大數據分析與資料探勘期末報告--第18組

台灣股市籌碼面之股票預期分析

組員名單

資源111-何昆錡

資源112-賴彥霖

資源112-劉居衛

大綱

1. 研究背景
2. 研究動機與目的
3. 研究方法
4. 資料來源與處理
5. 研究成果
6. 結果與討論

1 研究背景

根據證券期貨交易所提供之資料顯示：

1. 近年來開戶人數大幅增加，且大多數為年輕且投資經驗小於五年之族群。
2. 2021年平均每位股民賺了新台幣95萬元。

受惠於資本市場熱絡，台股從 2020 至今指數屢創新高，全民大舉瘋台股，開戶年齡層更開始反轉，年輕人買台股趨勢大增，但據 Money101 統計，5 年以下投資資歷的受訪者比例竟達近六成、高達 59%，剛進入股市的年輕人應該多吸收各種專業金融知識，千萬別當股市「韭菜」被收割。

根據證券期貨局最新統計數據顯示，截至 2021 年 5 月底，台股集中市場證券總開戶數為 2129.1 萬戶，相比去年增加 149.4 萬戶。此外，從 2019 年 10 月至 2020 年 10 月，有高達 66%開戶投資人年齡在 40 歲以下，年輕投資族群比例大幅提高。

參考網頁：https://news.cnyes.com/news/id/4665928

不甩疫情與通膨壓力，台股 2021 年仍由多頭主導格局，全年大漲 3486 點，全年漲幅約 24%，市值大增 11.38 兆元，如以開戶數 1196 萬戶計算，平均每位股民今年荷包約賺進 95 萬元。

參考網頁：https://news.cnyes.com/news/id/4794790

那為什麼我沒賺95萬呢？

因為股票終究是**本多終勝**的數學遊戲。

(不是這個)

簡單舉例：

A投了500萬元入股市，年投報率為10%，那他一年賺了50萬元。
B比較窮，只投了50萬入股市，若他想和A賺一樣的錢，那B所投資的所有標的都需要翻倍(年投報率100%)才可達成。

投資有哪些方式？

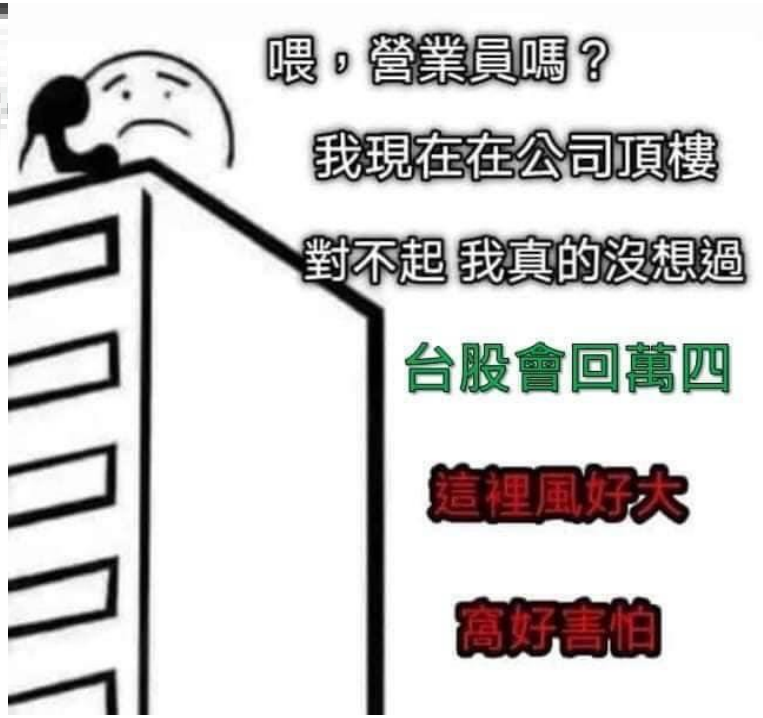
台股主要投資方法：

投資方法	參考依據	風險	最佳狀態之收益	投資方法及標的	投資週期	所需之最重要特質
基本面	年均線 公司財報和營收 EPS 股利發放狀況	低	低	大盤權值股 基金 ETF	半年~數年	**錢**
籌碼面	三大法人、主力 月均線 券資比 千張大戶比例	中	中	中小型股 權證 融資融券	1個月~1季	**大腦**
技術面	量能 5分K線 即時成交動態	高	高	當沖 期貨 權證	數秒~數日	**衝動**

衝動的結果？

開戶年齡大幅下降，許多窮小子抱著雄心壯志前往台灣股市，準備大顯身手，利用高報酬高風險的投資標的，計畫著自己能年投報率500%.....

