

CH6



陳奕利 吳于杏 劉祐辰 宋宜潔

kaggle

[Create](#)

[Home](#)

[Competitions](#)

[Datasets](#)

[Code](#)

[Discussions](#)

[Courses](#)

[More](#)

[Your Work](#)

[RECENTLY VIEWED](#)

- Rui Hachimura's game ...
- Tutorial: Accessing Dat...
- Titanic with AdaBoost
- Gaussian Naive Bayes ...
- Five ways to use value...

[View Active Events](#)

Search datasets

Filters

Computer Science Education Classification Computer Vision NLP Data Visualization

Do You Know Where America Stands On... · Yam Peleg · Updated an hour ago · Usability **10.0** · 3 kB · 2 Files (CSV, other)

1

Club Soccer Predictions · Yam Peleg · Updated an hour ago · Usability **10.0** · 3 MB · 5 Files (CSV, other)

2

Analytics Industry Salaries - 2022 (India) · Sourav Banerjee · Updated 5 hours ago · Usability **10.0** · 58 kB · 1 File (CSV)

8

Most Crowded Airports · khalid · Updated 14 hours ago · Usability **10.0** · 5 kB · 1 File (CSV)

11

Popular Datasets

[See All](#)

Heart Failure Prediction Dataset · fedesoriano · Updated 4 months ago · Usability **10.0** · 9 kB · 1 File (CSV)

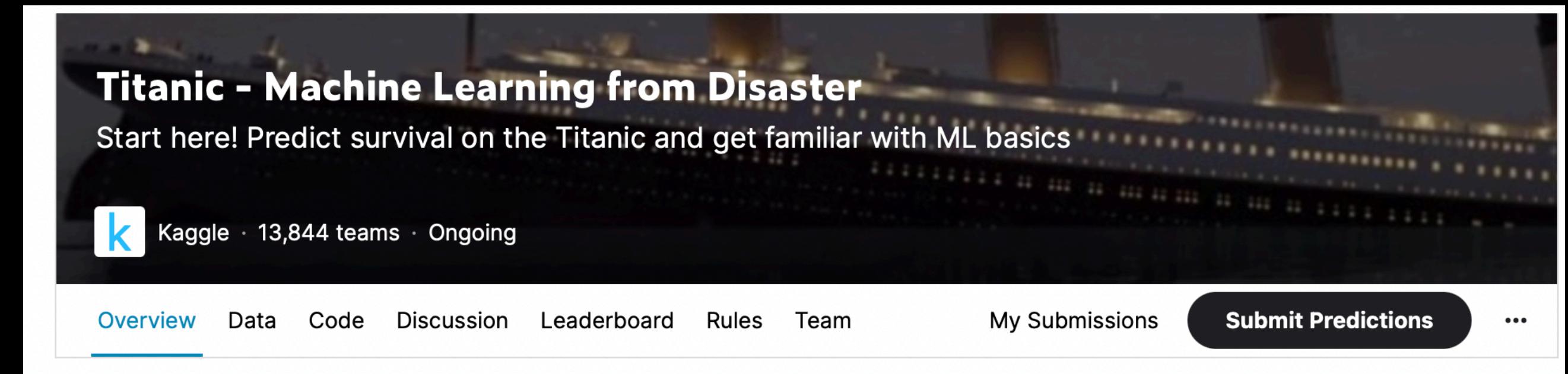
Hollywood Theatrical Market Synopsis 1995 t... · John Harshith · Updated 2 months ago · Usability **9.7** · 6 kB · 9 Files (CSV)

Big Data Certification KR · KIM TAE HEON · Updated a month ago · Usability 7.1 · 16 kB · 5 Files (CSV, other, JSON)

NIFTY-50 Stocks Dataset · Sourav Banerjee · Updated 14 days ago · Usability **10.0** · 3 kB · 1 File (CSV)

用了哪些東西？

資料集: Kaggle的經典項目 鐵達尼號生存分析



編輯器: Jupyter Notebook

語言及套件: Python sklearn, pandas, matplotlib, seaborn

STEP 1

Data Visualization 資料視覺化
&
資料預處理

填缺值

```
# fill in age value
def title_data():
    total_data['Title'] = total_data.Title.replace(['Miss', 'Mlle', 'Mme', 'Ms', 'Lady'], 'Mrs', regex=True)
    re_title_age = total_data[['Title', 'Age']].dropna().groupby('Title', as_index=False)['Age'].median()
    return re_title_age
def fillin_missing_age():
    re_title_age = title_data()
    for index in total_data[total_data['Age'].isnull()].index:
        title = total_data.iloc[index]['Title']
        age = re_title_age.loc[re_title_age['Title'] == str(title),
                               'Age']
        total_data.loc[index, 'Age'] = int(age)
fillin_missing_age()
```

