Supervised Learning Capstone

Digit Recognizer

Slava Sablin
issablin@gmail.com
April, 2019

Goals

- Explore different models for supervised learning
- Find out which one is the best performer on the classic machine learning problem
- Take part in a Kaggle competition in order to start a career in data science

Dataset

MNIST ("Modified National Institute of Standards and Technology") is the de facto “hello world” dataset of computer vision. Since its release in 1999, this classic dataset of handwritten images has served as the basis for benchmarking classification algorithms.

The training set contains 42 000 labeled gray-scale images of hand-drawn digits, from zero through nine. The testing set is 28 000 unlabeled images.

Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255.

To do or not to do?

Feature engineering

Since, a process of extracting meaningful features in this particular case is a kind of an art and can take a huge amount of time and resources, I decided not to perform this step and used the data exactly as they are.

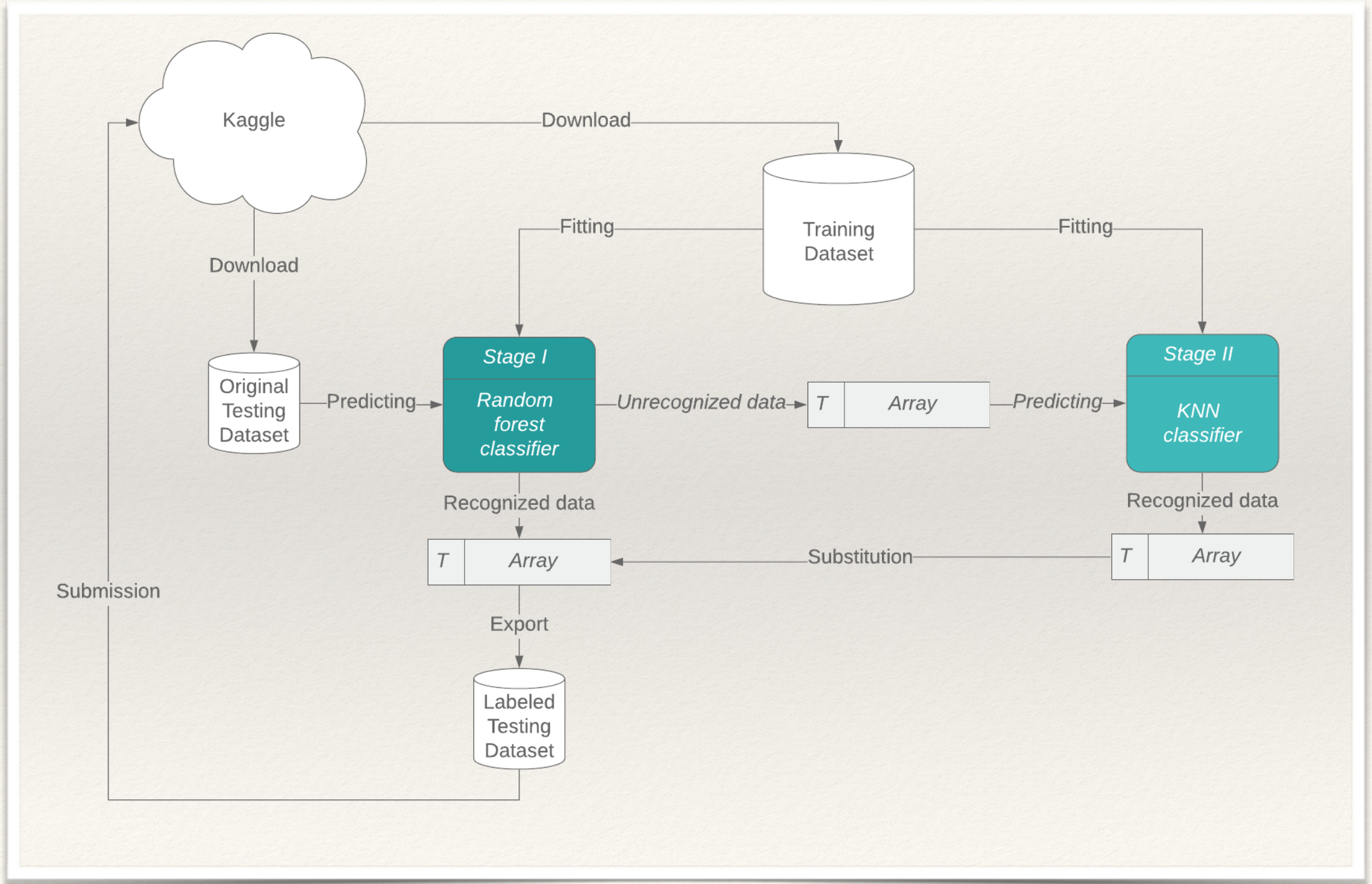
Models

- K Nearest Neighbors
- Random Forest
- Support Vector Machines
- Combinations of them

