1. **資料集特徵資料說明、屬性特性說明**

- no_times_pregnant: 懷孕次數，Interval
- glucose concentration: 血糖濃度，Ratio
- blood_pressure: Ratio
- skin_fold_thickness: 皮摺厚度，Ratio
- serum_insulin: 胰島素抗性，Ratio
- BMI: Ratio
- Diabetes pedigree: 糖尿病函數，Ratio
- Age: Ratio

2. **對特徵做甚麼樣的分析 ? 哪些前處理 ? 採用哪些特徵 ? 原因?**

我先補齊資料，然後對資料進行 feature selection. 嘗試採用了 Select From model(random forest) 和 RFE 兩種方法。不斷調整選擇的 feature 數目，發現當模型會在選擇 4 項特徵時得出最好的結果，
分別為: 'glucose_concentration', 'bmi', 'diabetes pedigree', 'age'

3. **基於什麼理由選擇哪個分類器?**

因為資料集要預測的是是 否為糖尿病患者，為一個分類問題，所以我用的都是 classification 的方法。嘗試了 Decision Tree, Random Forest, KNN, gradient boost. 最後選擇評分最高的當作 submission (Random Forest). 在過程中 如 gradien boost 我有嘗試去調整他的 n estimators, 有觀察到在 100 多的時候訓練效果最好，超過 200 左右就會開始 overfitting 了.

4. **採用的評估指標結果與觀察**

一開始我用的評比就是單純的 accuracy score, 發現不同的方法他們的 accuracy score 都差不多在 75-80 上下。但實際 submit 出去後正確率卻落差很大，從 60-75 都有。後來改採用 Kfold 來評估，不同模型的正確率排序就變得比較精準. Gradient Boost 跟 Random forest 在採用 4 個 feature 的時候訓練效果相對較好。
我也有調整前面 feature selection 所選的 feature 數量。當選擇 3,4,5 out of 7 features 時他的 accuracy score 都在 72-75 之間，沒有太大的變化。而超過或小於時對模型的準確度就有顯著的降低了。

5. **將預測結果上傳至 kaggle 並截圖測試的分數**