# 2020 데이터 크리에이터 캠프

데이터사나이
이재상
김윤환, 박준민, 박대한, 정규형

NIA 한국정보화진흥원

# Ⅰ. 소개

1. 문제 분석
2. 모델 소개

# Ⅰ. 소개

## 1 문제 분석

### 본선 5회차 출제문제

Q. 당신은 교통관리공단의 첨단 데이터 분석팀에서 근무하고 있습니다. 최근 과속으로 인한 교통사고 증가율이 늘어남에 따라 CCTV 성능을 높이기 위해 국내 차량 이미지 데이터를 활용해 차량 클래스를 구분하고자 합니다. 국산 차량 이미지를 통해 차량 클래스를 구분하는 모델을 만드시오. (데이터셋 별도 제공)

[공통사항]

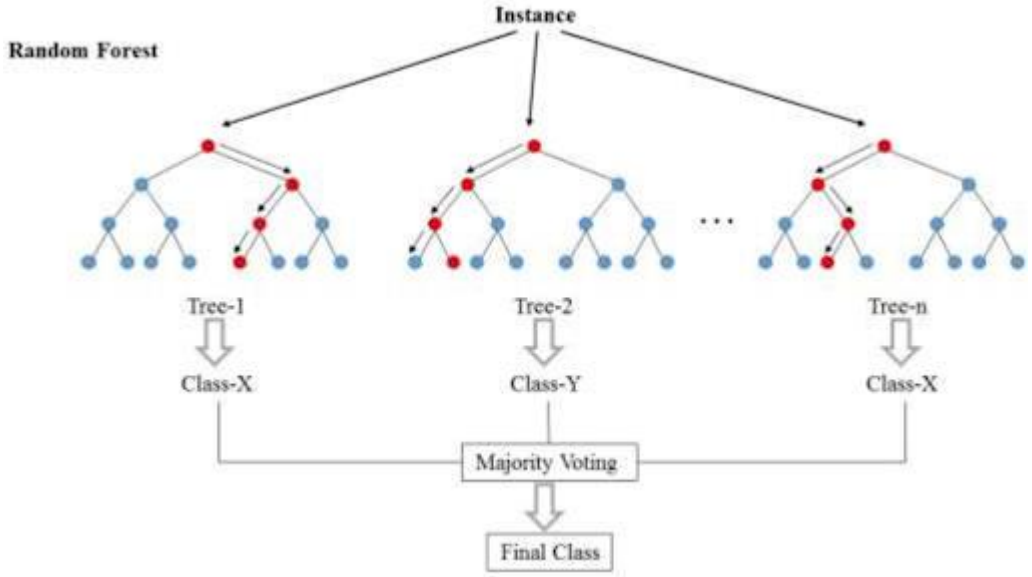
- 총 24,916개의 국산 차량 이미지 데이터(30 X 30 X 3 픽셀 값으로 들어가 있고, 마지막 칼럼이 차량 클래스)
- 데이터셋 파일명 : kcar.pkl (압축 형식 : Gzip) https://www.aihub.or.kr/aidata/130
- X 데이터와 Y 데이터가 하나의 파일(csv)에 있음(학습 데이터와 테스트 데이터의 비율은 80%:20%)
- 모든 데이터(칼럼, 로우)를 학습에 사용할 필요는 없습니다. 실제 모델 결과보다 모델을 만들기까지의 과정이 중요합니다.
- 어떠한 논리로 분석을 진행하였는지 설명을 반드시 세부적으로 발표자료에 적어서 발표해주시길 바랍니다.
- 머신러닝을 사용한 모델링 과정을 하나 이상 넣어 주시길 바랍니다.

