CLASSIFICATION METHODS

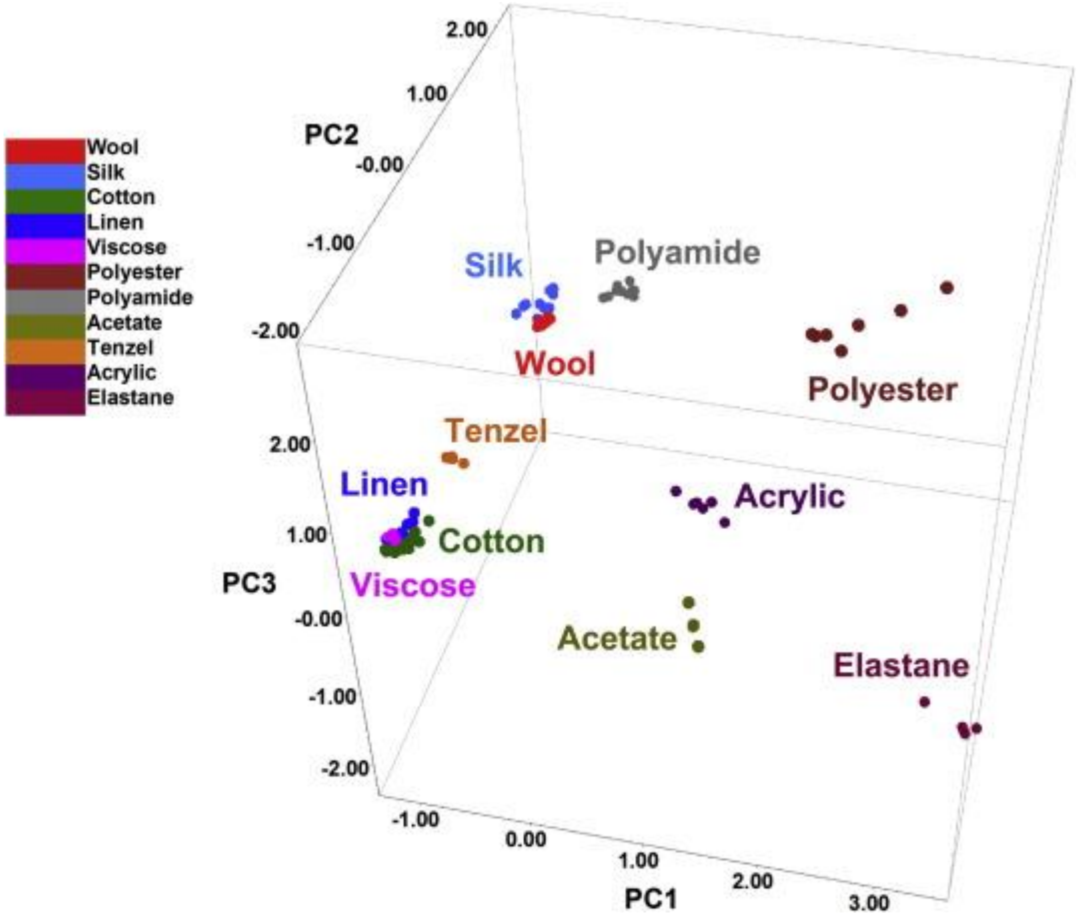
WHAT IS A CLASSIFICATION METHOD?

Mathematically are regression and classification very similar

Intuitively: Output is a qualitative not quantitative

EXAMPLES
Does the metabolite concentration in blood relate to sick or healthy?
Which dye has been used to colour the textile?
Which region is this wine from?
Which chromatographic system yields best separation and sensitivity?

