Data Structure

Homework: Linked List and Decision Tree

Report and discuss

B12508025 盧雅筠

1. 簡介

本作業利用此資料”Diagnosis_7features.csv” Total :640 datas，每個data包含七個特徵

1. p_bmi
2. personal_Hypertension
3. personal _Hypertension _Year
4. personal _CHF
5. personal_PepticUlcer
6. SBP_pre
7. eGFR_pre

與一個二標籤 class label: 0, 1 (1: positive, 0:negative)，採用Gini index、物件導向設計建構以linked lists的二元決策樹，同時透過設定每個leaf node至少包含 5 個受試者來避免overfitting。

---

2. 方法

2.1 `Load .csv`

- 程式中利用 `ifstream` 讀取 CSV 檔案，並分離出特徵與標籤。第一行為特徵名稱，後續每行分別存入 7 個數值（特徵）與一個標籤。

2.2 `Count the number of 1 and 0`

- 計算label == 0的數量count0=n2(negative), label == 1 的數量count1=n1(positive)

2.3 `Compute Gini Index`

$$Gini(t) = 1 - \sum_{j} [\, p\,(j\,|t)\,]^2$$

2.4 `Decision Tree`

- 節點結構：每個節點以 `Node` 類別實作，包含是否為葉節點、所使用的特徵索引、分割門檻、左右子節點（以 unique_ptr 管理）以及當前節點包含的資料索引。
- 樹的建構：
    - 利用遞迴方法 `buildTree`，對給定的資料索引集進行分割。當資料數小於最小樣本數（5）或該節點資料均屬單一類別時，將該節點設為葉節點。
    - 在每個非葉節點中，程式對所有特徵進行排序，並針對連續值的中點作為候選門檻，計算左右子集合的 Gini 指數，再以加權平均的方式評估分割品質，選擇使加權 Gini 指數最小的分割。
- 過度擬合的避免與樹的簡化
    - 為避免過度擬合，每個葉節點至少保留 5 筆資料。
    - 此外，若左右子節點均為葉節點且所歸類的標籤相同，則合併成一個葉節點，以達到樹的簡化。

2.5 `Train`

- 模型訓練提供兩種情境：
    - 訓練/測試分割：使用 450 筆資料作為訓練集，剩餘 190 筆作為測試集，建立模型後評估其在未見資料上的表現。
    - 全資料訓練：使用全部 640 筆資料建立決策樹，並計算所有葉節點中根據多數決原則正確分類的樣本數，進而計算整體準確率。

2.6 `Accuaracy`

- 預測時，從樹根開始，依據每個節點的分割條件，將樣本傳遞到左或右子節點，直到到達葉節點，其標籤即為預測結果。
- 準確率計算分為兩種情境：
    1. 使用 450 筆資料訓練，190 筆資料測試，計算測試集上正確預測的比例。
    2. 全部資料用於訓練，計算所有葉節點中正確分類的數量（對於正向葉節點累加 n1，對於負向葉節點累加 n2），並除以 640 以得到整體準確率。

2.7 `Print`

- 提供 printTree 函數，可在終端以樹狀結構形式輸出決策樹。
    - 每個節點顯示其分割依據（特徵名稱、門檻值 ）及該節點的 Gini 指數。
    - 葉節點同時顯示其分類結果與 n1、n2 統計數據。

