UNIVERSITY OF CAPE TOWN
DEPARTMENT OF STATISTICAL SCIENCES
Data Science – Supervised Learning 2019

ASSIGNMENT 2

Due date: Fridday, 2 May 2019 at 12:00
Late submissions will be penalised at 10% per day (pro rata)

INSTRUCTIONS:

- Present your final report as a pdf document. You may use any typesetting software you wish, but I would encourage you to use R-Markdown or LaTeX.
- Provide complete code for both questions under separate headings as an appendix to your write-up. Start each question on a new page.
- You may NOT provide R output interspersed between your answers! Please typeset relevant elements in the output either in-line, or tabulate results formally. Plots are very useful, but use them sparingly – make sure that a given plot is relevant to the question and pertains to text in your answer. Figures are meant to enrich your analysis, don't leave it to the reader to analyse. Provide captions for all figures and tables. Square figures only!
- When you typeset R code use `courier` or an equivalent 'typewriter'-like font.
- You are expected to work on this **on your own**. Please attach a plagiarism declaration to your report – a template is provided on Vula.

PAGE BLOCKS CLASSIFICATION

This problem concerns classifying all the blocks of the page layout of a document that have been detected by a segmentation process. This is an essential step in document analysis in order to separate text from graphic areas. The five classes are:

1 text

2 horizontal line

3 picture

4 vertical line

5 graphic

There are 4925 examples which come from 54 distinct documents; each observation concerns one block. The attributes (all numeric) are as follows:

- height: Height of the block.
- length: Length of the block.
- area: Area of the block (height $\times$ length).
- eccen: Eccentricity of the block (length / height).
- p_black: Percentage of black pixels within the block (blackpix / area).
- p_and: Percentage of black pixels after the application of the Run Length Smoothing Algorithm (RLSA) (blackand / area).
- mean_tr: Mean number of white-black transitions (blackpix / wb_trans).
- blackpix: Total number of black pixels in the original bitmap of the block.
- blackand: Total number of black pixels in the bitmap of the block after the RLSA.
- wb_trans: Number of white-black transitions in the original bitmap of the block.

Your assignment should implement "Support Vector Machines" and "Neural Networks" for classification. Write up a short report comparing the various strategies that you have used. Make sure to include:

- A short description of each technique.
- A comparison of their performance.
- Motivate your choice of regularisation mechanism and hyper-parameters.
- Do not forget to generate predictions for the test data! You will be required to hand in all materials (R-scripts, R-Markdown and/or write-up) and a .csv file giving your test data classifications.