중간 발표

7조

김희균, 송준규, 정승연, 정하연

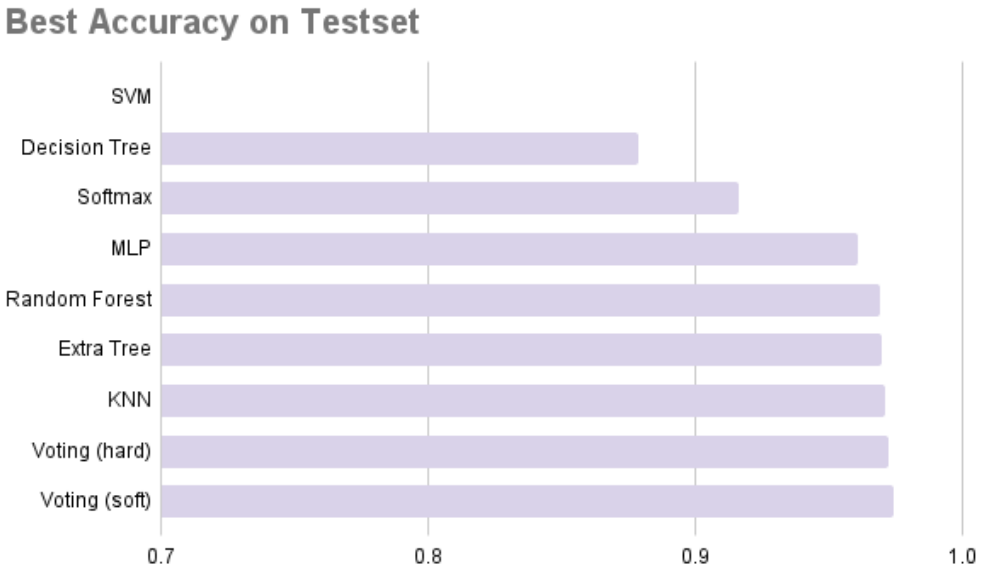
목차

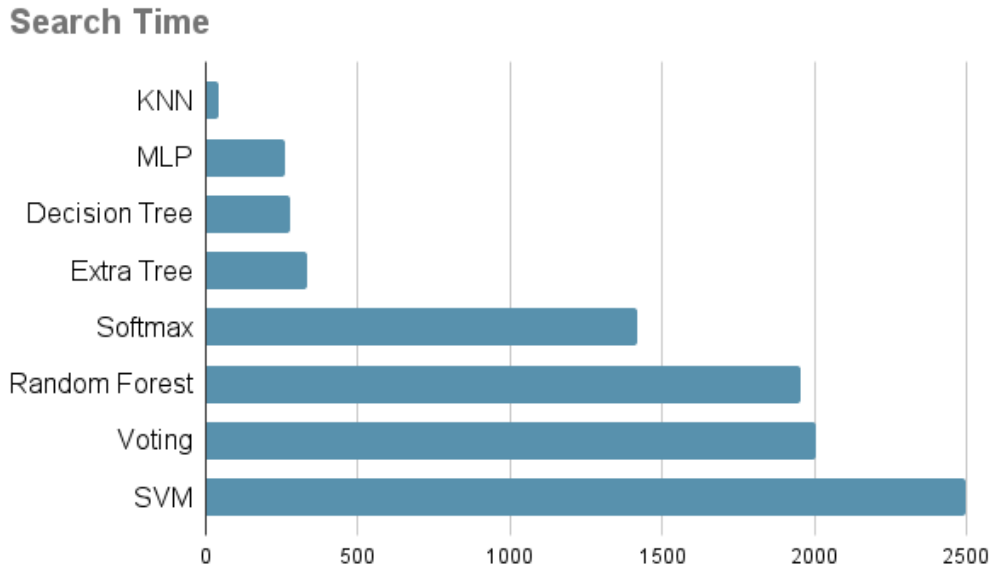
Original Dataset

- Testset의 비율은 전체 dataset의 20%로 설정함.
- GridSearch로 최적의 파라미터를 찾음.