結局



2 研究動機與目的

在如今這個高速發展的時代，科技日新月異、人民對物質生活越來越講究，頻繁的貿易和日漸提高的消費力，付出的代價是通膨率的不斷提升，我們手上的金錢隨著時間越來越貶值，因此我們必須學會投資（否則會破產，如下圖）。

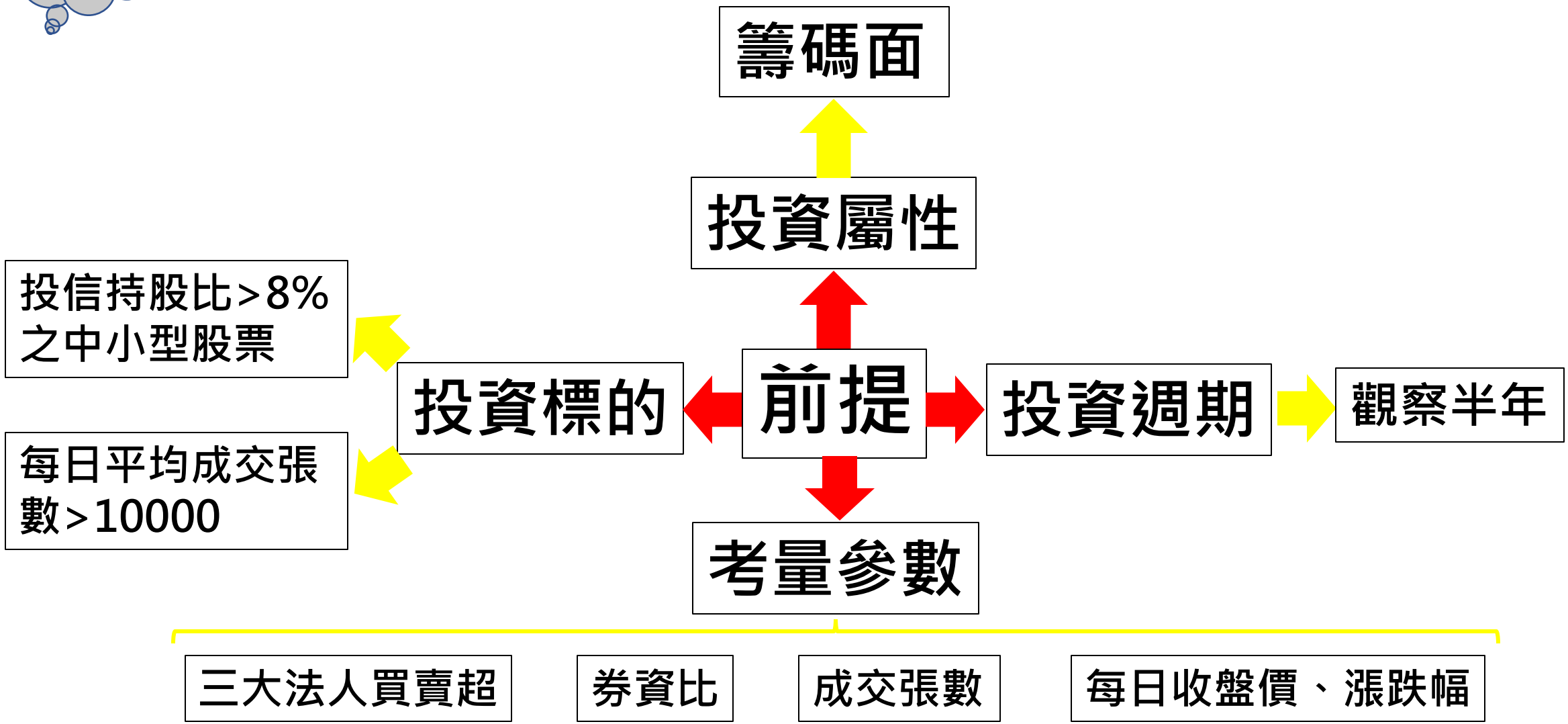
BUT...

