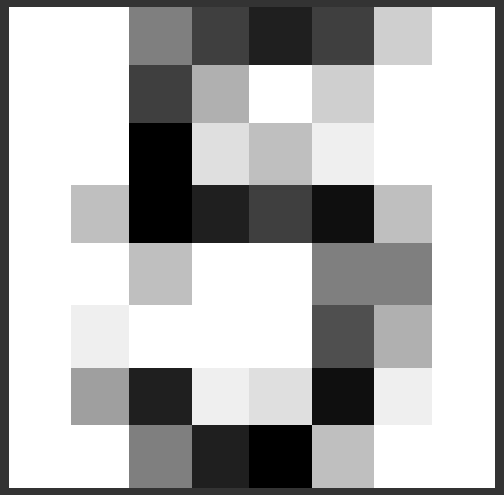
Short Presentation

on the process and results of the Machine Learning question
for the ECG internship

By Cengizhan Can

1. Data pre-processing

- Chosen dataset: The Digit Dataset
  - Made up of 1797 images of handwritten digits
  - The images are 8 × 8 pixels
  - Each image has a corresponding label
  - Easy to import through scikit-learn
  - No data pre-processing required



Example of dataset I have used in the past that required pre-processing of the data: Moneyball dataset

- Dealing with null values
- Some features contained an integer encoding instead of a string

2. Feature Engineering

- We use the 8 x 8 pixel information as a feature vector with length 64
- Rather straightforward since no other features to choose from really

```
In [16]:

print("Type:", type(digits.images[0]))
print("Size: ", digits.images[0].size)
print("Array: ", digits.images[0])
```

```
Type: <class 'numpy.ndarray'>
Size:  64
Array:  [[ 0.  0.  5. 13.  9.  1.  0.  0.]
 [ 0.  0. 13. 15. 10. 15.  5.  0.]
 [ 0.  3. 15.  2.  0. 11.  8.  0.]
 [ 0.  4. 12.  0.  0.  8.  8.  0.]
 [ 0.  5.  8.  0.  0.  9.  8.  0.]
 [ 0.  4. 11.  0.  1. 12.  7.  0.]
 [ 0.  2. 14.  5. 10. 12.  0.  0.]
 [ 0.  0.  6. 13. 10.  0.  0.  0.]]
```

Example of dataset I have used in the past that required more feature engineering: Moneyball dataset:

- Utilize one-hot encoding for categorical values

3. Supervised Learning Models

- Used three different models:
  - k-Nearest Neighbors, a non-parametric method
  - Logistic Regression, a regression analysis model
  - Random Forests, an ensemble learning method also non-parametric

