PAGE BLOCKS CLASSIFICATION

Presented by Shangzhi LOU, Wenbo SUI and Limin TIAN

CONTENT

DEFINITION OF THE PROBLEM

The problem consists of classifying all the blocks of the page layout of a document that has been detected by a segmentation process. The goal is to classify each block into one of five classes.

DEFINITION OF THE PROBLEM

The different variables

- height: Height of the block.
- lenght: Length of the block.
- area: Area of the block (height * lenght);
- eccen: Eccentricity of the block (lenght / 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.

RESEARCH METHODOLOGY

The description of the dataset

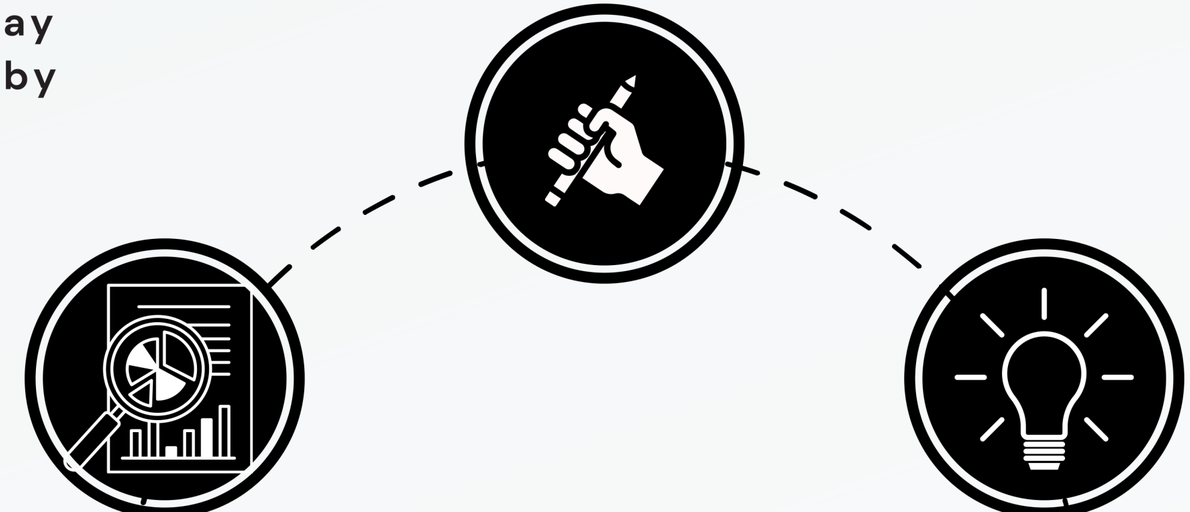
- 5473 examples,5 classes.
- All attributes are numeric.
- The data is formatted in a way that is readable by C4.5.

Dataset preprocessing

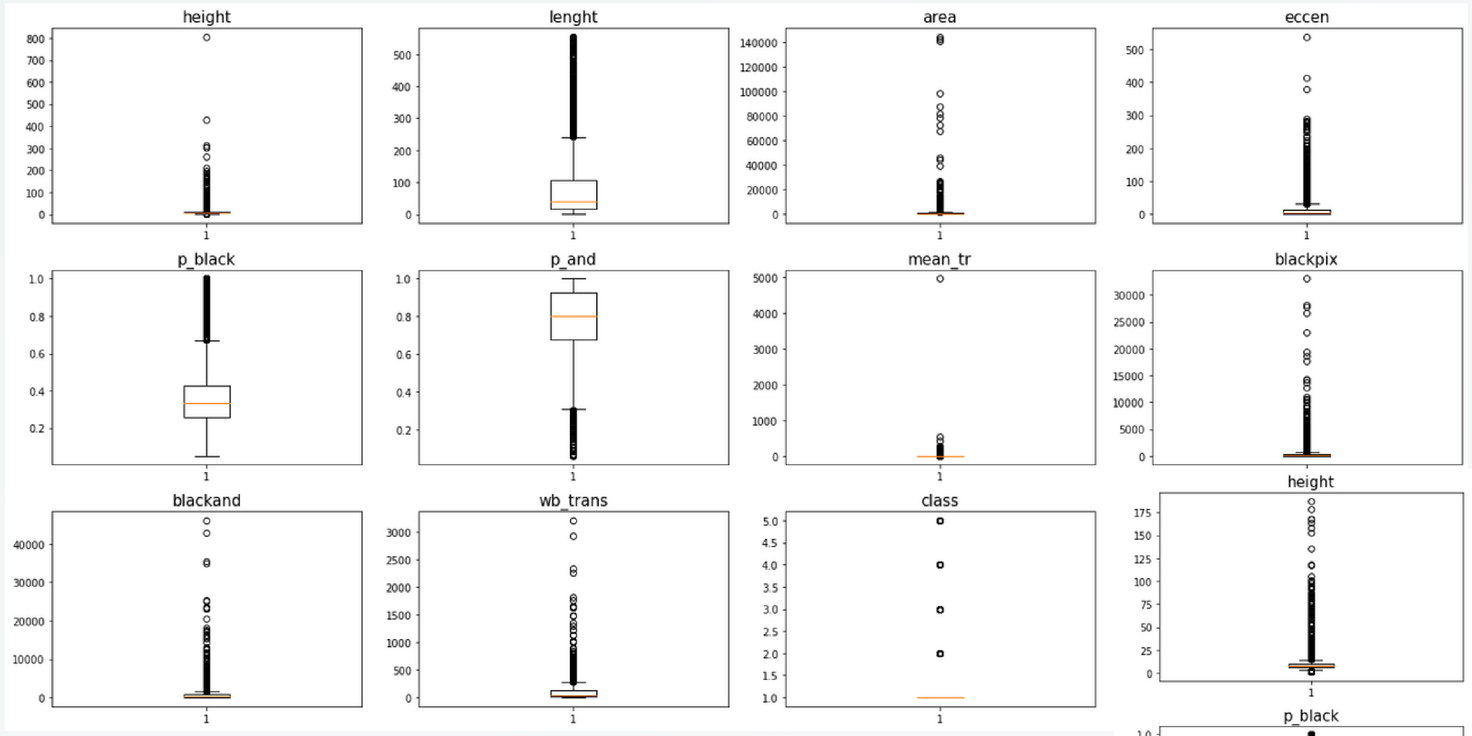
- Check the missing values
- Outlier Detection
- Normalization (z-score)