TEXTILE TYPE

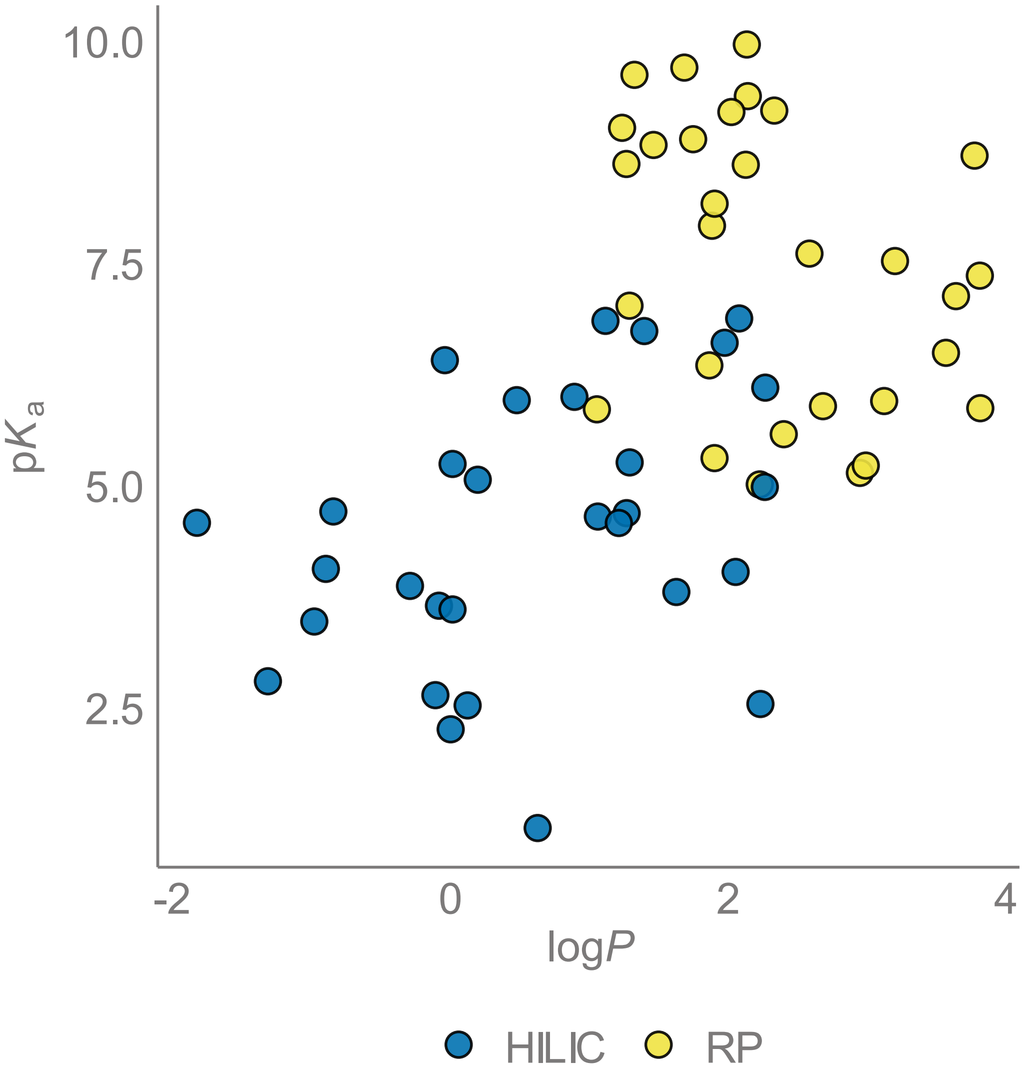


METHODS

k-Nearest Neighbours
Logistic Regression
Linear Discriminant Analysis (LDA)
Quadratic Discriminant Analysis (QDA)
Decision Trees
Random Forest

THE PROBLEM

CHOOSING SEPARATION MODE IN LC



Anneli Kruve

k-Nearest Neighbours

CHOOSING SEPARATION MODE IN LC

A new compound

- log*P* = 1.5 and p*K*a = 6.4

Should we prefer RP or HILIC?
Where to start?