人類的判斷通常伴隨著情感，導致無法在投資上做出正確判斷，再加上日常消費的不檢點，導致陷入尷尬局面（如右圖）

因此，本組希望能運用上課所學之資料分析技巧，
設計出一套提高勝率的股票投資策略，
幫助那些和營業員頻繁聯絡的社會新鮮人脫離險境。

//懂得運用科技，數錢數到心悸。//

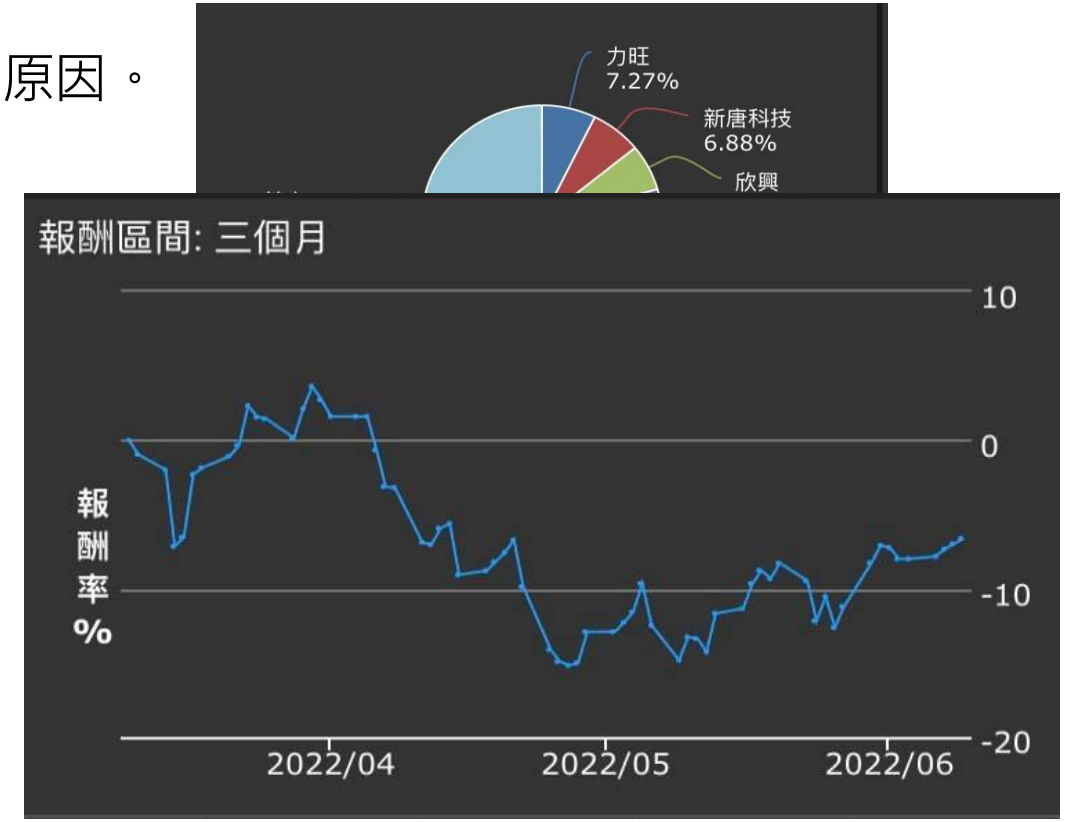
3 研究方法


籌碼面
投資屬性
投信持股比>8%之中小型股票
每日平均成交張數>10000
投資標的
前提
投資週期
觀察半年
考量參數
三大法人買賣超
券資比
成交張數
每日收盤價、漲跌幅

為何要選擇投信持股比>8%的投資標的？

外資，佔台股資金比例較投信高，但通常因避險(匯率)等原因進行買賣。
而**投信**，即是基金經理人，他們創辦基金讓股民投錢進去給他們玩股票，由於不同基金經理人會相互比較每季之績效，每個基金經理人會用盡渾身解數提高自己的投報率，而當一檔股票之投信持股比>8%，即代表**投信有足夠力道使股價波動**。
因此，這是我們認為投信比外資還更具參考價值的原因。

參考資料：三竹股市APP



4 資料來源與處理

先從三竹股市APP挑選投信持股比>8%的股票，再至台灣股市資訊網匯出每日資料。

所需資料：
收盤價、成交張數、三大法人買賣超、券資比

顯示範圍 三個月 匯出XLS 匯出HTML

交易日期	開盤	最高	最低	收盤	漲跌	漲跌(%)	振幅(%)	成交資料				法人買賣超(張)					融資(張)		融券(張)		券資比(%)
								張數	筆數	均張	億元	外資	投信	自營	合計	(%)	增減	餘額	增減	餘額	
06/07	34.55	34.8	34.4	34.75	+0.1	+0.29	1.15	26,233	10,751	2.44	9.08	-629	-35.2	+1,170	+506	19.5	-71	46,270	-92	583	1.26
06/06	33.75	34.65	33.65	34.65	+0.9	+2.67	2.96	28,129	11,110	2.53	9.63	+8,058	+149	+1,165	+9,371	19.5	-342	46,341	-52	675	1.46