Analysis methods

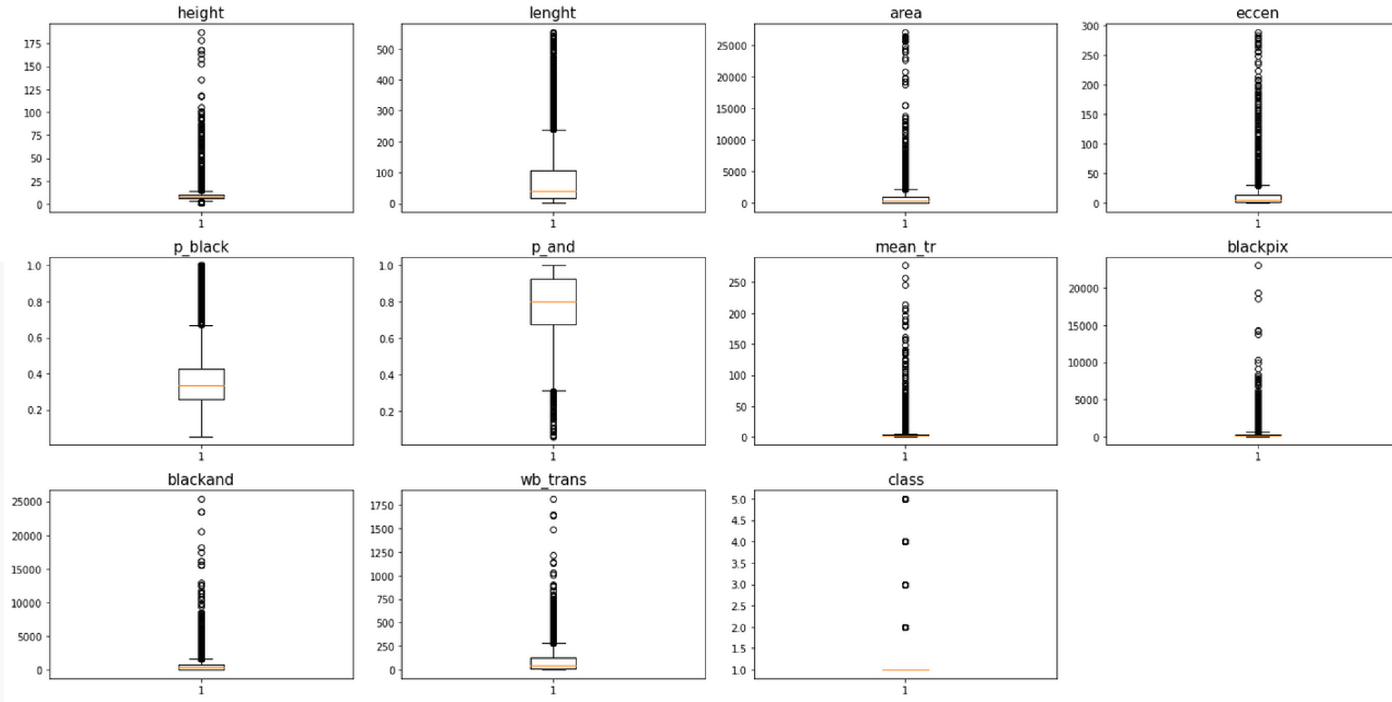
- GaussianNB
- StratifiedKFold



VISUALIZATIONS



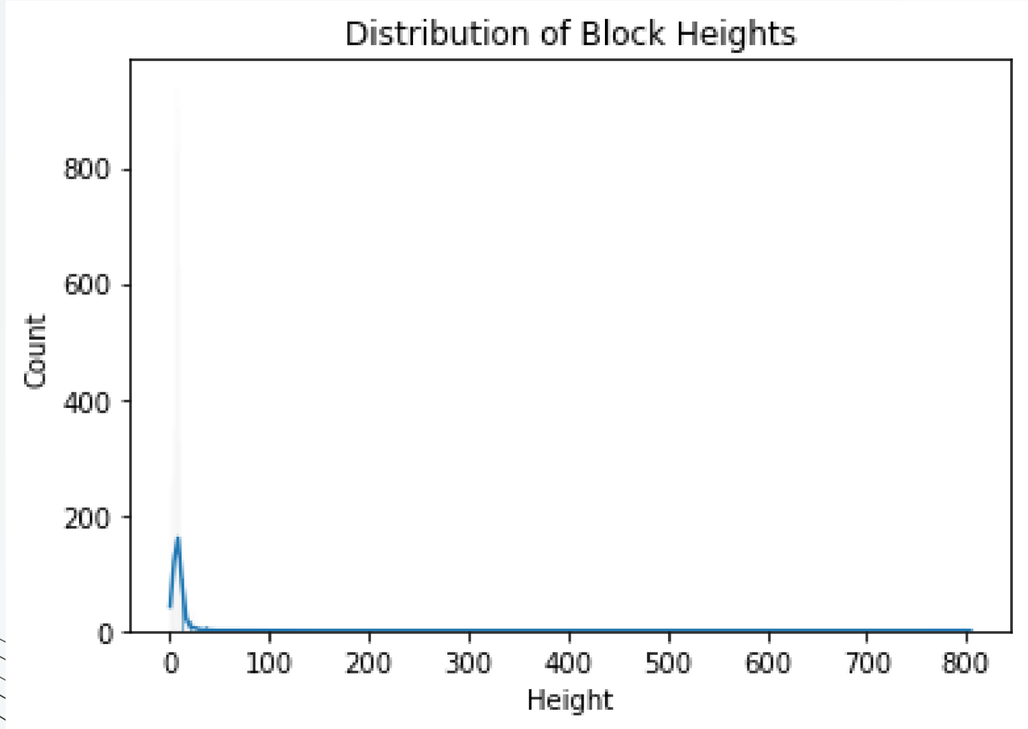
Outlier Detection

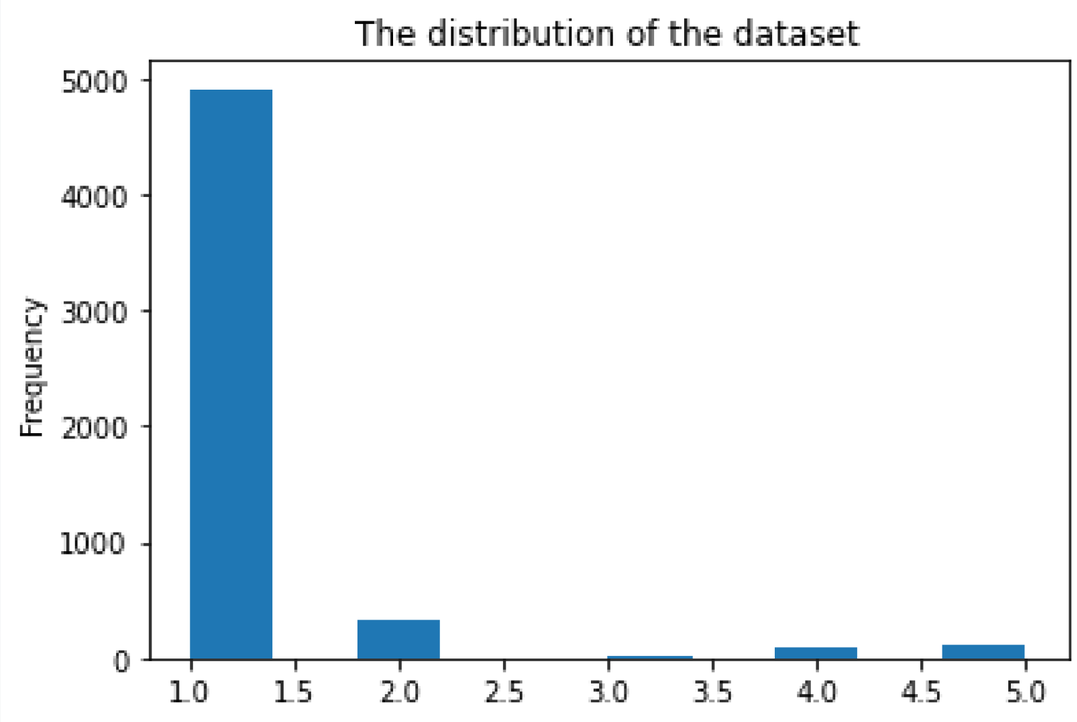


EXPLORATORY DATA ANALYSIS

