程式設計與資料科學導論期末專案

使用基因演算法和機器學習建構選股策略

工工所二 陳禎

中華民國 112 年 12 月

目錄

第一章　研究動機

股市八二法則中表示，股市中大約 80%的人在市場中是虧錢的，只有約 20%的人是賺錢的，因為主動投資大部分的人都是虧錢的，台灣又是 ETF 非常盛行的市場，所以大部分人都在提倡被動投資，主動投資往往被視為吃力不討好的事情，對於被動投資來說，只要每年穩定把錢財放到 ETF 中，幾年後就能享受資產上升的甜美果實，但事實真是如此嗎?難道主動投資真的贏不了大盤?

投資想要賺錢，不外乎在於買到會上漲的股票，而要贏過大盤，買到的股票不只要會漲，更要能漲贏大盤，0050 從 2003 年開始上市到現在的年化報酬率約為 8.52%，代表主動投資每年至少要賺 8.52%以上才能打敗大盤，而且是要連續 19 年，所以顯見要打敗大盤是真的有點難度的，那到底該怎麼做才有機會打敗大盤呢?

所以本研究想要站在初入投資市場的新手的角度，利用所有台股的歷史資料，包括股價和營收資訊，甚至歷年財報資訊來研究並打造一個能穩定打敗大盤的策略。

第二章　研究方法

本研究所使用的工具為 Python，使用的 module 分別為:

Deap、Finlab、Pandas、scikit-learn。

研究方法為先使用基因演算法來建構初步的策略，建構完初步策略後再使用機器學習來優化策略。

2.1 基因演算法(Genetic Algorithm)

定義與背景:

- 一種仿生元啟發式演算法(**metaheuristic)**。
- 模仿生物進化過程中的遺傳和自然選擇。

基因演算法組件:

- 族群(Population):族群是由多個可能解決方案組成的集合
- 基因(Gene):基因是構成染色體的基本單位，可以看作是染色體上的一個特定的屬性或變數。
- 染色體(Chromosome):染色體在基因演算法中代表一個完整的解決方案，它由一系列基因組成。

演算法過程:

- 初始化:隨機生成初始族群。
- 適應度評估:評估每個個體的適應度。
- 選擇:根據適應度選擇個體進行交配。
- 交配:選中的個體交換基因，產生新個體。
- 變異:隨機改變個體的某些基因。
- 迭代:重複適應度評估到變異過程，直到滿足條件結束。

