The Problem

The problem consists in classifying all the blocks of the page layout of a document that has been detected by a segmentation process.
In this case we use the Run Length Smoothing Algorithm (RLSA), a top-down approach to document segmentation.
It means that the RLSA first subdivide the area of a document into regions (blocks), and then calculate some of their basic features (attributes).
This is an essential step in document analysis in order to separate text from graphic areas.

The five classes of blocks are:

- Text
- Horizontal line
- Picture
- Vertical line
- Graphic

Attributes

Each block is defined by 10 different attributes :

- **height** (integer) : Height of the block
- **length** (integer) : Length of the block
- **area** (integer) : Area of the block (height * lenght)
- **Eccen** (continuous) : Eccentricity of the block (lenght / height)
- **p_black** (continuous) : Percentage of black pixels within the block (blackpix / area)
- **p_and** (continuous) : Percentage of black pixels after the RLSA (blackand / area)
- **mean_tr** (continuous) : Mean number of white-black transitions (blackpix / wb_trans)
- **blackpix** (integer) : Total number of black pixels in the original bitmap of the block
- **Blackand** (integer) : Total number of black pixels in the bitmap of the block after the RLSA
- **wb_trans** (integer) : Number of white-black transitions in the original bitmap of the block

My thoughts on the problem

The question of how to classify page blocks is really important in today’s world because of the important quantity of documents being digitized everyday.
It is thus impervious that we automatize their transformation into a collection of information to be stored, classified, retrieved, combined, and updated.

Variables created

- **ColumnsNames** and **DataBlocks** to stock the dataset
- **corr** to see the correlation between variables (given in the dataset description)
- **tab01**, **tab02**, **tab1** and **tab2** to compare the number and percentage of black pixels before (**tab01** and **tab1**) and after (**tab02** and **tab2**) the RLSA for each class
- **ax0**, **ax1**, **ax2** and **ax3** to help plot graphics
- **tab3** to see the number of black pixels for each white-black transition
- **X** and **y** to separate the dataset between the attributes (**X**) and the target (**y**)
- **X_train**, **y_train**, **X_test** and **y_test** to separate the training set and test set
- **param_grid1**, **grid1** and **model1** to make a model using grid search with KNN
- **param_grid2**, **grid2** and **model2** to make a model using grid search with decision tree
- **param_grid3**, **grid3** and **model3** to make a model using grid search with SVM
- **cf**, **b** and **t** to stock the confusion matrix (**cf**) and prevent the final graphic from being cut at the top and bottom (**b** and **t**).

How the problems fit in the context of the study

The study consists in finding the best methodology for document recognition.
The study detail several metodologies using top-down and bottom-up approach, like “projection profil cut”, “neiborhood line density” and “connected component analysis”.
The problem focuses only of the RLSA method of segmentation and discrimination.