model correlation pixel mean letter data problem content

LETTER RECOGNITION

model correlation pixel mean letter data problem

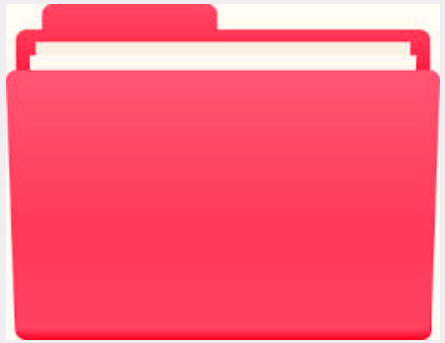
content

model
correlation
pixel
mean
letter
data

Letter Recognition

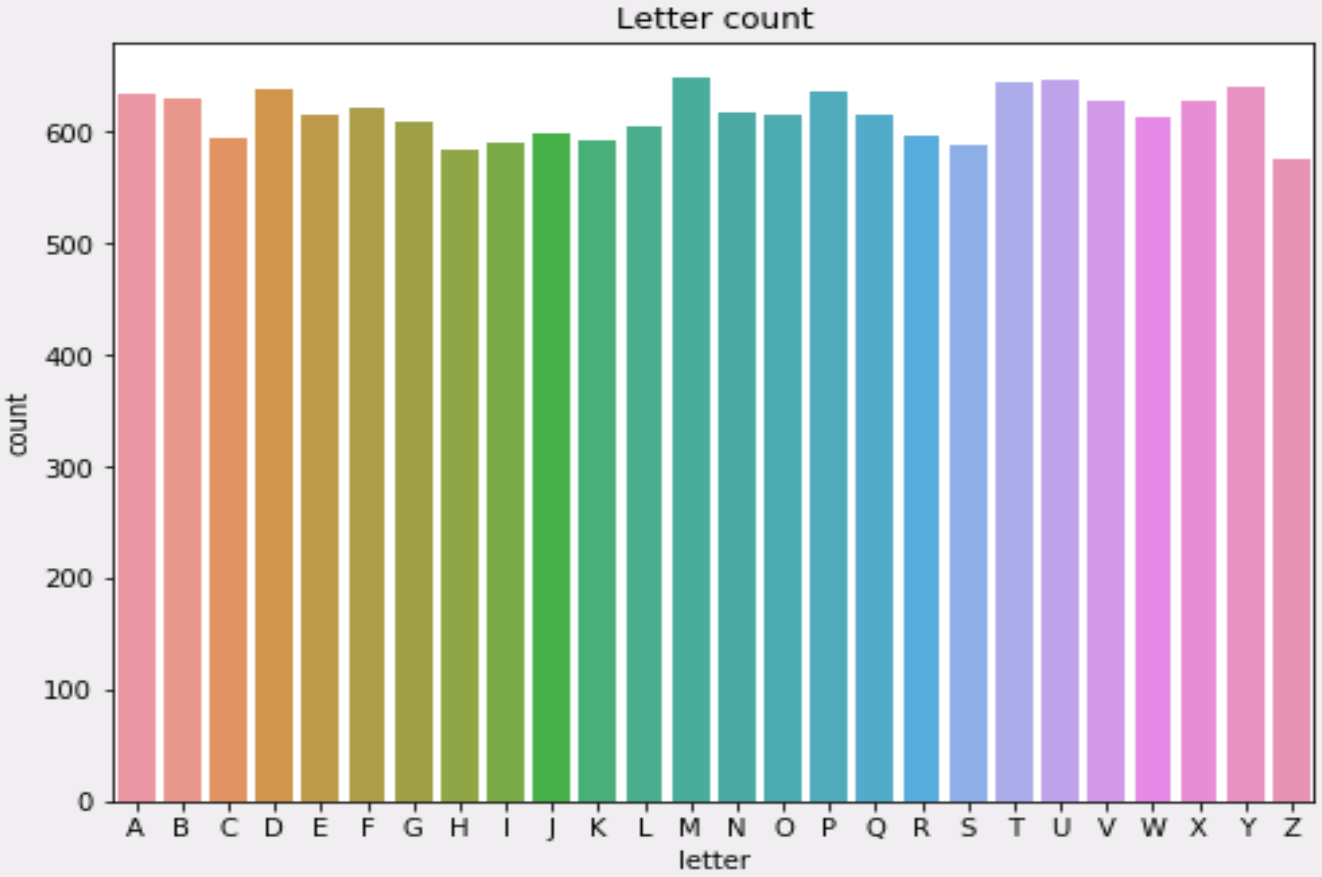
The objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet.

