流程

資料get > 清資料(缺值和異常值) > 了解資料分布特性 > 資料預處理(選特徵 、確認尺度) > 跑模型 > 調參數 > 測試

缺值處理:

移除(不推薦) or 填補

填補方法:

平均數:資料成連續性數值分布，且缺失值為隨機

mean()

中位數:資料包含極端(異常)值用

median()

眾數:類別特徵用

mode()

特徵選擇

1.特徵重要性:

feature\_importances----模型裡面有的方法

2.相關性分析:

----卡方檢定 for 分類

-----皮爾森 for 數值

--------------------------------------------------------------------------------

熵 (Entropy) 基尼不純度 (Gini impurity)

ID3演算法（Iterative Dichotomiser 3）：

無法處理缺失值。

可以處理多類別分類問題。

深而窄的樹，容易過度擬合。

CART演算法（Classification and Regression Trees）：

可分類、回歸。

使用基尼不純度（Gini impurity）或均方誤差（MSE）

能夠處理缺失值。

可以處理多類別分類問題。

深而寬的樹，容易過度擬合。

C4.5演算法：

Gain Ratio

能夠處理缺失值。

可以處理多類別分類問題。

深而窄的樹，但不容易過度擬合。

C5.0演算法：

分支策略：Gain Ratio or Entropy

能夠處理缺失值

可以處理多類別分類問題

深而窄的樹，也不容易過度擬合。

所以在使用ID3、C4.5、C5.0演算法時，分支策略 要用熵 (Entropy)

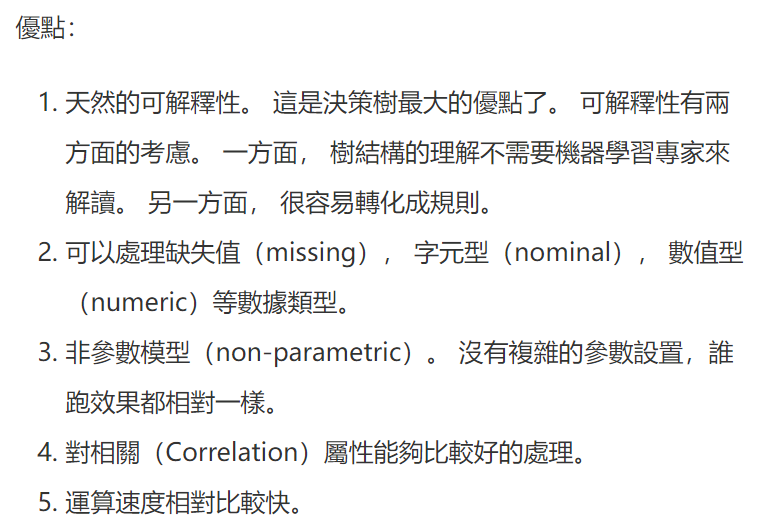
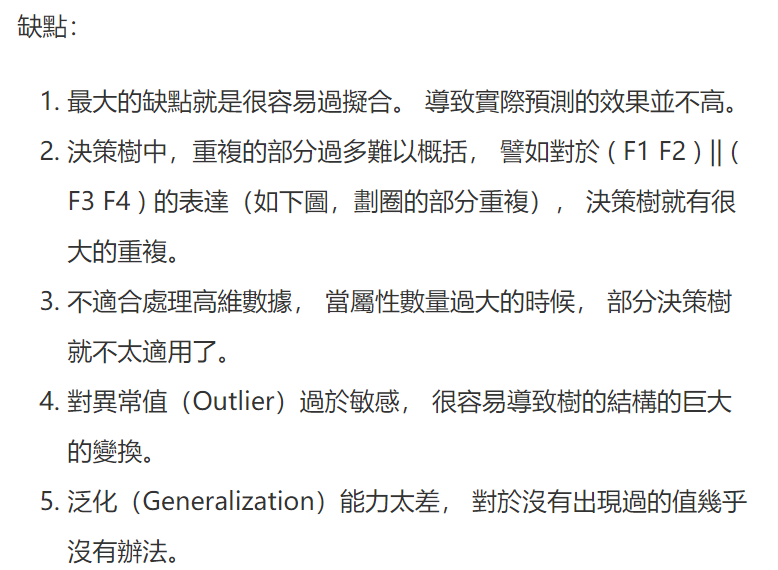
CART 用 基尼不純度（Gini impurity）

