**資料探勘期中報告--第五組**

邱冉彤 陳映安 葉瀞瓀 陳姿吟

學號: B103040002 B103040022 B103040049 B103040014

1. 摘要

近年來國人得到糖尿病的比例持續增加，位居十大死因中的第五名。糖尿病屬於慢性疾病，與家族遺傳、環境因素息息相關。本次實驗，我們將設計程式以找出以下因素與糖尿病的相關程度：１.懷孕次數、２.血液中葡萄糖濃度、３.舒張壓、４.三頭肌皮摺厚度、５.胰島素濃度、６.BMI值　７.糖尿病函數。使用ｋｎｎ、隨機森林、決策樹等機器學習方法去比對最終果和找出各個因素與糖尿病的相關程度，也可以此比較各個演算法的準確度。

1. 簡介

機器學習演算法是一種讓電腦自動學習的方法。透過大量數據找出規律，並利用此規律對測試資料進行預測。其中，機器學習演算法又可分為以下四類：監督式學習、非監督式學習、半監督式學習、強化學習。

為找出與糖尿病最相關的因素，本實驗建置三個機器學習演算法，分別是KNN、決策樹，與隨機森林，將懷孕次數、血液中葡萄糖濃度、舒張壓、三頭肌皮摺厚度、胰島素濃度、BMI值、糖尿病函數輸入至本系統，以利機器學習，找出各個因素與糖尿病的相關程度，也可以此比較各個演算法的準確度。

1. 相關研究

**KNN：**

KNN (K Nearest Neighbor)是一種資料挖掘分類的技術，而k的意思是「取K個離樣本最近的鄰居，即可代表此樣本」，也就是說，我們能以已分類的資料為基礎，決定尚未分類的資料。而決定的方式為「多數決」。

優點：簡單、容易理解及實現，適合處理多分類問題。

缺點：計算成本高、預測時間較長，當樣本不平衡時，可能導致預測不精準的情形。

**決策樹、隨機森林：**

決策樹會依據訓練資料，利用決策樹演算法，如：Information gain、Gini index……等，訓練出規則對測試資料進行預測。

隨機森林是由多棵決策樹所組成，而每棵樹都是從訓練集隨機抽出的n筆資料中隨機挑選m個特徵建立，最後進行多數投票進行預測。

優點：準確率高、能有效運行在大數據集上、能評估各個特徵的重要性且不須降維。

缺點：在噪音較大的分類上可能過度擬合。

1. 程式設計方式

**KNN（C）：**

**程式流程圖：** **資料前處理：**

一張含有 圖表 的圖片

自動產生的描述將訓練集的資料讀入陣列，並將為0的不合理資料

改為其所在類的算術平均數，測試集作法亦相同。

一張含有 文字 的圖片

自動產生的描述

**將測試集丟入 KNN 程式進行預測： 計算正確率：**

計算每個資料與其他值的距離（平方後開根號）， 用迴圈測試 K值不同時（1~300）所得出的

再計算有得糖尿病（ SICK = 1）的次數比較多， 結果，並分別與實際情況做比對（實際情況

還是沒得糖尿病（ SICK = 0）的次數比較多， 存在 correct\_outcome 陣列中），最後

return 比較多的 type 。 計算其正確率。

一張含有 文字 的圖片

自動產生的描述 一張含有 文字 的圖片

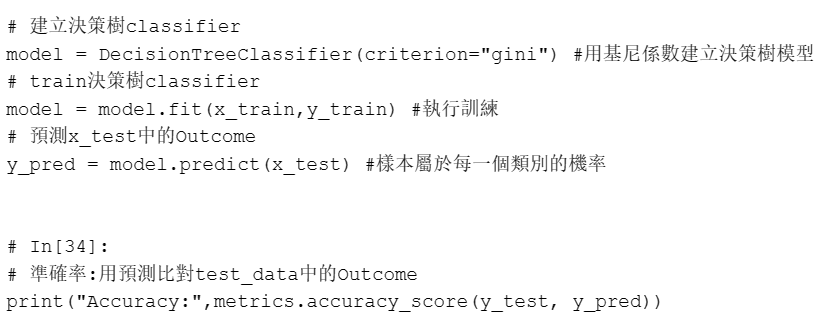
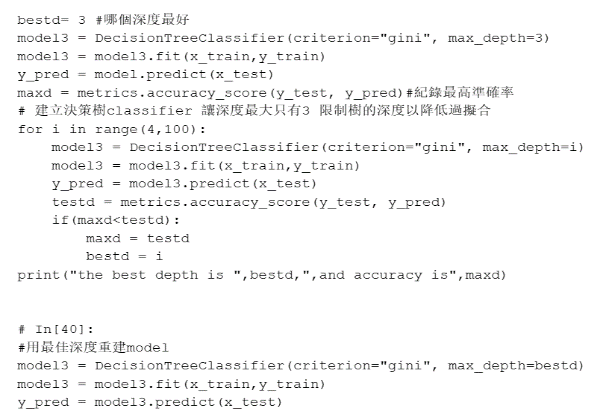
自動產生的描述