Model	KNN	SVM	Extra-tree	Softmax	Decision Tree	Random Forest	MLP	Voting
Search Time (sec)	43.30	10k개 30sec. -> too long	333.25	1415.07	275.32	1957.44	260.05	2005.04
Best Parameters	**n_neighbors: 3**	**x**	**max_depth : 20**	**C: 1, max_iter : 500**	**Criterion : entropy**	**n_estimators : 2000**	**alpha : 0.01**	**soft**
			n_estimators : 300	**multi_class : multinomial**	**max_depth : 100**		**max_iter : 500**	
Best Accuracy on Test Dataset	0.9713	x	0.9696	0.9163	0.8789	0.9690	0.9611	soft: 0.9743
								hard: 0.9726

Original Dataset





Handmade Dataset

- **데이터 클리닝 아이디어**
  - *숫자(10개) + 기호(5개) 에 속하지 않는 라벨을 가진 데이터는 모두 삭제한다.*
  - 0~9, 기호 순서대로 데이터를 정렬한 후, 잘못 라벨링된 데이터를 육안으로 판별 후 삭제하거나 수정한다.
    - Ex. +인데 x로 잘못 라벨링 됨.
    - 학습 전에 데이터를 shuffle한다.
  - 이미지 해상도를 조정하여 숫자와 기호의 굵기를 동일하게 바꿔본다.

Handmade Dataset

- 데이터 클리닝 진행 과정

	숫자	기호
Training	(15119, 28, 28)	(15329, 28, 28)
Test	(2160, 28, 28)	(2190, 28, 28)

〈 Data Cleaning 이전 dataset의 shape 〉

	숫자 + 기호
Training	(30448, 28, 28)
Test	(4350, 28, 28)

〈 15개 클래스 분류기를 위한 데이터 통합 〉

최종 데이터셋 shape	
Training	(26249, 28, 28)
Test	(3730, 28, 28)

〈 최종 데이터셋 〉

제거된 데이터셋 개수	
Training	4199 (13.8%)
Test	620 (14.3%)

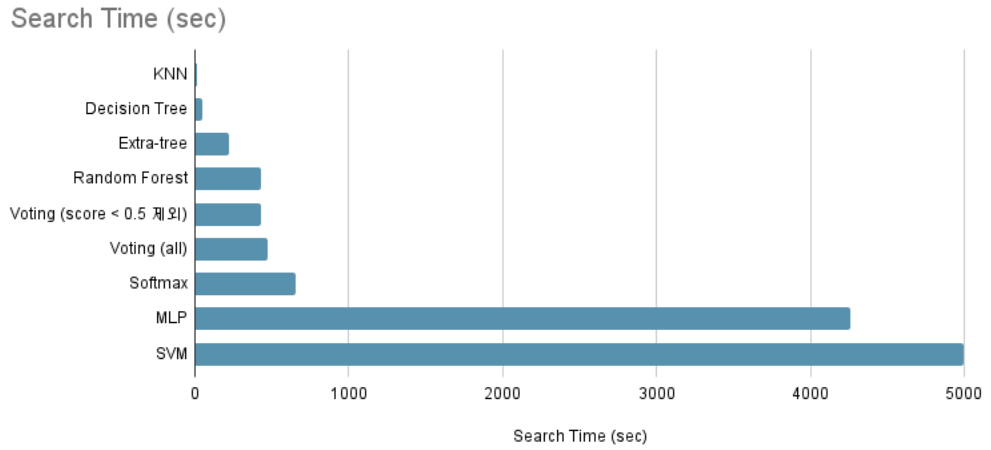
〈 15개 이외 라벨 제거 〉

Handmade Dataset

● 학습 결과

Model	KNN	SVM	Extra-tree	Softmax	Decision Tree	Random Forest	MLP	Voting (score < 0.5 제외)	Voting (all)
Search Time (sec)	16.10	takes too long	218.40	654.03	51.40	426.11	4255.92	431.	472.67
Best Parameters	**n_neighbors=3**		max_depth : 20 **n_estimators : 300**	C : 0.1 **max_iter : 500** multi_class : multinomial	Criterion : gini **max_depth : 50**	**n_estimators : 900**	alpha : 0.1 **max_iter : 500**	**soft**	**hard**
Best Accuracy on Test Dataset	0.6059		0.7263	0.3769	0.4903	0.7445	0.7370	soft: 0.7499, hard: 0.7458	soft: 0.7190, hard: 0.7357

Handmade Dataset



추후 계획

- Data Cleaning
  - 남은 데이터 클리닝 아이디어를 시도해볼 것.
  - 추가적인 데이터 클리닝 방안을 모색해볼 것.
- Training
  - Search Time과 Accuracy가 좋은 순서대로 다양한 앙상블 학습을 시도해볼 것