輔仁大學統計資訊學系

第十九屆專題研究成果報告

指導老師：李鍾斌教授

題目

「Winner winner, chicken dinner」

誰才是最後贏家—以資料探勘預測誰能吃雞

學生：劉律奇、曾士育、阮之群、邱邦旭

蔡宗宏、李奇學、游宏文 撰

中華民國 108 年 12 月

摘　要

題目：「Winner winner, chicken dinner」誰才是最後贏家─以資料探勘預測誰能吃雞

校（院）系所組別：輔仁大學管理學院統計資訊學系

學生：劉律奇、曾士育、阮之群、邱邦旭、蔡宗宏、李奇學、游宏文

指導教授：李鍾斌教授

論文頁數：34

關鍵詞：機器學習、決策樹、分類演算法

論文摘要內容：

隨著電腦與伺服器的配備的進步，要及時存取遊戲中玩家的遊玩資料不再是一件困難的事，而如何從大量的遊玩數據中掌握遊戲的致勝關鍵也一直是各路好手甚至是職業玩家想了解的，因此本研究將以遊戲「絕地求生」為例，透過 AWS 所提供的雲端計算服務處理龐大的遊戲數據，利用機器學習的方法找出其中獲勝的關鍵因素，其使用演算法包括 C.45 與 CART 決策樹、K-最鄰近法與最鄰質心分類法、支持向量機、線性判別比較分析遊戲中的遊玩資料，本研究從分析結果中發現變數 walkDistance、killPlace、matchDuration 對目標變數的影響程度相當大，除了可以做為遊戲獲勝的關鍵因素外，亦可成為電競比賽戰術的考慮要素，並進一步提升遊戲品質與玩家程度。

Abstract

Title of Thesis：”Winner winner, chicken dinner” Who will be the last one, use data mining predict who can be the winner

Name of Department: Department of Statistics and Information Science, College of

Management, Fu Jen Catholic University

Names of Students: Lu-Chi Liu, Shi-Yu Tseng, Chih-Chun Juan, Bang-Xu Qiu Zong-Hong Cai, Chi-Hsueh Lee, Hung-Wen Yu

Advisor: Jung-Bin Li

Total Pages: 34

Key Words: Machine learning, Decision tree, classification

Abstract:

With the advancement of computer and server equipment, it is no longer a difficult task to access the game information of the players in time. And how to master the game's winning key from a large amount of game data is always the thing that all the players or even professional players asked for. Therefore, our study will take the game "PUBG" as an example. Through the cloud computing service provided by AWS to calculate huge game data. Using machine learning method to find the key factors for winning. The algorithms we use including C.45, Cart, KNN, NCN, SVM, LDA to compare and analyze the game data in the game. From the analysis results finds that the three variables

“walkDistance”, “killPlace”, “matchDuration” have a considerable impact on the target variables. Not only identifying the key factors for the game to win, it can also be considered as a tactic factor of e-sports competition, and further improve the quality of the game and the level of players.

謝 辭

專題計畫不只是大學生涯中非常重要的一課，也是一項考驗學生如何活用大學時期所學到的知識與技能來做出成果的挑戰，在進行專題計畫的過程中同時也在考驗著學生的組織能力、溝通能力以及互相配合互相幫忙的責任心與同理心，在此我們特別感謝我們的指導教授李鍾斌教授的指導與協助，總是在我們遇到問題時有求必應，提供相關的資訊和資源以及幫助我們指點迷津，在此透過此專題計畫展現我們在這一年來的學習以及成長的成果。

目 錄

頁次

表　目　錄

頁次

圖 目 錄

頁次

第壹章、前言

遊戲在人類的歷史中一直存在著很重要的地位，從前簡易的遊戲演變到有繁複規則的線上遊戲。與各式球類運動相同，由單純的遊戲發展成為萬眾矚目的競技活動，線上遊戲也是如此，從而產生可觀的利潤。

本章節將說明本研究中的研究背景、研究動機和定義研究目的，利用各種機器學習方法分析遊戲數據，探討實際影響遊戲勝負的重要因素。

第一節、研究背景與動機

隨著網路發展的快速進步與電子設備日新月異的更新，網路遊戲在一般社會大眾的日常網路使用習慣中所佔的比例也越來越高。根據在 2018 年的統計顯示，國內 12 歲以上的人上網率達到了 82.1%而其中又有 41.6%的人會將大部分的時間花在線上遊戲的遊玩上，這些人平均每週遊玩線上遊戲的時數就長達 11 個小時，無論是遊玩的人數或時數上都有都佔有相當大的比例（李雅萍，2018）。在這個情況下，線上遊戲的產值相對的也就開始提升，現金儲值、限定虛寶以及實體週邊商品都成了遊戲公司主要營利來源，但是在發展電子競技之後，電子競技變成了遊戲公司另類的商業機會，企業紛紛投入電競產業，吸引了大量的人力物力進入該產業，企業開始培養選手，若是能讓選手在比賽中大放異彩，相當於在主要客群面前打響了知名度，進而提高公司利益。

本次專題研究的遊戲「絕地求生」吸引玩家的原因有以下幾點：

一、射擊遊戲已是當今最為熱門的遊戲之一，其中「絕地求生」這款遊戲加入了更多不確定的因素豐富了遊戲內容，例如競爭對手比起其他的射擊遊戲多出許多，多達九十多人。另外每位玩家在遊戲開始時身上的資源都相同，需要透過蒐集、戰鬥以增加身上的資源，增加了遊戲的可玩性。

二、「絕地求生」能吸引多種類型的玩家，以往的射擊遊戲更注重玩家自身的遊戲技術，此遊戲除了可以滿足喜歡追求刺激的玩家在一開始就能大殺四方；能讓喜歡運用戰術取勝的玩家能夠充分的展現運籌帷幄的組織能力；多樣化的造服裝造型也能滿足喜歡設計搭配的玩家表現自我。

三、「絕地求生」在公平性上面也下足了功夫，在遊戲中儲值的點數並不會影響玩家的能力。所有的裝備和武器都是在遊戲中拾取，更改武器造型以及服裝無法改變本身的數值，沒有儲值的玩家在此遊戲中與有儲值的玩家享有平等的待遇。

基於上述的條件，加上其特殊創新的競技玩法，讓這款遊戲迅速地吸引了大量的遊戲用戶，且成為了電子競技的新寵兒，許多企業不斷相繼投資成立戰隊與培養選手，大大小小的賽事如雨後春筍般地在世界各地舉辦，使得「絕地求生」在短時間內就發展出了完整的賽制與電子競技市

場，在比賽中獲勝便是每個隊伍的主要目標，此時將遊戲中的各項數據加以分析就成為了獲勝條件的其中一個重要因素。

第二節、研究目的

「絕地求生」是一個講究團隊精神的運動，每一局比賽玩家必須在手無寸鐵的情況下開始遊戲，從地圖中找尋各種資源，並必須與其他玩家互相爭鬥獲得生存，地圖上有著複雜的地形和建築、涵蓋範圍非常大，加上各種槍械、補給品等物資都是隨機在地圖上生成的，所以根據當下情況做出最佳決策，甚至在遇到敵人時都要權衡利弊後再決定是否射擊。

遊戲中可以單人遊玩或最多 4 位玩家組成一隊，每局遊戲最多會有 100 名玩家同時進行，每次進行遊戲都會產生大量資料。其中包含玩家的移動位置與距離、物件道具的消耗、擊殺與救援、遊戲進行時間與排名，我們將利用這些數據進行決策樹分析，找出影響遊戲獲勝的關鍵因素，並建立模型以預測比賽排名結果。