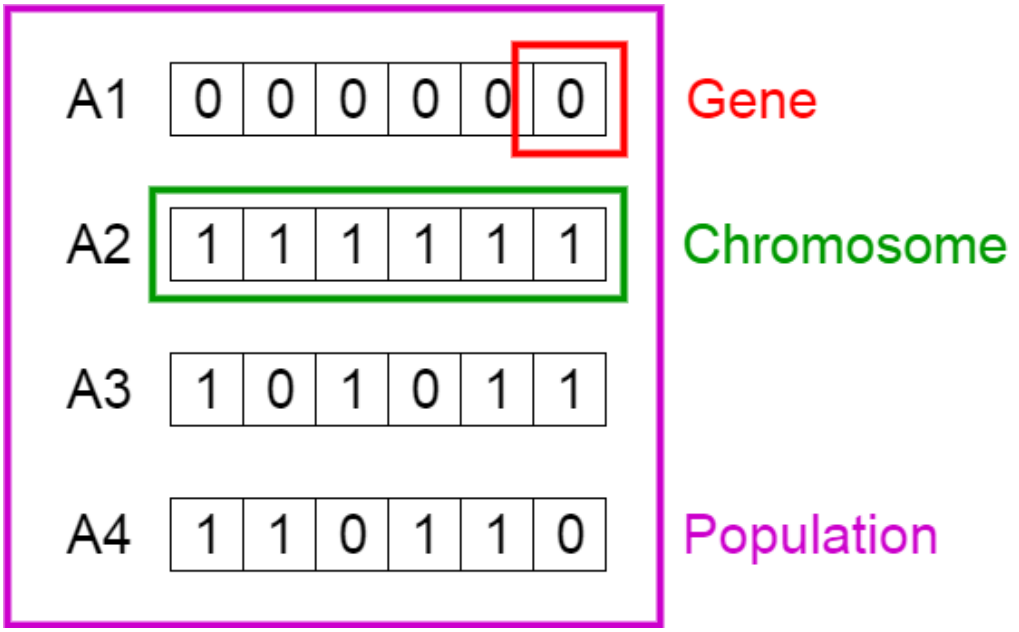

圖 2-1 基因演算法

2.2 機器學習

機器學習，作為人工智能的一個核心分支，正迅速改變著我們分析數據和做出決策的方式。它利用算法從大量數據中學習，不僅能發現數據中的模式和趨勢，還能預測未來事件和行為。從影像識別到自然語言處理，從市場預測到醫療診斷，機器學習正開啟著無限可能，引領我們進入一個數據驅動的未來，本研究使用了其中的 K-Means 聚類算法和決策樹算法。

K-Means 聚類算法：

在機器學習眾多算法中，K-Means 聚類以其簡潔而強大著稱。作為一種非監督式學習方法，K-Means 旨在將數據分成若干個群組，使得同一群組內的數據點相似度最高，而不同群組間的數據點則盡可能不同。這一過程無需事先對數據進行標記，因此非常適用於探索性數據分析，如市場細分、社交網絡分析等領域。

決策樹算法：

決策樹，作為機器學習中的另一種重要工具，提供了一種直觀的方法來進行分類和預測。這種算法通過樹狀結構模擬決策過程，每個內部節點代表一個屬性的檢測，每個分支代表一個決策結果，最終在葉節點給出預測結果。決策樹易於理解和解釋，適用於各種決策支持系統，如醫療診斷、信用評分等領域。

第三章　研究步驟

本研究的研究步驟為:

- 決定基因演算法細節
- 使用基因演算法產生策略
- 分析策略
- 使用機器學習優化策略

3.1 決定基因演算法細節

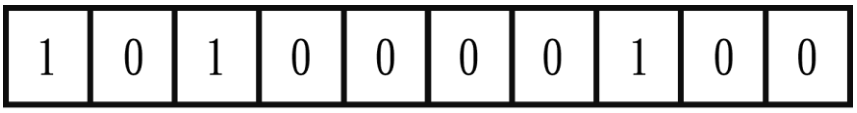

圖 3.1 一條染色體

每個基因都是一個選股條件，1 條染色體代表選股條件的組合，1 代表有，0 代表無，例如圖 3.1 這條然色體有第 1、3、8 個選股條件。

本研究設定此問題的基因演算法染色體長度為 10，代表需要 10 個選股條件，來讓基因演算法幫我找到最好的選股條件，或是最好的選股條件組合。

其他參數分別為:

- 族群大小為 50
- 交配率為 0.5
- 突變率為 0.2
- 演化代次為 20 次

因本研究想站在初入投資市場的新手的角度，所以初始的 10 個條件直接採取詢問 ChatGPT 的方式，直接請他產生 10 個簡易的進出場條件，以下即為 ChatGPT 提供的 10 種進出場條件:

- SMA 五日線向上穿越 20 日線則買入，向下穿越則賣出。
- EMA 五日線向上穿越 20 日線則買入，向下穿越則賣出。
- MACD 線向上穿越 MACD 訊號線則買入，向下穿越則賣出。

- RSI 指標小於 30 則買入，大於 70 則賣出。
- 股價低於布林通道下軌道則買入，股價高於布林通道上軌道則賣出。
- KD 指標都小於 20 則買入，都大於 80 則賣出。
- CCL 指標小於-100 則買入，大於 100 則買入。
- 威廉指標小於-80 則買入，大於-20 則買入。
- MFI 指標小於 20 則買入，大於 80 則買入。
- ROC 指標小於-10 則買入，大於 10 則買入。