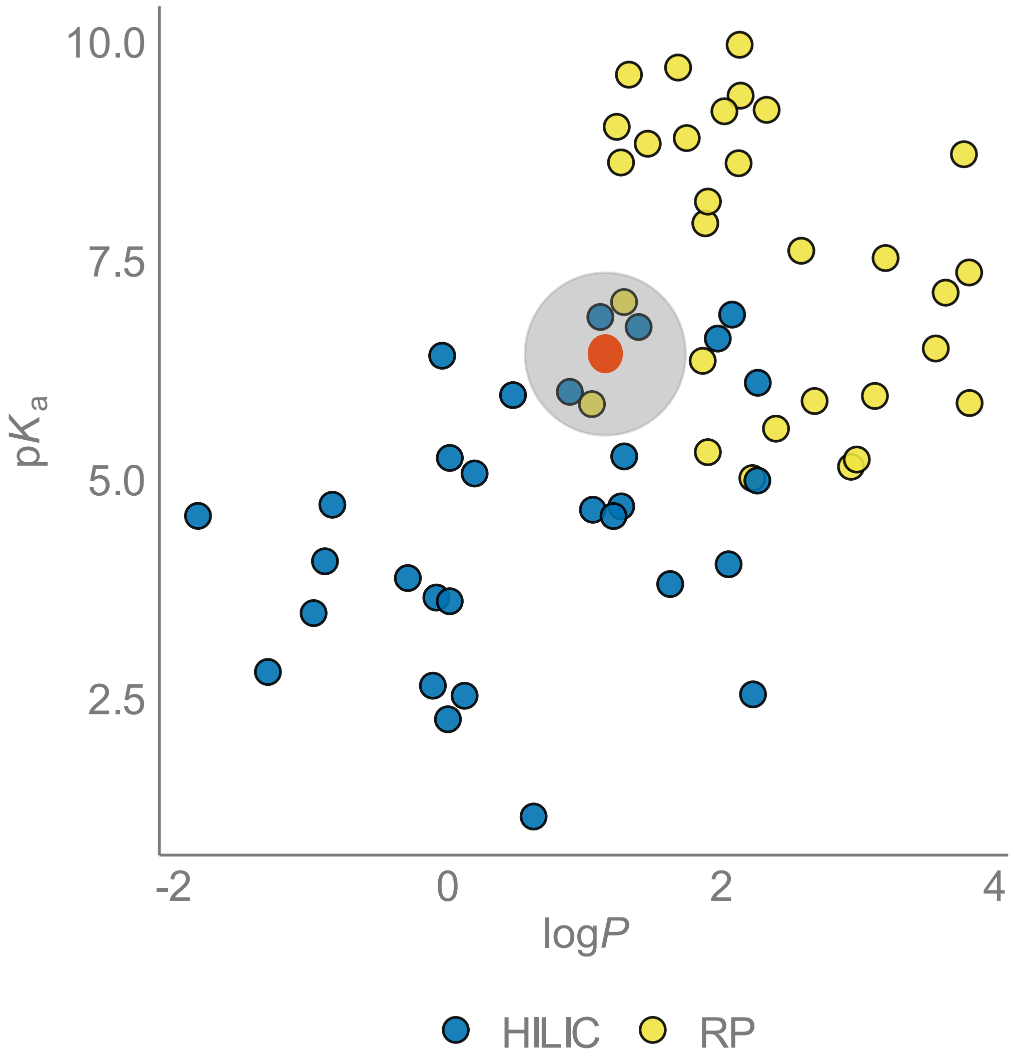
k = 1

- To which group does the nearest neighbour belong?

k = 5

- 3 measured compounds were better with HILIC
- 2 compounds were better with RP

$$\Pr(Y = j | X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j).$$



Anneli Kruve

Calculating the distance

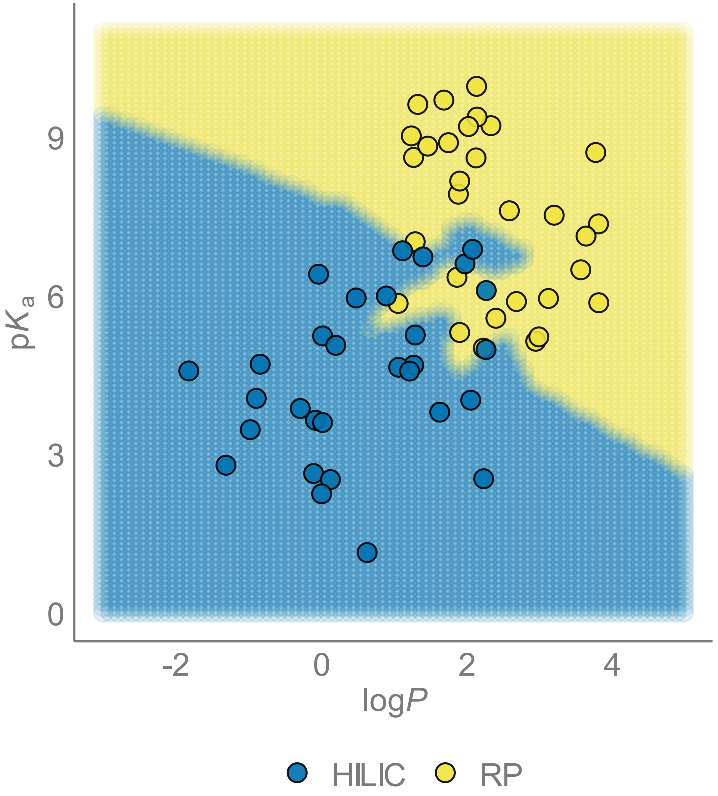
Euclidean

$$d_{a,b} = \sqrt{\sum_{i=1}^{m} (a_i - b_i)^2}$$

Manhattan

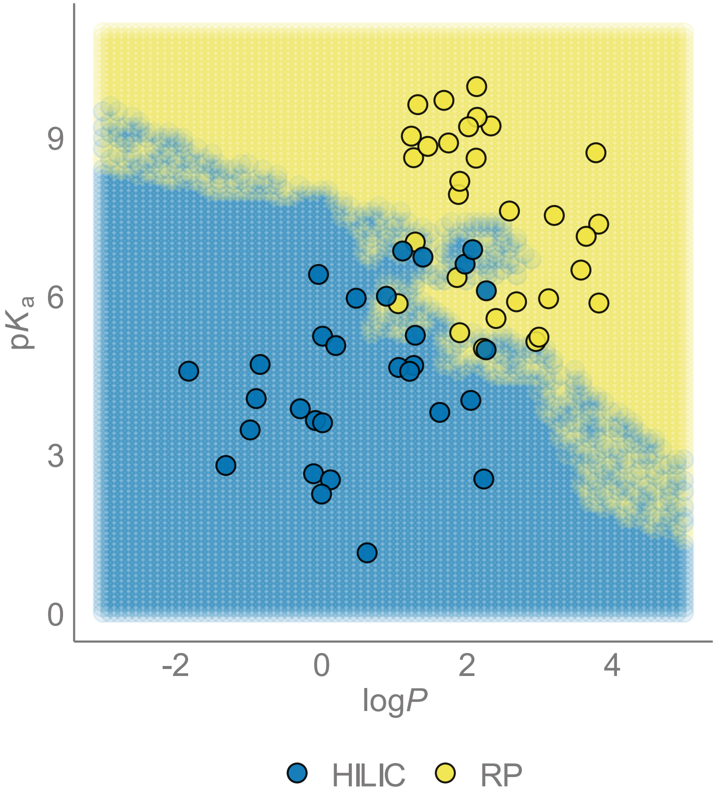
$$d_{a,b} = \sum_{i=1}^{m} abs(a_i - b_i)$$