參考資料：GOODINFO!台灣股市資訊網https://goodinfo.tw/tw/index.asp

本次挑選的標的：

1. 智原(3035) - 110/8 ~ 111/1

2. 創惟(6104) - 110/11 ~ 111/4

處理過程：

1. 架設環境

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.metrics import r2_score
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
stock = pd.read_excel("3035_Chart.xlsx")
stock
```

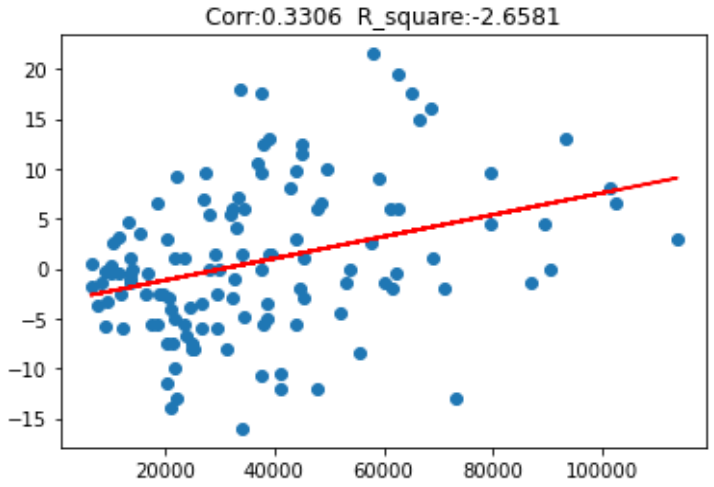
2.

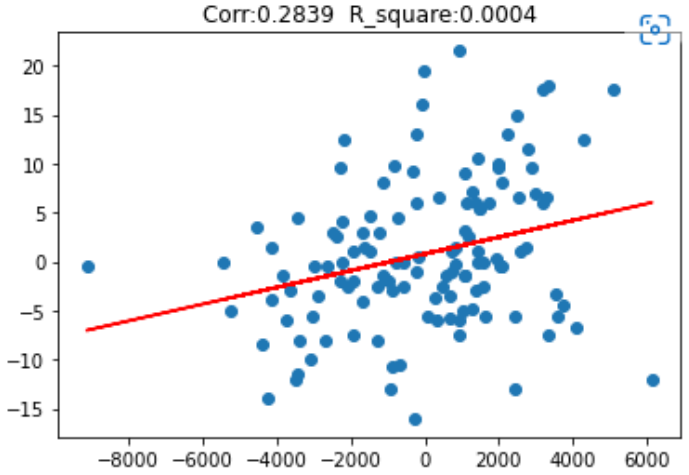
將x分別帶入成交張數、成交筆數、外資、投信、自營商、券資比等，y為漲跌，作圖並求出相關係數Corr與適合度R^2

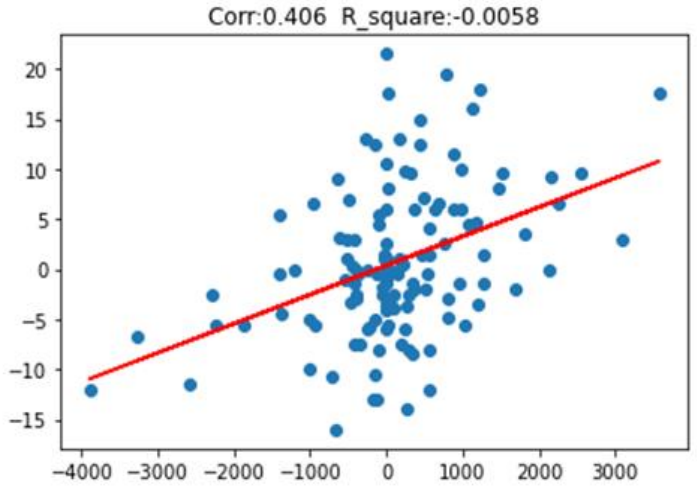
```
x = pd.Series(stock["成交張數"])
y = pd.Series(stock["漲跌"])
corr =round(x.corr(y), 4)
r_square=round(r2_score(x, y), 4)
z = np.polyfit(x, y, 1)
p = np.poly1d(z)
print("corr:", corr)
print("r_square:", r_square)
plt.scatter(x, y)
plt.plot(x, p(x), color="red")
plt.title("Corr:"+str(corr)+"  R_square:"+str(r_square))
plt.show()
```

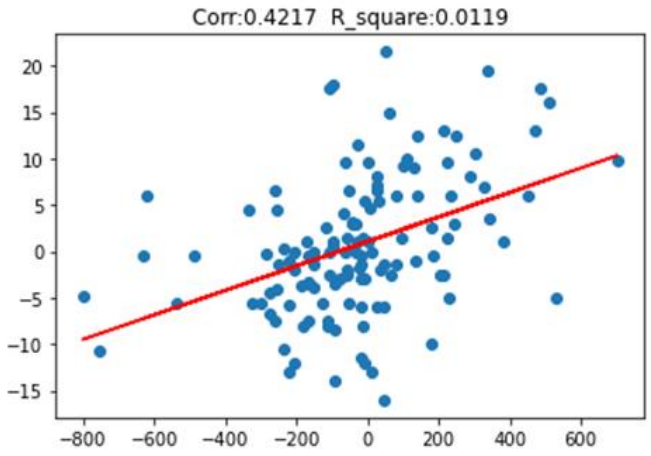
處理過程：

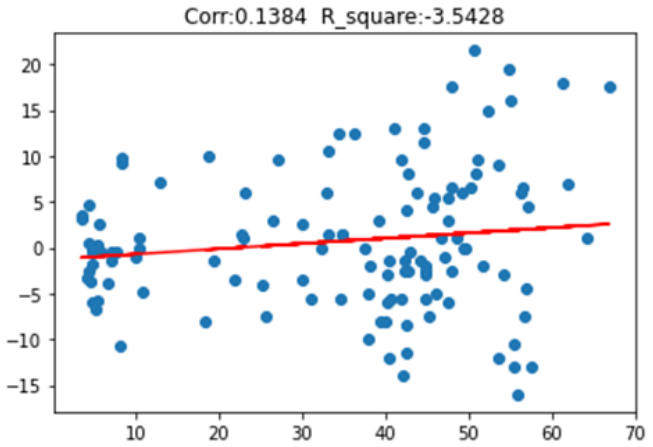
選擇參數由左至右由上而下分為成交張數、外資、投信、自營、券資比











處理過程：

3. 分別將漲跌與漲幅進行分類