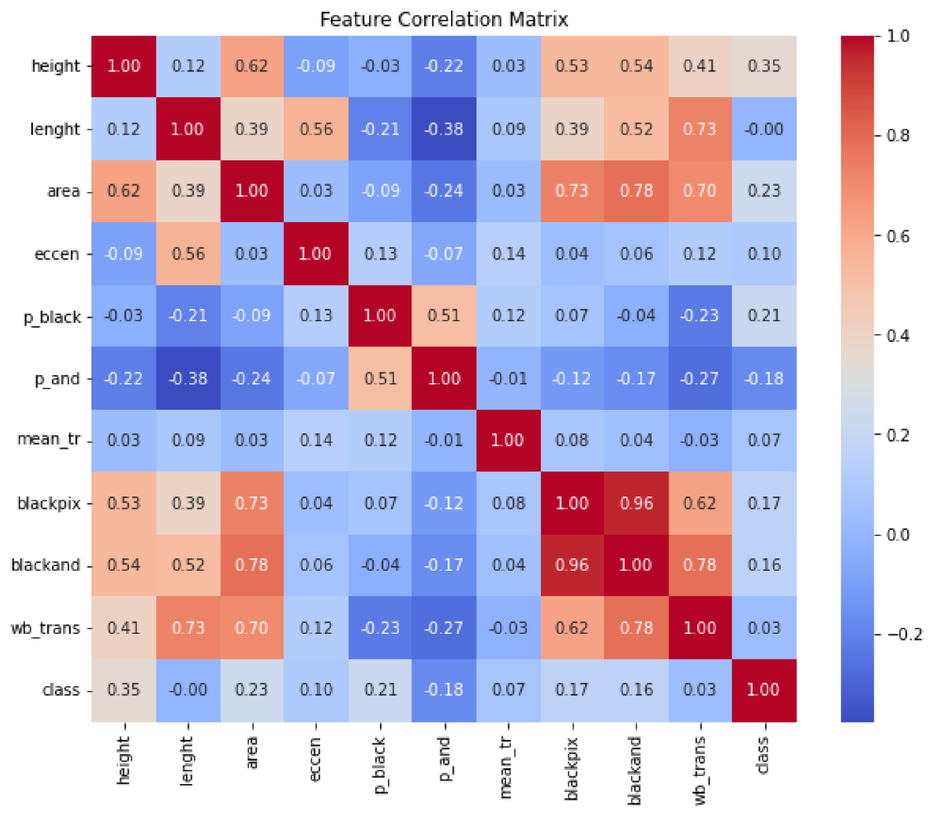
- Distribution of Block Heights
- The distribution of the dataset
- Correlation matrix heatmap between features
- Distribution of all attributes
- The relationship between attributes with target attribute