k = 1



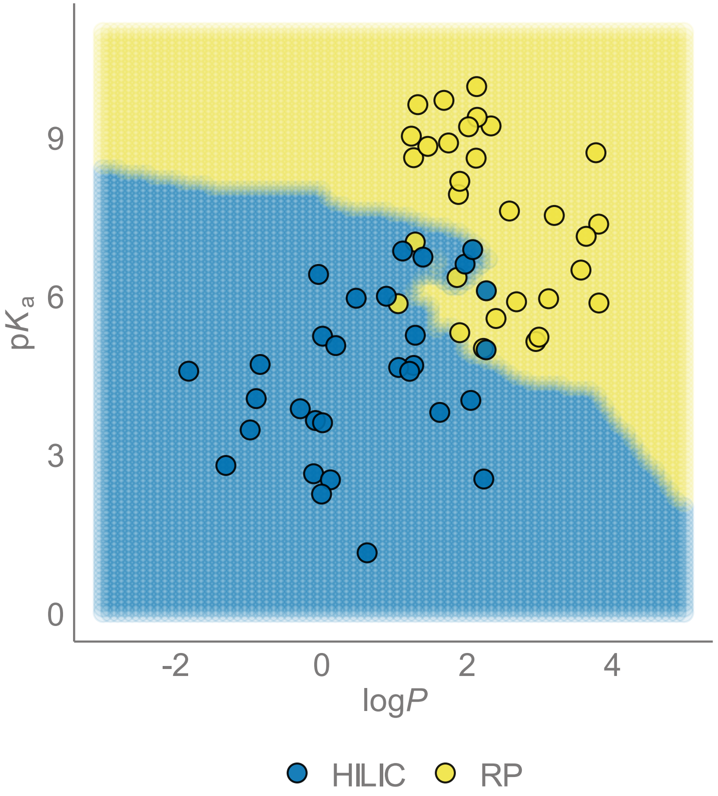
Anneli Kruve

k = 2



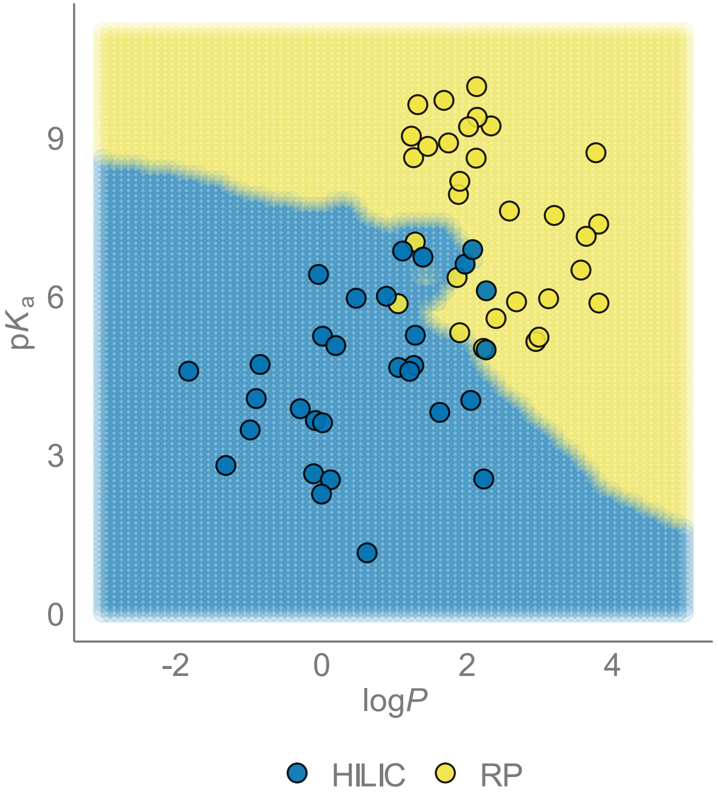
Anneli Kruve

k = 3



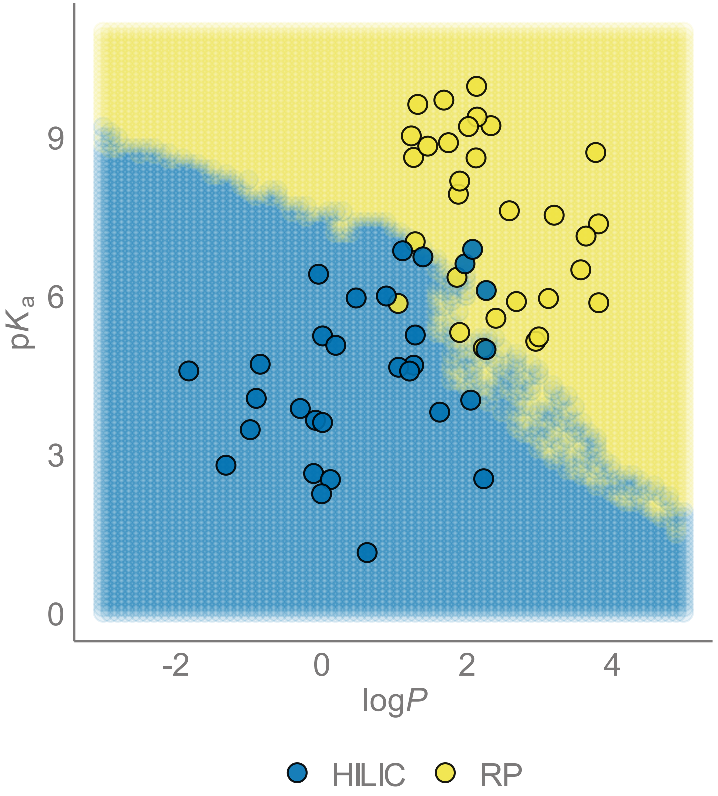
Anneli Kruve

k = 5



Anneli Kruve

k = 10



Anneli Kruve

KNN

Increase in *k* increases robustness & reduces flexibility.
…points far away from the new observation have too much weight.