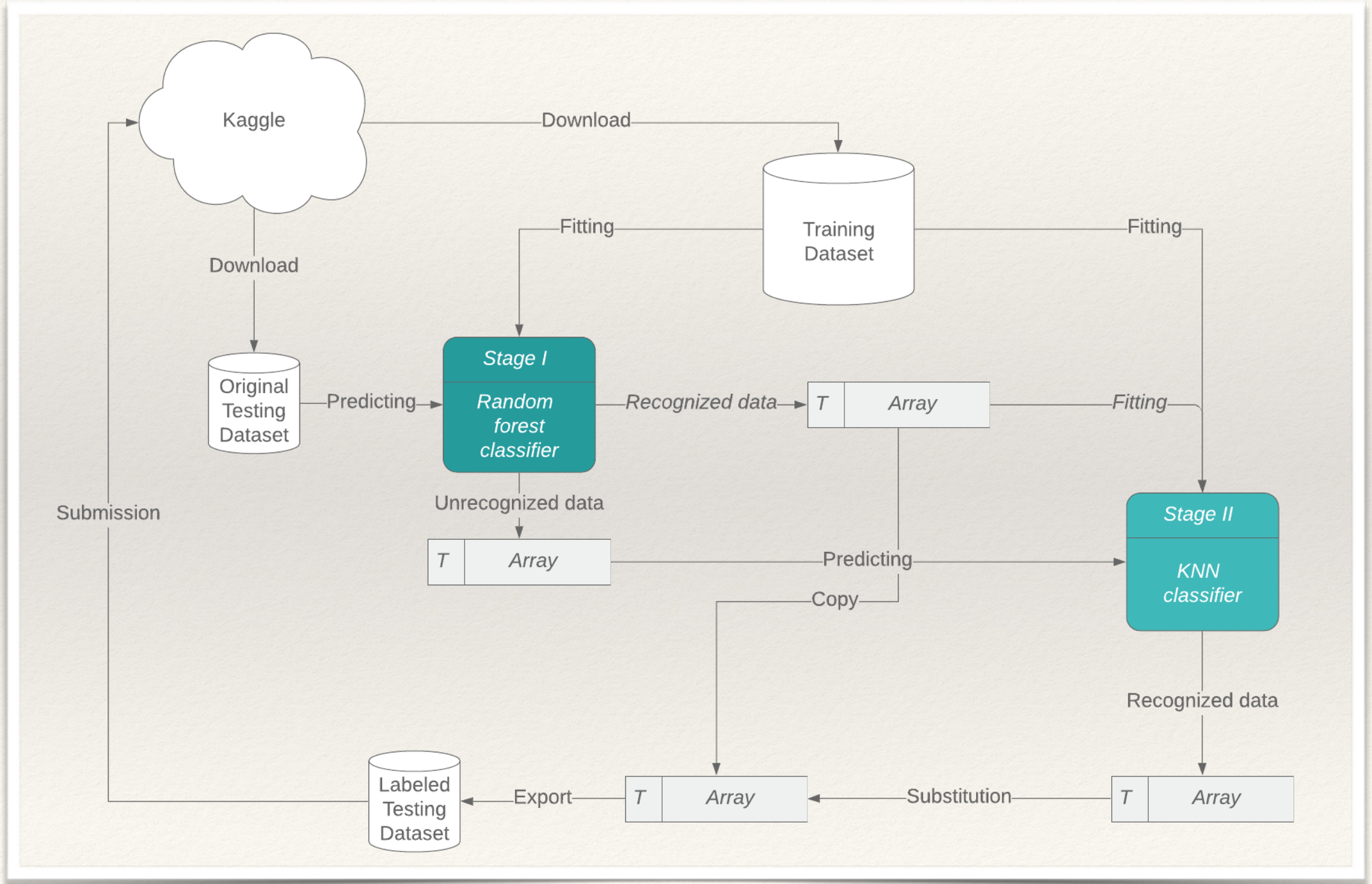
Results (training set)

Model	Score	Not Detected	Pure Accuracy
KNN	0.9669	0	0.9669
RF	0.8900	10.38%	0.9931
SVM	0.9771	0	0.9771

Combination (sample 1)



Combination (sample 2)



Results (testing set)

Model	Score by Kaggle (on 25% of submitted data)
RF -> RF+KNN	0.97528
RF+RF -> RF+KNN	0.97500
RF -> SVM	0.97242
RF -> RF+SVM	0.97200
SVM	0.97357
RF+SVM	0.97285
RF -> KNN	0.97328

Ideas for Future Research

- Feature engineering
- Neural Networks
- Team work
- Creative thinking

Inferences and Insights

- ML is fun
- VM on AWS or Oracle Cloud is recommended
- Reserve more time for research
- Progress indicators are useful
- Hardware to be upgraded
- Never give up!

https://www.kaggle.com/c/digit-recognizer/leaderboard#score

Overview | Data | Kernels | Discussion | Leaderboard | Rules | Team | My Submissions | Submit Predictions

1898	javad helali	0.97528	7	15d
1899	Alexey Kuznetsov	0.97528	2	1d
1900	Pierre-Adrien	0.97528	11	6h
1901	Slava Sablin	0.97528	7	~10s

Your Best Entry ↑

You advanced 53 places on the leaderboard!

Your submission scored 0.97528, which is an improvement of your previous score of 0.97357. Great job! Tweet this!

1902	YuchenWang	0.97514	1	1mo
1903	Hany	0.97514	6	19d
1904	Dinmukhamed Mailibay	0.97500	2	1mo
1905	Yevheniia Bu	0.97500	3	1mo

Thank you!

Slava Sablin
issablin@gmail.com