2.8 `Test`

- 利用 predict 函數，從根節點根據分割條件依次遞迴至葉節點，返回葉節點的最終分類標籤，以實現對新樣本的分類。

2.9 視覺化輸出

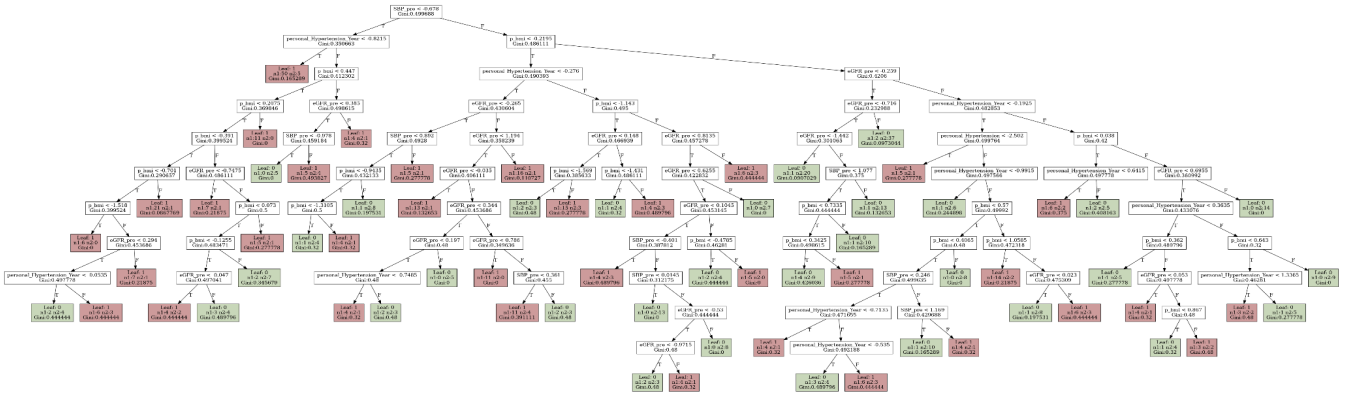
- 程式中提供導出 DOT 格式檔案的功能，透過生成節點與邊的描述，使得決策樹可以進一步轉換成圖形（如 PNG 格式）來觀察整體樹結構。

3. 決策樹結果討論

3.1 整個樹結構

- 視覺化 DOT轉PNG 格式 result

  （紅色代表leaf node label==1==positve, 綠色代表leaf node label==0==negative)