特徵編碼

```
def encoded_age():
    for index in total_data.index:
        mark = 0
        age = total_data.iloc[index]['Age']
        if (age < 17):
            mark = 0
        elif (17<= age <= 33):
            mark = 1
        else:
            mark = 2
        total_data.loc[total_data.index==index, 'Age'] = int(mark)
encoded_age = pd.get_dummies(total_data['Age'], prefix='Age')
data = pd.concat([total_data, encoded_age], axis=1)
return data
```

Age
1
1
2
3
2

獨熱編碼

one-hot-encoding



	Age_1	Age_2	Age_3
1	1	0	0
1	1	0	0
2	0	1	0
3	0	0	1
2	0	1	0

STEP 2

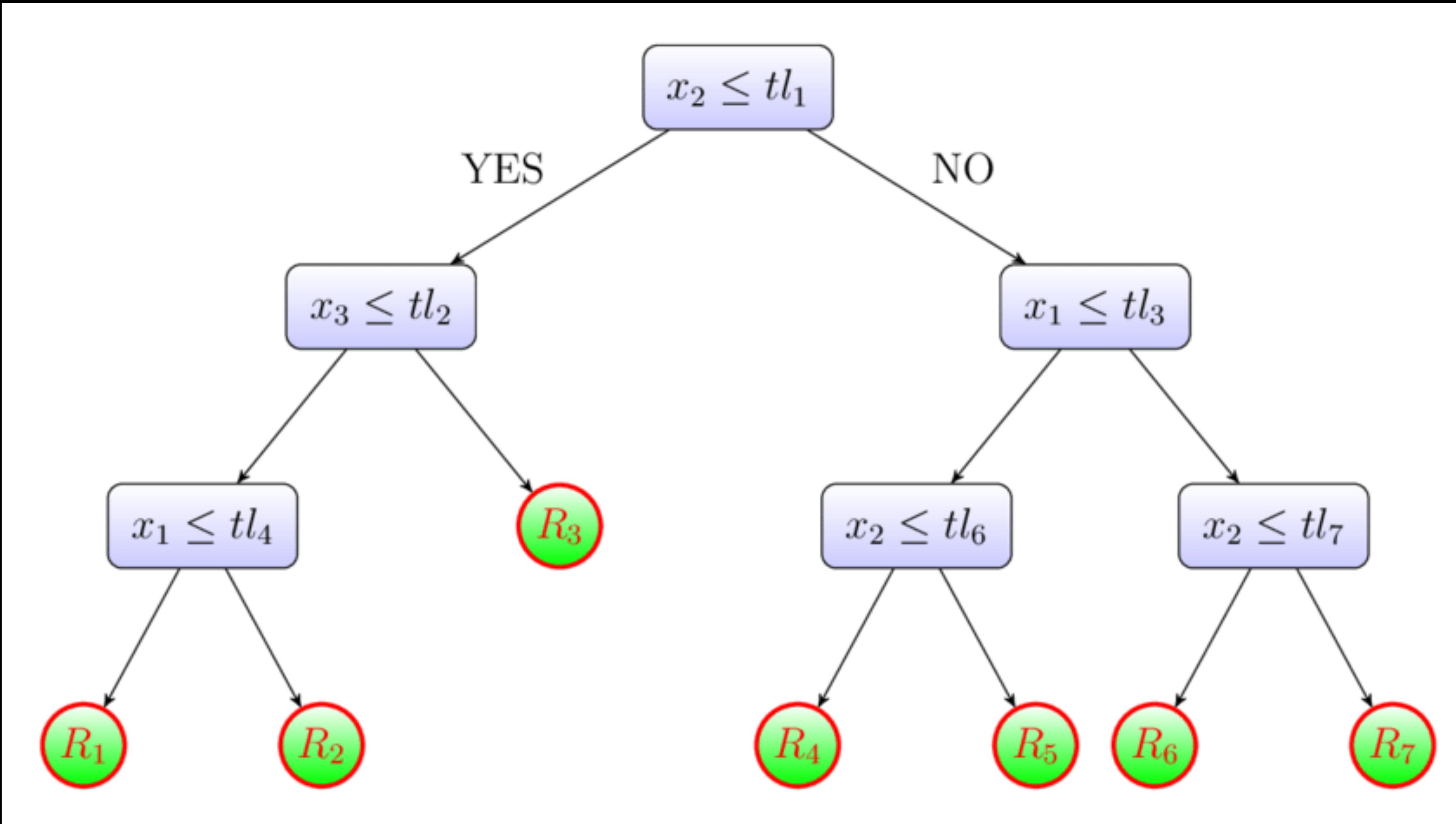
監督式學習

- 決策樹及其衍生
- KNN
- SVM

非監督式學習

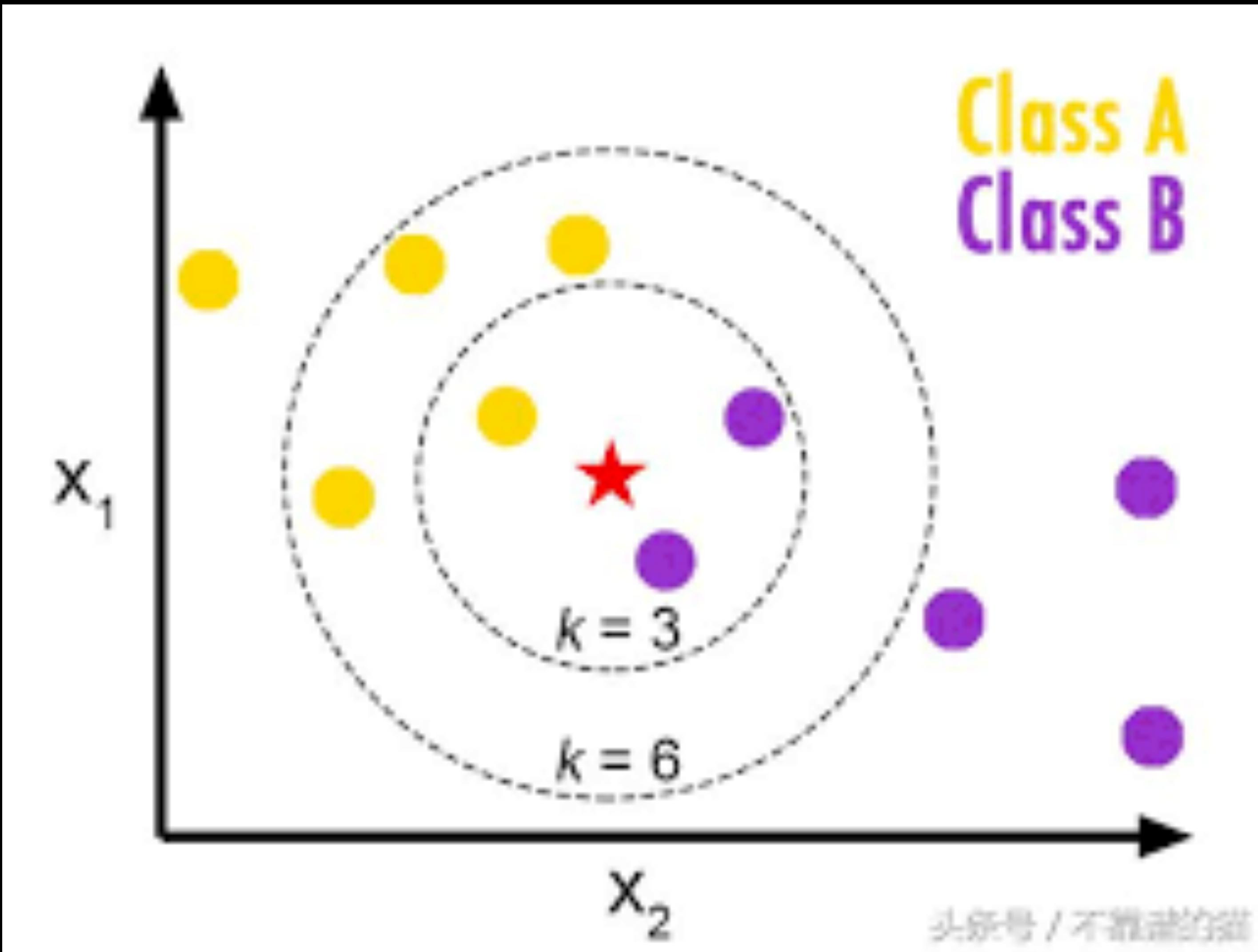
- 集群分析 K-MEANS
- 混合模型

決策樹



KNN k-nearest neighbors

‘從鄰居判斷你是誰’



SVM

