Teaching Computers to Read

Diego Wahl

The EMNIST Dataset

The EMNIST Dataset

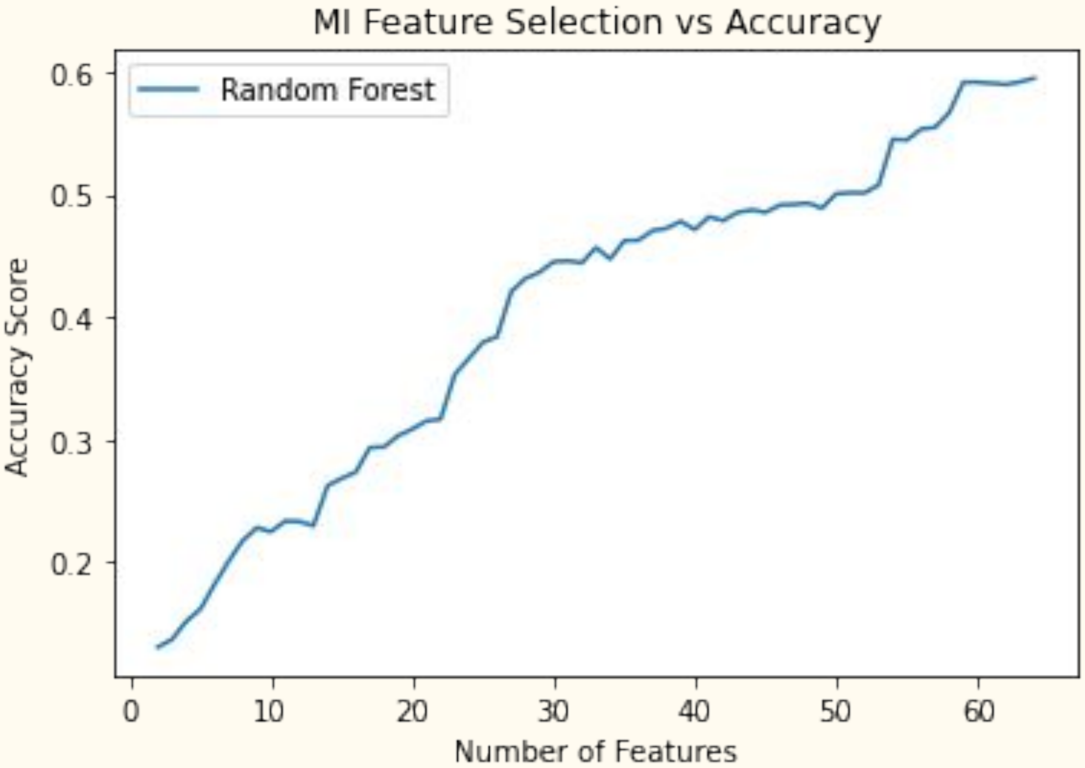
- Images are in black and white
- Centered and pre-processed
- Each feature represents a pixel and the intensity (brightness) of that pixel
- 28x28 pixels for a total of 784 features
- 814,255 rows, 62 **unbalanced** classes

How does one handle a dataset with this many features?

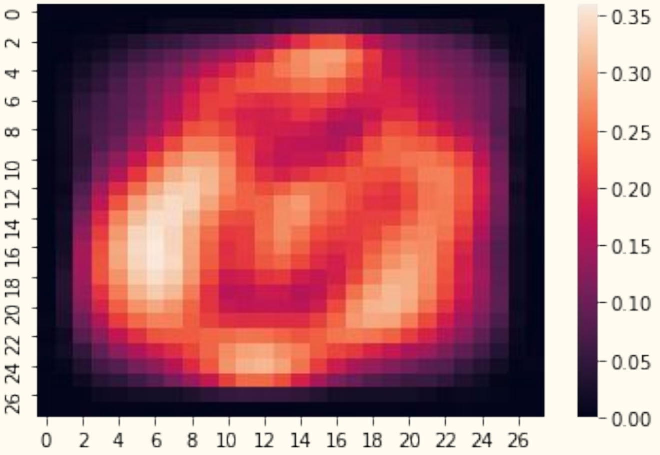
Feature Selection: MI

Mutual Information

- Native Feature Selection package in SKLearn
- Basically, a measure of the mutual dependence of two variables.
- In SKLearn implementation, get a value on [0,1]. Two independent variables have an MI of 0