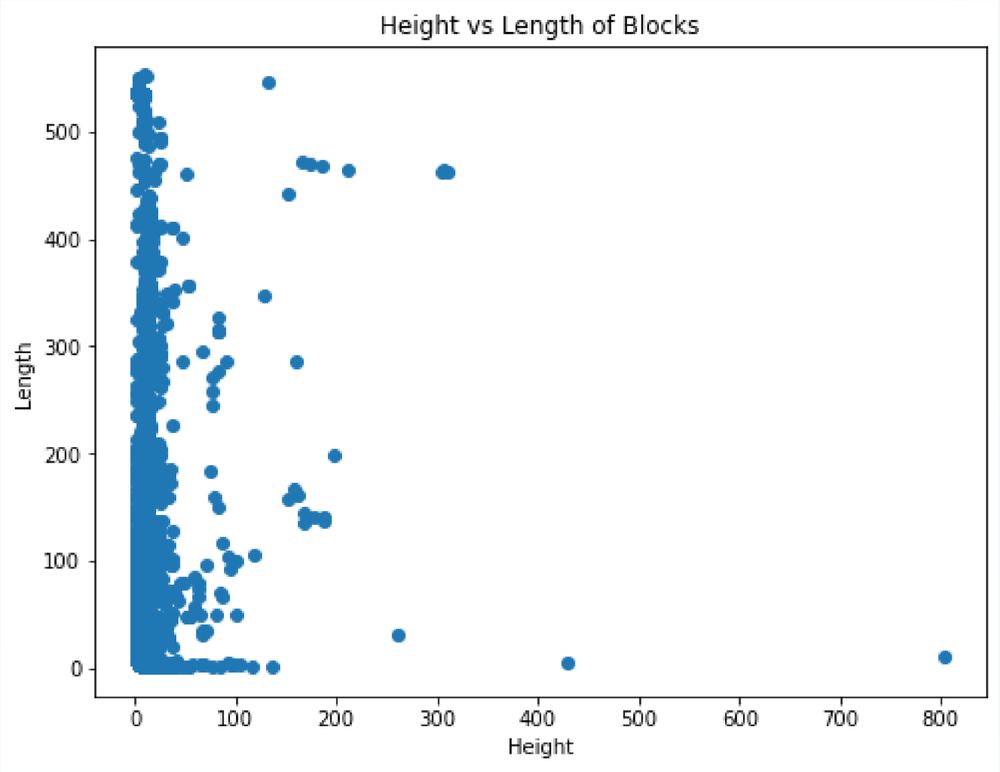
VISUALIZATIONS



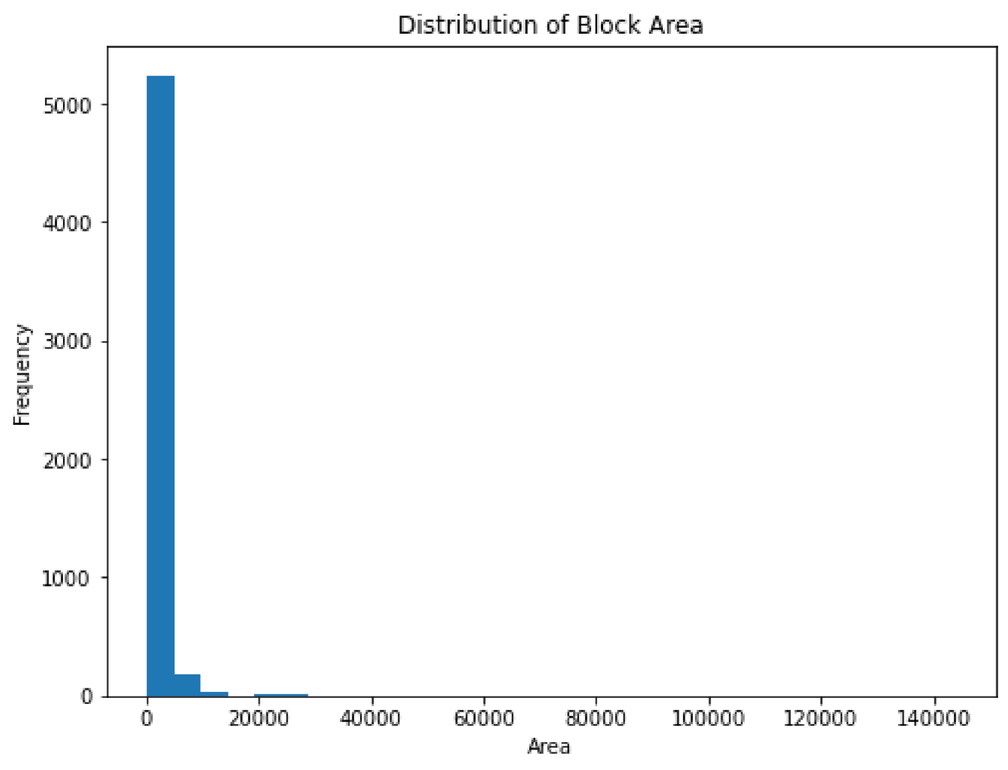


VISUALIZATIONS

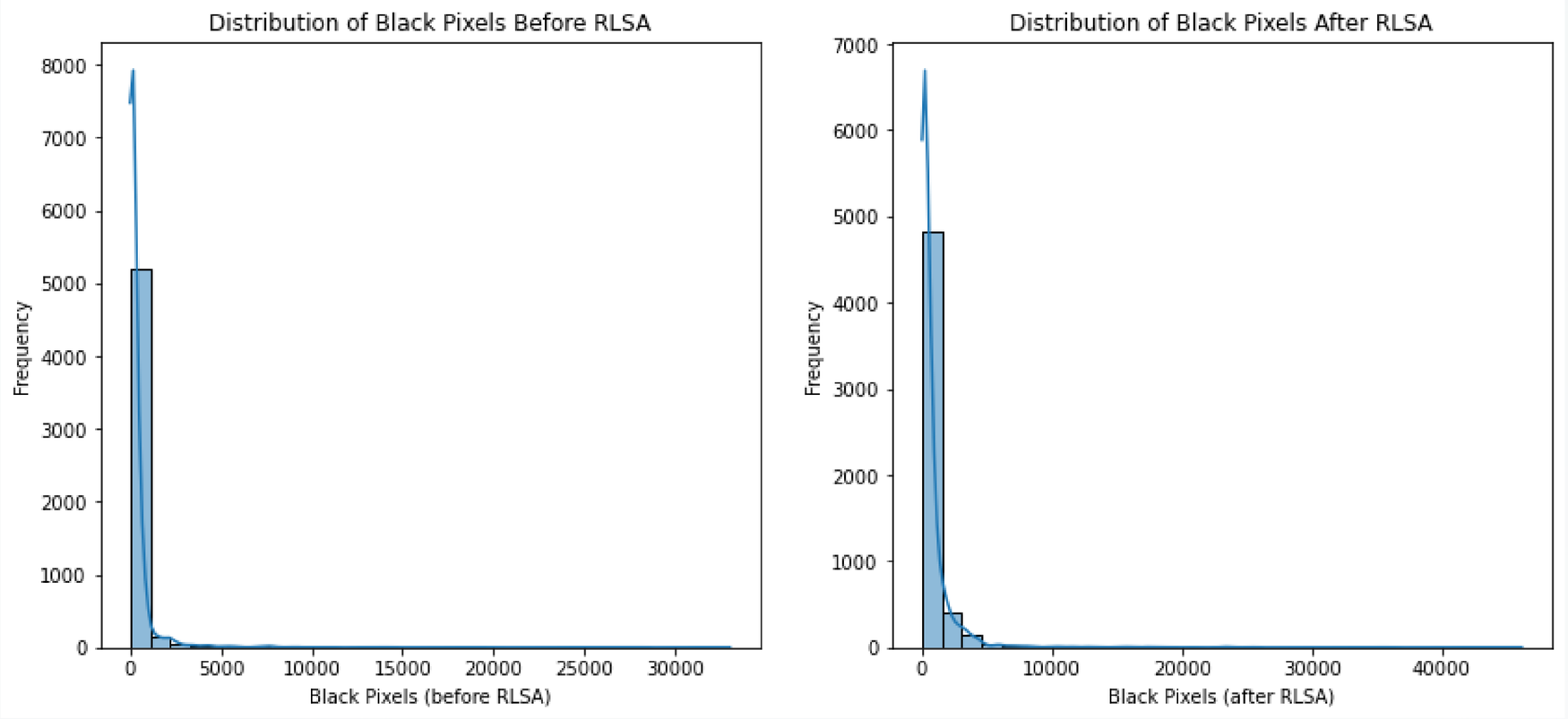




VISUALIZATIONS

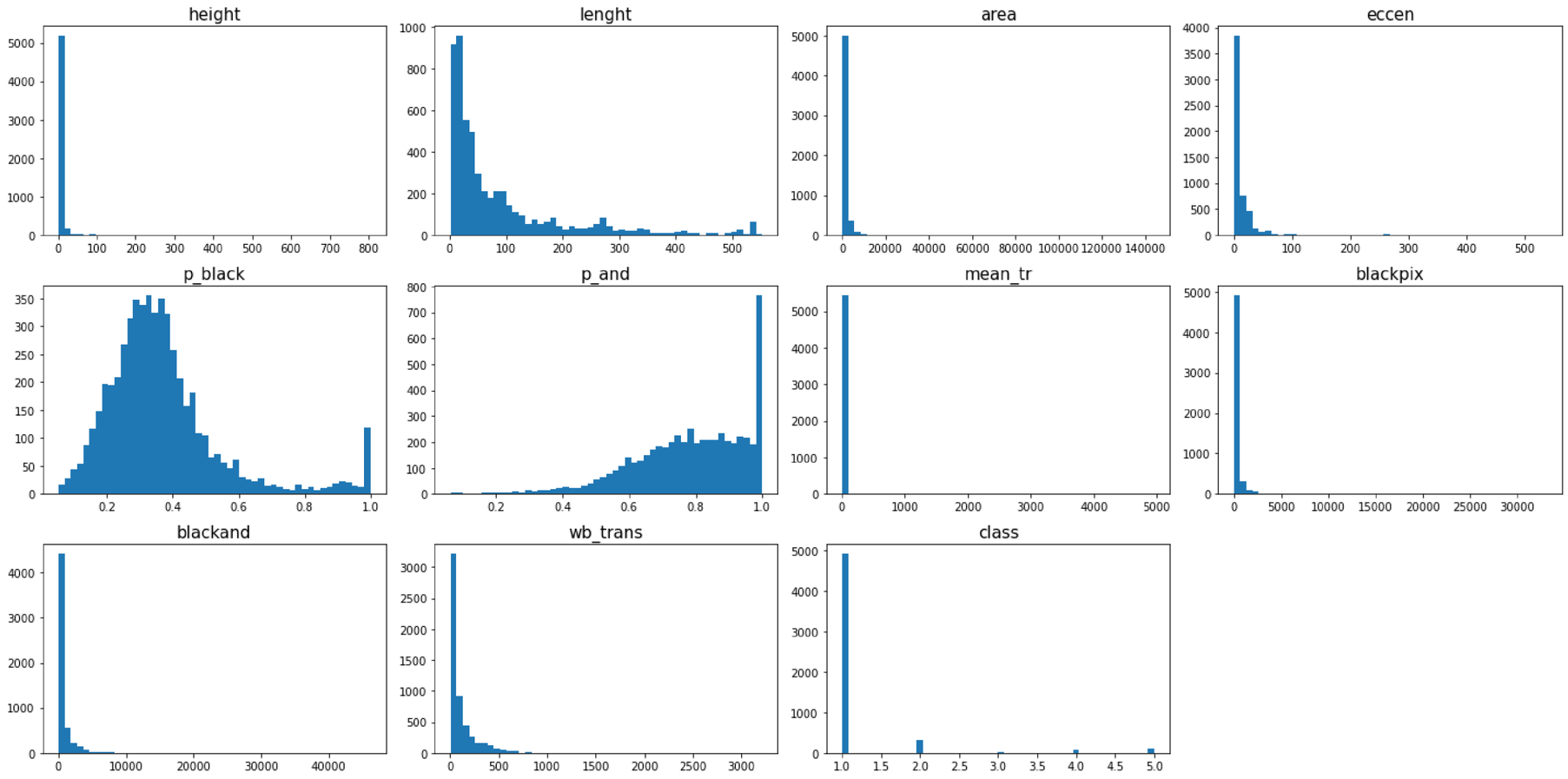


VISUALIZATIONS



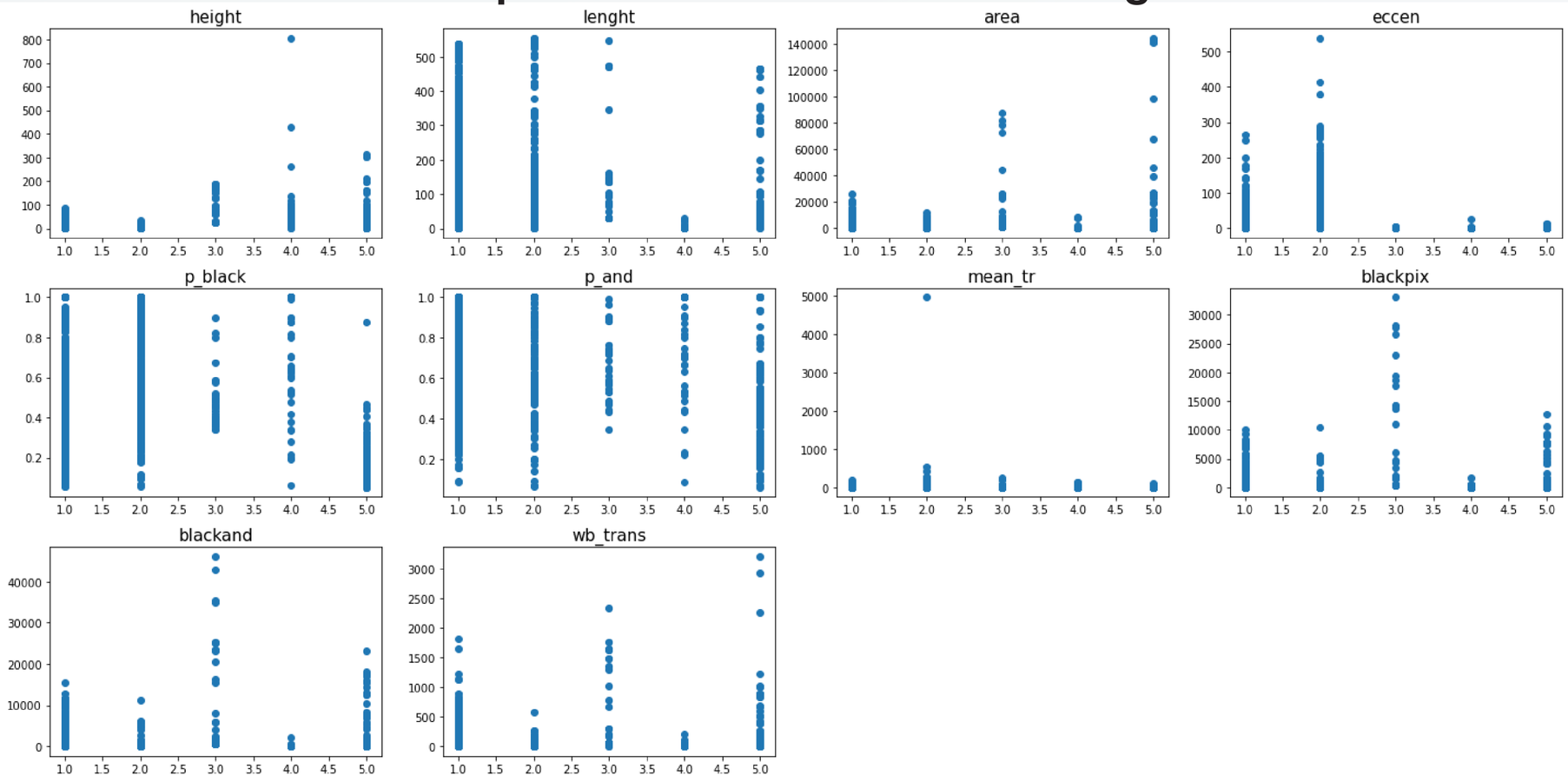
