Homework 1

電機系 傅立威 C24061088

1. 從Kaggle中找一個資料集，描述此資料相關資訊，包含：

• 問題定義

• 潛在問題

• 分析與預測難度

• 價值

資料集：FIFA 19 complete player dataset

網址：<https://www.kaggle.com/karangadiya/fifa19>

• 問題定義：

FIFA 19是一款由EA加拿大、EA羅馬尼亞開發由美商藝電發行的足球電子遊戲，和現在也很多人在玩的2K19類似。因為這個遊戲是可以連上網的，所以可以利用爬蟲程式取得相當多有關遊戲的資料。

這一個資料集有遊戲中每個球員的ID、Name、Age、Height、Weight……等基本資料（基本上是固定的），也有Value、Wage、Skill Moves……等和玩家遊玩過程中產生的資料。我們可以透過這一些資料來進行資料探勘，看看能不能找出一些有用的資訊。

• 潛在問題：

因為這個網站一次給了所有種類的資料，用這一些資料去做分析的話可能會遇到上課教授講到的過度擬合的問題（overfitting），可能最後會產生一個荒謬的模型，像是今天上課教授提到的，我們架一台攝影機去拍所有去買彩券的人當時的狀況，再去和買彩券中獎人的資料進行分析，可能會得出一些奇怪的結果，例如：分析發現，中獎的人當天幾乎都是穿紅色的衣服，用左手拿出錢包買彩券。所以我們推論，只要穿紅色的衣服、用左手拿出錢包買彩券就有很高的機率中獎這種奇怪的結論。

• 分析與預測難度：

我還沒學到分析與預測的方法，不過因為這個網站上了非常充足且詳細的資料，所以我推測要用這些去做資料探勘、資料挖掘應該算是容易的。不過資料要小心的處理，才不會有潛在問題裡講的過度擬合的問題（overfitting）。

• 價值：

我們可能可以利用這一些世界各地玩家產生的資料，去製造一個超級強的AI，強到幾乎沒有人能打得贏電腦，這一些AI可能也可以當成是電競選手訓練的對手；或者，我們可能可以從這一些資料去找哪一類型的角色比較受到玩家喜愛，因為裡面的所有角色都是現實生活中真有其人，我們可以去比對是不是這個人在現實中也這麼被看好，有什麼原因。

1. 描述一個在你學習領域的資料或是日常生活中可取得的資料, 說明可以用來產生何種價值（預測，分析，別人不知道的事…）

我可以和歷屆的學長要大量的、各種科目的、各個教授的考古題，同時也可以上網去查每一次考試全班的學生成績，去計算全班的平均值、標準差、最高分和最低分……等，然後有那一些人被當掉，被當掉的人平常的表現如何。

然後利用這一些的資料去分析看看，每個教授的出題模式，那一些教授特別喜歡出哪一類型的問題，或者是每個教授出考題的難易度得變化，如果期中考全班考得怎樣，期末考的難度可能會是怎樣等等。

最後可能就可以告訴接下來的學弟妹們，依據他們的學習方式，這學期要選哪一些教授的課會比較容易過，比較輕鬆，考試要怎麼準備，預測你這學期的分數、排名大概會落在哪一個區間之類的，讓他們在選課的時候不要踩到地雷。