problem
content

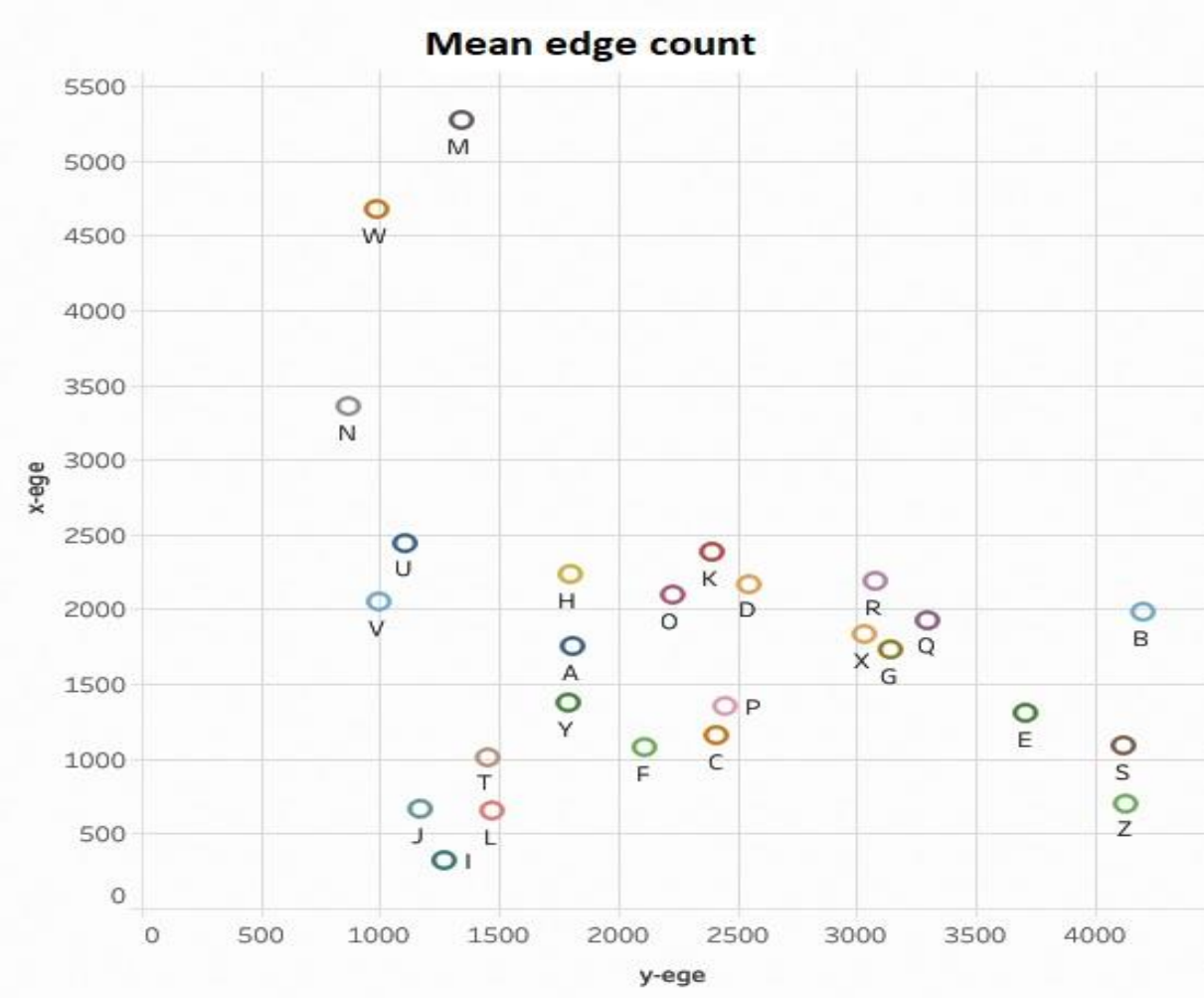


This dataset was created to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15. We typically train on the first 16000 items and then use the resulting model to predict the letter category for the remaining 4000.

Letter M(648) has got the highest number of observations followed by U(646) and T(644). Letter Z(576) is with the least number of samples.

