109 學年度第一學期

物聯網與巨量資料分析書面報告

主題：Assignment 3 作業

指導老師：王文彥 副教授

組長：4070E021 蘇宇祥

組員：4070E003 唐靖凱、4070E019 徐江弘、4070E022 沈明楷

中　　華　　民　　國　　１０９年１０月１２日

目錄

壹、工作分配

<table>
<tr><th rowspan="2">組員</th><th colspan="2">工作分配</th><th rowspan="2">貢獻度</th></tr>
<tr><th>第(1)題</th><th>第(2)題</th></tr>
<tr><td>蘇宇祥</td><td>B：詳細描述此整合的目的
C：請給對應的 open data 之 url(註:此 url 須能夠被找到，且與你們的動機目的相關，且合乎邏輯！)</td><td>C：而這些變數與比較中的數值是獨立關係或相依關係？為甚麼？</td><td>25%</td></tr>
<tr><td>唐靖凱</td><td>E：你會使用決策樹方法(Decision Tree)的哪一個 Algorithm？為甚麼？</td><td>A：你若是令狐沖老師會認為那一班成績考的較好？為什麼？(需詳細描述)</td><td>25%</td></tr>
<tr><td>徐江弘</td><td>A：請詳細描述此整合中有那些屬性與其整合的動機
B：詳細描述此整合的目的</td><td rowspan="2">B：另對比較中，對統計而言，你覺得要如何控制那些影響變數，比較起來會更合理？為甚麼？</td><td>25%</td></tr>
<tr><td>沈明楷</td><td>D：若使用人工智慧的決策樹方法(Decision Tree)，你的目標值屬性會是那一個？為甚麼？</td><td>25%</td></tr>
</table>

貳、題目與解答

1)(人工智慧/資料探勘學前工作) 請尋找 open data 中可用的屬性(註：可複數個) ，並結合此三屬性：溫度、溼度、感測時間的數值，來做合理的整合。

A：請詳細描述此整合中有那些屬性與其整合的動機。

解：

時間(須了解一年 12 個月裡每天的數據)，溫度(須了解溫度對於種植品質的影響)，濕度(棉花適合溫度為比較乾燥的環境)，棉花是世界上最重要的紡織纖維，全球約 65% 的棉花產區在北緯 30～37 度間，這個範圍包括了美國、舊蘇聯、中國的大部份生產地區，目前有四種棉花在商業上用途最多，其中最大的棉花出口國為美國，主要以高地棉為主，一年大概產出五百多萬公噸，棉花適合溫和跟氣候適宜的環境，適合於美國的南部地區與多個地區。而我們便是要統計出美國南部城市奧斯丁[1]的棉花產能，在一年中哪些月份能使種植品質達到優良，和哪些月份種出來成效比較不佳並了解原因，以及在不同的月份裡的溫溼度環境對於當下種植的影響，為此能讓棉花產能達到最大且最優良的產值。

B：詳細描述此整合的目的。

解：

因應美國為世界最大的棉花出產國，如果只靠傳統的種植產業規模來因應現在的市場的話因為產量不足可能會少賺很多錢，隨著資訊時代數位科技的發展，這種問題可能會慢慢地解決，資訊可使我們大規模生產、掌控種植管理技術，了解市場即時動態，而我們所做的目的便是偵測一整年的溫濕度，透過部屬在農業生產現場的感測器，查看在一年不同的月份裡，不同的溫溼度對於種植棉花的品質，而區分出品質不好與品質優良的產品，依照這些數據推算出棉花最適合大量種植的時段，並通過農業物聯網，可以在需要的時段配合農場上的自動灌溉、施放肥料，並精確的控制水量及肥料的取捨，一方面能省錢環保又能大量種植出最優質的產品。

C：請給對應的 open data 之 url(註：此 url 須能夠被找到，且與你們的動機目的相關，且合乎邏輯！)

