# Assignment 2: Regression and classification

1. The dataset associated to this assignment was downloaded from [www.hemnet.se](http://www.hemnet.se/) on 2020-10-18. The data contains information about selling prices of villas in Landvetter that were sold in the past 12 months.
2. Find a linear regression model that relates the living area to the selling price. If you did any data cleaning step(s), describe and explain why you did that or describe why it was not necessary.
3. What are the values of the slope and intercept of the regression line?
4. Use this model to predict the selling prices of houses which have living area 100 m2, 150 m2 and 200 m2. (1.75p a-c)
5. Draw a residual plot. (0.75p)
6. Discuss the results, and what changes could be done to create a model with improved performance. (1.5p)
7. In this question, you will use the Iris data set (“from sklearn.datasets import load\_iris”). [4.5p]
   1. Use a confusion matrix to evaluate the use of decision trees to classify the iris data set. Should you use a train/test set? (1p)
   2. Use k-nearest neighbours to classify the iris data set with some different values for k, and with uniform and distance-based weights. What will happen when k grows larger for the different cases? Why? How about if k gets very large? (2p)
   3. Compare the classification models for the iris data set that are generated by k-nearest neighbours (for the different settings from question b) and by decision trees. Compare confusion matrices for these models and discuss the performance of the various models. (1.5p)
8. Explain why it is important to use a separate test (and sometimes validation) set. (1.5p)

## Submitting work

In each file that you submit, give the names of the people submitting the work. On the first page of the report state how many hours each person spent working on the assignment.

If you upload a zip file, please also upload any PDF files separately (so that they can be viewed more conveniently in Canvas).

Remember, we check for plagiarism and we are obliged to report suspected cases.

## Self-check

Is all the required information on the front page? Have you answered all questions to the best of your ability? Anything else you can easily check? (details, terminology, arguments, clearly stated answers etc.?)

It is important to:

1. Present clear arguments
2. Present the results in a pedagogical way
   1. Should it be table/plot? What kind of plot? Is everything clear and easy to understand?
3. Show understanding of the topics
4. Give correct solutions.
5. Make sure that the code is well commented.
   1. Important parts of the code should be included in the running text and the full code uploaded to Canvas.