ASSIGNMENT – 2 ANALYSIS

The problem statement asked us to make the predictions of the test data by implementing K-Means clustering using map-reduce.

Below is the summary of classification accuracies resulted for various values “n” which is the training data percentage.

Training Data (%)	Classification accuracy
n=5 (5% training data)	25.56%
n=10 (10% training data)	***27.62%***
n=25 (25% training data)	27.33%
n=40 (40% training data)	26.05%

The champion model among the above four data partitions performed is 27.62%. However, the model accuracies are very low.

A. Train size =5%:

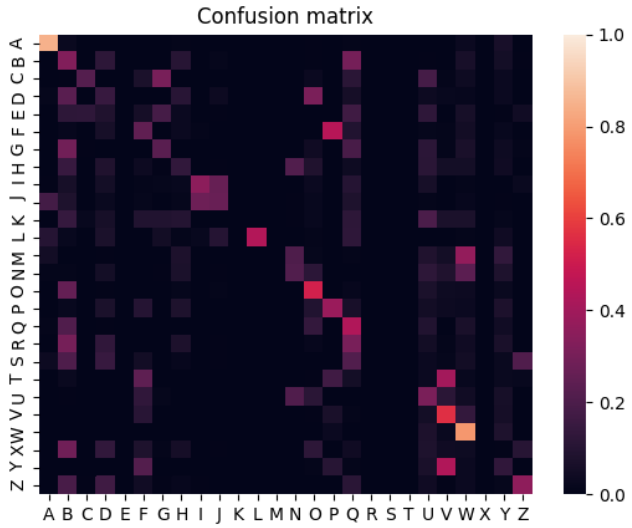
Train data dimension: (961, 17)

Test data dimension: (19039, 17)

Target variable distribution in the training dataset is as follows:

	Column0
D	40
P	39
X	39
M	39
T	39
U	39
Q	39
V	38
F	38
W	37
Y	37
N	37
E	37
R	37
G	36
K	36
L	36
O	36
S	36
C	36
B	36
A	36
H	35
I	35
J	34
Z	34

Almost all the letters have similar frequency in the train data.



Classification Accuracy: 26.566192301633144%

We can infer from the heatmap that only the letter A's actual labels and preicted label match well, compared to other set of letters.

B. Train size =10%

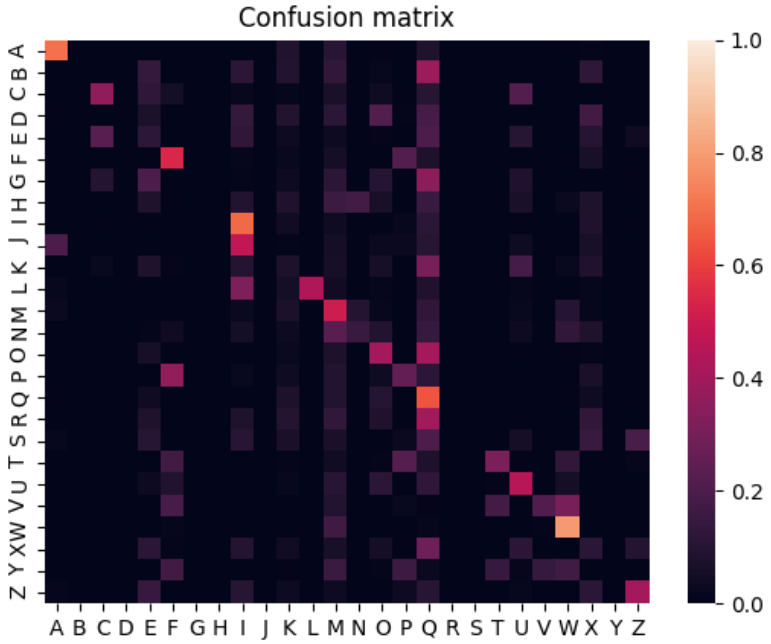
Train data dimension: (1894, 17)
Test data dimension: (18106, 17)

Target variable distribution:

	Column0
U	78
M	76
D	75
F	75
V	74
A	74
X	74
P	74
T	74
O	74
G	74
Y	74
E	73
Q	73
L	73
N	72
I	72
S	72
B	72
W	72
R	71
J	70
Z	70
C	70
H	69
K	69

The most frequent letter is **U** and least frequent letter observed in this train data set is **K.**

Classification Accuracy: 27.62787617944049%



In this dataset as well, the letter A's actual labels and predicted labels match better, compared to other set of letters.

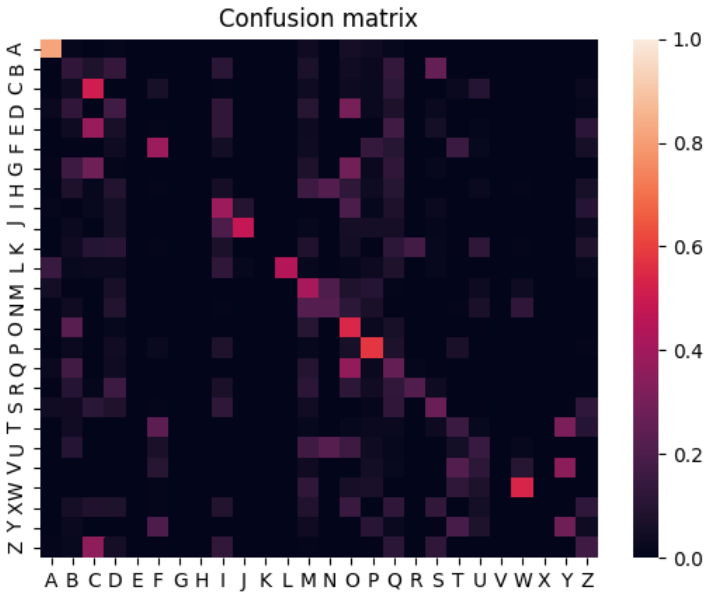
C. Train size =25%

Train data dimension: (4436, 17)
Test data dimension: (15564, 17)

	Column0
U	184
Y	181
N	181
A	175
T	175
D	175
Q	174
W	174
G	174
B	173
M	173
P	173
O	173
X	172
R	171
I	170
F	170
L	170
E	169
V	168
S	165
K	163
C	163
H	159
J	157
Z	154

Most frequently observed in the target variable is letters {U,Y,N,A,T} and the letters {J,Z} are least frequent in this train data set relatively.

Classification Accuracy: 27.33964921264883%



In this dataset as well, the letter A's actual labels and predicted labels match better, compared to other set of letters. And, moreover it is also among the top4 most frequent letter in the target variable.

D. Train size =40%

	Column0
B	267
T	267
A	265
D	264
M	264
Q	263
U	261
X	261
N	258
P	257
G	256
F	255
W	254
J	253
V	251
Y	250
R	248
C	247
L	247
E	247
I	246
H	246
O	244
K	243
Z	242
S	242

Most frequently observed in the target variable are letters {B, T, A } and the letters {S,Z} are least frequent in this train data set relatively.

Classification Accuracy: 26.057856771026994%