(二) 缺失值處理 :

為什麼要處理缺失值? 1 .缺失值會降低模型的功效&擬合度或導致模型出現偏差

2 .因為我們沒有正確分析行為和與其他變量的關係，它可能導致錯誤的預測或分類

為什麼我的數據會有缺失值?

數據提取：數據提取過程可能有問題 :1. 與數據監護人仔細檢查數據是否正確

2. 一些散列程序也可以用來確保數據提取是正確的

數據收集：錯誤發生在數據收集時

1. 完全隨機缺失：對所有的觀察值，缺失變數的機率都相同
2. 隨機缺失：隨機缺失變數

缺失率因輸入變量的不同，造成值或水平變化

例：收集年齡數據，與男性相比，女性的缺失值更高

3. 取決於未觀察到的預測變量的缺失：

缺失值不是隨機的，與未觀察到的輸入變數相關

例：如果特定診斷引起不適，則該變數缺失值高

4. 缺失取決於缺失值本身：是缺失值的機率與缺失值本身直接相關

例：收入較高或較低的人可能對他們的收入不作回應

處理缺失值的方法有哪些?

1. 刪除：列表刪除 : 刪除缺少任何變量的觀察結果

會降低模型功效，因為它減少了樣本量，整行刪

成對刪除 : 對感興趣變量的所有情況進行分析

保留盡可能多的案例進行分析，只刪缺失部分

缺點 : 對不同的變量使用不同的樣本量

使用時機 : 缺失數據性質為「完全隨機缺失」，否則會使模型輸出產生偏差

2. 平均/眾數/中位數插值：用數據集中已識別的關係來估計缺失值

定量屬性 : 所有已知值的平均值或中位數

定性屬性 : 眾數，替換給定屬性的缺失數據-

廣義插補：計算該變量的所有非缺失值的平均值或中位數，替換缺失值

相似情況插補：依據類別，各別計算，根據類別替換缺失值

3. 預測模型 : 訓練集 : 不含缺失值的數據集；測試集 : 含缺失值的數據集

目標變量 : 缺失值的變量

根據訓練集的其他屬性預測目標變量並填充測試集的缺失值

方法 : 回歸、ANOVA、邏輯斯回歸和各種建模技術

缺點：1. 模型估計值通常比真實值表現得更好

2. 數據集屬性和缺失值屬性沒關係，將不能精確地估計

4. KNN 插值 : 使用與缺失值屬性最相似的給定數量的屬性來插補缺失值

使用距離函數確定兩個屬性的相似性

好處：1. KNN 可以預測定性和定量屬性

2.不需要為缺少數據的每個屬性創建預測模型

3.可以輕鬆處理具有多個缺失值的屬性

4.考慮了數據的相關結構

壞處：1. KNN 在分析大型數據庫時非常耗時，它搜索所有數據集

2. k 值選擇非常關鍵 : 較高的 k 值將包含我們需要的顯著不同的屬性

較低的 k 值意味著缺少重要的屬性。