解：

https://reurl.cc/2g692m

D：若使用人工智慧的決策樹方法(Decision Tree)，你的目標值屬性會是哪一個？為甚麼？

解：

挑出具最大資訊獲利，選擇最大的資訊獲利值，利用資訊獲利來衡量屬性於分類資料的能力。以本組種植棉花為主，在美國五月到七月份是最適合種植的期間，該期間溫濕度恰好

是棉花最佳身長的環境，溫溼度即為所謂的分類資料，於是最大資訊獲利就是依照棉花適合的時機種植，透過分類資料的走向與判斷，選擇最佳種植環境，以致可以獲取最大產量。

E：你會使用決策樹方法(Decision Tree)的哪一個 Algorithm？為甚麼？

解：

使用 ID3(Iterative Dichotomiser 3)演算法(如圖 1)。ID3 演算法[2][3]是一種使用資訊獲利概念的貪心演算法，選擇資訊獲利最大的一個屬性作為決策樹的根節點。而種植棉花決策樹中資訊獲利最大的兩者分別為溫度與濕度，這兩個因素是關鍵，所以在決策樹中利用 ID3 演算法，挑選對於棉花最合適的溫度與濕度，也就是說此演算法選擇對自己最佳利益的因子，產生最大的效益。

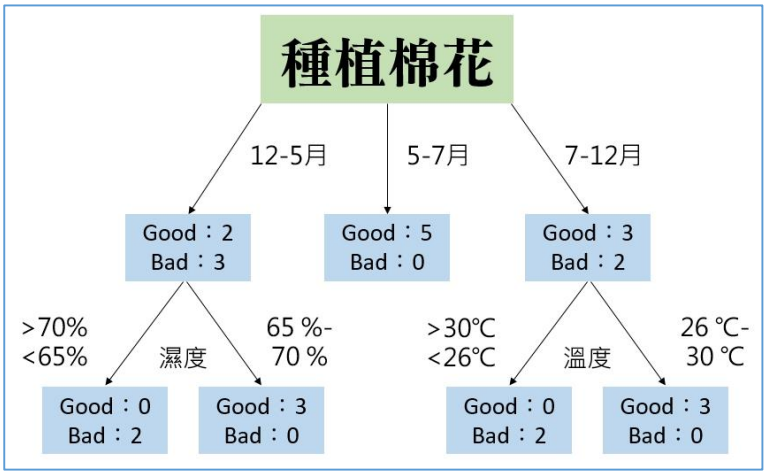

圖 1 種植棉花決策樹[4]

2)(統計應用學前工作) 令狐沖老師在快樂天堂學院教授兩班，分別為忠班與 孝班. 此次期中考成績，忠班學生：平均成績 81 分，標準差 10 分.孝班學生： 平均成績 81 分，標準差 5 分.
A：你若是令狐沖老師會認為那一班成績考的較好？為什麼？(需詳細描述)
解：

孝班成績較好。因為孝班之標準差較小，而標準差是利用每個人的分數減去平均後平方的值再平均。所以代表若標準差的值越大，表示大部分的數值和其平均值之間差異較大(即資料越分散)；標準差的值越小，代表這些數值較接近平均值(資料越集中)。

B：另對比較中，對統計而言，你覺得要如何控制那些影響變數，比較起來會更合理？為甚麼？
解：

要控制標準差[5][6]較合理。因為標準差是利用每個人的分數減去平均後平方的值再平均，因此標準差跟個人成績有關聯，這裡的變數就是每個人的成績，如果每個人的考試成績可以提高，那班級整體標準差數值會越小。

C：而這些變數與比較中的數值是獨立關係或相依關係？為甚麼？
解：

是相依關係。標準差是根據每個人的分數減去平均後平方的值再平均，每個人的分數就是此題的變數，標準差是此題比較中的數值，所以當變數(每個人分數)再次有變動時，比較中的數值(標準差)亦會隨著改變。

叁、參考資料

1. kaggle. (2017). Austin Weather. https://reurl.cc/2g692m(2017)
2. 鵝廠優文|決策樹及 ID3 演算法學習，程式前沿(2018 年 05 月 17 日)，https://reurl.cc/Q35rR9
3. 決策樹學習，陳士杰，國立聯合大學資訊管理學系，https://reurl.cc/14X6vV
4. 決策樹，維基百科(2020 年 9 月 30 日)，https://reurl.cc/3L6gX8
5. 一文看懂決策樹，區塊鏈遊戲研究院，鏈聞 ChainNews(2019 年 9 月 17 日)，https://reurl.cc/zzp3D7
6. 真希望老師這樣教統計，Pizza，Medium(Jun,17,2019)，https://reurl.cc/Gr58pA