**7조 머신러닝 프로젝트 제 3차 회의지**

**[Hand-made MNIST dataset을 이용한 머신러닝 모델 최적화 및 분석]**

| **작성자: 정하연**  **날짜: 2023-11-16**  **회의 참석자: 20236079 김희균, 20201590 송준규, 20192447 정승연, 20201610 정하연**  **회의 장소: Google Meet** |
| --- |
| **회의 내용:**   1. **original data vs handmade data 자체 비교 분석** 2. **original dataset으로 학습한 다양한 model의 성능 비교 및 논의**   **random forest // ensemble**  **extra tree classifier**  **knn**  **mlp**  **voting classifier**  **svm**  **softmax**  **decision tree**   1. **handmade dataset - cleaning**    1. **label 잘못된 것까지 확인**    2. **숫자인데 기호라고 파일명을 잘못 지정해서, 숫자데이터에서 기호 데이터를 빼는 작업이 필요함.** 2. **handmade dataset으로 다양한 모델 학습** 3. **모델 최적화** 4. **분석** 5. **최종 선정된 모델 및 파라미터** |
| **추진 계획:**   1. **금요일까지 *data cleaning 방안 모색 및 공유하기*** 2. **토요일까지 handmade dataset으로 학습 및 비교 분석, 최적화** 3. **일요일까지 발표자료 제출, 발표 대본 작성 및 리허설** |
| **결정 사항:**   * **진행 일정 위 참고** * **중간 발표 내용 목차 아래에 정리함.** |

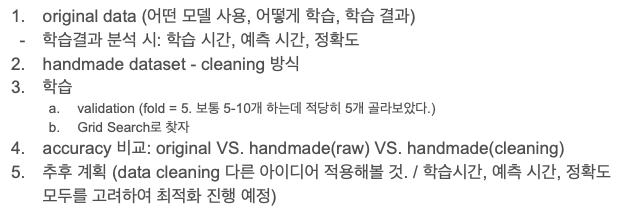
**[중간 발표에 들어가야 할 것 같은 내용 정리]**

1. 사용할 데이터 결정 기준 [1] 1~9,+,-,=,/,x 제외 버리기 or 흡수하기

[있는 거 없는 거 비교]

1. 결정 기준 [2] 학습을 따로 따로 할지 데이터 [합쳐서 할지]
2. original data로 학습한 결과, handmade(raw), handmade(cleaning) 비교
3. ~~학습시간, 예측시간 (inference time), Accuracy~~
4. 최종 선정 모델 및 데이터
5. 추후 계획

**[ 중간발표 자료 - 목차 ]**



**[ data cleaning - idea ]**

1. ***15개 클래스에 속하지 않는 라벨은 다 버림 (11/17까지 게시판 답변 보고 논의 후 진행).*** 0~9, 기호 순서대로 정렬 후 육안으로 제거?
2. 틀린 정답으로 나온 데이터 다시 확인 필요(1,-./) (3,5)

**[ 최종 ipynb에 들어갈 내용 참고 (02\_end\_to\_end\_machine\_learning\_project) ]**

1. Get the Data
   1. download the data
   2. take a quick look at the data structure
   3. create a test set
2. Discover and visualize the Data to Gain Insights
   1. Visualizing Geographical Data
   2. Looking for Correlations
   3. Experimenting with Attribute combinations
3. Prepare the data for machine learning algorithms
   1. data cleaning
   2. handling text and categorical attributes
   3. feature scaling
   4. Transformation pipelines
4. Select and Train a Model
   1. Training and Evaluating on the Training set
   2. better evaluation using cross-validation
5. fine-tune your model
   1. grid search
   2. randomized search
   3. analyze the best models and their errors
   4. evaluate your system on the test set
   5. model persistence using joblib