Good for solving complex non-linear tasks

Variables need to be scaled

Provides NO understanding

Problem if too many variables…
… and if insignificant are in the dataset

Logistic regression

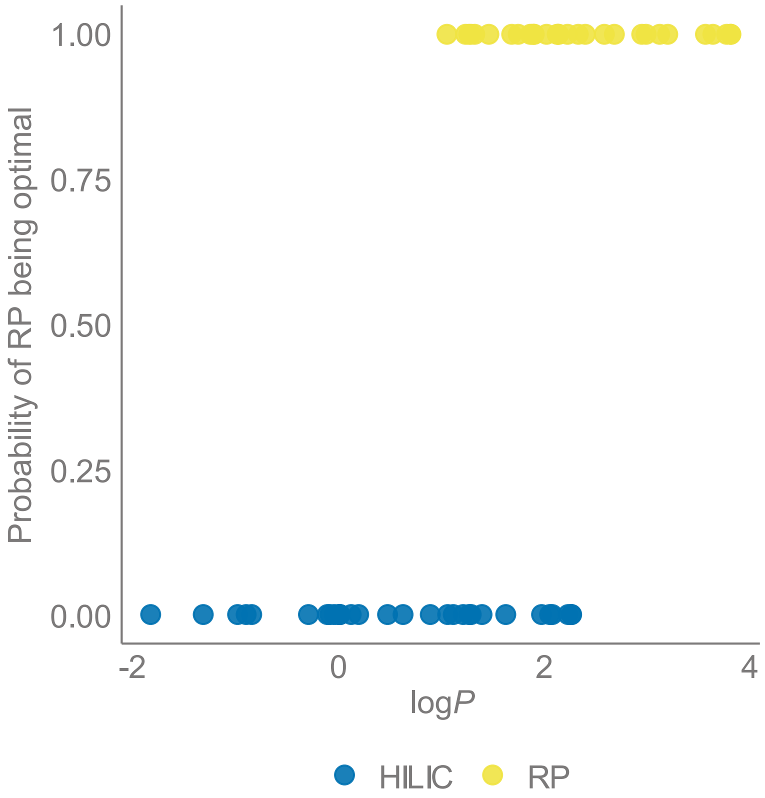
We can simplify the classification problem into a linear regression:
Convert classes to numeric variables