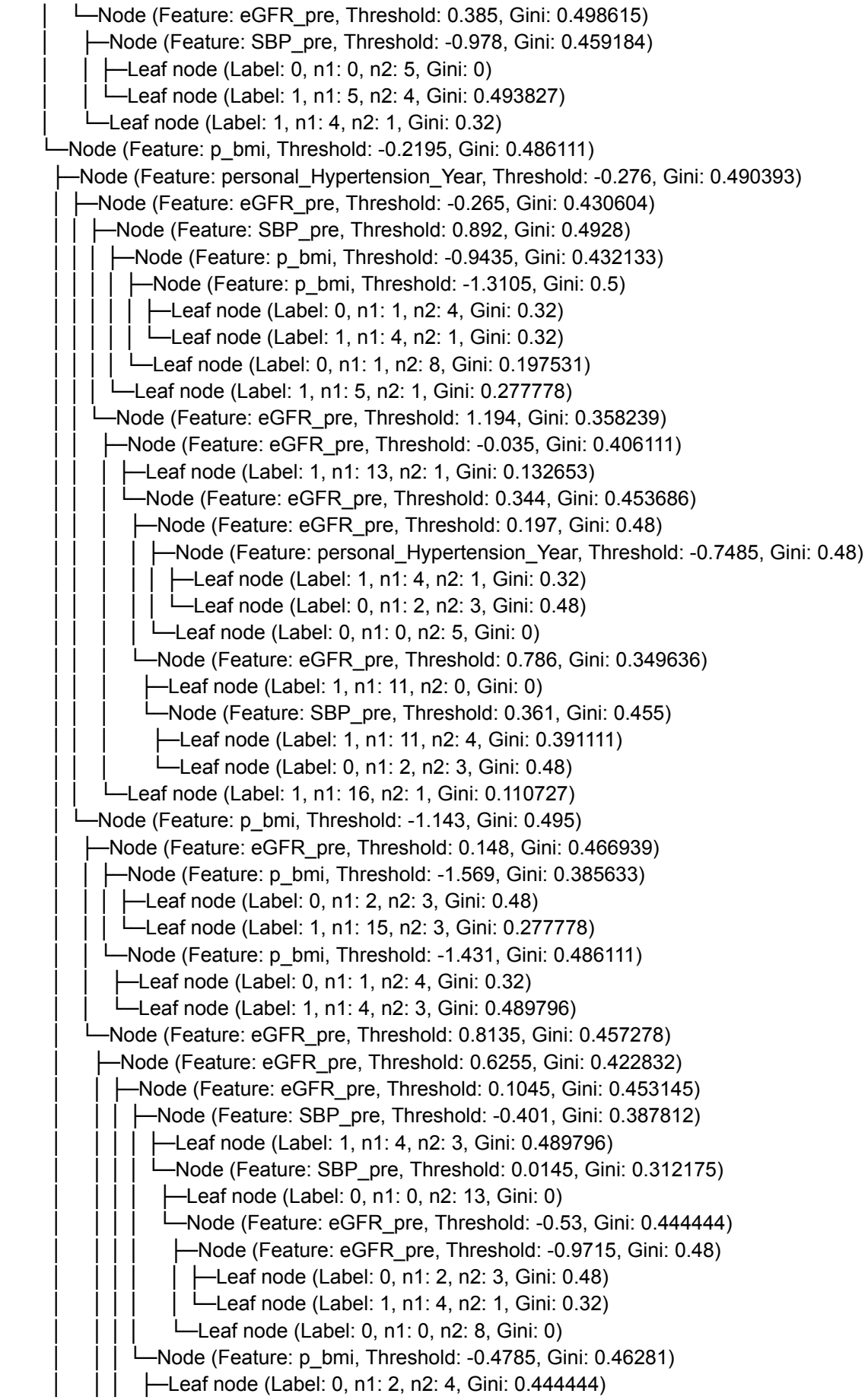


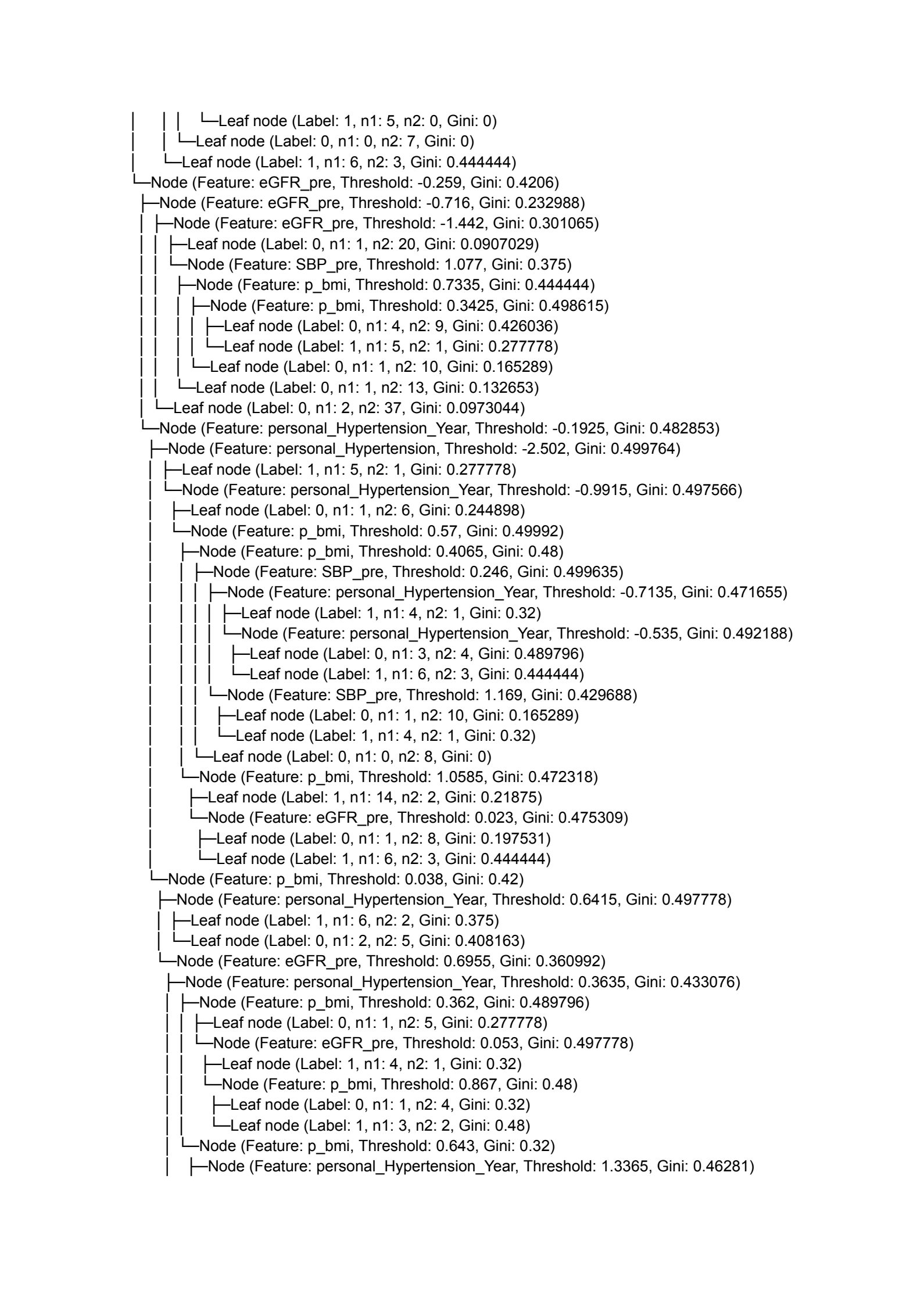
- Terminal Result

Case1: 450 datas for train; 190 datas for test
Test set accuracy: 58.9474%

Case:2 All datas go trainning!
Decision Tree:

```
└─Node (Feature: SBP_pre, Threshold: -0.678, Gini: 0.499688)
  ├─Node (Feature: personal_Hypertension_Year, Threshold: -0.8215, Gini: 0.350663)
  │ ├─Leaf node (Label: 1, n1: 50, n2: 5, Gini: 0.165289)
  │ └─Node (Feature: p_bmi, Threshold: 0.447, Gini: 0.412302)
  │   ├─Node (Feature: p_bmi, Threshold: 0.2075, Gini: 0.369846)
  │   │ ├─Node (Feature: p_bmi, Threshold: -0.391, Gini: 0.399524)
  │   │ │ ├─Node (Feature: p_bmi, Threshold: -0.701, Gini: 0.290657)
  │   │ │ │ ├─Node (Feature: p_bmi, Threshold: -1.518, Gini: 0.399524)
  │   │ │ │ │ ├─Leaf node (Label: 1, n1: 6, n2: 0, Gini: 0)
  │   │ │ │ │ └─Node (Feature: eGFR_pre, Threshold: 0.294, Gini: 0.453686)
  │   │ │ │ │   ├─Node (Feature: personal_Hypertension_Year, Threshold: -0.0535, Gini: 0.497778)
  │   │ │ │ │   │ ├─Leaf node (Label: 0, n1: 2, n2: 4, Gini: 0.444444)
  │   │ │ │ │   │ └─Leaf node (Label: 1, n1: 6, n2: 3, Gini: 0.444444)