分割方式:隨機森林用隨機分割

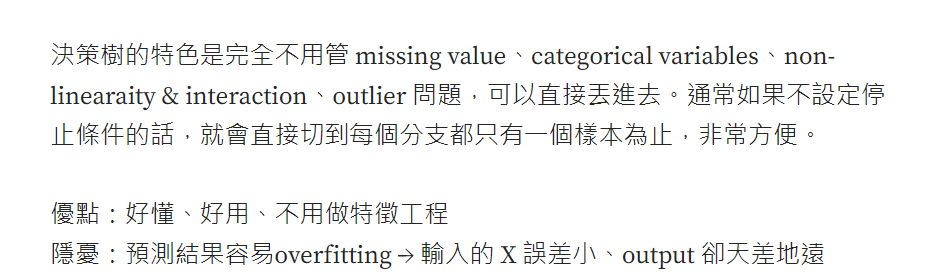
最佳分割(貪婪分割) for 一般樹演算法 但要注意過擬合

參考文章:

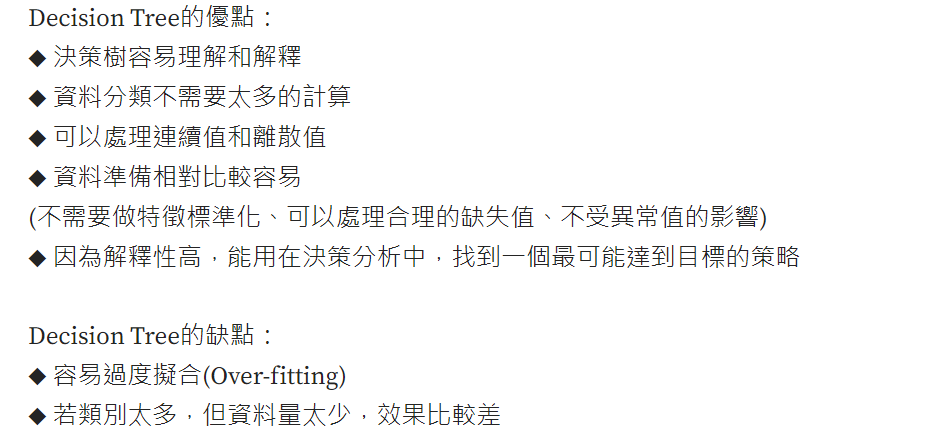
<https://www.getit01.com/p20180102558230411/>

<https://ntudac.medium.com/python-%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E8%B3%87%E6%96%99%E5%89%8D%E8%99%95%E7%90%86%E8%88%87%E6%A8%B9%E6%A8%A1%E5%9E%8B-410cf4f777b0>



<https://medium.com/@imirene/python%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E5%A4%9A%E5%85%83%E5%88%86%E9%A1%9E%E7%9A%845%E7%A8%AE%E6%A8%A1%E5%9E%8B-f7b6026c5ce6>



<https://medium.com/%E4%BA%BA%E5%B7%A5%E6%99%BA%E6%85%A7-%E5%80%92%E5%BA%95%E6%9C%89%E5%A4%9A%E6%99%BA%E6%85%A7/decision-tree-%E6%B1%BA%E7%AD%96%E6%A8%B9-%E5%96%AE%E7%B4%94-%E5%BF%AB%E9%80%9F-%E8%A7%A3%E9%87%8B%E6%80%A7%E9%AB%98%E7%9A%84%E6%B1%BA%E7%AD%96%E8%A1%93-ef28e0e75a55>