$$Y = \begin{cases} 0\ if\ HILIC \\ 1\ if\ RP \end{cases}$$

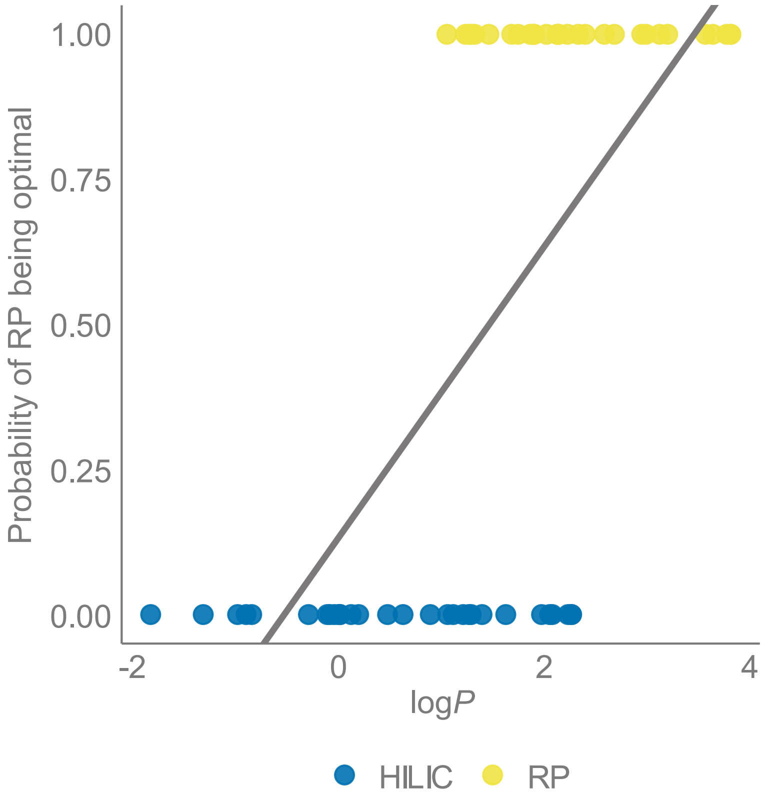
Carry out linear regression to binary response.

Suits well only binary data.

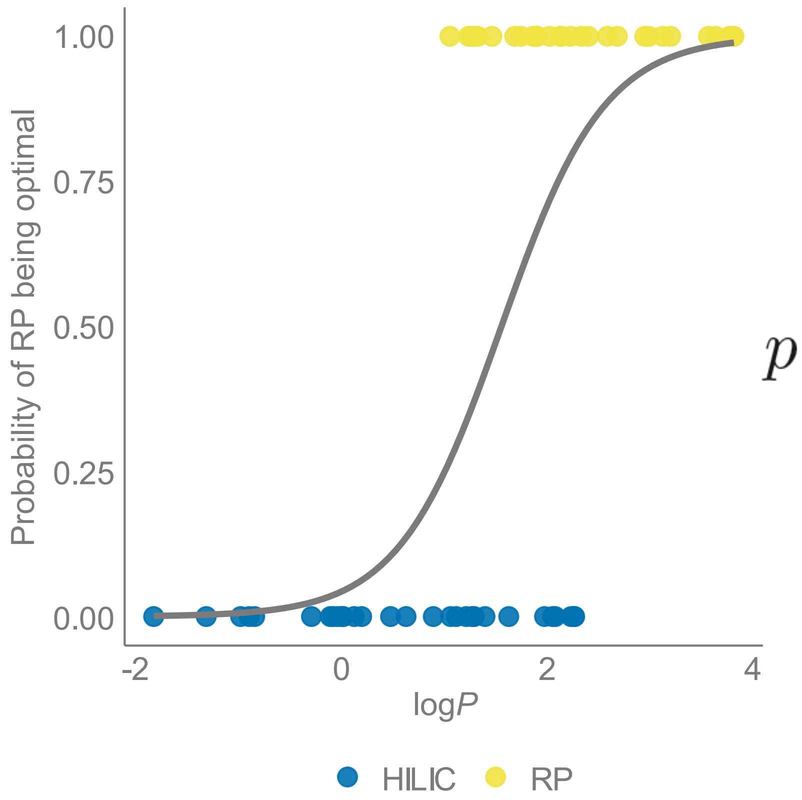
Let's look at the data!



Linear regression



Logistic regression



$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

Anneli Kruve

Making prediction

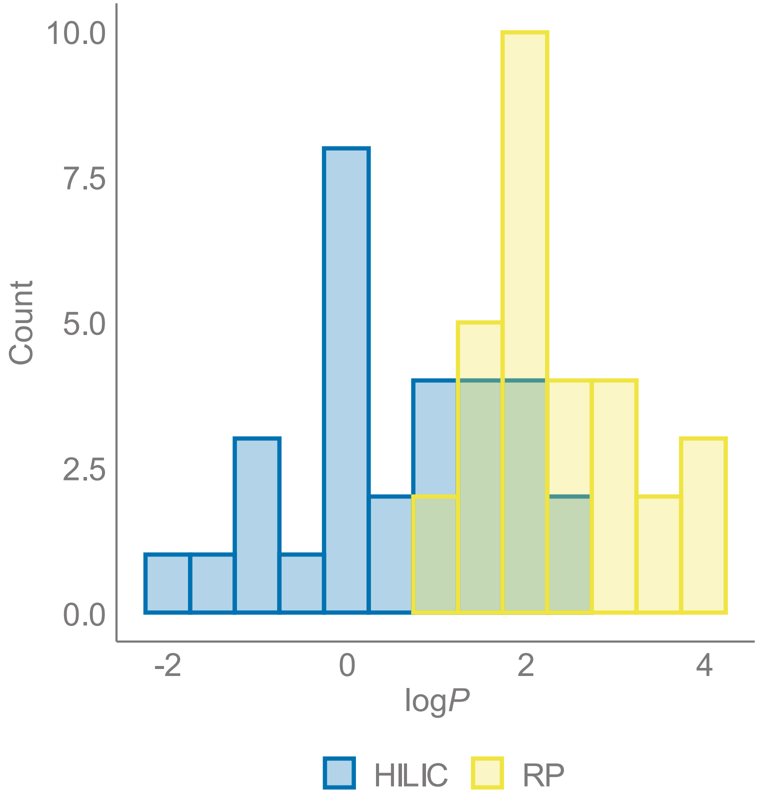
We receive a probability

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1,000}}{1+e^{-10.6513+0.0055\times 1,000}} = 0.00576$$

Probability needs to be converted to the Class!