本研究的目的如下：

一、多種資料採礦方法來建立預測模型。

二、比較出最優秀的模型。

三、利用模型預測比賽排名與結果。

使用機器學習模型如下：

C4.5、Cart、KNN、NCN、LDA、SVM

第三節、研究架構

本研究分五個章節，各個章節安排為界定研究主題與目的、探討相關文獻、擬定研究方法、資料分析與解釋、結論。

一、界定研究主題與目的：

確定本研究的問題與目的，作為研究的方向

二、探討相關文獻：

依據研究的主題與目的，進行蒐集與研究相關之理論，整理相關的研究方法與結果。

三、擬定研究方法：

參考現有的原理、原則、經驗法則或研究結果，建立適當的研究模式，以符合研究問題及目的。

四、資料分析與解釋：

將蒐集到的資料加以整理、彙整，並運用適當的分析工具和技術，進行資料的處理以及分析。

五、結論與建議：

根據資料分析、解釋以及發現，做成研究的結論，依據結論，針對現狀做成具體的建議，並提出後續的研究方向。

第貳章、文獻探討

本章節中討論的內容包括電競產業對於商業發展的潛力及了解各種分析演算法，針對這部分相關文獻的研究，並將這些研究文章加以整理，做為本論文研究的參考基礎。

第一節、電競產業

近年來，電子競技（以下簡稱電競）逐漸被各國重視，在中國及美國更有許多電競賽事已發展成如同籃球的 NBA 一樣的聯盟制度，也就是主客場制度及更具規模性的組織發展電競賽事。根據 PWC（資誠聯合會計師事務所）2017 的報告指出，電競產業的產值從 2012 年的 12 億新台幣逐年增加，2017 年已正式突破 120 億新台幣（見下圖）；同一份報告更指出，台灣 2016 在電玩遊戲的收入高達 510 億新台幣，排名世界第十。由此可知，台灣電競產業發展的走向與全球快速成長的趨勢大致相同，顯示台灣電競產業的未來發展具有相當大的可看性。

圖 2-1-1 電競產業之產值估計，資料來源：整理自資誠台灣（2017）

2012 年 10 月 14 日，台北暗殺星（TaipeiAssassins）在北美取得英雄聯盟第二季世界錦標賽冠軍（中華民國電子競技運動協會，2017），這振奮台灣許多遊戲玩家，更使得「英雄聯盟」這款遊戲在台灣迅速地打響名號，也讓「電競」在台灣開始被重視。在 2017 年 11 月 7 日，立法院正式三讀通過將「電子競技」列入運動產業，但是相關的法律細節等皆尚未修改完成，因此許多長期關心及推動電競產業的人們都十分期盼政府的下一步。（朱泓任，2017）

在賽事方面，遊戲公司為了推廣，通常電競賽事舉辦時亦會有實況轉播。根據電競實況平台 Twitch 所公布的資訊指出，台灣在電競實況收看流量位居全球第五（4GANERS，2015）；從這則報告中可以知道在台灣有大批的人群是不斷的關注電競。任何比賽有足夠觀眾流量，就會有廠商願意投資，就算是遊戲比賽也是一樣，雖然電競並不是傳統的運動比賽，可背後所表示的商業原理卻是一樣的，只要廠商願意投資，就能持續舉辦比賽，進而再吸引更多贊助商投資隊伍，這也使得電競產業有了正向的循環能夠成長茁壯（羅悅軒、曾凌軻，2017）。

隨著觀看電競賽事流量不斷增加，投資人進入電競產業的意願就開始提升，例如周杰倫、黃立成等藝人紛紛成立自己的戰隊（何世昌，2017）。除此之外，韓國及香港的投資者也前來台灣與企業合作籌組戰隊，更引進韓國電競相關的制度以培訓人才，希望透過這樣的合作共同發展台灣的電競產業（王昱澄，2017）。除了「英雄聯盟」這項遊戲作為比賽，早期的「星海爭霸」、「跑跑卡丁車」、「爐石戰記」和近期竄升的「絕地求生」也在電競產業中佔有一席之地，台灣的電競產業鏈中不乏選

手以外的各項人才，例如主播、賽評等專業人才，更有練習生制度培訓新秀選手。

放眼世界的電競產業，也興起了一股企業紛紛開始投資此產業；例如：美國 Riot 公司於 2017 年 9 月宣布於韓國建造直營且專屬的電競場館作為英雄聯盟賽事使用。在中國，騰訊公司推出「五年計畫」，與政府合作共同建立電競規範及培養人才，更計畫建設十個的電競園區。此外，騰訊已與蘇湖市政府簽約共同打造電競小鎮，包含主題公園、電競大學、文創園區、科技創業社區及雲計算中心等多項建設結合，顯示騰訊公司對於電競產業的積極投入（林宸誼，2017）。

作者	論文名稱	結論
涂國濠(2017)	電競遊戲觀賞動機、體驗、價值對行為意圖之研究-以 DOTA 2 電競為例	目前觀眾還是以年輕人為主。
宋傑恩(2019)	電競賽事對消費者購買相關週邊產品意願之研究	電競賽事是可以促進消費者對於電競週邊產品的購買意願。

表 2-1-1 電競研究實例

綜上所述，近來各國企業紛紛投入電競產業，可以發現電競產業有潛在的發展可能，並有著極大的商機，亦被各政府機關重視。

第二節、資料探勘

2019 年，我們的生活早已處在一個資訊爆炸的年代，巨量的資訊量遠遠超過人類所能處理的能力。為了能從這些龐大的資料中，找出有利的資訊，資料探勘隨之出現。資料探勘被稱為探究的資料分析，可以研究分析減少及重複使用特定資料庫中所產生的大量資料。搜尋議題可以是不同模

式的銷售預測、市場反應及利潤。（David Olson & Yong Shi，2008），1996 年 Fayyad 認為資料探勘為依照使用者需求，自資料庫中選擇合適的資料，加以處理、轉換、探勘至評估的一連創過程。2014 年 C.L. Philip Chen 與 Chun-Yang Zhang 認為利用資料探勘技術，能夠有效的處理大量的數據，進而找出能提高競爭力的方式。2002 年蘇詠翔認為在商業性資料探勘的過程中，應用最成功的地方就是在行銷學上，也就是所謂的資料庫行銷，亦即利用資料探勘所挖掘出來的資料，分析顧客偏好並依此方向發掘潛在客戶或顧客需求。2014 年 Nawsher Khan et al.認為在有效的分析技術中，包含了資料探勘、視覺化方法、統計分析及機器學習。1997 年 Berry & Linoff 以資料探勘是利用自動或半自種的方式來探索並分析大量資料，已從中挖掘出有意義的樣型或規則。用於資料探勘的技術有非常多種，例如：決策樹分析、類神經分析、菜籃分析、群集分析、迴歸分析、基因演算法。不同的方法可以處理不同的目的，不同的資料類型，各有各自的優缺點。其中的決策樹分析，算的上是最早的機器學習演算法之一。

一、決策樹

早在 1966 年，Hunt、Marin 和 Stone 提出的 CLS 學習系統中就有了決策樹演算法的概念，1979 年，J.R Quinlan 提出了 ID3 演算法雛形，1986 年，Schlimmer 和 Fisher 在這基礎上又更進一步進行改造，提出了 ID4 演算法，1993 年，Quinlan 改進了 ID3 演算法，提出了 C4.5 演算法（鄭捷，2017）。2002 年 Moran & Bui 說明決策樹分析屬於監督性的分類演算法，由使用者決定要分類哪一個欄位，以及利用其他欄位做分類之解釋與描述，將混雜的資料做有效的分類，整理出有結構化的架構知識。2003 年 Pal & Mather 決策樹分析 是一個類似樹狀結構的流程圖，層級性的方式，將知識做有結構化的表現，並且依據不同的樹狀流程，整理出不同的決策規則。決策樹具有監督式的特徵萃取與描述的功能，將輸入變數根據目標設定來選擇分支變數與分支方式，並以樹枝狀的層級架構呈現， 以萃取分類規則，經過修整後的決策樹模型可以做為資料探索或預測。

陳樹衡教授、郭子文教授與棗厥庸教授提到決策樹之概念是依事物的特徵，將事物區分為不同的種類，每一種類再對應不同的決策模式。例如在文中的房地產模型，即根據房屋的特徵來區分不同類別，各類別中再建立本身價格模型。在現有的決策樹分類器中，以 ID3 (Quinlan, 1986)、

C4.5 (Quinlan, 1993)、CART (Breiman et al., 1984)最廣為採用。（陳樹衡、郭子文、棗厥庸，2007）。決策樹分析被應用於多種領域且都有不錯的表現。

演算法步驟如下

1. 將樣本分成兩組，訓練組資料與測試組資料
2. 使用訓練組資料建立決策數
3. 使用測試組資料進行修剪
4. 持續上述第 2~3 步驟，直到所有新內部節點都是樹葉節點為止。

(一)、C4.5 決策樹

由 Quinlan 於 1993 年提出，為他先前提出之 ID3 的延伸，改良了 ID3 產生過多的子集合，而每個子集合僅有少數資料的問題，且更具有處理連續型變數、雜訊及決策樹修剪的能力。

