LETTER RECOGNITION

In order to preserve the valuable historical information contained in hand-written documents, recent efforts have focused on scanning these historical documents in order to "digitize" their content. After scanning, however, oftentimes modern computing is unable to recognize the writing (due to poor handwriting, faded ink or complicated font type) which prevents the content from being completely digitized. One labor-intensive solution to this problem is to have people read the document and type the content (think "indexing" for family history). A better approach would be to train a computer to recognize the letters and then have people proofread the computer version of the document for content.

The dataset *letter-recognition.txt* contains 16 attributes of writing in a historical document along with the "person-verified" letter. Your task is to build a statistical model (or algorithm) that the computer can use to appropriately digitize the historical document. The variables in the dataset are as follows:

Column	Variable Name	Description
1.	lettr	capital letter (26 values from A to Z)
2.	x-box	horizontal position of box (integer)
3.	y-box	vertical position of box (integer)
4.	width	width of box (integer)
5.	high	height of box (integer)
6.	onpix	total # of pixels (integer)
7.	x-bar	mean x of pixels in box (integer)
8.	y-bar	mean y of pixels in box (integer)
9.	x2bar	mean x variance (integer)
10.	y2bar	mean y variance (integer)
11.	xybar	mean x y correlation (integer)
12.	x2ybr	mean of x * x * y (integer)
13.	xy2br	mean of x * y * y (integer)
14.	x-ege	mean edge count left to right (integer)
15.	xegvy	correlation of x-ege with y (integer)
16.	y-ege	mean edge count bottom to top (integer)
17.	yegvx	correlation of y-ege with x (integer)

1. How well are you able to classify letters using the given features?
2. What factors are useful in classifying letters?
3. What letters are commonly mistaken?