주제 소개

학회장팀

 주혜인

 이승우

Reference

Kaggle의 Instant Gratification 대회를 레퍼런스로

데이터 생성 및 평가 지표를 결정

https://www.kaggle.com/c/instant-gratification

데이터 생성

```
NUM_SUB_DATASETS = 32
NUM_SAMPLES = 2048
NUM_FEATURES = 255
MAX_SEED = 2**32 - 1
```

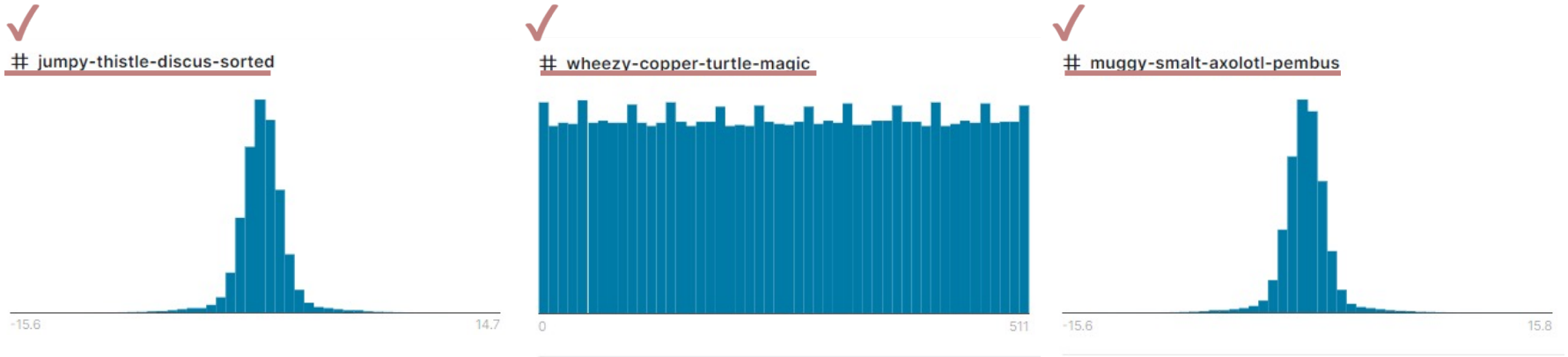
```
def create_dataset(random_seed):
    random.seed(3 + random_seed
    X,y = make_classification(n_samples=NUM_SAMPLES,
                              n_features=NUM_FEATURES,
                              n_informative=random.randint(33,47),
                              n_redundant=0,
                              n_repeated=0,
                              n_classes=2,
                              n_clusters_per_class=3,
                              weights=None,
                              flip_y=0.05,
                              class_sep=1.0,
                              hypercube=True,
                              shift=0.0,
                              scale=1.0,
                              shuffle=True,
                              random_state=random_seed)
    df = pd.DataFrame(X, columns=['V' + str(x) for x in range(0,len(X[0]))])
    df['magic'] = random_seed
    df['target'] = y
    return df
```

아래 링크를 참고로 scikit-learn 라이브러리의

make_classification() 함수를 이용해 데이터 생성

https://www.kaggle.com/c/instant-gratification/discussion/96519

데이터 생성



기존 대회의 경우 변수 이름이 복잡했지만 큰 의미가 없어

불필요한 데이터 탐색으로 인한 시간 낭비를 막고자 변수 명 단순화

데이터 생성

```
seed_list = random.sample(range(1,MAX_SEED), NUM_SUB_DATASETS)
df = pd.concat([create_dataset(s) for s in seed_list], axis=0, sort = False).reset_index(drop=True)
df = df.sample(frac=1, random_state=9726).reset_index(drop=True)

le = preprocessing.LabelEncoder()
le.fit(df['magic'])
df['magic'] = le.transform(df['magic'])
```

```
train = df.groupby(['magic']).sample(frac=0.75, random_state=17)
test = df.drop(train.index, axis=0)
```

```
train = train.sample(frac=1).reset_index(drop=True)
test = test.sample(frac=1).reset_index(drop=True)

```

32개의 데이터셋을 이어 붙여 최종 데이터셋 생성

THANK YOU