# Tumor detection assignment

## Background

We will look at the data of the paper "Detection and localization of surgically resectable cancers with a multi-analyte blood test" https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6080308/ This paper describes a blood test developed by the authors to detect cancer.

## Data

You will use the table S6 from the paper. The dataset is available as a CSV file on canvas.

## Tasks

Complete the following tasks in R:

1. Make a ROC plot to evaluate the performance of the CancerSEEK model using the columns Tumor.type and CancerSEEK.Logistic.Regression.Score from the data provided by the authors. (Make sure the steps in threshold are 0.01 or smaller.)

Hint: You don't need any additional functions for this. You can build it with the basic R and plot functions that you already learned in the first tutorials. Just look up the definitions of a ROC plot and True and False Positive rate and expand on that.

2. Build your own classifier using K-nearest neighbors. Think about properly scaling and centering your data. Split your data into a training set of 80% of the data and a test set of 20% of the data. Train your model on the training set and use the trained model to predict the class, ie Positive or Negative, of the test set. Report the number of True Postives, False Positives, True Negatives and False Negatives in a confusion table (you can make the table in word using the values from R). Compare your results with the results of the CancerSEEK model.

Hint: You can use the R function knn from the library class.

3. Perform a Principal Component Analysis (PCA) on (the numeric columns of) the data frame. Plot the first and second principal components in a scatter plot. Color the points by whether the samples are healthy (blue) or have a tumor (red).

Hint: You can use the built in R function prcomp to do a PCA.

## Hand in

1. All the R code that you have used (Hint: work in a script, not in the command line).
2. A CSV file containing your training and test set and your prediction for the test set.
3. A written report (500-1000 words) containing all the figures and tables that you made for the tasks, make sure they all have a proper caption. Feel free to add one or two extra figures for potential bonus points (if they are interesting). Describe in the report

the steps that you have taken and what is seen in the figures. Additionally discuss the following points in your report:

1. What type of machine learning do the authors use (what kind of algorithm? And is that supervised or unsupervised machine learning)? Do you approve of this choice?
2. What is overfitting? Do you think there is a big chance of overfitting on this data? What measures could be taken to prevent overfitting?
3. The current models focus on binary classification (Positive or Negative)? Why not create a model that classifies the exact tumor type and stage. What would be the challenges in this case?

## Grading

| | |
|---|---|
| Task 1 | 20% |
| Task 2 | 30% |
| Task 3 | 10% |
| Report | 30% |
| Bonus points | 10% |