**決策樹、隨機森林:（python）（使用sklearn套件）**

**程式流程圖：**



**Decision Tree的建立、訓練、預測： 對Decision Tree限制最佳深度：**

Decision Tree採用gini係數來建立模型 從深度=3開始到深度=100，建立

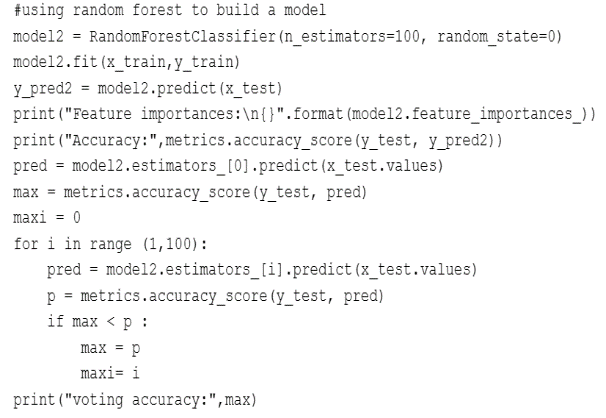
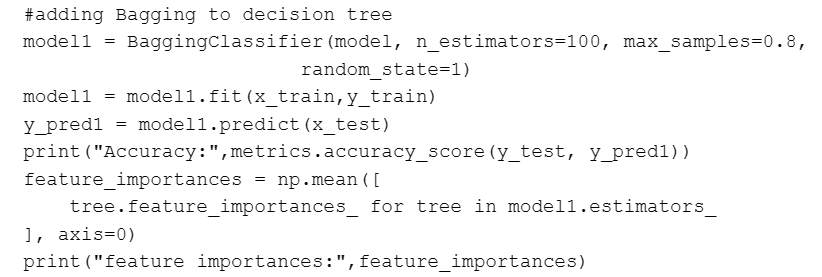
經過訓練後做預測，accuracy可檢測模型 Decision Tree(同樣採用gini係數)

