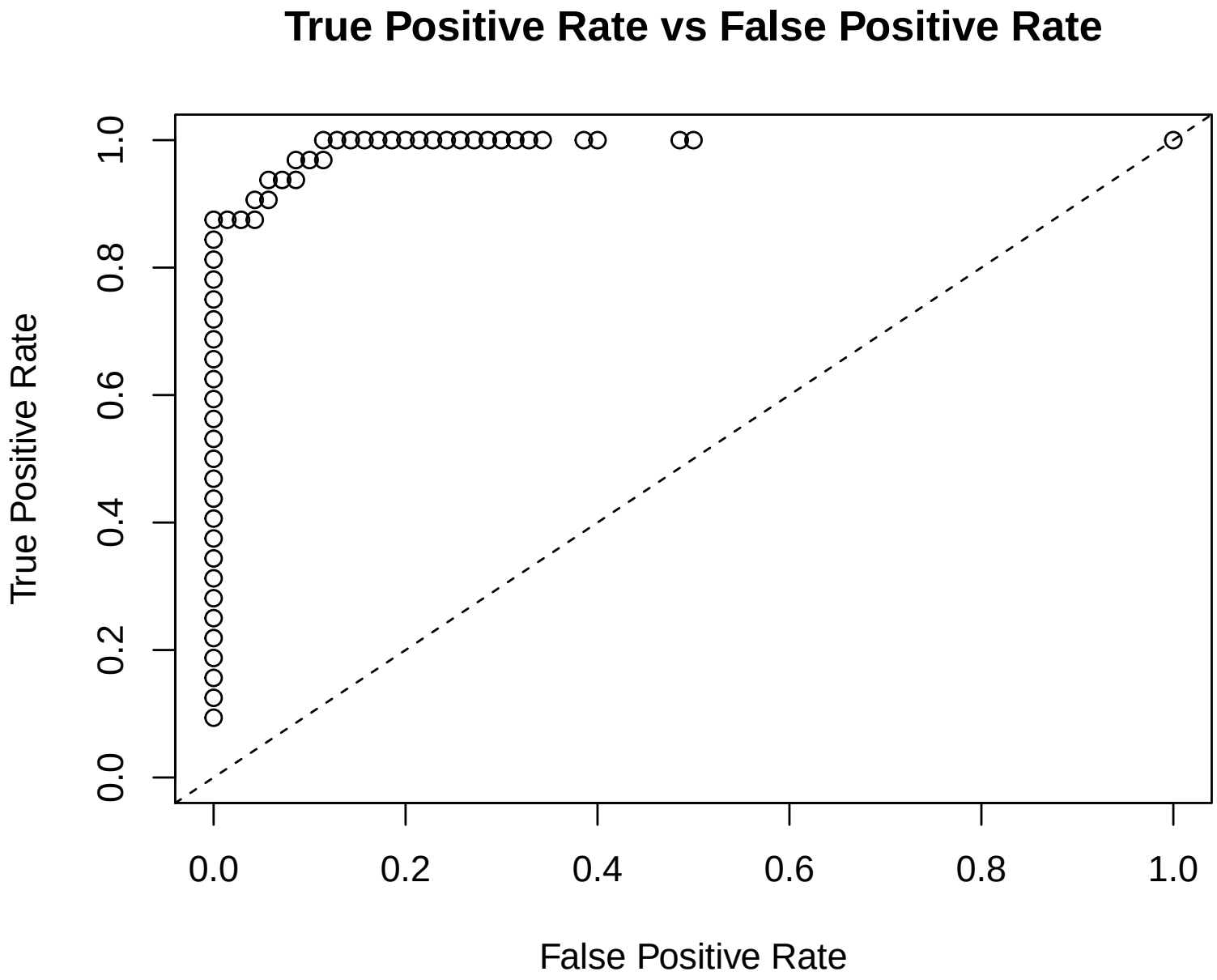
Regression with (Generalized) Linear Models