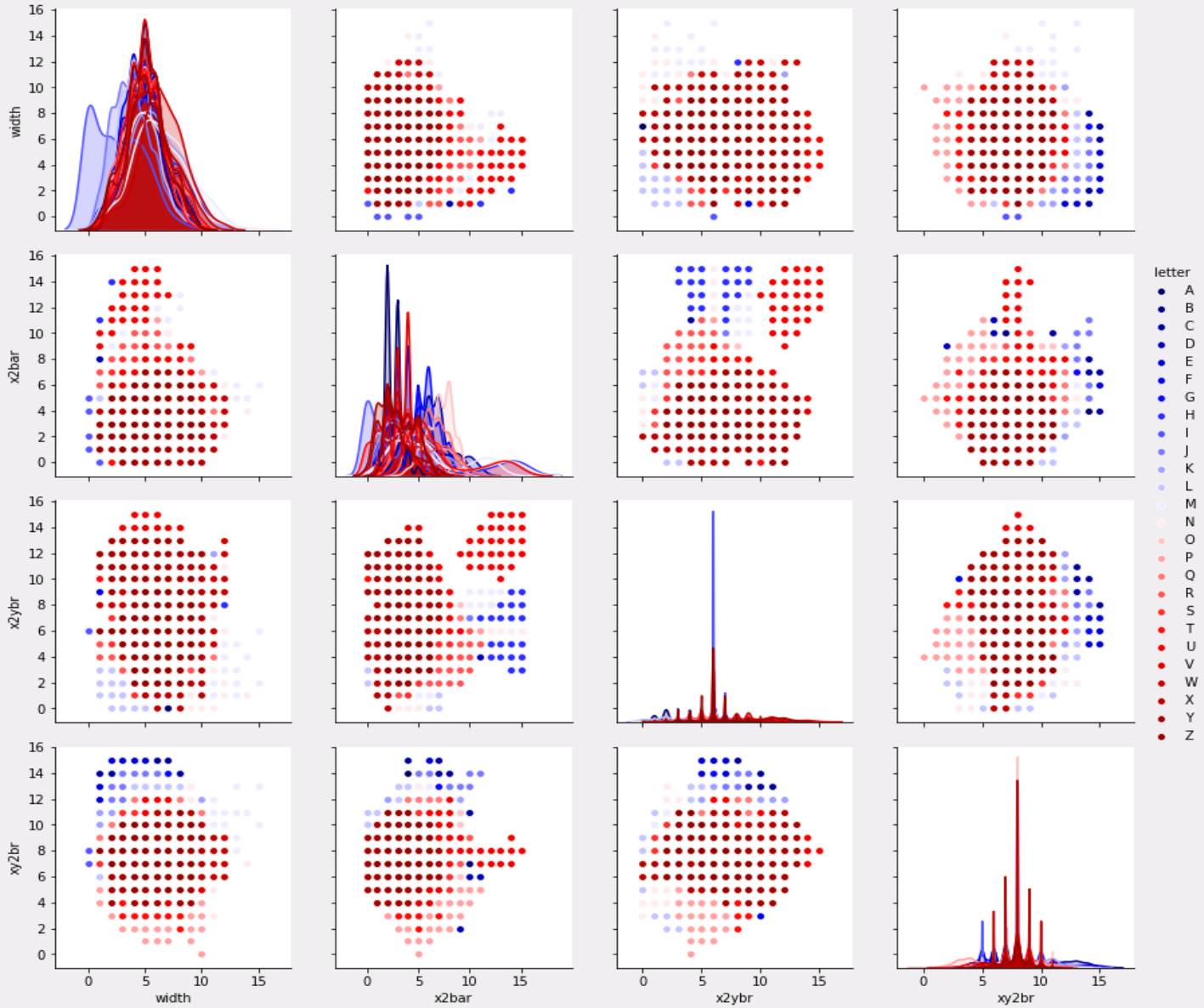


Every variable of the dataset is contributing it's bit towards the letter classification. Here is, x-ege (mean edge count from left to right) and y-ege (mean edge count from bottom to top) showing their bit in classifying the 26 letters.



model correlation pixel mean letter data problem content

The diagonal graph helps us understand Width, x2bar, x2ybr and xy2br variables contribution in the letter classification.



