**國立成功大學理學院模組化課程**

**問題導向之資料科學與機器學習應用HW1**

**N07041334 許嘉倪**

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: zomato.csv**

<https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants/downloads/zomato.csv/1>

**問題定義:**

了解影響Bengalur地區，各種類型餐廳建立的因素，每個餐廳的總體評級。因Bengalur城市，有超過12,000家餐廳及提供來自世界各地的菜餚。隨著每天新開業的餐館已經飽和，需求日益增加。儘管需求不斷增加，但新餐館難以與老牌餐廳競爭。他們中的大多數都提供同樣的食物。Bengalur是印度的IT首都。這裡的大多數人主要依賴餐館的食物，因為他們沒有時間自己做飯。由於對餐館的壓倒性需求因此研究一個地點的人口統計變得很重要、什麼樣的食物在當地更受歡迎。

**分析與預測難度:**

此dataset的資料到2019年3月15日。數據將從兩個階段中刪除，每個街區都有6-7類餐廳，包括自助餐，咖啡廳，外食店，甜點店，只提供外送的店面，夜店，和酒吧。

在分析時，則分成2個階段:

第一階段:

在提取的第一階段中，僅提取在首頁上可見的餐館的URL，名稱和地址。 zomato上每個餐館的URl都記錄在csv文件中，以便以後可以為每個餐館單獨提取數據。 這使得提取過程更容易，並減少了機器的額外負載。 每個社區和每個類別的數據都可以在這裡找到。

第二階段:

在第二階段中，讀取每個餐館和每個類別的記錄數據，並且一一刪除部份餐館的數據。 在這個階段共刪除了15個變量。 對於每個社區和每個類別，他們的online\_order，book\_table，rate，votes，phone，location，rest\_type，dish\_liked，cuisines，approx\_cost（兩個人），reviews\_list，menu\_item被提取。

預測難度應為無法用自己的觀念區別哪些餐廳的變量數據是不重要的，有可能因此刪除了重要的變量，使得數據分析結果大為不同。

**價值:**

倘若在Bengalur城市整個地方都喜歡吃素食。假如是，那麼該地區會由特定人群居住，例如。耆那教、古吉拉特人，他們大多是素食主義者。這些分析可以使用數據進行，通過研究因素，例如: 餐廳的位置、食品的大致價格、主題餐廳與否、該城市的哪個地方提供最多餐廳的美食等等。以上的分析可以提供在地餐廳經營人了解當地人民的喜好，以進一步朝此方向努力，可也可以提供愛好美食的人獲得哪個區域能提供自己喜好的美食等。

**(b)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)。**

最近我在學校的實驗室與台船合作，取得台船的資料，裡頭包含名有質量式流量計資料、航行數據記錄儀資料、以及馬力計資料。

**質量式流量計資料:** 精確測量流量儀器，可用來監測船舶油耗等。裡面資料共有50個欄位，包括溫度、質量流率、密度、鍋爐密度等等。

**航行數據記錄儀資料:** 用來記錄船舶航行資料、狀態、指令等。裡面資料共有11個欄位，包括緯度、經度、航向、船速、水深、風速、風向等。

**馬力計資料:** 量測船舶馬力。裡面資料共有3個欄位，包括扭轉力矩、引擎轉速、軸馬力。

以上3種資料可以用來分析油耗的相關因子、什麼因素是船舶航行的最佳化狀態，才能有最少的油耗、預測什麼情況下的船舶的油耗量、船舶的最佳航行路徑、船舶故障分析診斷及預測等等。由於剛加入實驗室團隊內，希望藉由此次模組化課程中，學習如何作數據分析。