3.2 使用基因演算法產生策略

0	1	0	0	1	0	0	0	0	0

圖 3.2 基因演算法跑出的最佳解

圖 3.2 為基因演算法跑出的最佳解，代表同時有第 2(EMA 五日線向上穿越 20 日線則買入，向下穿越則賣出)和第 5 個(股價低於布林通道下軌道則買入，股價高於布林通道上軌道則賣出)選股條件。

年化報酬率	夏普值	最大回撤
8.52%	0.42	-55.75%

圖 3.3 0050 歷史績效

年化報酬率	夏普值	最大回撤
3.06%	0.15	-78.81%

圖 3.4 基因演算法最佳解的歷史績效

圖 3.3 為 0050 的歷史績效，圖 3.4 為基因演算法最佳解的歷史績效，經過比對能發現基因演算法產生的最佳解在各項指標中都還是輸給大盤，最主要的原因為基因演算法的最佳解會一次買入所有符合條件的股票，可能會買到許多不賺錢且較沒有未來性的公司，而股價上漲的本質為公司未來性佳，獲利持續成長，所以這邊再多加入一個條件，條件為只買入去年同月營收增加最多的前 20 檔股票，績效就變為:

年化報酬率	夏普值	最大回撤
28.62%	1.34	-56.82%

圖 3.5 加條件後的基因演算法最佳解

能發現這時策略基本上已經打敗大盤了，年化報酬率和夏普值都非常優秀，只有最大回撤還小輸一點點，但此策略是否還能繼續優化?

3.3 分析策略

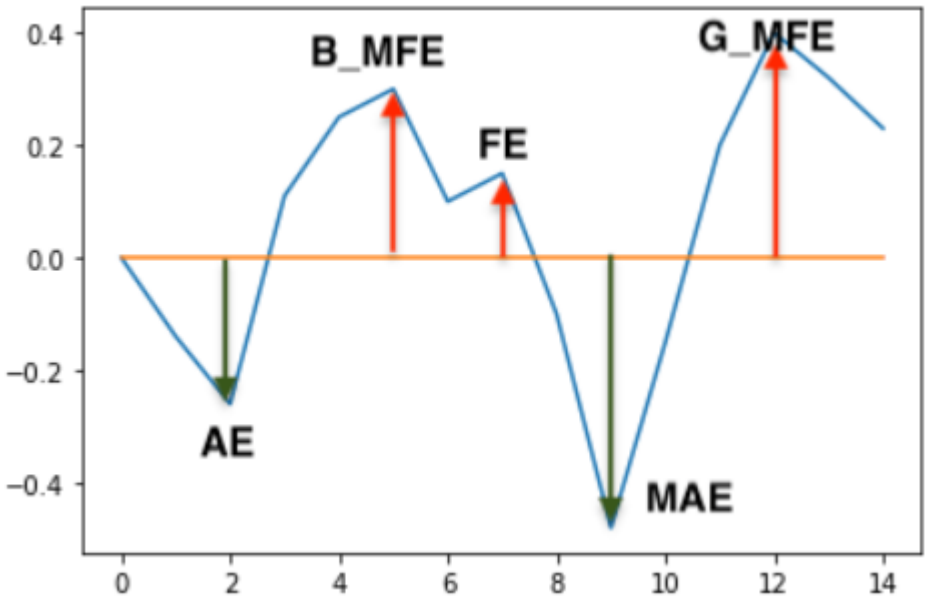

圖 3.6 MAE/MFE 範例

首先介紹 MAE/MFE

- AE (adverse excursion)：不利方向幅度，做多的話，就是下跌的波段。
- MAE ：最大不利方向幅度，做多的話，就是持有過程中的最大累積跌幅。
- FE (favorable excursion)：有利方向幅度，做多的話，就是上漲的波段。
- BMFE ：MAE 之前發生的最大有利方向幅度。若 BMFE 越高，越有可能在碰上 MAE 之前，先觸及停利出場。
- GMFE (Global MFE)：全域最大有利方向幅度。若發生在 MAE 之前，則