決策樹學習的關鍵在於，在每個分裂節點處如何選擇最優劃分屬性。一般而言，隨着劃分過程不斷進行，我們希望決策樹的分支節點所包含的樣本儘可能屬於同一類別，即節點的「純度」越來越高。C4.5 主要是透過增益率(Gain Ratio)來劃分屬性，先從劃分屬性中找出信息增益高於平均水平的屬性，再從中選擇增益率最高的。

信息增益(Information Gain)：

在訓練過程中決策樹會問出一系列的問題，像是年齡是否>30 歲，年收入是否<100 萬之類的是非問題。由最上方的樹節點開始用資料的特徵將資料分割到不同邊，分割的原則是：這樣的分割要能得到最大的資訊增益(Information gain)。由於我們希望獲得的資訊量要最大，因此經由分割後的資訊量要越小越好，信息增益的資訊量最常使用的是熵(Entropy)，其公式如下方程式(1)：

(1)

p(i | t) ：屬性的機率，c：分類類別，i：第幾個類別的資料總數

增益率(Gain Ratio)：

實際上，信息增益準則對可取值數目較多的屬性有所偏好，為減少這種偏好可能帶來的不利影響，C4.5 決策樹算法不直接使用信息增益，而是使用增益率(Gain Ratio)來選擇最優劃分屬性。因為增益率偏向選擇可取值數目較少的屬性，C4.5 算法並不是直接選擇增益率最大的候選劃分屬性，而是使用了一個啟發式的方法：先從候選劃分屬性中找出信息增益高於平均水平的屬性，再從中選擇增益率最高的。

(2)

GA 即為 information gain ratio 之縮寫。Gain(A)表示已屬性 A 作為分支屬性對資訊的貢獻程度，Info(D)為原資料的 Entropy，k 為該節點內的樣本數，N 為樣本總數。Split Info(A)指分支 A 屬性的資訊量，l 為該節點內的樣本數，N 為樣本總數。(葉子維 2016)

(二)、CART 決策樹

分類與迴歸樹（Classification And Regression Tree，CART）CART 演算法是建構決策樹時最常用的演算法之一。自從 1984 年布里曼（L. Brieman）與其同僚發表這種方法以來，就一直機械學習實驗的要素。(CH.Tseng,2019) 如果目標變數是標稱的，並且是具有兩個以上的類別，則 CART 可能考慮將目標類別合併成兩個超類別（雙化）;如果目標變數是連續的，則 CART 演算法找出一組基於樹的迴歸方程來預測目標變數。

CART(Classification and Regression Tree)演算法中利用基尼指數構造二叉決策樹。Gini 係數是一種與資訊熵類似的做特徵選擇的方式，可以用來衡量資料的不純度，當每個預測值在節點中出現頻率都一樣，則值越小；而當個節點指涵義個預測值，則值為最大。不純度越小，即表示此屬性越適合當作分枝。

Gini 係數的計算方式如下：

Gini(D)= 1-j=1kPj2

其中，D 表示資料集全體樣本，pj 表示每種類別出現的概率。取個極端情況，如果資料集中所有的樣本都為同一類，那麼有 p0=1，Gini(D)=0，顯然此時資料的不純度最低。

吉尼獲利 GiniGain(A,S)可表示為：

GiniGain(A,S)=Gini(S)−Gini(A,S)

對於連續型節點:

對於連續值處理引進“分裂點”的思想，假設樣本集中某個屬性共 n 個連續值，則有 n-1 個分裂點，每個“分裂點”為相鄰兩個連續值的均值。

表 2-2-1 連續型節點

先把連續屬性轉換為離散屬性再進行處理。雖然本質上屬性的取值是連續的，但對於有限的取樣資料它是離散的，如果有 N 條樣本，那麼我們有 N-1 種離散化的方法：<=vj 的分到左子樹，>vj 的分到右子樹。計算這 N-1 種情況下最大的 Gini。

（一）對特徵的取值進行升序排序

（二）兩個特徵取值之間的中點作為可能的分裂點，將資料集分成兩部分，計算每個可能的分裂點的 Gini。優化演算法就是隻計算分類屬性發生改變的那些特徵取值

（三）選擇 Gini 最小的分裂點作為該特徵的最佳分裂點（N 是連續特徵的取值個數，D 是訓練資料數目）

(三)、過度配適(Overfitting)

決策樹是一種分類器，通過 ID3，C4.5 和 CART 等算法可以通過訓練數據構建一個決策樹。但是，算法生成的決策樹非常詳細並且龐大，每個屬性都被詳細地加以考慮，決策樹的樹葉節點所覆蓋的訓練樣本都是“純”的。因此用這個決策樹來對訓練樣本進行分類的話，你會發現對於訓練樣本而言，這個樹表現完好，誤差率極低且能夠正確得對訓練樣本集中的樣

本進行分類。訓練樣本中的錯誤數據也會被決策樹學習，成為決策樹的部分，但是對於測試數據的表現就沒有想象的那麼好，或者極差，這就是所謂的過度配適(Overfitting)問題。

產生過度配適(overfitting)的原因有兩個，第一個是物體本身的屬性 太多，有些和種類相關，有些和種類不相關，所以，決策樹演算法容易選用到和種類不相關的屬性，換句話說，就是自由度太大。要避免這種問題的發生，就是要盡量減少物體描述裡可能和種類不相關的屬性。但是，這並不容易，因為在一般的情況下，可能不會有足夠的資訊讓我們清楚地了解哪些屬性和種類相關，哪些和種類不相關，因為這就是我們希望決策樹幫我們找到的。第二個原因是偏見(bias)，每個屬性選擇演算法在尋找測試屬性時，都有自己的偏好，所以非常有可能會找到演算法所偏好，但不是真正和種類相關的屬性。 (2002,廖雅郁) 在決策樹方面對於過度配適(Overfitting)，我們可以使用預剪枝(Pre-Pruning)的方法來避免

預剪枝(Pre-Pruning)：預剪枝是根據一些原則及早的停止樹增長，如樹的深度達到使用者所要的深度、節點中樣本個數少於使用者指定個數、不純度指標下降的最大幅度小於使用者指定的幅度等。

其中最常使用的兩項門檻值參數為「節點容錯率」與「節點最小樣本數」，節點容錯率是在當節點內的樣本分錯率小於該值的時候即停止分割，而節點最小樣本數則是在限制該節點內的樣本數必須多於多少，如果節點內的樣本數少於某值，即停止分割；另外根據國外資料探勘專家 Michael J. A. Berry 所著作的書 Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management 中提到，最小門檻值最好不要低於總案例數的千分之一，且不要使用預設值 2 當門檻值。

另外，在資料量較大、變數要為多樣的情況下，建議設定樹的「最大深度」當作門檻值，可以有效控制過度配適的問題發生，具體的取值取決於資料的分佈，常用的可以取值 10 到 100 之間。在應用決策樹分析探討植群空間分布特徵-以曾文水庫集水區為例中作者提到最小門檻值應為 100，深度為 13 為參數(2017,陳元豪)，以決策樹模型探討未開立慢性病連續處方之影響因子中提出 30(2011,蔡佳玲)。

下表 2-2-2 為決策樹應用實例

作者與時間	研究題目	結論
周佩蓁(2005)	以決策樹分析顧客滿意度之研究	主要以顧客關係中管理網路消費、消費者行為模式及後續發展研究。透過 C4.5, CART, CHAID 三種決策樹演算法比較分析顧客滿意度並探討研究結果 ,使廠商瞭解現今與顧客之間的互動及後續發展機會
呂文吉(2016)	應用決策樹資料探勘模式於醫療院所資訊設備故障排除之研究—以北部某區域醫院為例	實際採用某區域醫院現有的資訊室報修紀錄資料，將該醫院的設備異常通報，針對其報修屬性進行 C4.5 決策樹分析，找出經常故障的硬體項目，並找出故障項目所對應之排除的方法，進一步建置一套有規則的資訊設備故障排除之模型

表 2-2-2 決策樹應用實例

二、K-最鄰近法(K-Nearest Neighbor, KNN)

K-最鄰近法(KNN)是一種監督分類演算法，也是以案例學習(Instance-based learning)為基礎之機器學習演算法。此分類法主要運用向量空間模型，將每份文件以多維度特徵表示，並根據距離度量函式計算待分類樣本 x 和訓練集中的每個樣本的距離，對計算出的距離排序，選擇與待分類樣本最近的 k 個訓練樣本作為 x 的 k 個最近鄰，如果這 k 個最近鄰中屬於某一類的樣本佔多數，則分類樣本 x 將歸為該類別。具體的 KNN 演算法大致可分為以下六個步驟：