Deep learning with neural networks

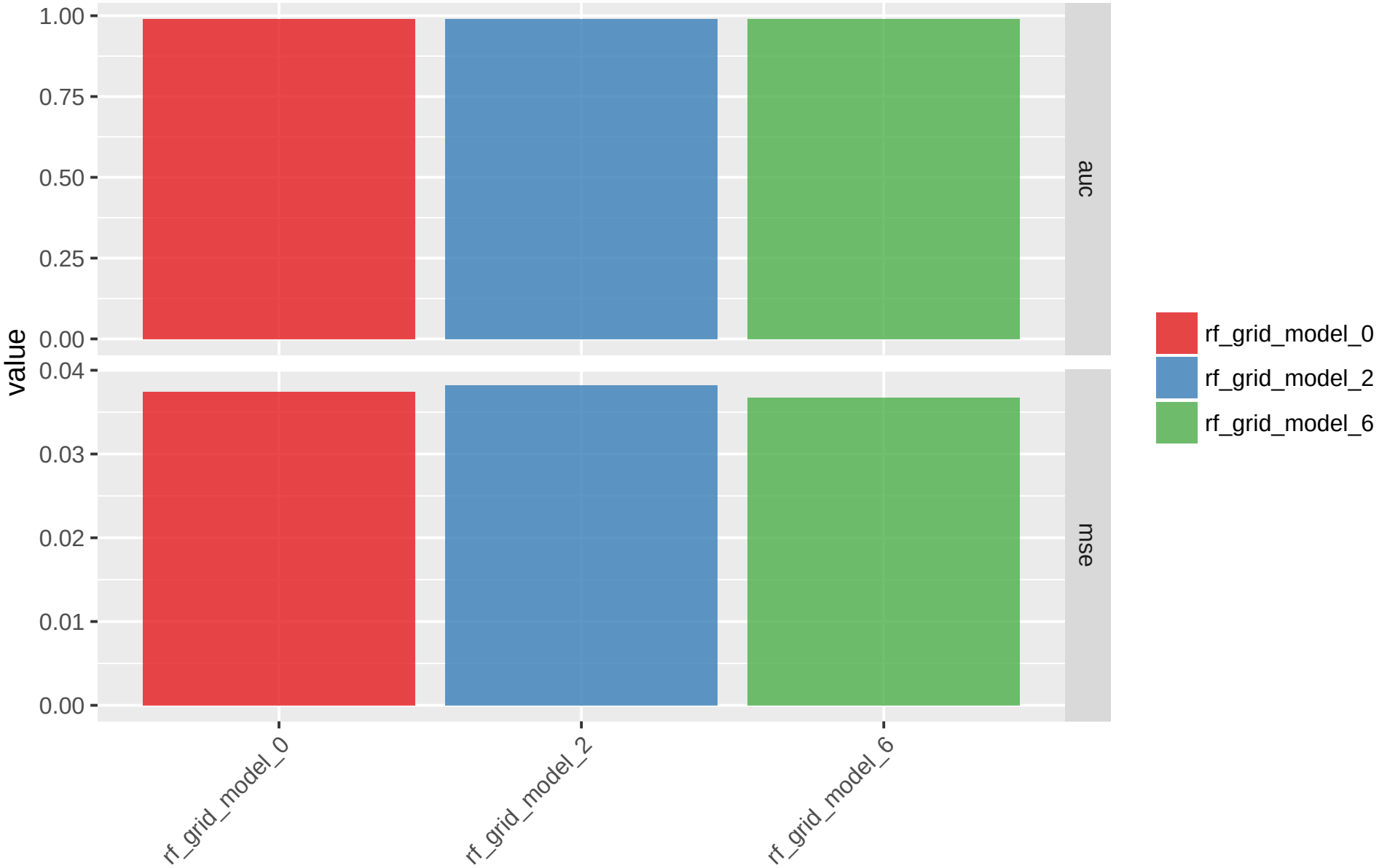
h2o
Grid Search

Evaluating model performance

Area Under the Curve (AUC



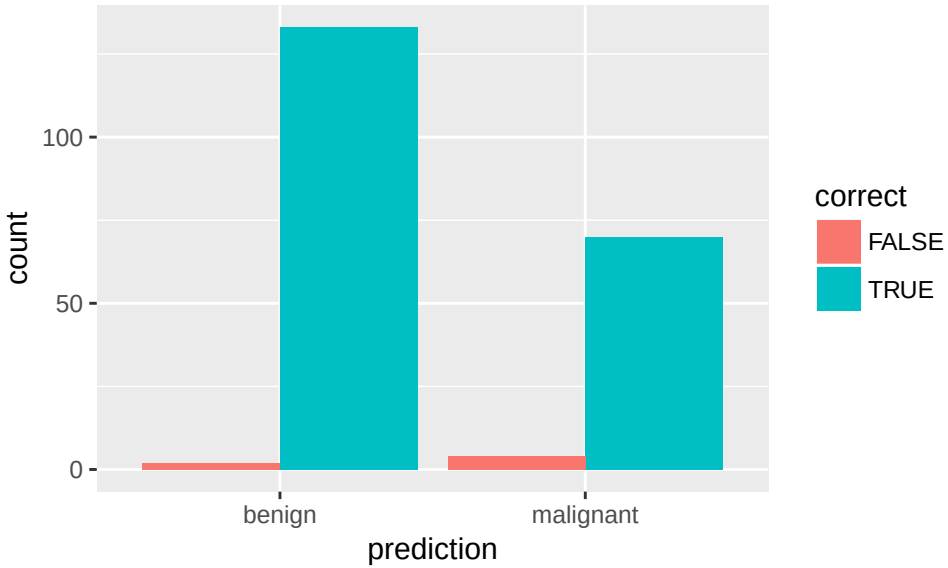
AUC and mean squared error (MSE)