BMFE 等於 GMFE。若在 MAE 之後，則代表要先承受 MAE 才可能吃到較高的獲利波段。

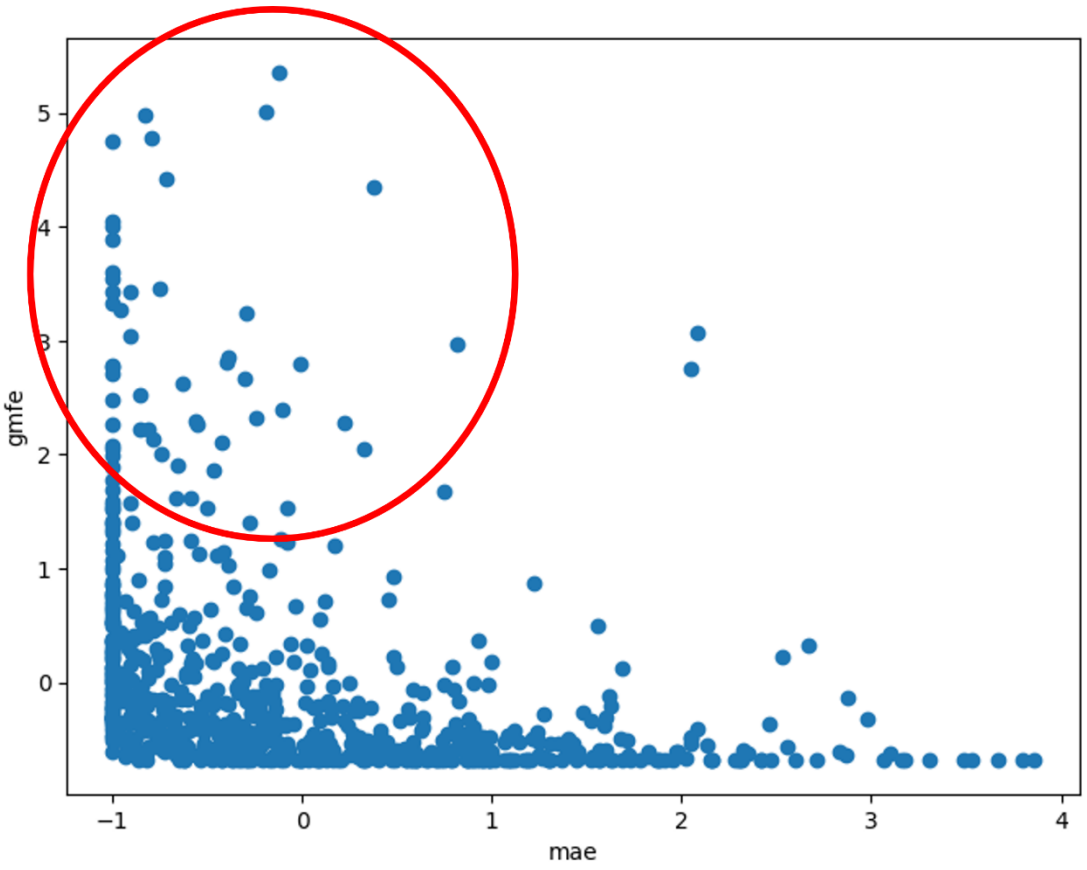

圖 3.7 加條件後的基因演算法最佳解的 GMAE/MFE 圖

想買入的股票為高 GMAE 低 MFE 的區域，也就是左上角那一塊，低 GMAE 高 MFE 的區域是我們絕不想碰的股票。

3.4 使用機器學習優化策略