一、建立一個訓練樣本集和測試樣本集。

二、設定 k 值。一般先確定一個初始值，然後根據實驗結果反覆調整至最優。

三、計算測試樣本和每個訓練樣本的歐氏距離。

四、選擇 k 個近鄰樣本。將計算出的距離降序排列，選擇距離相對較小的 k 個樣本作為測試樣本的 k 個近鄰。

五、找出主要類別：根據 k 個近鄰的類別並應用最大概率對所查詢的測試樣本進行分類。所用概率指的是每一類別出現 k 個近鄰中的比例，其根據每一類別出現在 k 個近鄰中的樣本數量除以 k 來計算。那麼擁有最大概率的類別記為主要類別。

六、將待測樣本歸至所屬類別。

其公式如下：

(3)

score(d,ci)表示待分類文件 d 屬於 ci 的可能性分數；KNN(d)表示 K 個最鄰近的訓練文件，dj 屬於 KNN(d)；Sim(d,dj)表示待分類文件 d 與 dj 的相似度；δ(dj ,ci)表示 dj 是否屬於 ci ，屬於 ci 則為 1，反之為 0。(2014,吳俊穎)

KNN 普遍使用歐氏距離，最常見的兩點之間或多點之間的距離表示法，又稱之為歐幾里得度量，它定義於歐幾里得空間中，如點 x = (x1,...,xn) 和 y = (y1,...,yn) 之間的距離為：

(4)

另外，在調整 K 值時必須特別注意，當 K 值設定過小會降低分類精準度；若設定過大，且測試樣本屬於訓練集中包含資料較少的類，則會增加噪聲，降低分類效果。K 值的設定通常會採用交叉檢驗的方式（以 K=1 為

基準)，根據經驗法則，K 值一般會低於訓練樣本數的平方根。(K 近鄰 k-Nearest Neighbor(KNN)演算法的理解-程式前沿,2018)

下表 2-2-3 為 KNN 應用實例

作者與時間	研究題目	結論
楊帆(2016)	基於改進 KNN 演算法的室內 WIFI 定位技術研究	通過從無線地圖中過濾掉到標籤處沒有相似 RSS 向量的 RP 來尋找最近鄰,以降低 KNN 演算法的時間和計算複雜度 ,並增加定位精確度。
張碩(2017)	基於 KNN 演算法的空間手勢識別研究與應用	在研究機器學習分類算法和 HTC 的 VIVE 設備應用系統的基礎之上,提出基於 KNN 算法的空間手勢識別研究並且在 VR 環境下進行了實際應用。

表 2-2-3 KNN 應用實例

三、最鄰質心分類法(Nearest Centroid Neighbor)

此方法與 k-近鄰分類算法相似而又有所不同：相似之處是都要通過最大的投票原則來確定測試樣本所屬的類別，不同的是 NCN 利用了 k 個近鄰與測試樣本點的相似性和空間分布特點。NCN 分類步驟如下：

給定 m 維特征空間的一組訓練樣本集 T=xi∈ Ri=1mn，該樣本集 T 有 n 個訓練樣本，有 m 個類，其類別的標簽為 c1,…,cM。對一個測試樣本點 x，NCN 通過如下步驟確定 x 的類標簽：

(1)選定測試樣本集合和訓練樣本集合；

(2)確定質心近鄰的個數 k 的初始值；

(3)按照式(1)的計算方式，從 T 中查找測試樣本點 x 的 k 個近質心近鄰點，記作質心近鄰點的樣本集為

(5)

(4)根據得到 k 個近質心近鄰點，計算每個類別所包含的近鄰質心近鄰點的個數，將含有最大近質心近鄰點個數的類別標簽分配給待測樣本點。

(6)

其中：

(7)

方程式(7)中 yincn 是第 i 近質心近鄰點 xincn 的類別標簽。

通過以上描述，NCN 和 KNN 都是通過投票原則，從訓練樣本集中查找待測樣本點的 k 個近鄰，但是確定 k 個近鄰的方法是截然不同的過程，這 k 個近鄰點也不完全相同。

KNN 算法通過距離確定訓練樣本和測試樣本的相似性，而 NCN 不僅通過距離衡量 2 個樣本的相似性而且還強調選取的 k 個近鄰點應盡可能地分布在測試樣本點的周圍。(謝紅，趙洪野，解武，2015)

下表 2-2-4 為 NCN 應用實例

作者與時間	研究題目	結論
謝紅，趙洪野，解武 (2014)	基於局部權重 k - 近質心近鄰算法	LWKNCN 有效地克服了 KNN 和 KNCN 算法上的不足，在近鄰選取上不同於 KNN 和 KNCN，更符合實際分類情況。

Jamshid Tamouk, Farshid Allahakbari (2012)	A comparison among accuracy of KNN, PNN, KNCN, DANN and NFL	本文比較了 KNN，PNN，KNCN，DANN 和 NFL 方法之間的結果和準確性。最終，KNCN 方法在其他提到的方法中效率最高。

表 2-2-4 NCN 應用實例

四、線性判別分析(Linear Discriminant Analysis)

LDA 的全稱是 Linear Discriminant Analysis(線性判別分析)是一種監督式學習，是一個模型,不需要去通過概率的方法來訓練、預測資料。LDA 的原理是,將帶上標籤的資料(點),通過投影的方法,投影到維度更低的空間中,使得投影後的點,會形成按類別區分,一簇一簇的情況,相同類別的點,將會在投影后的空間中更接近。要說明白 LDA,首先得弄明白線性分類器(Linear Classifier):因為 LDA 是一種線性分類器。對於 K-分類的一個分類問題,會有 K 個線性函式。

線性模型:

(9)

其中 為樣本的第 n 種特徵。線性模型的形式為： ，其中 為每個特徵對應的權重生成的權重向量。

LDA 的基本思想是：訓練時，將訓練樣本投影到某條直線上，這條直線可以使得同類型的樣本的投影點儘可能接近，而異型別的樣本的投影點儘可能遠離。要學習的就是這樣的一條直線。

預測時，將待預測資料投影到上面學習到的直線上，根據投影點的位置來判斷所屬於的類別。

圖 2-2-3 為同樣使用鳶尾花數據集來做訓練，標準化後的數據結果如下:

圖 2-2-3 LDA 鳶尾花實驗

可以看出，LDA 除了上述特性外，還能夠實現降維。假設存在 M 個類，則多分類 LDA 可以將樣本投影到 M-1 維空間。

下表 2-2-5 為 LDA 應用實例

作者與時間	研究題目	結論
王俊興(2017)	以粗集合理論和線性判別分析在半導體產業之篩選股票投報率的可行性研究	研究以蒐集不同研究公司之投資條件屬性資料來分析，並將粗糙集理論搭配之線性判別分析(LDA),能找出長期投報率及整體評估。
黃建齊(2014)	應用線性判別分析法改善衛星資料估算颱風生成潛勢指標之研究	研究成果顯示:若設定 60%發展機率為門檻值，本方法預報發展個案命中率達 96.4%，且平均可早於 JTWC 發佈 TS 警報前 53.1 小時發佈颱風生成預警。

表 2-2-5 LDA 應用實例

五、支持向量機(Support Vector Machine, SVM)

支持向量機，Support Vector Machine，簡稱 SVM，是一種分類演算法， 由 Vapnik 等根據統計學習理論提出的一種機器學習方法，在解決小樣本、非線 性及高維度模式識別問題中有相當強大的優勢。

圖 2-2-1 二分概念圖

如圖 2-2-1 所示，分類時我們希望兩筆數據分割的間隙越大越好，能夠把兩個 類別的點分的越開越好。兩個類別之間的間隙越大能夠讓分類器更精確的進行判斷、分類。在 SVM 中，這個最大間隙稱為 Maximum Margin，間隙的最外圍有兩個與最優分隔超平面(Optimal separating

hyperplane，簡稱 OSH)等距，支撐著中間分界的超平面稱為支持超平面 (Support hyperplane)，而支持超平面上必定會有一些數據點(如圖中三個圈起來的點)，這些數據點就是所謂支持向量 (Support Vector)。