的準確度。 模型。用bestd去紀錄找到可以得到

最佳準確度的深度。maxd則紀錄最佳

的準確度。

**Random Forest的建立、訓練、預測：**  **對Decision Tree加上Bagging增加正確率：**

 Random Forest與Decision相似， Decision tree模型最後的準確率為

同樣需要建立模型、訓練最後預測。 0.72。為了提高準確率，我們使用

Random forest中有許多tree，在畫 BaggingClassifier()。他的原理是

Forest的圖時，只能畫出一顆，於是 從數據中重複抽樣，然後建立很

我們先從random forest中的100棵樹 多個模型再一一訓練，最後投票看哪

去找出其中準確度最高的那一棵樹， 個預測結果好。經過bagging後的模

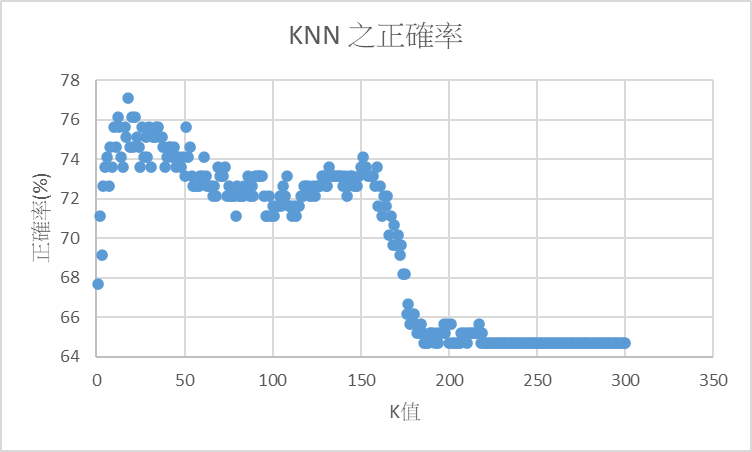
並用maxi記錄下來。後續再使用 型可以避免過度擬合，成功的提升準

Export\_graphiz。 度到0.81。

1. 分析、結論

還未將為0的不合理資料改為其所在類的算術平均數前，當K值為17時，有最高的正確率(76.6)。

將為0的不合理資料改為其所在類的算術平均數，當K值為18時，有最高正確率(77. 11)，整體正確率有所提升。

一張含有 圖表 的圖片

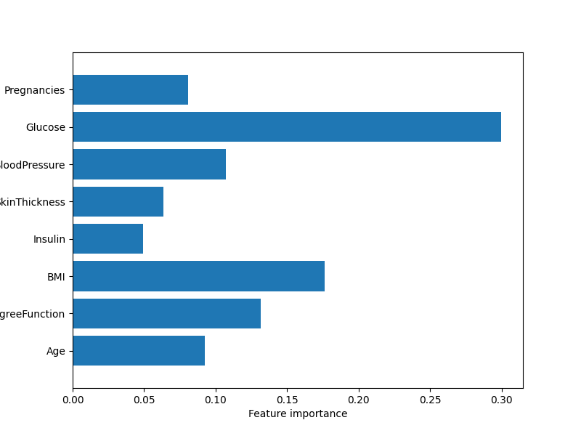
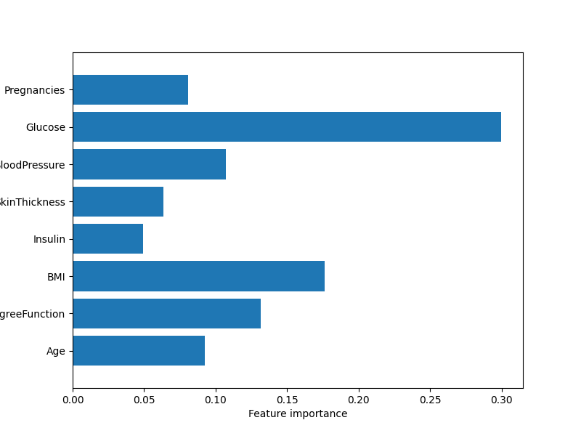
自動產生的描述

隨機森林的正確率略高於決策樹，決策樹加上Bagging後會增加其正確率。

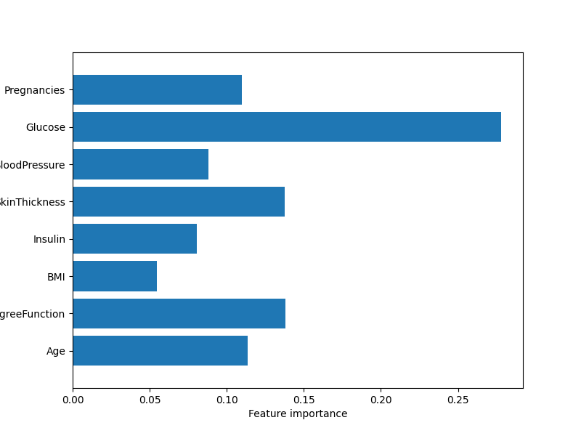
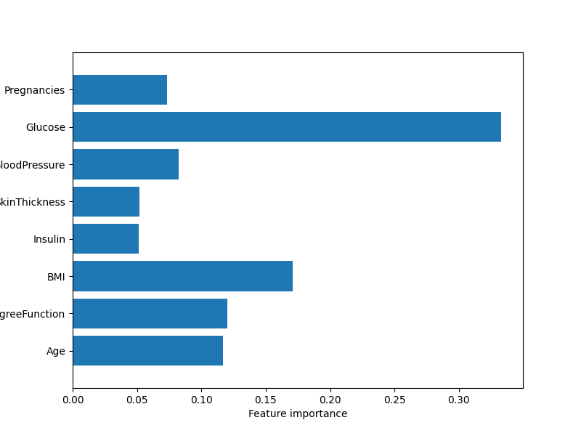


**七項因素與糖尿病相關性：**

Decision tree Decision tree (best depth)

Random forest Decision tree (Bagging)

1. 參考文獻

1, <https://pyecontech.com/2020/04/19/knn/>

2，[[Day 12] 決策樹 (Decision tree) - iT 邦幫忙::一起幫忙解決難題，拯救 IT 人的一天 (ithome.com.tw)](https://ithelp.ithome.com.tw/articles/10271143)

3，[機器學習 - 維基百科，自由的百科全書 (wikipedia.org)](https://zh.wikipedia.org/zh-tw/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0#%E5%88%86%E7%B1%BB)

4，[【機器學習懶人包】 10種演算法圖解-從監督式到非監督式學習 | 達內科技報 | 接軌AI世代的數位觀點 (tedu.tw)](https://www.tedu.tw/blog/10-machine-learning-algorithm-html.html)