model

correlation

pixel

mean

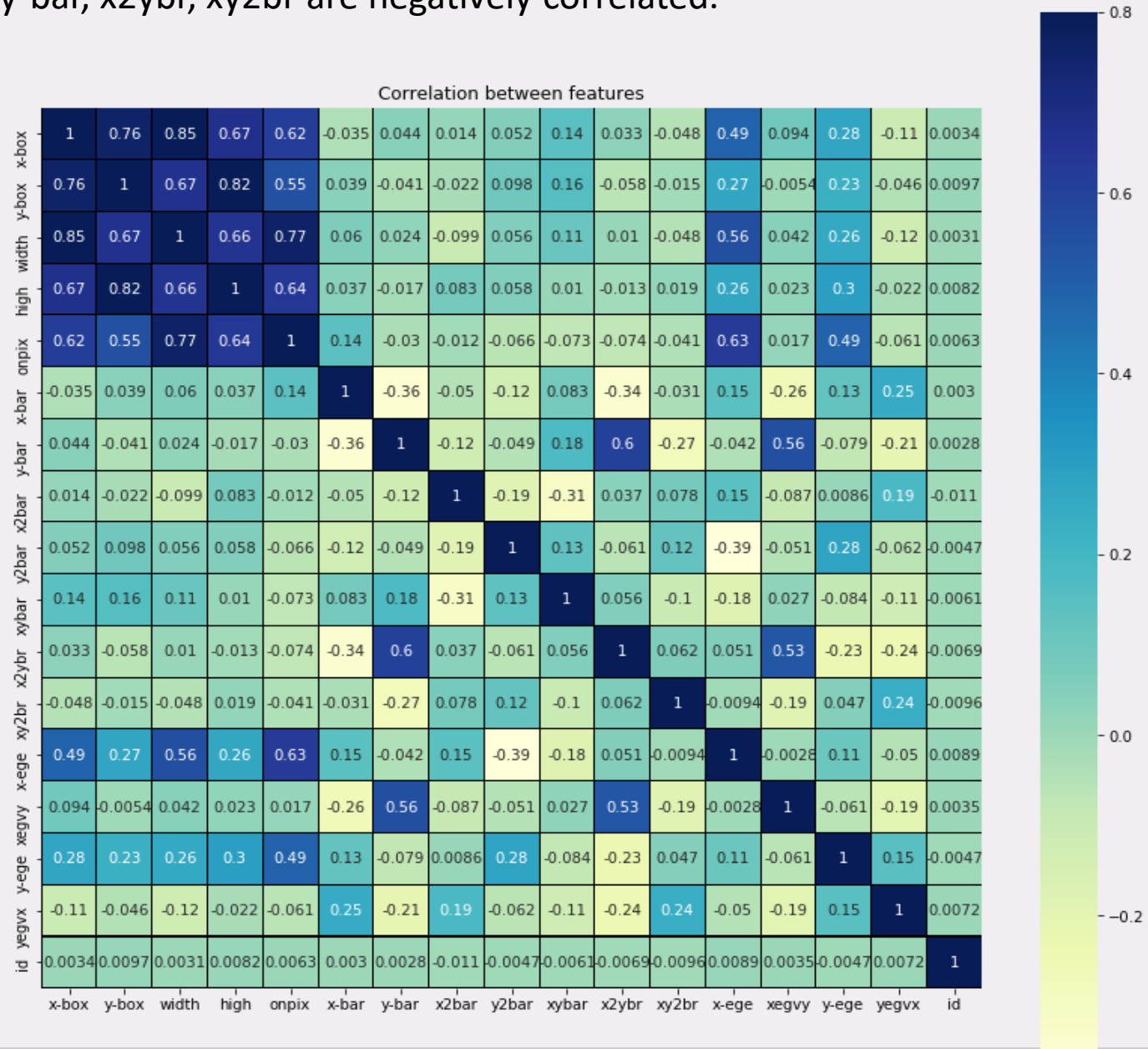
letter

data

problem

content

x-box, y-box, width, high, onpix are positively correlated. x-bar, y-bar, x2ybr, xy2br are negatively correlated.



model

correlation

pixel

mean

letter

data

problem

content

Various classifier models were used to predict the letters. These models were evaluated on different metrics to know their performance. Listed here are the results:

ALGORITHM	ACCURACY	F1_SCORE
Decision tree	0.999875	0.999876
Naïve bayes	0.651125	0.646085
Random forest	0.999812	0.999816
KNN	0.960187	0.960086
SVM	0.860625	0.859713
SGD	0.602125	0.600893
Ensemble	0.999812	0.999816

From the above, we observe that **Decision tree** model has performed well with highest scores in all the metrics.