LOVELY PROFESSIONAL UNIVERSITY

INT-247 CA-2 PROJECT –

(PAGE BLOCKS CLASSIFICATION)

Name – Navneet Bhargava
Reg_no:11804667
Section – KM007
Data set - Page Blocks Classification

Introduction:-

Study of structure and content of documents is done by using document analysis. Document Analysis uses a technique named text Segmentation. Text segmentation is the technique of dividing the document on the basis of words, topics or sentences.. Hence there is a problem to classify all the blocks of page layout of document that was detected by segmentation process. This is the main part of document analysis in order to separate text from graphics. In this study, we used rough set approach to classify all the blocks in a page.

.

Problem statement:

The problem consists in classifying all the blocks of the page layout of a document that has been detected by a segmentation process
This is an essential step in document analysis in order to separate text from graphic areas. Indeed, the five classes are:
text (1)
horizontal line (2)
picture (3)
vertical line (4)
graphic (5)
Relevant Information Paragraph: The 5473 examples comes from 54 distinct documents. Each observation concerns one block. All attributes are numeric

Cleaning and feature selection of Data-set

Firstly I started with renaming the columns with proper names after that I checked for null values in the data-set. Since there is no null values, I checked for special characters and I found that in some of the columns. After that I assigned NaN to all the special characters and later on I dropped them from the datset.

PAGE BLOCK CLASSIFICATION

The problem related to text segmentation is to classify
all blocks of page that has been detected by segmentation process. On some documents, text lines are associated together by the block segmentation procedure due to minor line spacing.

Data Set Used

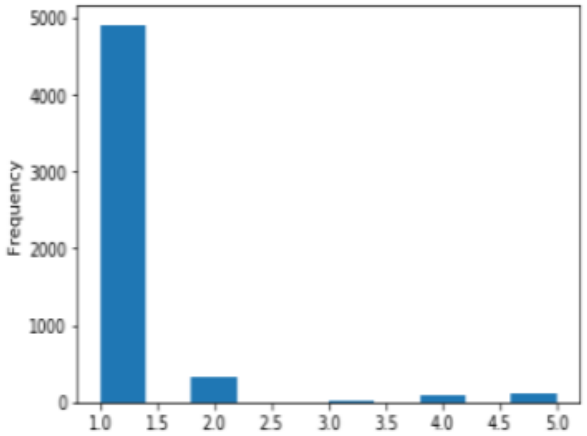
Number of Attributes :

height: integer. | Height of the block.
lenght: 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 application of the Run Length Smoothing
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.

Data Visualization

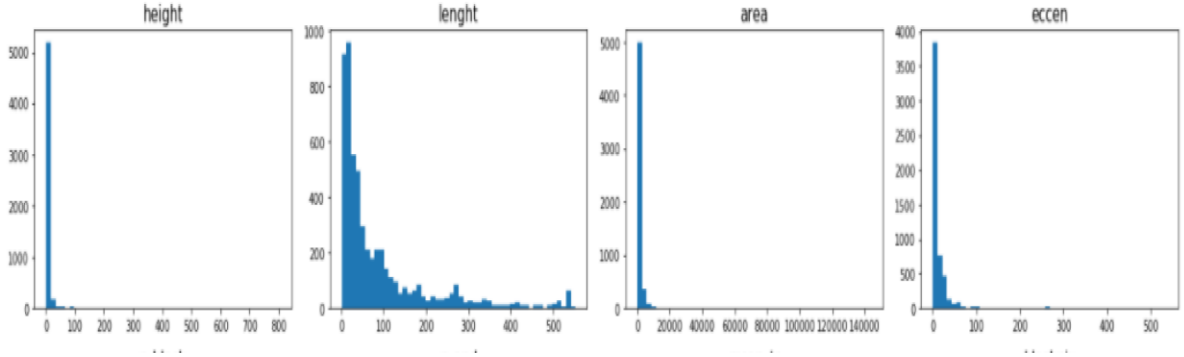
```
In [22]: df['class'].plot(kind='hist') # highly imbalanced dataset
```

```
Out[22]: <matplotlib.axes._subplots.AxesSubplot at 0x1d8d10810b8>
```



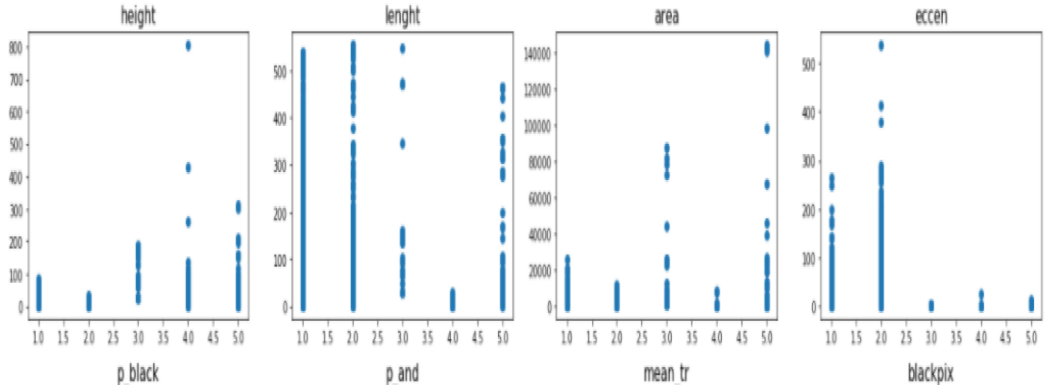
Distribution of all attributes

```
[23]: i=1
plt.figure(figsize=(20,10))

for col in df.columns:
    plt.subplot(3,4,i)
    plt.hist(df[col],bins=50)
    plt.tight_layout()
    plt.title(col,fontsize=15)
    i+=1
```