```
stock["漲跌"] = np.where(stock["漲跌"] > 0, 1, 0)
stock
```

```
stock.loc[stock.漲跌趴數 >= 5 , "漲幅"]="0"
stock.loc[stock.漲跌趴數 < 5 , "漲幅"]="1"
stock.loc[stock.漲跌趴數 < 0 , "漲幅"]="2"
stock.loc[stock.漲跌趴數 <= -5 , "漲幅"]="3"
```

4. 進行機器學習，分割資料集

```
x = pd.DataFrame(stock[["成交張數", "外資", "投信", "自營", "券資比"]])
y = pd.DataFrame(stock["漲跌"])
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=10)
```

處理過程：

5. 分別進行資訊增益(entropy)，以及gini係數分析

```
from sklearn.tree import DecisionTreeClassifier
tree = DecisionTreeClassifier(criterion = "entropy", max_depth=5, random_state=0)
tree.fit(X_train, y_train)
```

```
from sklearn.tree import DecisionTreeClassifier

tree = DecisionTreeClassifier(criterion = "gini", max_depth=5, random_state=0)
tree.fit(X_train, y_train)
```

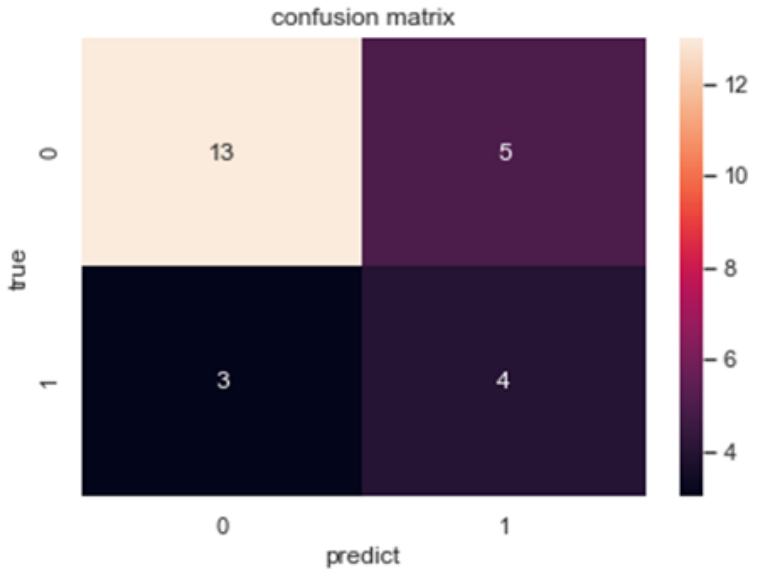
6. 用graphviz做出Decision Tree

```
from sklearn.tree import export_graphviz
class_names = ["跌", "漲"]
export_graphviz(tree, out_file="tree.dot", feature_names=["成交張數", "外資", "投信", "自營", "券資比"], class_names=class_names)
import matplotlib.image as mpimg
import matplotlib.pyplot as plt
get_ipython().run_line_magic('matplotlib', 'inline')
img = mpimg.imread("tree.png")
fig = plt.figure(figsize=(15,10))
plt.imshow(img)
```

處理過程：

7. 做出Confusion Matrix找出易誤判的類別與準確率，再進行模型的調整