# Ⅰ. 소개

## 2 모델 소개

### ① Random Forest

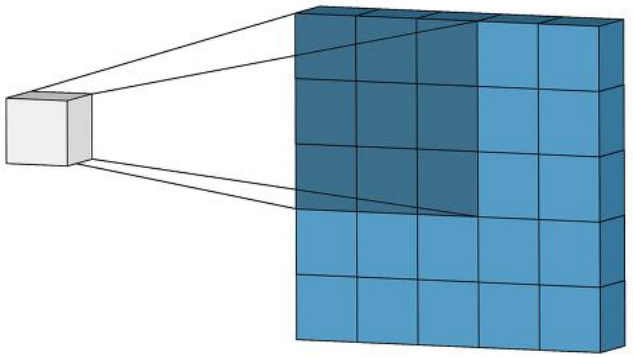
**랜덤 포레스트**는 분류, 회귀 분석 등에 사용되는 앙상블 학습 방법의 일종으로, 훈련 과정에서 구성한 다수의 결정 트리로부터 부류(분류) 또는 평균 예측치(회귀 분석)를 출력함으로써 동작한다.



## 2 모델 소개

### ② CNN

**합성곱 신경망**(Convolutional neural network, CNN)은 시각적 영상을 분석하는 데 사용되는 다층의 피드-포워드적인 인공신경망의 한 종류이다.

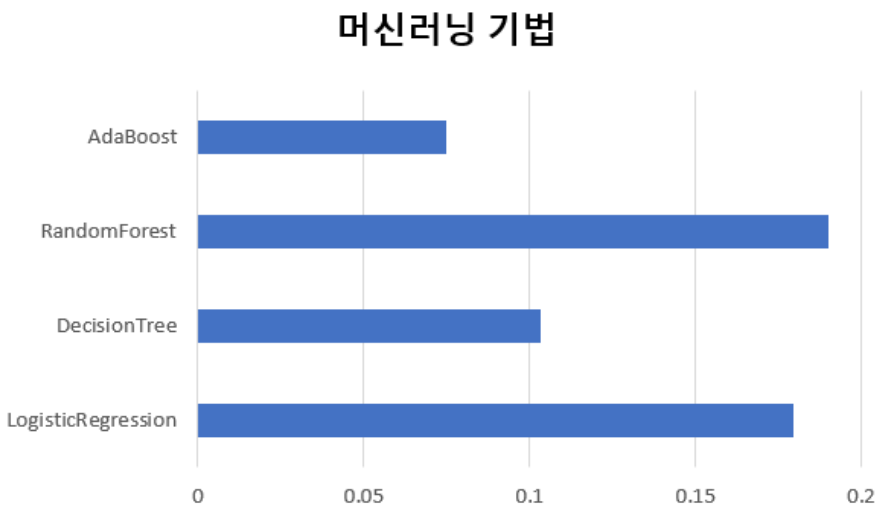


# Ⅱ. 모델링 및 분석결과

1. 머신러닝
2. 딥러닝
3. 분석결과

# Ⅱ. 모델링 및 분석결과

## 1 머신러닝




머신러닝 기법

AdaBoost, RF, Decsion Tree, Logistic Regression
4 개의 기법을 이용하여 정확도를 비교
(데이터는 3000 rows 만 사용)

RandomForest의 정확도가 가장 높게 나타남

# Ⅱ. 모델링 및 분석결과

## 1 머신러닝

```
from sklearn.model_selection import GridSearchCV

params = { 'n_estimators' : [200, 100],
           'max_depth' : [10, 12],
           'min_samples_leaf' : [6, 8],
           'min_samples_split' : [8, 10]
            }

# RandomForestClassifier 객체 생성 후 GridSearchCV 수행
rf_clf = RandomForestClassifier(random_state = 0, n_jobs = -1)
grid_cv = GridSearchCV(rf_clf, param_grid = params, cv = 3, n_jobs = -1)
grid_cv.fit(X1_train, y_train)

print('최적 하이퍼 파라미터: ', grid_cv.best_params_)
print('최고 예측 정확도: {:.4f}'.format(grid_cv.best_score_))
```

```
최적 하이퍼 파라미터:  {'max_depth': 12, 'min_samples_leaf': 8, 'min_samples_split': 8, 'n_estimators': 200}
최고 예측 정확도: 0.1829
```