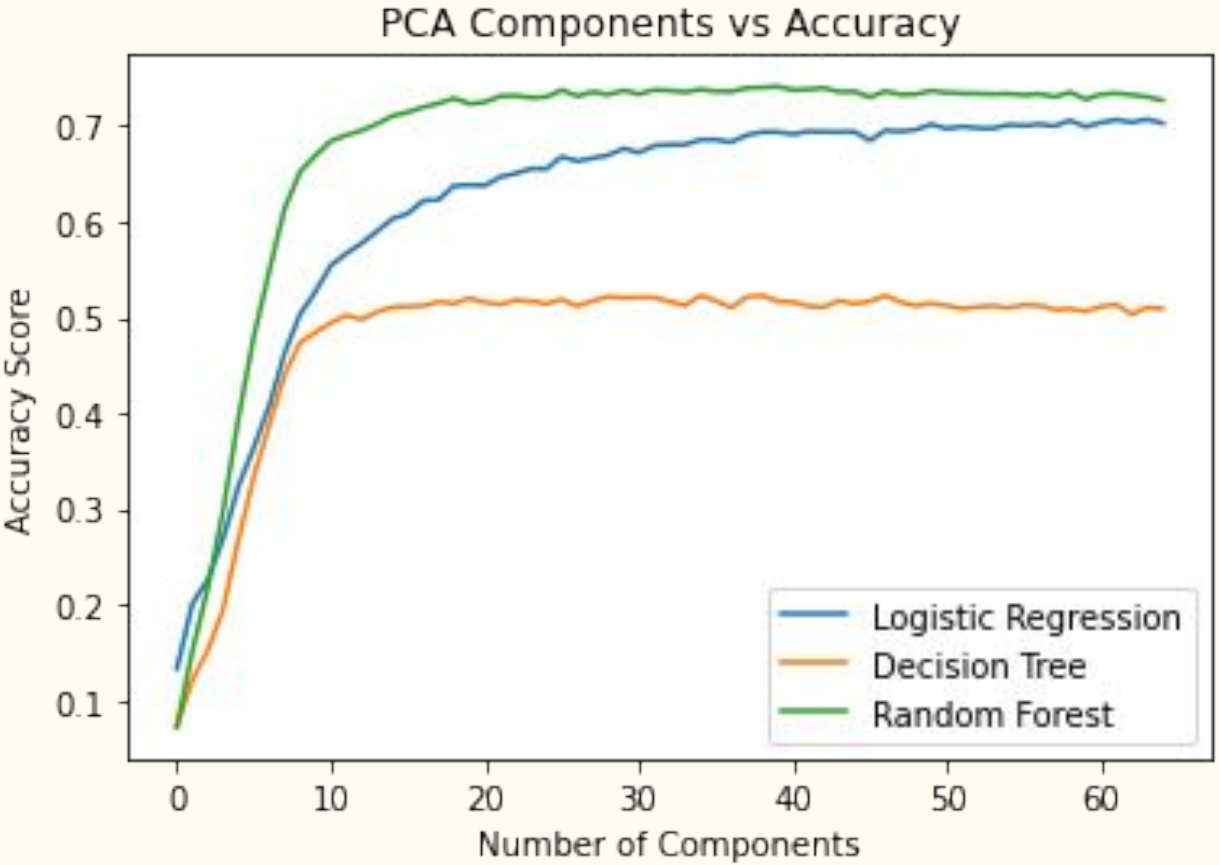
MI Feature Selection vs Accuracy



Dimensionality Reduction: PCA

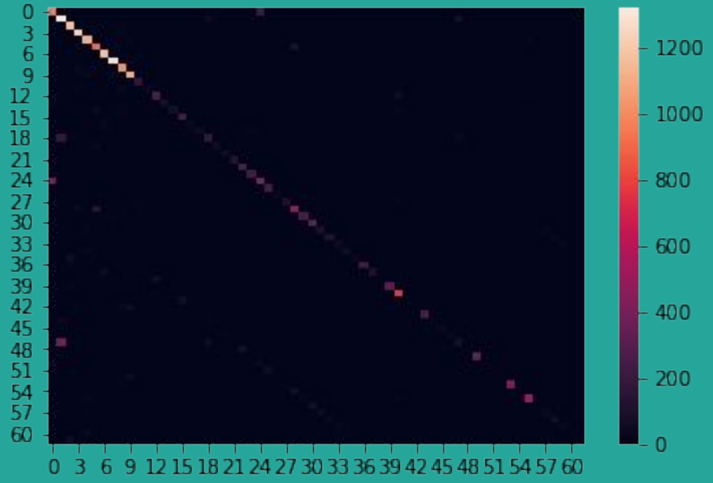
Principal Component Analysis

- Dimensionality Reduction Approach
- Eigenvectors of the covariance matrix with the greatest eigenvalues selected as the ‘primary components’
- Does require scaling
- Lose interpretability


PCA Components vs Accuracy
Accuracy Score
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Number of Components
0
10
20
30
40
50
60
Logistic Regression
Decision Tree
Random Forest

Final Model

- Random Forest
- Accuracy as the scoring metric
- ~77%
- I,i,1 o,O,0



THANKS