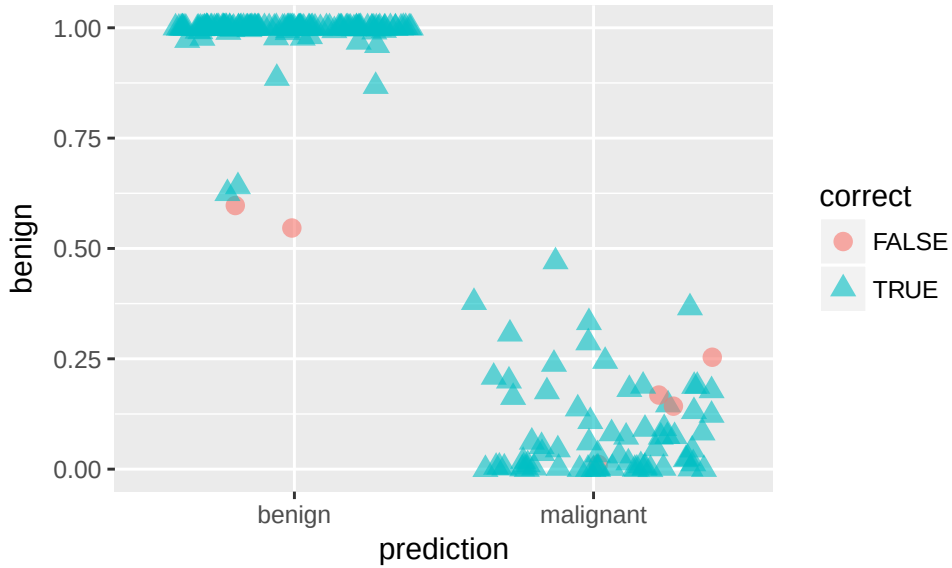


Predictions on test data

Classification

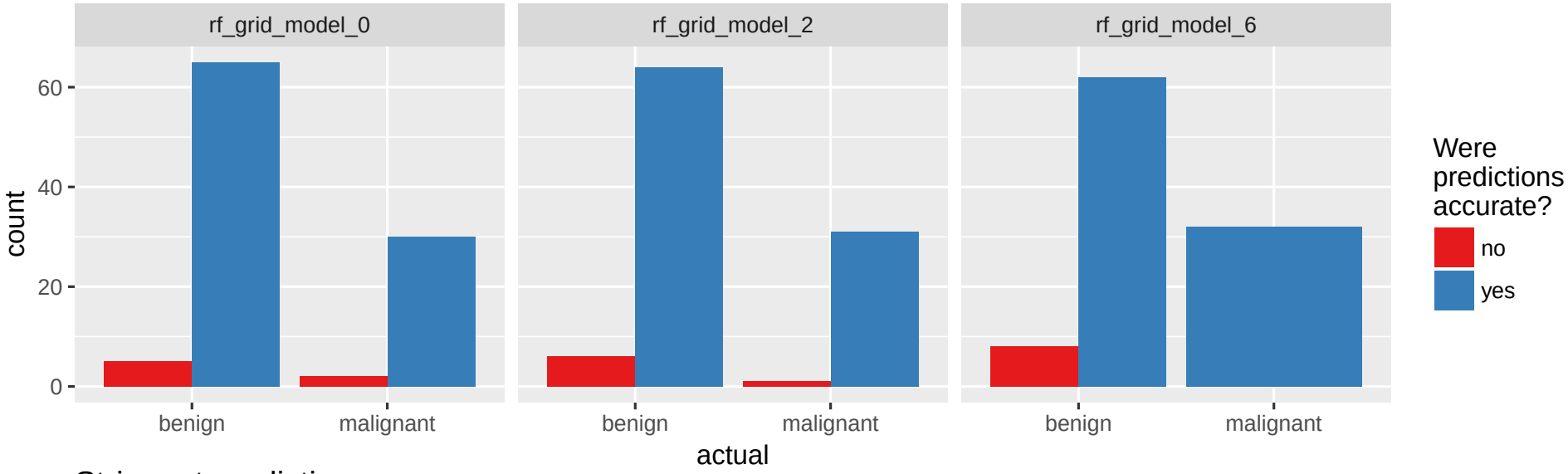