Linear Discriminant Analysis (LDA)

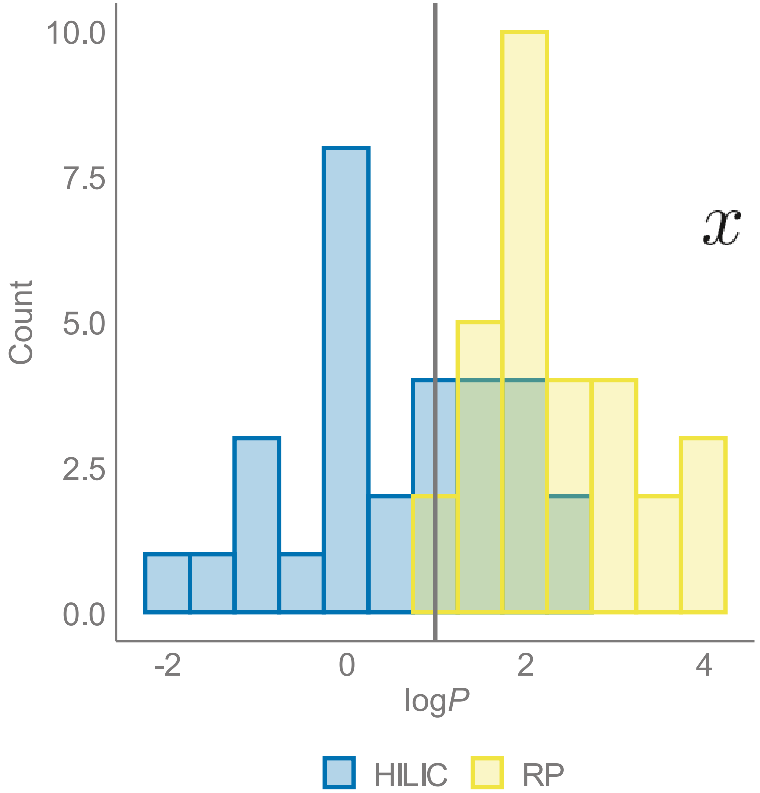
ASSUME WE HAVE ONLY ONE PREDICTOR log*P*



Anneli Kruve

Linear Discriminant Analysis (LDA)

WE CAN INTRODUCE A DECISION BOUNDARY



$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$

Anneli Kruve

LDA

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

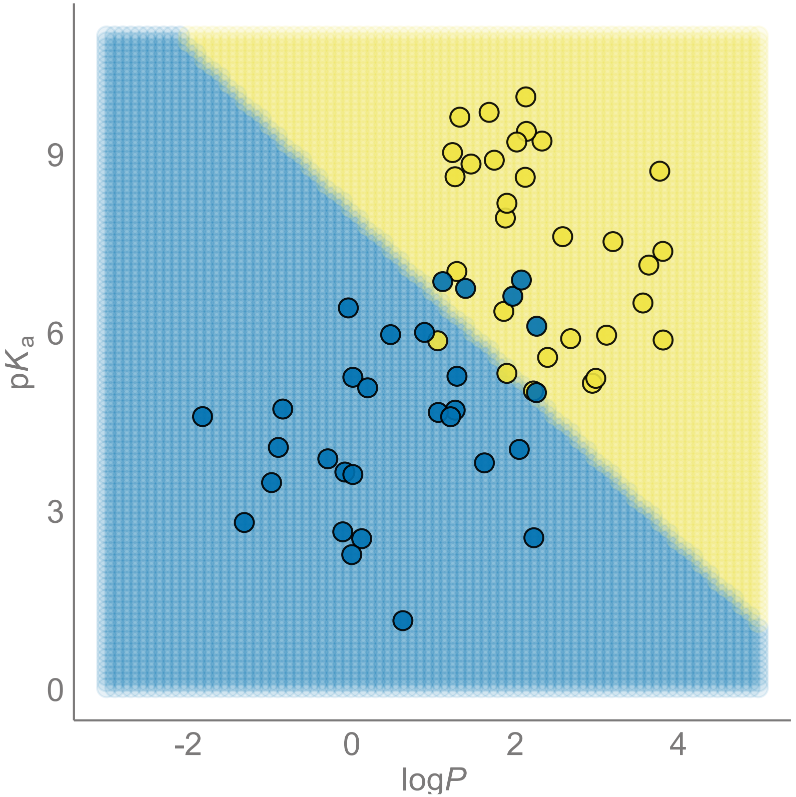
$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

Assigning the observation to the class for which

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

is largest!
We end up with a linear deviation of the data!