首先使用非監督式學習 Kmeans 來分群，讓模型自動幫我們藉由高 GMFE 低 MAE 兩項特徵分出三個群集

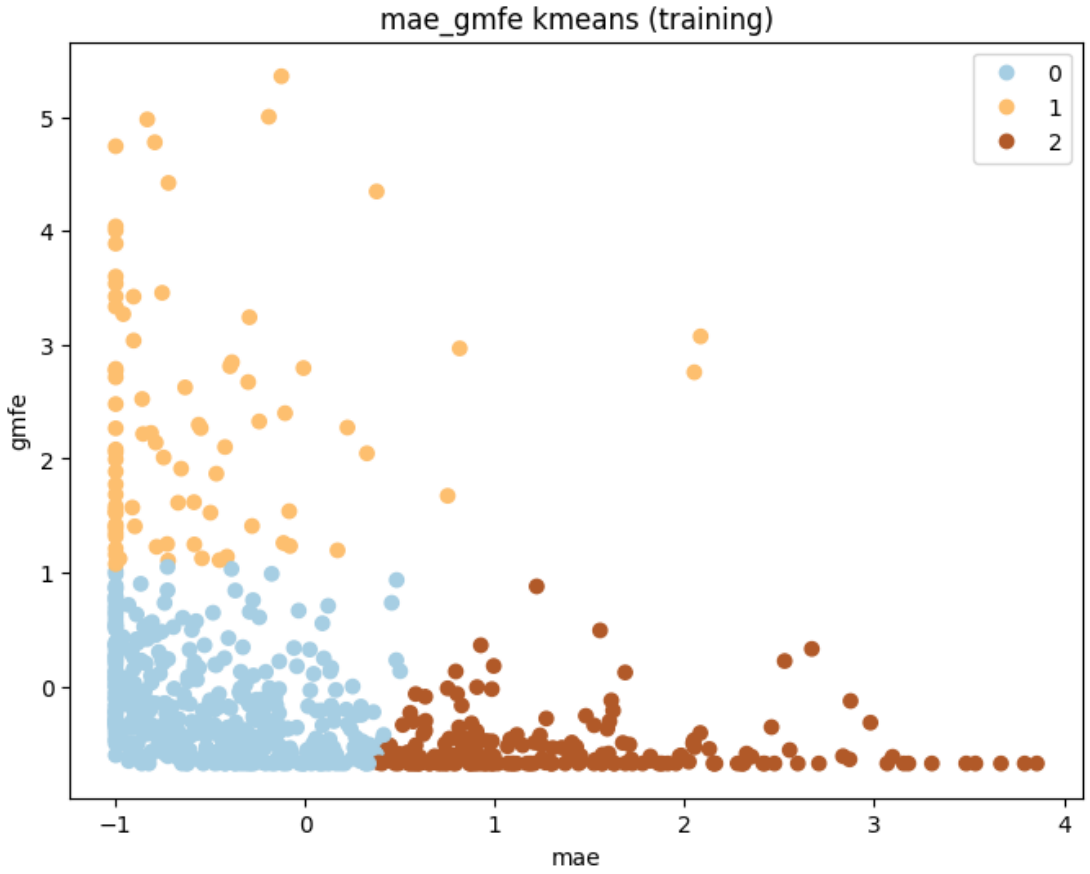

圖 3.8 分群後加條件後的基因演算法最佳解的 GMAE/MFE 圖

進行下一步，尋找有沒有選股條件能辨識 label ，使用監督式學習決策樹。

本研究使用以下資料來製作 features:

- 負債比率
- 存貨週轉率
- ROE 稅後
- 業外收支營收率
- 營業毛利率
- 營業利益率
- 營收成長率
- 現金流量比率
- 融資使用率

- 融券使用率
- 進場時的波動率

將這些資料做成 dataframe 準備進入模型訓練，並將資料以 2019 年為分界點，切成訓練及測試資料集。

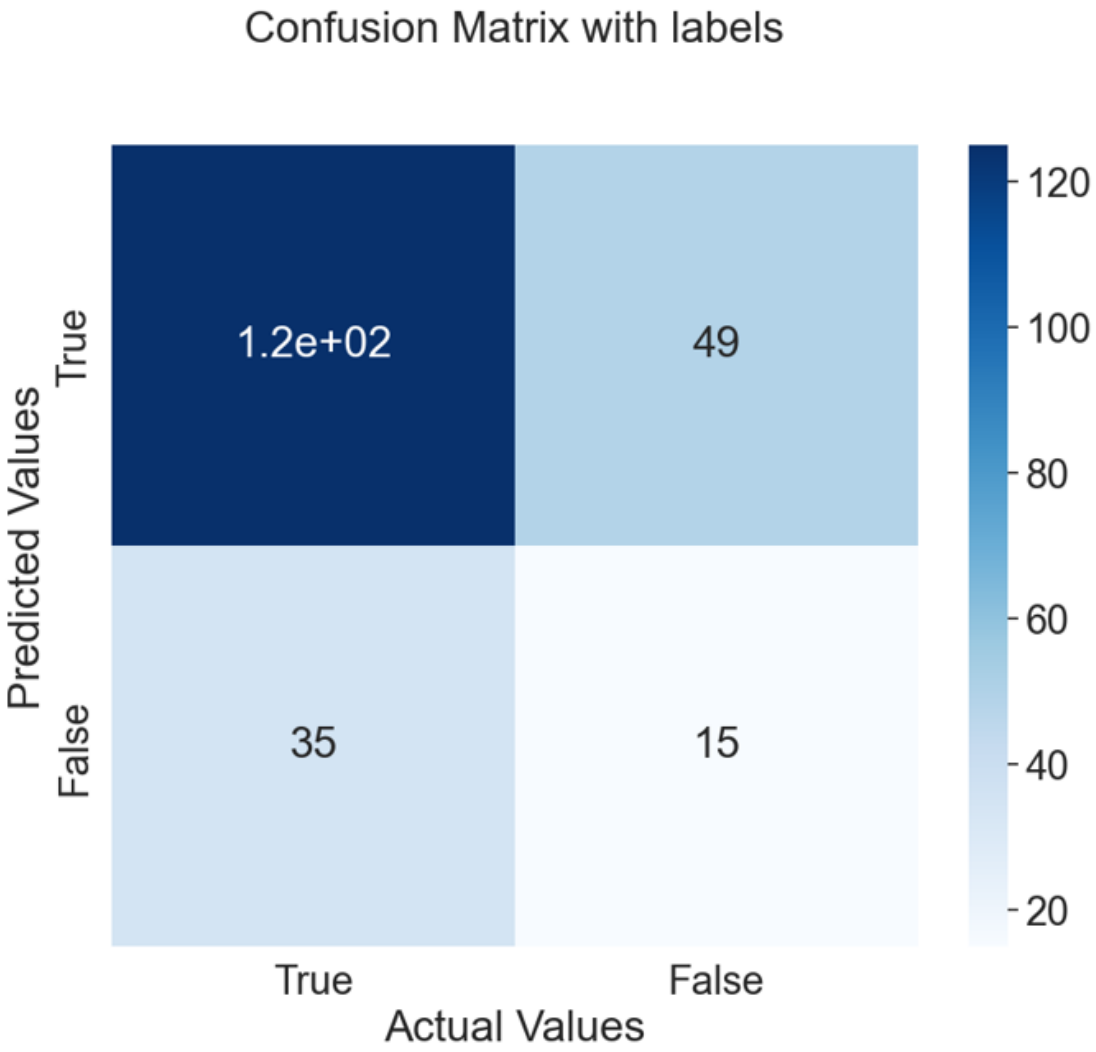

圖 3.9 confusion-matrix 驗證結果

將資料集套入 sklearn 決策樹，驗證目標為 label=2 的集群(低 GMFE 高 MAE)，若找到機率越高，能有效躲避波動。
測試結果放入 confusion-matrix 驗證結果，抓出「低 GMFE 高 MAE 」的準確度達 72%，預測 169(120+49) 次內有 120 次正確。