add confusion matrix verbatim



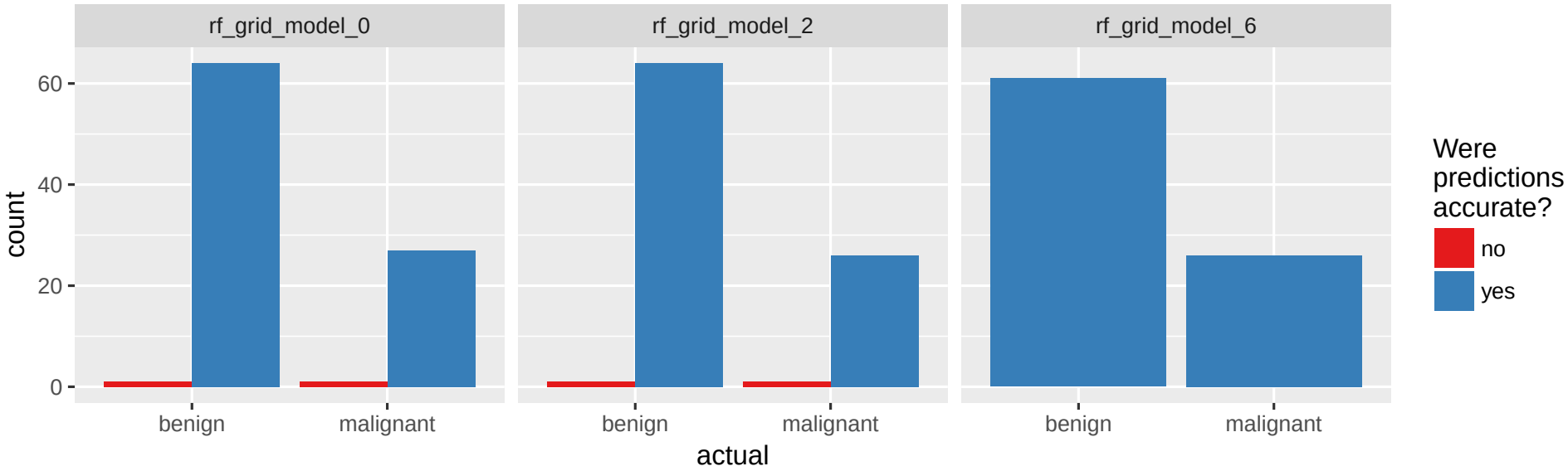


Predictions on test data

Default predictions

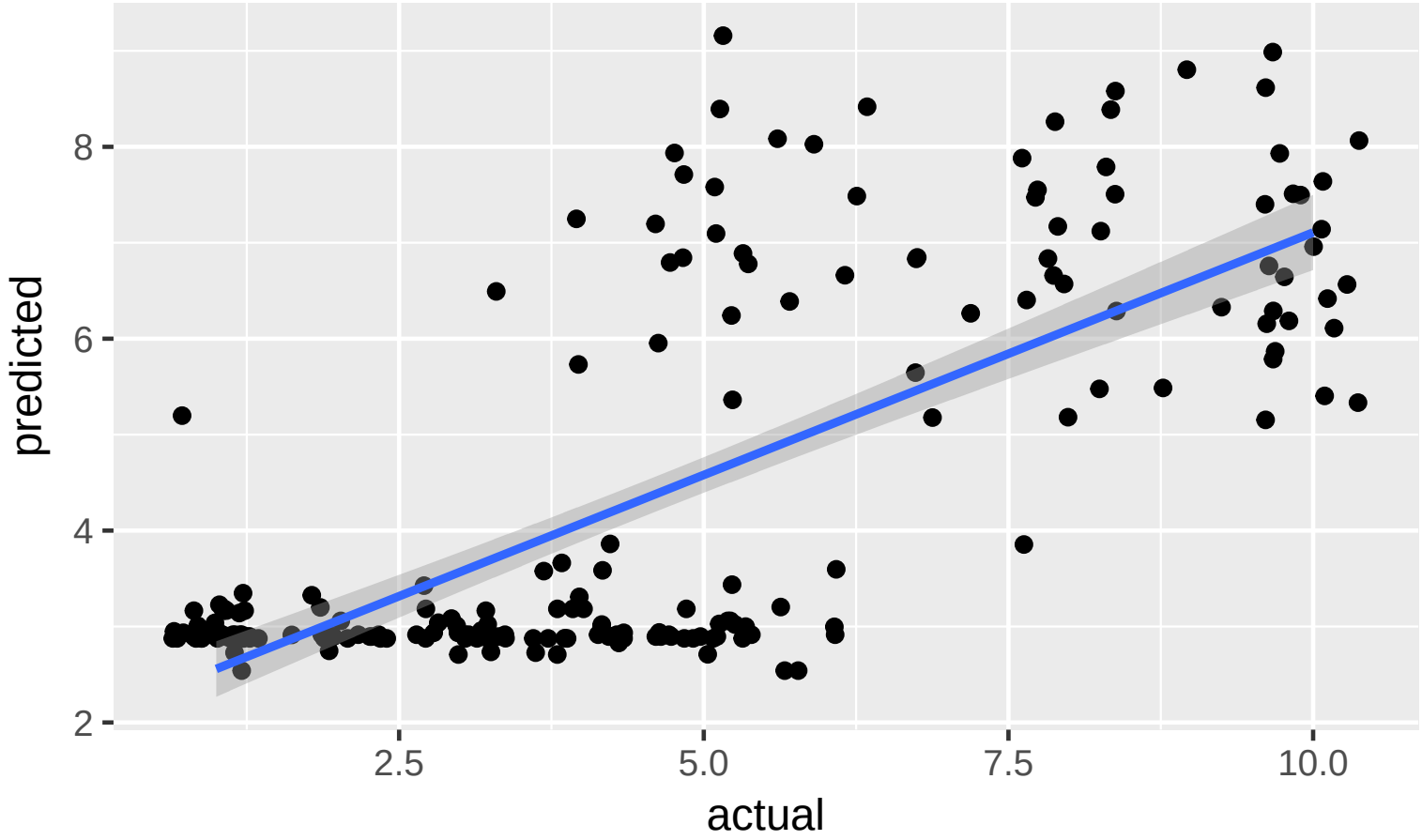


Stringent predictions



Predictions on test data

Regression



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