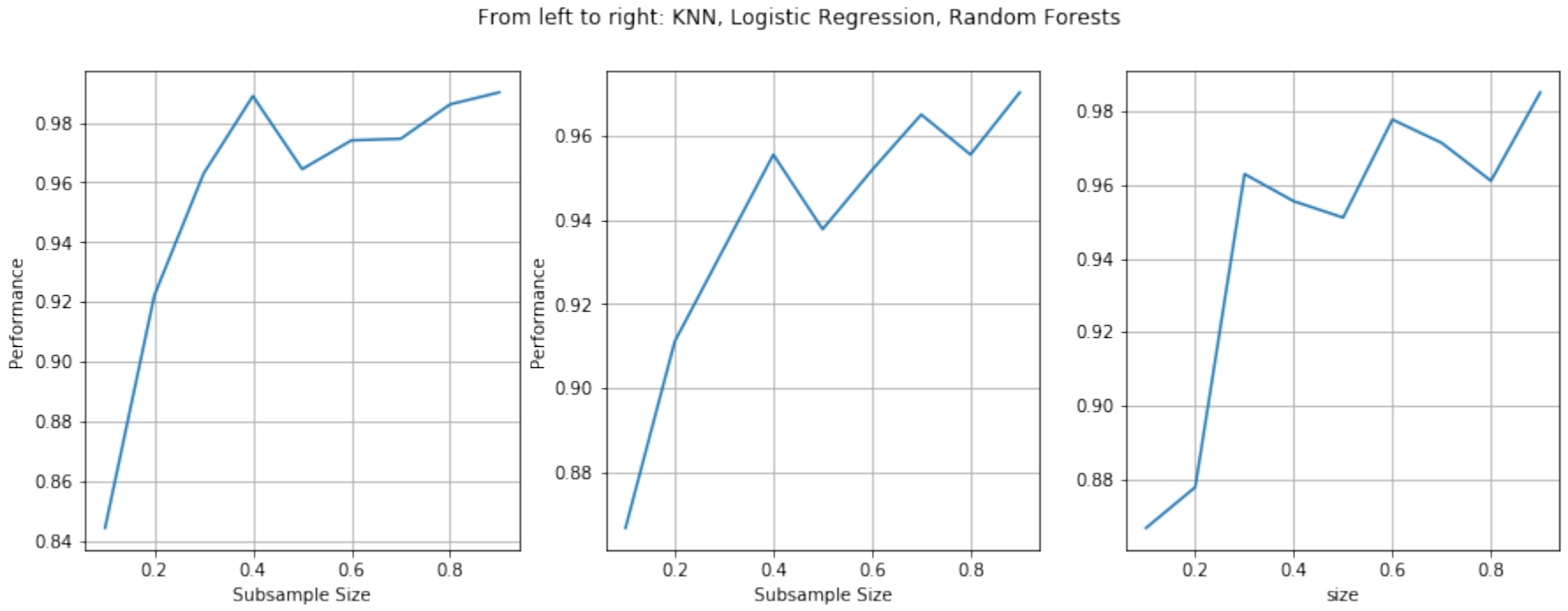
4. (Dis)advantages of each model

- k-Nearest Neighbors :
  - Supports non-linear classification problems, not a lot of hyperparameters
  - Selecting the right k might be difficult, getting slower when the sample size grow
- Logistic Regression:
  - Easy and simple, can be used for multiclass classifications
  - Cannot be used for non-linear classification problems, getting slower when the sample size grow
- Random Forests:
  - Makes no assumptions on the distribution of data
  - Somewhat slower runtime

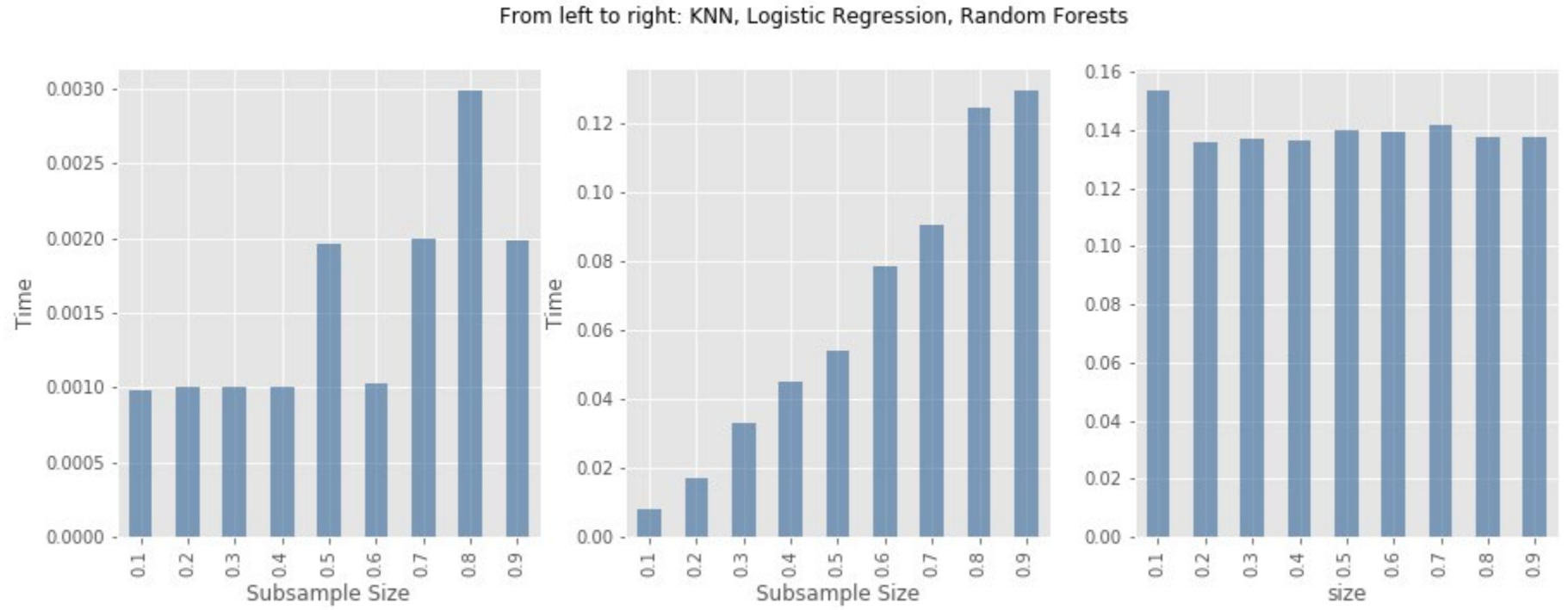
4. Evaluation

- Metrics to evaluate the machine learning models:
  - **Classification Accuracy (implemented)**
  - **Confusion matrix (implemented)**
  - Logarithmic Loss
  - Area under curve (AUC)
  - F-Measure

5. Evaluation



5. Evaluation



5. Evaluation