在 SVM 中參數□用來平衡分類器的複雜度和不可分數據點的數量，它可被視為一個由用戶依據經驗或分析選定的正則化參數，在 SVM 的程式封包中，參數□代表著訓練中允許著多少個錯誤出現，□的值 越大意味著訓練數據中允許的誤差越少，但要注意的是最優分隔超平面傾斜的 程度是會隨著□值的不同而改變的，過高的□值可能造成過擬合(Overfitting) 的發生，如圖 2-2-2 所示：

圖 2-2-2 □值差異

在□ = 10 中，分類器嘗試將圖中右下角的數據盡可能的分出來，雖然在訓練上 表現出的精確性更高，但事實上□ = 0.01 更好地抓住了普遍的趨勢。

在 SVM 中需要計算的數據向量總是以內積的形式出現，因此可以透過核函數能簡化映射空間中的內積運算，核(Kernel)是一個函數 K，對於所有的 x, □ ∈ X，K(x, z) = (□(x),□(□)) = □(x) □□(□)。在 SVM 中最常用的核函數為逕向基函數核(Radial basis function，簡稱 RBF，又名高斯核)：

(8)

RBF 之所以被廣泛使用，是因為若 σ 選的好，有著能夠將原始空間應射程無窮為空間的能力，但若 σ 過大，高次特徵上的權重衰減的非常快，實際上相當於低維的子空間；反之若 σ 過小，則可以將任意的數據應設為線性可分，這並不一定是好的，隨之而來的可能是過擬和(Overfitting)的問題。因此通過調整 σ，RBF 擁有相當高的靈活性。（支持向量機通俗導論(理解 SVM 的三層境界)，2012）

下表 2-2-6 為 SVM 應用實例

作者與時間	研究題目	結論

陳俊瑋(2006)	一個擷取 Gabor 特徵的 SVM 人臉辨識方法	研究中提出了結合 Gabor 特徵擷取、SVM 分類辨識以及雙重門檻值之概念，建構出門禁管制系統平台。採用 SVM 一對多的分類方式，以模擬現實情況中的情形。
許哲銘(2002)	不完整資料學習演算法使用支援向量機器	研究中提出對於 SVM 核函數及高斯模型之建議，並對於遺失值使用非條件均值和條件均值的比較。

表 2-2-6 SVM 應用實例

第三節、Amazon Web Services(AWS)

面對巨量的資料，一般的個人電腦已經無法滿足處理的條件，需要使用強力運算能力的電腦，但是高效能的電腦當然需要的成本會比一般的電腦高上許多，不是所有的企業願意承擔的，如果可以不那麼仰賴更高等級的硬體設備，事情就會更加完美了，為了解決這樣的問題我們可以使用叢集計算(Cluster Computing)技術(黃羿凱,2007)，但對於不了解如何建立電腦叢集的人來說，還是相當不方便，幸好 2006 年 Amazon 開始以 Web 服務的形式為企業們提供 IT 基礎架構服務，也就是雲端運算平台，並且成立了 Amazon Web Services(AWS)這個品牌(李政霖,2019)。它為一般企業提供了極為穩定、可擴展、成本低廉的網架構平台，如其中 Amazon EMR 提供了一個托管集群平台，可簡化在 AWS 上運行大數據框架（如 Apache Hadoop 和 Apache Spark）以處理和分析海量數據的操作，Amazon EMR 支援 Apache Spark，讓我們可以從 AWS 管理主控台、AWS CLI 或 Amazon EMR API，輕鬆快速地建立受管的 Apache Spark 叢集。借助這些框架和相關的開源項目 (如 Apache Hive 和 Apache Pig)。可以處理用於分析目的的數據和商業智能工作負載(什麼是 Amazon EMR,2019)。讓一般企業不需要了解如何架設電腦叢集也可以快速上手。

Hadoop 和 Apache Spark 兩者都是大數據框架，但是各自存在的目的不盡相同。 Hadoop 實質上更多是一個分佈式數據基礎設施: 它將巨大的數據集分派到一個由普通計算機組成的集群中的多個節點進行存儲，意味著您不需要購買和維護昂貴的服務器硬件。同時，Hadoop 還會索引和跟踪這些數據，讓大數據處理和分析效率達到前所未有的高度。 Spark，則是那麼一個專門用來對那些分佈式存儲的大數據進行處理的工具，它並不會進行分佈式數據的存儲。(TiBaMe,2016)

下表 2-2-7 為 AWS 應用實例

作者	論文名稱	結論
紀柏安 (2015)	雲端錄影系統基於 AWS 之研究與實作	本論文提出以 AWS 所提供的雲端運算、儲存服務,建置一個雲端錄影儲存系統，此系統能夠讓網路攝影機透過網路將影片儲存至雲端空間,而這個儲存的空間是非常穩定不容易有資料流失、毀損。
李政霖 (2019)	即時空氣品質動態監測系統結合 LSTM 模型預測 PM2.5 濃度之應用	本論文以 LoRa 無線通訊技術開發一套低功耗的即時空氣品質動態監控系統並結合 LSTM 模型和 AWS 服務以預測未來 PM2.5 濃度。該系統可即時偵測當前位置的空氣品質動態。

表 2-2-7 AWS 應用實例

第參章、研究方法

本章節要討論的是電腦叢集架構及研究流程，本研究以決策樹分析挑選出影響遊戲勝負的重要因素，並利用這些重要變數帶入 Cart、C4.5、Knn、NCN、SVM、LDA 方法中建立模型以預測玩家之勝負。

第一節、分析方法

1.預處理	1.建立 AWS EMR 叢集 2.資料處理 (1)刪除以第三人稱視角遊玩的資料 (2)以 matchtype(個人模式/雙人模式/四人模式)分割資料集 (3)取相同 groupID 的資料並取其平均值
2.目標變數二元化、決定評估模型的標準	使用三種比例方法進行分割： (1)50:50 屬於前 50%勝利玩家(win)or 後 50%失敗玩家(lose) (2)10:90 吃雞玩家(win)or 其餘 90%一般玩家(lose) (3)35:65 官方比賽中能夠獲得晉級積分的前 35%勝利玩家(win)or 後 65%失敗玩家(lose)
3.決策樹分析及挑選變數	以 CART 和 C4.5 分別建立決策樹，找出重要變數，並比較其中差異，對比正確率已經實際需求選擇最好的目標變數切割方式。
4.第二次決策樹分析	使用重要變數重建 CART 與 C4.5 決策樹。
5.KNN 分析	將重要變數帶入 KNN 模型中以建立模型，並將模型作為對照組與重建的決策樹比較正確率。

6. NCC 分析	將重要變數帶入 NCN 模型中以建立模型，並將模型作為對照組與重建的決策樹比較正確率。
7. LDA 分析	將重要變數帶入 LDA 模型中以建立模型，並將模型作為對照組與重建的決策樹比較正確率。
8.SVM 分析	將重要變數帶入 SVM 模型中以建立模型，並將模型作為對照組與重建的決策樹比較正確率。
8.各分析方法之正確率比較	

表 3-1-1 分析流程表

註: "吃雞" 絕地求生遊戲中最終只會有一名玩家存活下來。最後獲勝者的畫面會出現“ Winner, Winner, Chicken Dinner.”，直白的中文翻譯就是「大吉大利，晚上吃雞。」隨著遊戲的火紅，「吃雞」便成為玩家間溝通的默契。

> **排名擊殺分：吃雞 8 分，第二名 4 分，第三、四名 2 分，後面沒有積分，(https://kknews.cc/zh-tw/game/9zlmlrl.html)

一、預處理：

(一)、建立 AWS EMR 叢集：

在 AWS 中建立 EMR 叢集(1 台 master，8 台 slave)，並在建立過程中添加引導文件(bootstrap action file)，用於安裝 Anaconda 和 Jupyter Spark，並針對 Spark 進行預先設定，讓我們在 EMR 叢集上可以使用 PySpark 進行資料處理和分析。

(二)、資料處理：

由於 kaggle 蒐集了四百萬多筆的資料，大致上可將資料依照遊玩視角分為兩類，第一人稱視角與第三人稱視角，但在目前的絕地求生(PUBG)國際賽事中是採用第一人稱視角進行，所以我們決定將第三人稱視角的資料刪除。隨後依照組隊模式(個人模式/雙人模式/四人模式)分割資料集，並在相同的比賽中合併相同隊伍的數據，使用隊伍平均數據代表該支隊伍的表現。