```
import seaborn as sns
from sklearn.metrics import confusion_matrix
import matplotlib.pyplot as plt
sns.set()
f,ax=plt.subplots()
y_pred = tree.predict(X_test)
a = confusion_matrix(y_test, y_pred)
sns.heatmap(a, annot=True, ax=ax)
ax.set_title("confusion matrix")
ax.set_xlabel("predict") #x軸
ax.set_ylabel("true") #y軸
```

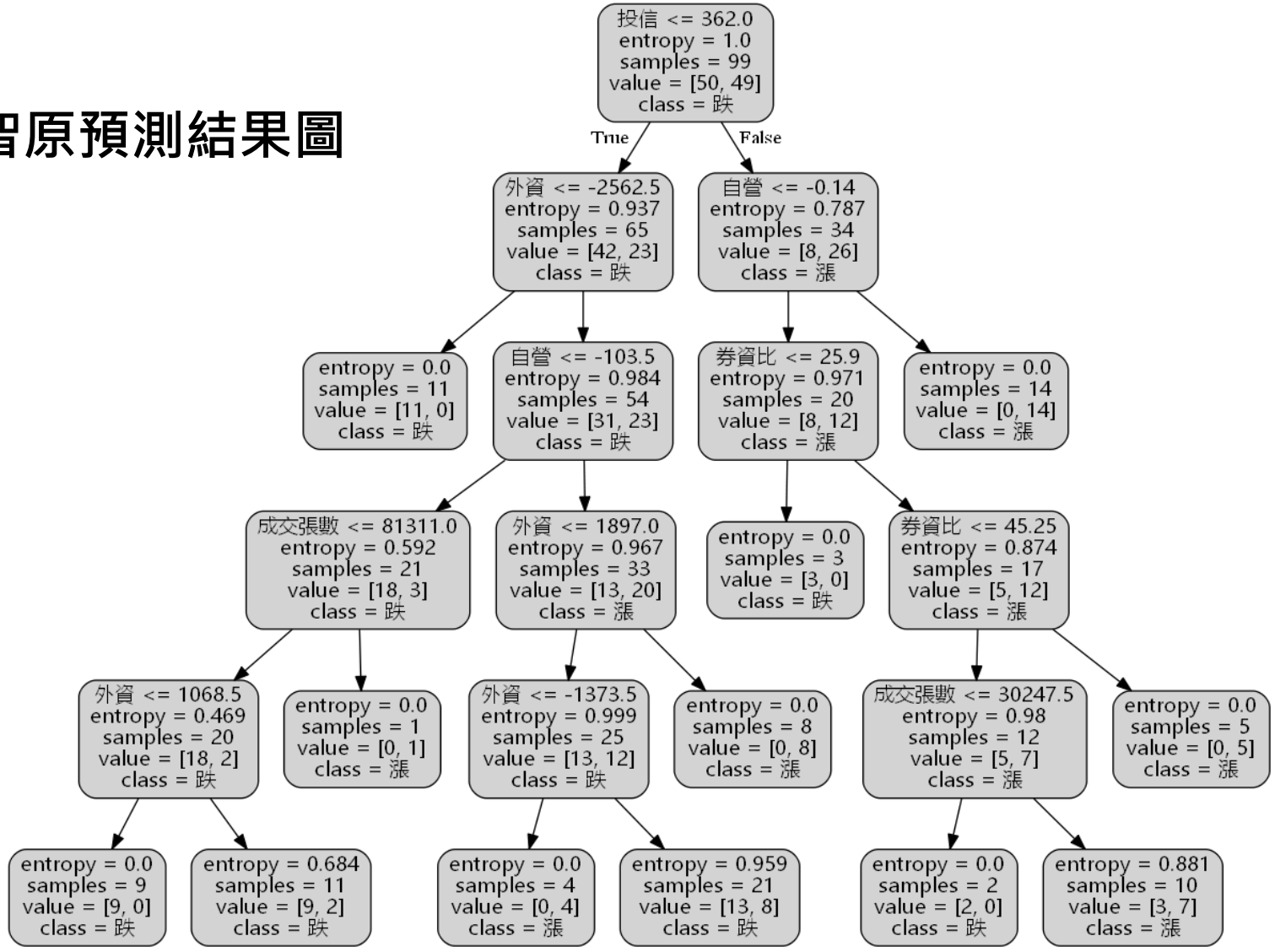