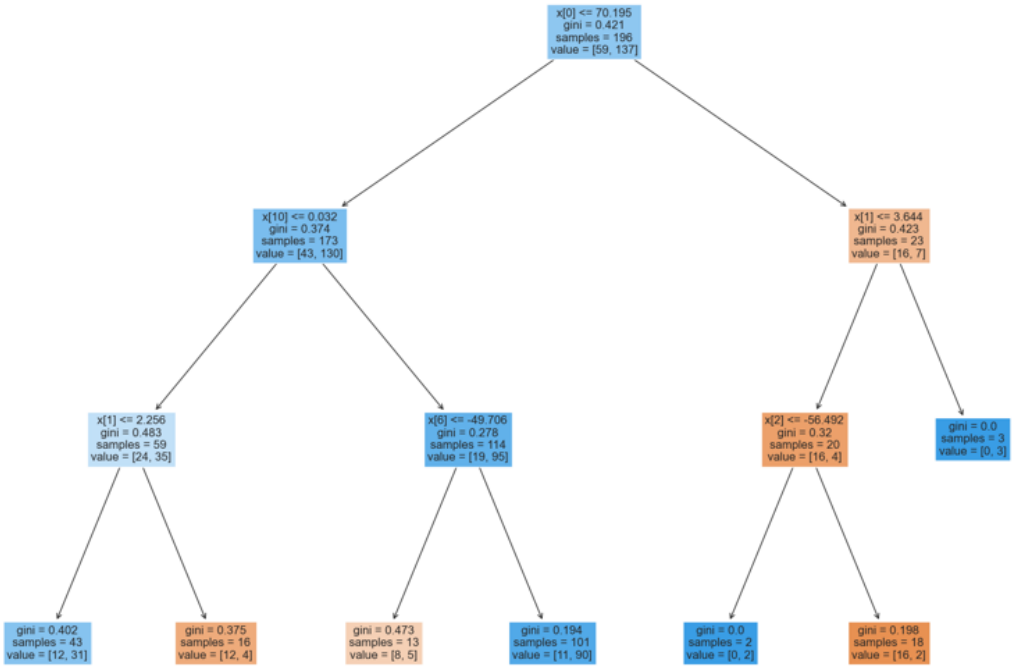

圖 3.10 決策樹決策流程

繪製出決策樹的機器學習選股決策流程，最上頭的 X[0]為負債比率，使用 70.195 當數值分界點，X 序列等同 features 排序。

經測試後，最佳條件為加入 **進場時的波動率**<=0.029，以下為最終策略績效。

年化報酬率	夏普值	最大回撤
21.6%	1.87	-13.61%

圖 3.11 最終策略績效

第四章　研究結果

	年化報酬率	夏普值	最大回撤
0050	8.52%	0.42	-55.75%
基因演算法最佳解	3.06%	0.15	-78.81%
加條件後的基因演算法最佳解	28.62%	1.34	-56.82%
最終策略	21.6%	1.87	-13.61%

圖 4.1 績效對比

圖 4.1 為各績效對比，能發現最終策略在各項數據確實都顯著優於大盤，也代表主動投資絕對是可行的，本研究簡單使用基因演算法加上機器學習的雙流分工，就能輕鬆做出穩定打敗大盤的策略，但以上都是回測資料，過去不代表未來，此策略也只是一個概念，策略和整個方法都還是有很多地方可以優化的地方!

第五章　參考資料

https://alankrantas.medium.com/kmeans-%E8%83%BD%E5%BE%9E%E8%B3%87%E6%96%99%E4%B8%AD%E6%89%BE%E5%87%BA-k-%E5%80%8B%E5%88%86%E9%A1%9E%E7%9A%84%E9%9D%9E%E7%9B%A3%E7%9D%A3%E5%BC%8F%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%E6%BC%94%E7%AE%97%E6%B3%95-%E6%89%80%E4%BB%A5%E5%AE%83%E5%88%B0%E5%BA%95%E6%9C%89%E5%95%A5%E7%94%A8-%E4%BD%BF%E7%94%A8-scikit-learn-%E8%88%87-python-5dd8c0c8b167

https://medium.com/jameslearningnote/%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90-%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E7%AC%AC3-5%E8%AC%9B-%E6%B1%BA%E7%AD%96%E6%A8%B9-decision-tree-%E4%BB%A5%E5%8F%8A%E9%9A%A8%E6%A9%9F%E6%A3%AE%E6%9E%97-random-forest-%E4%BB%8B%E7%B4%B9-7079b0ddfbda

https://deap.readthedocs.io/en/master/

https://www.finlab.tw/display_mae_mfe_analysis/

https://doc.finlab.tw/getting-start/