分割後資料集如下：

個人模式(資料集)	雙人模式(資料集)	四人模式(資料集)

二、目標變數二元化：

(一)、使用以下三種方式將目標變數二元化：

1. 50:50，屬於前 50%勝利玩家(win)or 後 50%失敗玩家(lose)：

由於絕地求生的同時遊玩人數最多可以達到 100 人，所以能常常在遊戲中成為名次前 50%的玩家也可以算是大眾公認脫離新手的參考依據之一，因此研究中使用 50:50 的比例分割當作其中一樣實驗方法。

2. 10:90，吃雞玩家(win)or 其餘 90%一般玩家(lose)：

每一場遊戲中，只有第一名玩家能獲得「吃雞」的榮譽，在比賽中生存到最後也是玩家們遊玩此遊戲的最終目標。由於並不是每一場遊戲皆參加滿 100 名玩家，因此資料集中 winplace 前 10%的遊戲玩家具備能在大部分的單場遊戲中獲得「吃雞」表現的遊戲能力，所以我們使用此方法進行實驗。

3. 35:65，根據官方比賽規則而定：

官方所舉辦的絕地求生比賽規則採用積分制，每場比賽由同賽區的 16 支隊伍同時進行比賽，單場比賽只有前 8 名的隊伍可以獲得積分，且第七名與第八名的隊伍獲得的積分相同，而只有積分前六名的隊伍可以晉級到季後賽，因此我們最後決定使用前 35%勝利玩家(win)or 後 65%失敗玩家(lose)比例分割來進行實驗。

分割後資料集如下：

個人模式(資料集)			雙人模式(資料集)			四人模式(資料集)		
50:50	10:90	35:65	50:50	10:90	35:65	50:50	10:90	35:65

表 3-1-2 資料分割表

(二)、評估模型方法

評估分類結果的指標很多，這些指標皆源自混淆矩陣（Confusion Matrix）

	預測		
真實		正確	錯誤
	正確	True Positive (TP)	False Negative(FN)

	錯誤	False Positive (FP)	True Negative(TN)

表 3-1-3 混淆矩陣

1. 正確率(Accuracy)：基本上就是模型的整體判斷的正確率，所以有時候也稱為 overall accuracy，越高越好。

公式：準確率=(TP+TN)/(TP+TN+FP+FN)

2. 精準率(Precision)：又叫查準率，是指在所有被預測為正的樣本中實際為正的樣本的概率。

公式：精準率=TP/(TP+FP)

3. 召回率(Recall)：又叫查全率，是指在實際為正的樣本中被預測為正樣本的概率。

公式：召回率=TP/(TP+FN)

4. F1-score：前面提到，召回率和精準率，也稱查全率和查準率，這兩個指標，若希望他們同時都很高，但是事與願違，他們是對立的、矛盾的，這就要我們去取捨，找到一個平衡點，這就是 F1 分數。

公式：F1 = 2 *精準率 * 召回率 / (精準率 + 召回率)

在本研究中將使用正確率來評估模型好壞的標準。

三、決策樹分析：

分別使用 CART 和 C4.5 決策樹以及兩組參數組合的表現，找出重要的變數，對比正確率及實際所需選出最好的目標變數切割方式，實驗出最好的變數組合。

依照文獻探討的章節中所得出的結果研究中在預剪枝的門檻值上的設定使用以下兩組參數組合。

	max_depth	mini_sample_leaf
第一組參數組合	13	100
第二組參數組合	5	1000

表 3-1-4 參數組合

四、第二次決策樹分析

使用第三點所提到的四個重要變數，重新建立 CART 與 C4.5 決策樹預測模型並測試。

五、KNN 分析

在使用決策樹分析挑選出重要變數之後，帶入 KNN 模型中測試，K 值的選擇參照目標變數的切割方法，分別使用對應 100、1000 當作 K 值參數，並使用正確卻率最高的。

六、NCN 分析

使用決策樹分析挑選出重要變數，帶入 NCN 模型中進行預測，使用的距離計算方式為歐式距離，觀察其正確率並與其他方法進行對照其差異。

七、LDA 分析

利用 python sklearn 套件內建函數來完成 LDA，並將處理後的資料丟入 LDA 的機器中執行預測模型並測試，觀察正確率是否與其他方法有不同並探討方法結果。

八、SVM 分析

利用 python sklearn 套件內建函數來完成 SVM，並將處理後的資料丟入 LDA 的機器中執行預測模型並測試，觀察正確率是否與其他方法有不同並探討方法結果。

八、各分析方法之正確率比較

依照三種組隊模式(個人/雙人/四人)。並以 KNN、LDA 與 NCN 模型作為對照組與重建的決策樹比較正確率。

圖 3-1-1 研究流程圖

第二節、研究限制及範圍

本研究是分析來自 kaggle 的 PUBG 玩家資料，可能會影響玩家技術表現的因素像是場地、設備狀況、環境等因素均不相同、多排（雙排、三排、四排）時隊友的實力衡量並沒有相應的資料顯示。在多排玩家的情況下使用隊伍的平均數據進行分析與研究。本研究無法加以掌握， 無法把以上因素考慮進去。

「絕地求生」是一個講究團隊精神的運動，每一局比賽玩家必須在手無寸鐵的情況下開始遊戲，從地圖中找尋各種資源，並必須與其他玩家互相爭鬥獲得生存，地圖上有著複雜的地形和建築、涵蓋範圍非常大，加上各種槍械、補給品等物資都是隨機在地圖上生成的，所以根據當下情況做出最佳決策，甚至在遇到敵人時都要權衡利弊後再決定是否射擊。

遊戲中可以單人遊玩或最多 4 位玩家組成一隊，每局遊戲最多會有 100 名玩家同時進行，每次進行遊戲都會產生大量資料。其中包含玩家的移動位置與距離、物件道具的消耗、擊殺與救援、遊戲進行時間與排名，我們將利用這些數據進行決策樹分析，找出影響遊戲獲勝的關鍵因素。

本論文主要探討「絕地求生」比賽名次之預測。藉由每場比賽結束後所有玩家的統計資料，使用決策樹中進行訓練，得到預測模型，找出影響比賽結果的關鍵因素，並以之推算最後比賽所有玩家之預測排名。

第肆章、實證分析

第一節、第一次決策樹分析結果

一、個人模式

在單人模式中，最重要變數為 walkDistance、killPlace、matchDuration，這三個變數在 Cart，C4.5 的兩個參數組合的十二個表格中都有出現，即機率 p=1，次重要變數為 rideDistance(p=7/12），NumGroups(p=6/12），kills（p=6/12），boosts（p=6/12），killStreaks（p=4/12），winpoints（p=2/12），maxPlace(p=2/12）。

選擇這些變數的原因是因為在所有的決策樹中，walkDistance(行走距離) 與 killPlace(擊殺排名) 在前三層中幾乎都有出現，是最重要的分界點。matchDuration 指的是玩家在當次比賽中存活的時間，此變數主要出現在以前 35%為勝者的分類決策樹中的前三層，但在往下挖的過程中，發現它在所有分類都有出現。

單人模式中，rideDistance 這個變數，在以前 35%為勝者的分類決策樹和以前 50%為勝者的分類決策樹都有出現，而在以前 10%為勝者的分類決策樹中沒有出現，無論是在 CART 還是 C4.5 都是如此。所以本研究認爲如果參賽者想在一場比賽中取得較好成績（前 50 或前 35），他可能就要注意你 rideDistance。而如果參賽者想取得更好的名次，可能就要把注意力放在其他變數。

本研究認爲單人模式中的 kills 這個變數，大部分（p=6/12，6 次中的 4 次出現在）以前 10%為勝者的分類決策樹中，説明這個變數對於取得好成績（吃鷄）有重要影響。同樣的，單人模式中的 boosts 這個變數，大部分（6 次中的 4 次出現在）以前 10%為勝者的分類決策樹中，説明這個變數對於取得好成績（吃鷄）有重要影響。

單人模式中，本研究發現 winpoints 是一個高階玩家才會去 care 的變數，因爲他只出現與以前 10%為勝者的分類決策樹中。本研究認爲 maxPlace 是一個中階玩家數據中特有的變數，因爲他只出現與以前 35%為勝者的分類決策樹中。