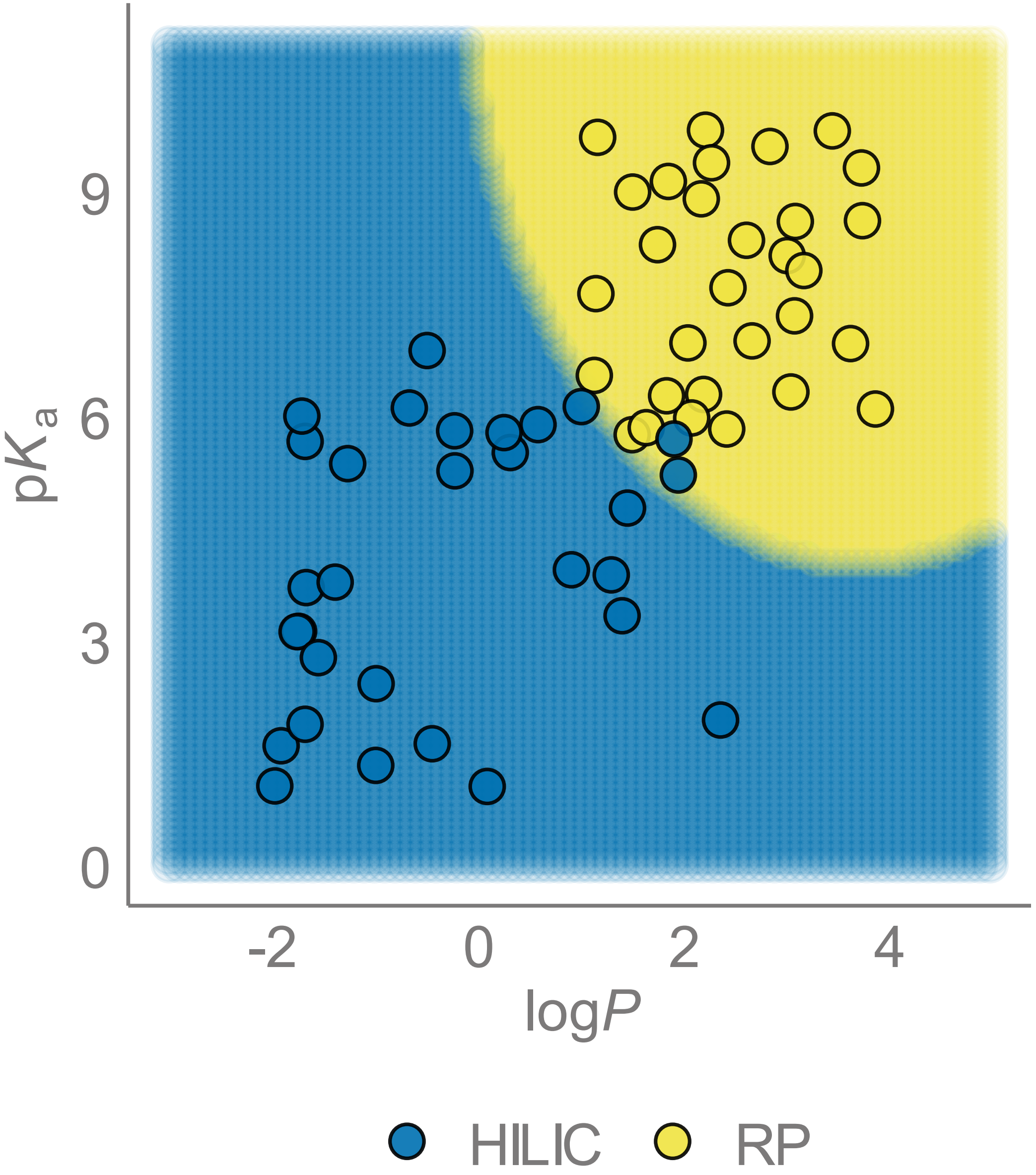
LDA



$$x^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2}\mu_k^T \mathbf{\Sigma}^{-1} \mu_k = x^T \mathbf{\Sigma}^{-1} \mu_l - \frac{1}{2}\mu_l^T \mathbf{\Sigma}^{-1} \mu_l$$

Anneli Kruve

Quadratic Discriminant Analysis



Anneli Kruve

Imbalanced dataset

A dataset that contains significantly more instances from one class then from another

One class may become ignored! Model plays it safe and predicts that all instances come from the over dominated class.

Overcoming:
Obtain more data for the underrepresented class
More measurements of one class?
Throw out data for over represented class (if you have many datapoints)
Multiply the datapoints from the underrepresented class. What are the drawbacks?)