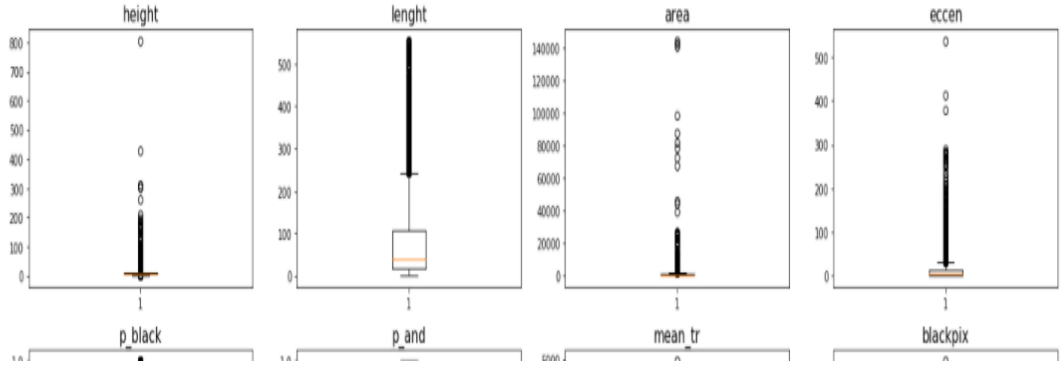
Relationship between other attributes with target attribute

In [24]:
```
i=1
plt.figure(figsize=(20,10))

for col in df.drop(columns='class').columns:
    plt.subplot(3,4,i)
    plt.scatter(df['class'],df[col])
    plt.tight_layout()
    plt.title(col,fontsize=15)
    i+=1
```



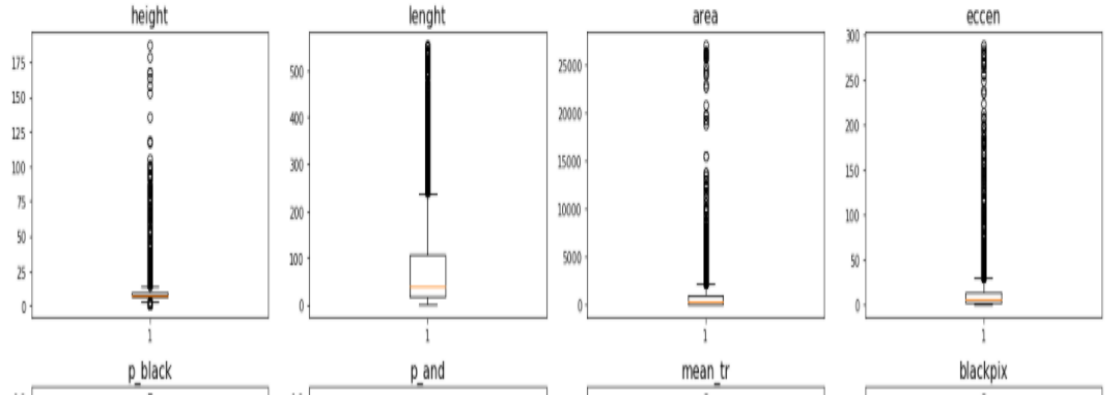
Boxplot of all attributes (Outlier Detection)

In [25]:
```
i=1
plt.figure(figsize=(20,10))

for col in df.columns:
    plt.subplot(3,4,i)
    plt.boxplot(df[col])
    plt.tight_layout()
    plt.title(col,fontsize=15)
    i+=1
```



After Removing Outliers

In [27]:
```
i=1
plt.figure(figsize=(20,10))

for col in df.columns:
    plt.subplot(3,4,i)
    plt.boxplot(df[col])
    plt.tight_layout()
    plt.title(col,fontsize=15)
    i+=1
```



Gaussian NB

linkcode

K-Fold :

That method is known as “k-fold cross validation”. It’s easy to follow and implement. Below are the steps for it:

1. Randomly split your entire dataset into k”folds”
2. For each k-fold in dataset, build model on k – 1 folds of the dataset. Then, test the model to check the effectiveness for kth fold
3. Record the error on each of the predictions
4. Repeat this until each of the k-folds has served as the test set
5. The average of k recorded errors is called the cross-validation error and will serve as performance metric for the model

Stratified K-Fold

Stratification is the process of rearranging the data so as to ensure that each fold is a good representative of the whole.
It is generally a better approach when dealing with both bias and variance. A randomly selected fold might not adequately represent the minor class, particularly in cases where there is a huge class imbalance.

Splitting Data and Model selection

After identifying feature that to be trained and tested I have splitted the data in the manner that 70% to be trained and 30% to be tested.

RESULTS

Using rough set, in page block classification, number of rules generated for objects are 95.It means only 95 rules are sufficient to classify page blocks generated the process. Table shows the quality of classification results. Quality of classification is 0.9959.The existing result has proved the suitability of rough set approach for the analysis of information
system having different domains

After Traning and testing the data-set on three different models we got accuracy 82%, 75% and 71% respectively. Hence, our first model Random forest classifier is best for our data-set.

.

Prediction on Unknown data

After Doing all the things from scratch , I have tested My best model which is Random Forest Classifier on unknown data. In this I will take data from user,

ADVANTAGES

Gives possibility for data analyst to control data analysis process in simpler way.

Results are not adulterated by subjective indirect evaluation or arbitrary chosen operators.

Conclusion

In final words, firstly I fetch the data-set cleaned it and made that perfect, later on trained on three different models and got 82%, 75% and 75% accuracy respectively.

i applied these methods on a dataset consisting of 5473 example coming from 54 distinct documents. I found that only 95 rules are sufficient to classify page blocks generated by segmentation process and the quality of classification is 0.9959.