二、雙人模式

在雙人模式中，最重要變數為 walkDistance、killPlace、matchDuration，boosts 這四個變數在 Cart，C4.5 的兩個參數組合的十二個表格中都有出現，既 p=1。次重要變數為 rideDistance(p=8/12)，kills（p=6/12），killStreaks（p=6/12），NumGroups(p=4/12），damageDealt（p=2/12）。

雙人模式中，boost 這個變數在以前 10%為勝者的分類決策樹中的重要性甚至高過了本研究通常認爲最重要的 walkDistance 和 killPlace，其對於能否吃鷄的重要性可見一斑。

在雙人模式 killStreak 出現了 6 次，在以前 10%為勝者的分類決策樹中出現了 4 次，在以前 35%為勝者的分類決策樹出現了 2 次，本研究認爲參賽者要想取得好成績（吃鷄），killStreak 是一個不可忽視的變數。另外 rideDistance 這個變數，在以前 35%為勝者的分類決策樹和以前 50%為勝者的分類決策樹都有出現，而在以前 10%為勝者的分類決策樹中沒有出現，無論是在 CART 還是 C4.5 都是如此。所以本研究認爲如果參賽者想在一場比賽中取得較好成績（前 50 或前 35），他可能就要注意你 rideDistance。而如果參賽者想取得更好的名次，可能就要把注意力放在其他變數。而 NumGroups 出現的四次都是在以前 50%為勝者的分類決策樹中，本研究認爲如果參賽者想取得較好的成績（前 50%)，他會需要注意 NumGroups 這個變數。另外在雙人模式，本研究認爲 Damagedealt 是一個高階玩家數據中特有的變數，因爲它只出現與以前 10%為勝者的分類決策樹中。

三、四人模式

在四人模式中，最重要變數為 walkDistance、killPlace、matchDuration、boosts，這四個變數在 Cart，C4.5 的兩個參數組合的十二

個表格中都有出現，既 p=1，次重要變數為 kills（p=6/12），killStreaks（p=6/12），rideDistance(p=4/12)，assists（p=4/12)，longestKill(p=4/12)，damageDealt(p=2/12)，maxPlace(p=2/12），NumGroups(p=2/12）。

與單人模式、雙人模式相同，walkDistance(行走距離) 與 killPlace(擊殺排名) 在前三層中幾乎都有出現，是最重要的分界點。longestKill(最長擊殺距離) 和 assists (助攻數)個在表格中出現了 4 次，相較 NumGroups 這兩個變數對於四人模式的遊戲更加的重要。assists 對於以前百分之十為目標的隊伍來說是很好的指標，喜歡單打獨鬥的玩家在此模式中可能不容易有吃雞的機會。

四、第一次決策樹的小結

總體來說，walkDistance、killPlace 和 matchDuration 是在每一棵決策樹的前三層都有出現，是最重要的三個變數。再來對於 boost 這個變數，在雙人模式和多人模式中，boosts 在表格中都出現了 12 次，而在單人模式這個數字只有 6，且其中 4 次出現在以前 10%為勝者的分類決策樹中，說明在單人模式，參賽者想要取得好成績（吃鷄），需要對照 boosts（飲料）的使用量。值得一提的是四人模式中，出現了其他兩個模式所沒有的 longestKill(最長擊殺距離) 和 assists (助攻數)，可以猜測在四人模式中及早發現敵人與團隊合作是影響勝負的重要因素。且每個模式都有一些固定的變數 t 影響著中端玩家和高端玩家，像是單人模式的 winpoints 是影響高階玩家（前 10%）的變數，雙人模式的 damageDealt 是影響高階玩家（前 10%）的變數等等。

選取參數：在 pubg 個人模式、雙人模式、多人模式中，選出同樣的重要變數，經過 CART、C4.5 三種演算法跑出來的正確率，經過對比本研究發現經過對比，cart 兩個算法中準確率較高的參數組合是第一組參數組合（max_depth：13，mini_sample_leaf：100），c4.5 算法的結果也和 cart 相似。所以本研究決定刪除第二組參數組合。從另一個角度看，目標分類是三種 50：50，35：65，10：90，三種分類的準確率相差不多，而 35：65 這個分類指的是是官方比賽中能夠獲得晉級積分的前 35%勝利玩家 or 后 65%失敗玩家。因此本研究選取三個目標分類中的(35：65)這個目標分類，來避免繁雜的重複動作。

在研究的進行過程中，本研究使用了其他對照組來對比現有的 cart 和 c4.5 算法，於是選取 knn，LDA，NCN 這三個演算法來與現有算法進行比較，將四個重要變數（walkDistance、killPlace、matchDuration，boosts)和第一組參數組合（max_depth：13，mini_sample_leaf：100）以及目標分類按照 35：65 分割，得出準確率，研究結果如下表所示

第二節、分組比較：

在個人模式中使用的重要變數為 walkDistance、killPlace、matchDuration，雙人模式和四人模式中使用的重要變數皆是

walkDistance、killPlace、matchDuration、boost 這四個項目。

表 4-2-1 為列出各模型在不同模式下的正確率匯總：

模型名稱	cart	c4.5	knn	LDA	NCN	SVM
個人模式	0.940	0.939	0.922	0.914	0.905	0.658
雙人模式	0.930	0.930	0.923	0.921	0.906	0.643
四人模式	0.909	0.909	0.902	0.905	0.883	0.636

表 4-2-1 正確率總表

由以上表格得知，以上六種方法中，SVM 模型正確率表現不佳，故本研究以下將討論決策樹(Cart,C4.5),KNN,LDA,NCN。這五種方法的正確率幾乎都落在 90%至 94%之間，因此我們以運算時間與重要變數做討論：

(一)、運算時間

本研究所使用的分析演算法，皆是處在相同的電腦叢集之下，如下圖 4-2-1 所示。

圖 4-2-1 電腦叢集設定

可以發現到 NCN 分析演算法擁有最快的速度，故在相同的正確率之下，NCN 是較為適合的。

演算法	Cart	C4.5	Knn	LDA	NCN
運算時間(ms)	3231.61	3243.48	3733.71	3412.87	2437.33

表 4-2-4 演算時間總表

(二)、重要變數

在單人模式中，使用決策數分析中會發現變數 walkDistance(步行距離)、killPlace(本場比賽殺敵排行)、matchDuration(比賽時長)是具有較高影響力的變數。在此三項變數的影響下，大部分的分析方法正確率皆在 90%以上，在 walkDistance(步行距離)這項變數表現上取得較高的玩家，意味著，能夠蒐集到的資源也越多，但是同時也會增加遇見敵方的機會，killPlace(本場比賽殺敵排行)數據上取得較高的玩家往往會直接影響自己在本場比賽中成績，但是同樣的自身被擊殺的風險也會較高，matchDuration(比賽時長)遊戲技術較好的玩家的比賽時長，往往會較高，但並不是絕對，或許有兩名技術相近且具有一定程度的玩家，在比賽初期就因為爭奪資源而死亡，導致較差的排名，綜合以上三點來看在遊戲中在這三項數據表現突出的玩家往往可取得較好的成績。

在雙人、四人模式有第四項重要變數 boosts(使用能量飲料的次數)，此樣道具能在短時間內增強各項表現，但有一定的施法時間，將導致玩家出現破綻，所以在有隊友的雙人、四人模式中使用才能夠有較高的安全

性，換言之，在隊友的掩護下，相對於單人模式有更多的機會可以安全地使用能量飲料。

第伍章、結論

在台灣電競產業蓬勃發展的現在，越來越多人會去關心怎麼樣的玩家比較容易勝利;什麼樣的配裝是主流以及現代忙碌的生活讓更多的人投入遊戲中忘卻一切世俗的煙硝。也因此近年來直播電競聯賽吸引了大眾的眼球，而選手們精彩的表現使得大眾的模仿，使得一場又一場的遊戲遊玩次數增加，一筆又一筆的資料存入了遊戲數據庫。龐大的數據庫必須靠著數據分析專業的人士將其的價值最大化，也就是說從數據中找出關聯性，找出因為使用某裝備某特質所以能夠贏得比賽;而這樣的結論使得玩家們快速的進步實力，使得原先彼此冷漠的人們增添了交流聊天的話題。

本研究利用線上遊戲「絕地求生」玩家在單場次中的遊戲數據作為研究對象，透過 C4.5 與 CART 兩種決策樹演算法分析影響遊戲勝敗的主要因素，並將觀察值的目標變數用下列三種方式進行二元化：（1）50:50、（2）10:90、（3）35:65 三種比例皆為勝利玩家：失敗玩家，其中第三種是依據電競規則進行分割，並透過 1 人、2 人小組與 4 人小組的遊戲模式分別進行分析。