```
tree.score(X_test, y_test["漲跌"])
```

```
0.68
```

5 研究成果

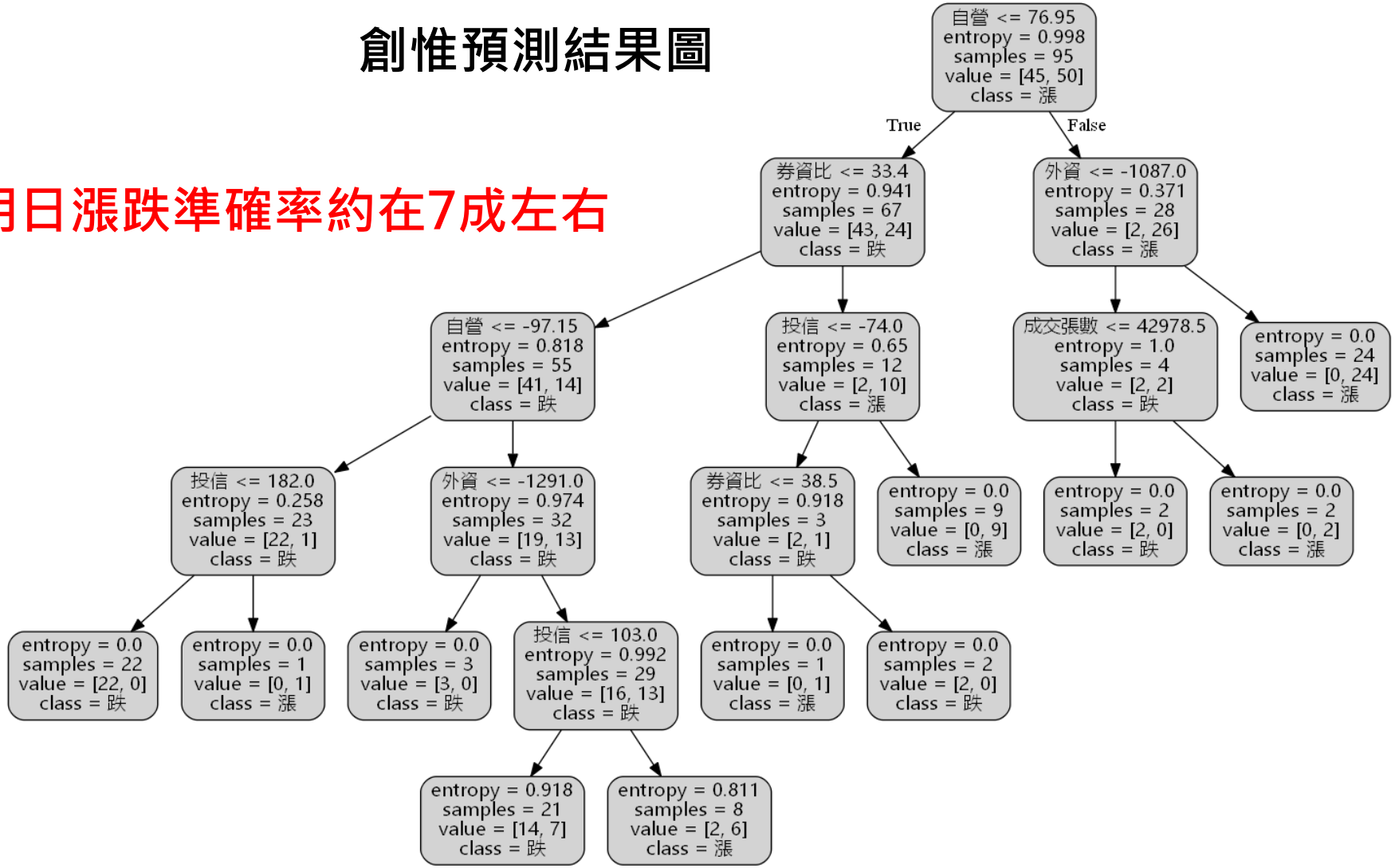
智原預測結果圖



5 研究成果

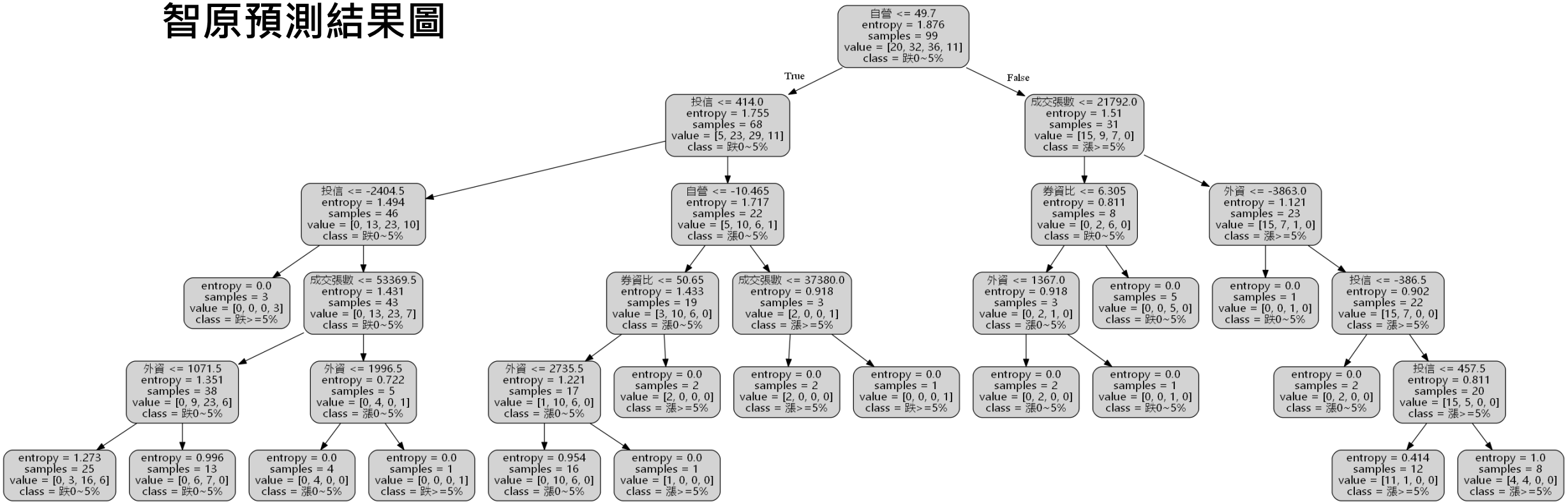
創惟預測結果圖

兩者在預測明日漲跌準確率約在7成左右



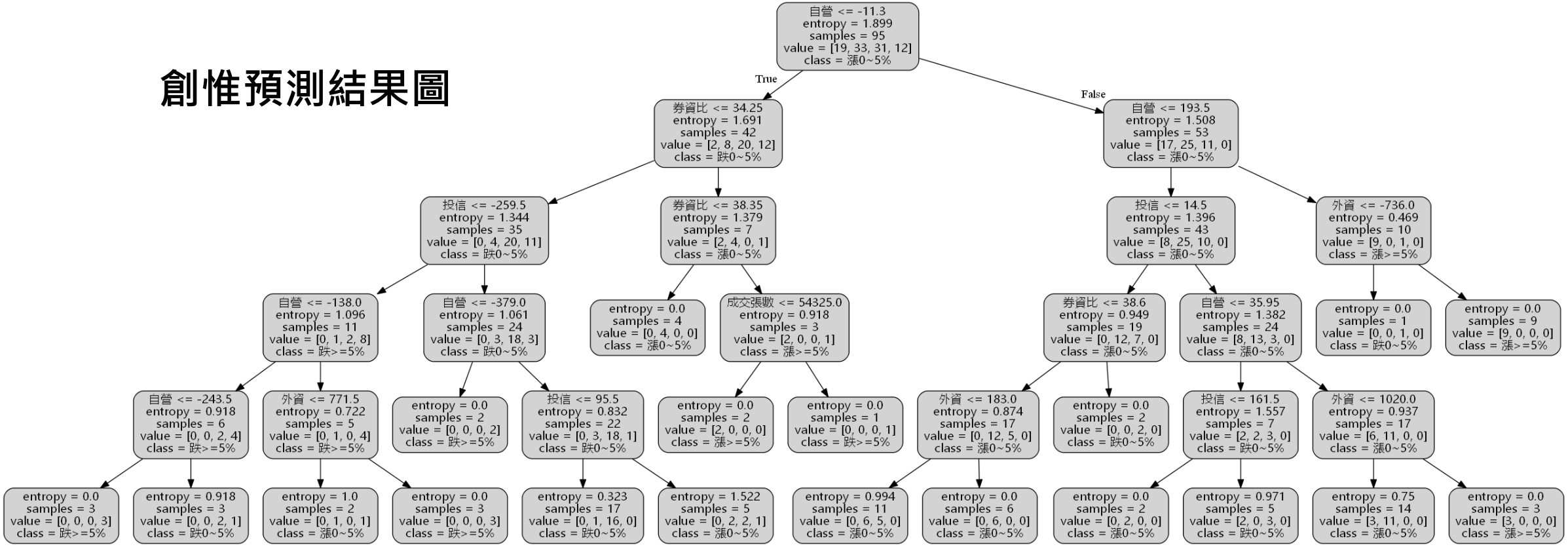
5 研究成果

智原預測結果圖



5 研究成果

創惟預測結果圖



兩者在預測明日漲跌幅度準確率只剩5成左右，較不具參考意義

6 結果與討論

1.兩支股票中，不論預期漲跌或漲跌幅度，都以資訊增益的結果較優。
2.在預測漲幅的成果不佳，可能受到資料組數的不足或應加入如技術分析(KD指標、MACD指標)的參數。
3.由於股票所牽扯的部分過於廣泛，在此模型當中，無法做到相當準確的預測。
4.未來能試著改用神經網路(RNN)的方式去進行深度學習，調配所丟入的參數權重。

祝福大家都能

窮得只剩錢

報告結束

歡迎各位提點與指教
也歡迎志同道合之人能
與我們共同討論
敬祝大家歐趴