VISUALIZATIONS

Distribution of all attributes



VISUALIZATIONS

The relationship between attributes with target attribute



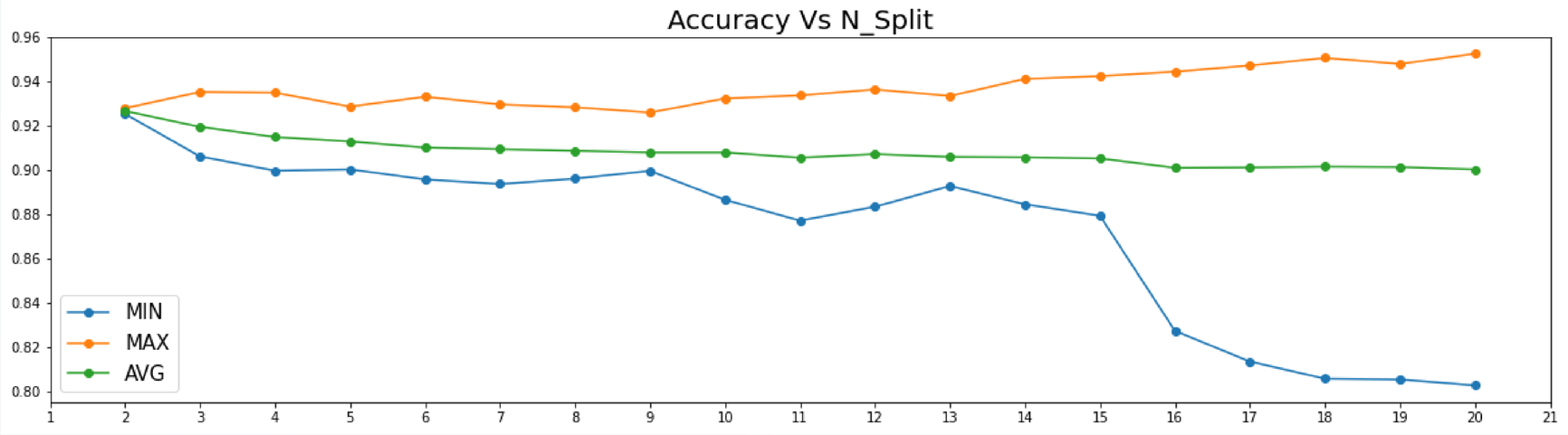
OBSERVATION OF THE DATASET

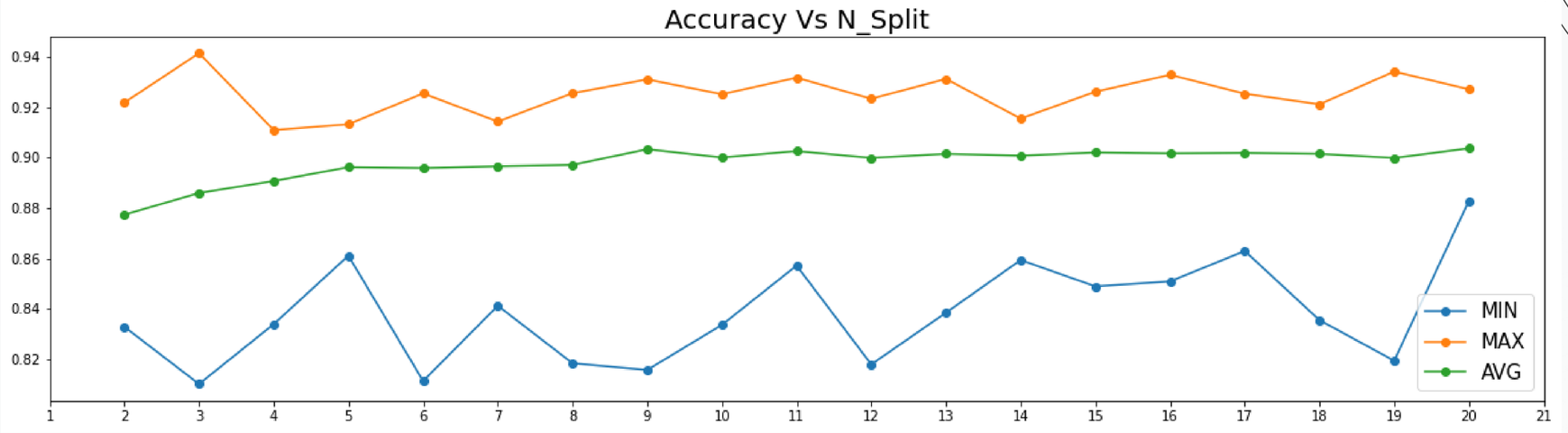
- The dataset is highly imbalanced,most of examples belong to class(1).
- Most blocks have a small height, suggests that blocks of smaller height are more common.
- The area of the block has a high correlation with height, length, blackpix, and blackand.
- The blackpix has a high correlation with blackand.

MODEL BUILDING AND EVALUATION

VISUALIZATIONS

CALCULATE THE CROSS VALIDATION SCORE ON DIFFERENT SPLIT POINT IN K-FOLD





RESULTS AND DISCUSSION

- Using GaussianNB, the model achieves an average accuracy of approximately 88%.
- In the 5-fold and 10-fold cross-validation, the interval between the minimum and maximum accuracy values is not large, indicating that the model has a certain stability.
- Try to create new features or transform existing features to improve the performance of the model.
- Handle imbalanced data sets through oversampling or undersampling

THANK'S FOR WATCHING