從分析結果中發現在三種遊戲模式中 C4.5 與 CART 兩種方法所算出的正確率非常接近，而從另一個角度看，三種目標變數分類的正確率差異不大。並且從結果中得知，在任何遊戲模式下，變數 walkDistance、killPlace、matchDuration 對目標變數的影響程度相當大。其中在雙人模式與四人模式的遊戲裡，除了以上三個變數，還多了一個變數 boost 也具有相當的影響程度。

此外本研究將上述的重要變數挑選出來，並將 knn、LDA、NCN、SVM 四種分類法作為對照組來對比現有的 CART 和 C4.5 演算法，而決策樹則是使用正確率較高的參數組合且目標變數的二分是使用 35:65，從對照結果中得出，除了 SVM 模型正確率表現不佳以外，其餘五種方法的正

確率幾乎都落在 90%~94%，以此驗證本研究的資料集擁有很好的解釋力，可以被多數演算法較好的解釋，其中並發現到 NCN 分析演算法擁有最快的速度，故在相同的正確率之下，NCN 是較為適合的，而變數 walkDistance、killPlace、matchDuration、boosts 能作為玩家評估實力或預測表現時的重要依據。

參考文獻

中文部分

1. 朱泓任(2017)。運產條例有助電競正名 40 校聯手業界打造培育環境。新頭殼 newtalk。2017 年 11 月 9 日。取自：http://newtalk.tw/news/view/2017-11-09/103302。

2. 4GAMERS(2015)。久等了，Twitch 完成台灣資料中心暨伺服器架設工程。2018 年 07 月 25 日。取自：https://www.4gamers.com.tw/news/detail/32314/twitch- bring-own-taiwan-data-center-and-sever。

3. 羅悅軒、曾凌軻（2017）。新時代的運動-電子競技概論。臺北市：中華民國電子競技運動協會。

4. 何世昌(2017)。＜封面故事＞明星瘋電競 周董、JJ 組隊參戰。自由時報。2017 年 11 月 20 日。取自：http://news.ltn.com.tw/news/weeklybiz/paper/1153178。

5. 王昱澄(2017)。台韓聯手 普儸電競宣佈進軍台灣電競產業。新頭newtalk。2017 年 12 月 15 日。取自：http://newtalk.tw/news/view/2017-12-15/107166。

6. 林宸誼(2017)。騰訊發布「五年計劃」 打造千億規模電競霸業。遊戲角落。2017 年 6 月 18 日。取自：https://game.udn.com/game/story/10446/2531065。

7. 涂國濠(2017)。電競遊戲觀賞動機、體驗、價值對行為意圖之研究-以 DOTA 2 電競為例。朝陽科技大學碩士

8. 宋傑恩(2019)。電競賽事對消費者購買相關週邊產品意願之研究。世新大學碩士。

9. David Olson and Yong Shi，郭志隆、張芳菱譯(2008)。資料探勘(Introduction to Business Data Mining)。台北：高立圖書有限公司，第 26 至 30 頁，第 39 至 42 頁。

10. 鄭捷(2018)。今天不學機器學習明天就被機器取代從：Pyhon 入手+演算法。臺北市：佳魁資訊。

11. 陳樹衡，郭子文，棗厥庸(2007)。以決策樹之迴歸樹建構住宅價格模型—台灣地區之實證分析。住宅學報，第十六卷第一期，民國 96 年 6 月。

12. 葉子維(2016)。顧客消費行為分析及行動銀行使用預測-決策樹、隨機森林與判別分析之比較。國立臺北大學統計碩士。

13. 陳元豪(2017)。應用決策樹分析探討植群空間分布特徵-以曾文水庫集水區為例。中國文化大學地學研究所未出版之碩士論文。

14. 蔡佳玲(2011)。以決策樹模型探討未開立慢性病連續處方之影響因子。資訊管理學報，18 卷 4 期，第 139 至 164 頁。

15. 周佩蓁(2005)。以決策樹分析顧客滿意度之研究。育達商業技術學院 資訊管理所碩士論文。

16. 呂文吉(2016)。應用決策樹資料探勘模式於醫療院所資訊設備故障排除之研究—以北部某區域醫院為例。元培醫事科技大學資訊管理系數位創新管理碩士班。

17. 吳俊穎(2014)。利用相似比例加權改善 KNN 演算法於偏斜資料之分類效能。中華大學。

18. K 近鄰 k-Nearest Neighbor(KNN)演算法的理解-程式前沿(2018)。取自：

https://codertw.com/%E7%A8%8B%E5%BC%8F%E8%AA%9E%E8%A8%80/635163/

19. 楊帆(2016)。基於改進 KNN 演算法的室內 WIFI 定位技術研究。西北工業大學碩士。

20. 張碩(2017)。基於 KNN 演算法的空間手勢識別研究與應用。吉林大學碩士。

21. 謝紅，趙洪野，解武(2015)。基於局部權重 k-近質心近鄰算法。哈爾濱工程大學信息與通信工程學院。

22. 機器學習: 降維(Dimension Reduction)- 線性區別分析(Linear Discriminant Analysis)。2018 年 5 月 15 日。取自 https://medium.com/@chih.sheng.huang821/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E9%99%8D%E7%B6%AD-dimension-reduction-%E7%B7%9A%E6%80%A7%E5%8D%80%E5%88%A5%E5%88%86%E6%9E%90-linear-discriminant-analysis-d4c40c4cf937

23. 機器學習:線性判別分析（LDA）。主成分分析(PCA)-PYTHON 教程(2018-10-04)。取自：https://www.itread01.com/p/484997.html

24. 王俊興(2017)。以粗集合理論和線性判別分析在半導體產業之篩選股票投報率的可行性研究。嶺東科技大學資訊管理系碩士。

25. 黃建齊(2014)。應用線性判別分析法改善衛星資料估算颱風生成潛勢指標之研究。國立中央大學大氣物理研究所碩士論文。

26. 支持向量機通俗導論(理解 SVM 的三層境界)。2012 年 6 月 1 日，取自： https://blog.csdn.net/v_july_v/article/details/7624837。

27. 陳俊瑋(2006)。一個擷取 Gabor 特徵的 SVM 人臉辨識方法人臉辨識方法。國立台灣科技大學。電機工程系碩士學位論文。

28. 許哲銘(2002)。不完整資料學習演算法使用支援向量機器。交通大學資訊科學研究所碩士論文。

29. 黃羿凱(2007)。使用個人電腦叢集計算環境之高解析度印刷標籤檢查系統。國立海洋大學資訊工程研究所碩士學位論文。

30. 什麼是 Amazon EMR。2019 年 9 月 18 日。取自：https://docs.aws.amazon.com/zh_cn/emr/latest/ManagementGuide/emr-what-is-emr.html

31. TibaMe。10 分鐘弄懂大數據框架 Hadoop 和 Spark 的差異。2016 年 1 月 5 日。取自：：https://blog.tibame.com/?p=1752

32. 李政霖(2019)。及時空氣品質動態監測系統結合 LSTM 模型預測 PM2.5 濃度之應用。台北科技大學碩士學位論文。

33. 紀柏安(2015)。雲端錄影系統基於 AWS 之研究與實作。

國立臺北科技大學電機工程研究所。

英文部分

1. Berry, M. & Linoff, G. (1997). Data mining techniques: For marketing, sales and marketing support.

2. Chen, C.L.P, Zhang, C.Y. (2014). “ Data-intensive applications, challenges, techniques and technologies: A survey on Big Data,” Information Sciences, Volume 275, pp. 314-347.

3. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine, 17*(3), 37.

4. Khan, N., Yaqoob, I., Hashem, I.A., Inayat, Z., Ali, W.K., Alam, M., Shiraz, M., Gani, A. (2014). Big data: survey, technologies, opportunities, and challenges. The Scientific World Journal.

5. Pal, M. & Mather, P.M. (2003), An assessment of the effectiveness of decision tree methods for land cover classification, Remote Sensing of Environment, 86(4) , 554-565.

6. Tamouk, J. and Allahakbari, F. (2012) “A comparison among accuracy of KNN, PNN, KNCN, DANN and NFL.IJCSI International Journal of Computer Science Issues, vol. 9, no. 3 No. 1, pp. 319-322, 2012.