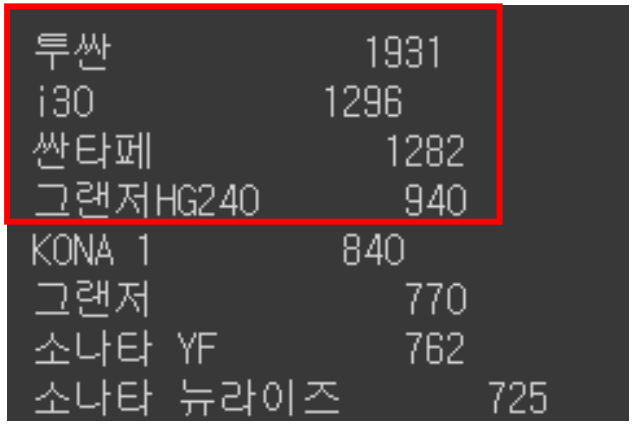
```
y_pred = grid_cv.predict(X1_test)
accuracy_score(y_test, y_pred)
```

```
0.19
```

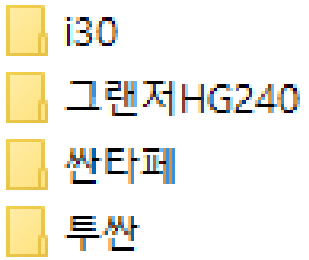
GridSearchCV를 수행하여 0.19까지 정확도를 높임

# Ⅱ. 모델링 및 분석결과

## 2 딥러닝




데이터 수가 많은 차량 모델 4개를 대상으로만 진행
(rows : Train : 4359 / Test : 1090)




이미지 원본 파일을 활용하여 CNN(goolglenet) 진행

# Ⅱ. 모델링 및 분석결과

## 2 딥러닝

```
Epoch: 1/10 Loss: 1.1275     4359 / 4359
Test Acc: 890/1090 (81.65%)
Epoch: 2/10 Loss: 0.4208     4359 / 4359
Test Acc: 980/1090 (89.91%)
Epoch: 3/10 Loss: 0.2317     4359 / 4359
Test Acc: 981/1090 (90.00%)
Epoch: 4/10 Loss: 0.1580     4359 / 4359
Test Acc: 1019/1090 (93.49%)
Epoch: 5/10 Loss: 0.1020     4359 / 4359
Test Acc: 1030/1090 (94.50%)
Epoch: 6/10 Loss: 0.0788     4359 / 4359
Test Acc: 1047/1090 (96.06%)
Epoch: 7/10 Loss: 0.0804     4359 / 4359
Test Acc: 1040/1090 (95.41%)
Epoch: 8/10 Loss: 0.0502     4359 / 4359
Test Acc: 1055/1090 (96.79%)
Epoch: 9/10 Loss: 0.0375     4359 / 4359
Test Acc: 1061/1090 (97.34%)
Epoch: 10/10 Loss: 0.0302    4359 / 4359
Test Acc: 1059/1090 (97.16%)
Training completed in 4m 3s
Best test accuracy: 97.339450
```

- 4개의 차량 분류에 대해 최종적으로 97.33%의 정확도를 보임

## 3 분석 결과

- 딥러닝 모델이 이미지 분류에 있어서 훨씬 정확도를 보인다.
  (실제로 34개의 Class 분류에 있어서는 비교적 낮은 정확도를 보일 것으로 예상됨)

# Ⅲ. 향후 발전 방향

# Ⅲ. 향후 발전 방향

- 딥러닝 모델을 통해 구체적인 차량을 파악할 수 있을 것이며 이는 CCTV의 성능 향상 혹은 자율 주행에 있어 주위 차량의 탐지하는데 유용하게 쓰일 것이라 예상된다.

- 추가적으로 전체 데이터에 대한 모델링을 다시 시도해봐야 할 것 같다.

# Q & A