موسسه آموزش عالی آزاد توسعه

برگزار کننده دورههای تخصصی علم داده

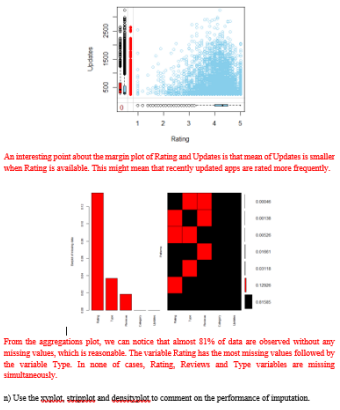
Homework 7: Summer 2020

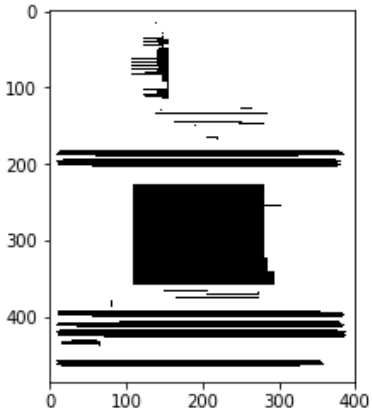
Due date: ۵ مرداد

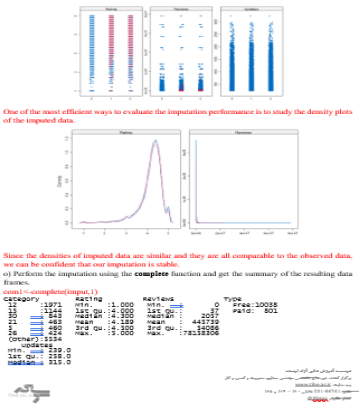
Please email your HWs to y.zerehsaz@gmail.com
Please append all your codes to your response

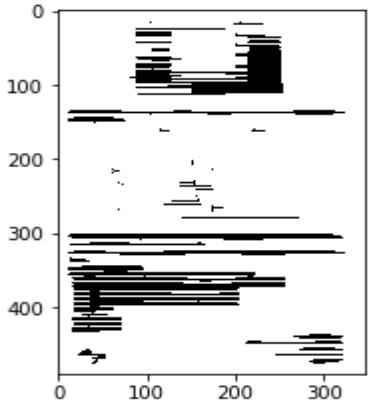
This HW must be completed in **Python**.

The dataset blocks.csv contains various measurements on blocks that are part of a page layout in a document. The blocks are the output of a segmentation algorithm which divides up the pages into relatively homogeneous images. Classifying such blocks is an essential step in document analysis, allowing one to separate text from graphics. There are several approaches to perform image segmentation. The run length smoothing algorithm is one of the classic methods to segment document layouts. The method tries to draw some blocks (ROIs) around different segments in a page. For your information, the function **rlsa** in package **pythonRLSA** can be used to obtain these blocks. As an example, I used this package to segment the page layout of one of your homework solutions, and the results are shown below (YOU DO NOT NEED TO DO SEMENTATION. IT IS DONE FOR THIS DATASET):









After obtaining the blocks, a classification method is applied to the data to determine the label of each of these blocks. The five classes (types of blocks) are: (1) text, (2) horizontal line, (3) picture, (4) vertical line, (5) graphic. The features that are used for classification are: (1) height of the block (2) length of the block (3) area of the block (height*length) (4) eccentricity of the block (length/height) (5) % of black pixels within the block (6) % of black pixels after smoothing the block (7) mean number of black-white transitions in the block (8) total number of black pixels in the original bitmap of the block (9) total number of black pixels in the smoothed bitmap of the block (10) number of black-white transitions in the original bitmap of the block. Bitmap means that the images should be presented by 0 and 1 binaries and smoothing means denoising the images.

Note: The first column of the dataset is the response variable which needs to be predicted. However, the response variable in the dataset is non-numerical and we, first, need to transform it to a numerical variable.

Apply LDA, QDA, Naïve Bayes, logistic regression and KNN methods on the data.

a) Leave 10% of the data for testing and use 90% for training the models.

b) For KNN, you need to perform a ten-fold cross-validation and select the best k. (repeat only once).

c) Compare all methods using the average test accuracy (repeat 1000 times and compute the average test accuracy).

d) Compute the confusion matrix and comment on the results. Which method gives you the best result? (Repeat only once)

e) Based on the results in Part c, select your best method. Apply PCA on the data. Use cross-validation to select the best value for the number of PC scores (components). Compare the performance of the method without PCA with the PCA-based approach using the average test accuracy (1000 replications).