M11115Q20 羅述慈 hw1

1. Data Prepare

1. 將資料有缺失的列都刪除，避免資料不足難以預測結果

```
# drop null data
train = train.dropna(how='any')
```

2. 因為欄位很多，我認爲日期、地點不是影響明日將與因素，因此將之移除
3. 風向難以用文字描述，不考慮也將之移除

```
# drop location, date, wind
drop_columns_list = ['Attribute1', 'Attribute2','Attribute8','Attribute10']
train = train.drop(drop_columns_list, axis=1)
test = test.drop(drop_columns_list, axis=1)
```

4. 降雨與否是使用yes/no組成，將其轉換成數字1/0

```
# change yes/no to 1/0
train['Attribute16'].replace({'No':0,'Yes':1},inplace=True)
train['Attribute17'].replace({'No':0,'Yes':1},inplace=True)

test['Attribute16'].replace({'No':0,'Yes':1},inplace=True)
```

2. Model - Decision Tree

決策數使用greedy方法來決定每一層要問哪些問題並且分類過後能夠明確知道是屬於哪一個類別，有測試過depth設定1~7所產生不同結果的正確率，最終選擇depth=5

```
dtree=tree.DecisionTreeClassifier(max_depth=7)
dtree=dtree.fit(train_x,train_y)
dot_data = tree.export_graphviz(dtree,
                                filled=True,
                                feature_names=list(train_x),
                                class_names=['No rain','rain'],
                                special_characters=True)
graph = graphviz.Source(dot_data)
```

```
predict_y = dtree.predict(test_x)
```

3. Result

上傳到public test最終結果為0.68012