  │   │ │ │ │   └─Leaf node (Label: 1, n1: 7, n2: 1, Gini: 0.21875)
  │   │ │ │ └─Leaf node (Label: 1, n1: 21, n2: 1, Gini: 0.0867769)
  │   │ │ └─Node (Feature: eGFR_pre, Threshold: -0.7475, Gini: 0.486111)
  │   │ │   ├─Leaf node (Label: 1, n1: 7, n2: 1, Gini: 0.21875)
  │   │ │   └─Node (Feature: p_bmi, Threshold: 0.073, Gini: 0.5)
  │   │ │     ├─Node (Feature: p_bmi, Threshold: -0.1255, Gini: 0.483471)
  │   │ │     │ ├─Node (Feature: eGFR_pre, Threshold: -0.047, Gini: 0.497041)
  │   │ │     │ │ ├─Leaf node (Label: 1, n1: 4, n2: 2, Gini: 0.444444)
  │   │ │     │ │ └─Leaf node (Label: 0, n1: 3, n2: 4, Gini: 0.489796)
  │   │ │     │ └─Leaf node (Label: 0, n1: 2, n2: 7, Gini: 0.345679)
  │   │ │     └─Leaf node (Label: 1, n1: 5, n2: 1, Gini: 0.277778)
  │   │ └─Leaf node (Label: 1, n1: 11, n2: 0, Gini: 0)
```


└Node (Feature: eGFR_pre, Threshold: 0.385, Gini: 0.498615)
├Node (Feature: SBP_pre, Threshold: -0.978, Gini: 0.459184)
├Leaf node (Label: 0, n1: 0, n2: 5, Gini: 0)
└Leaf node (Label: 1, n1: 5, n2: 4, Gini: 0.493827)
└Leaf node (Label: 1, n1: 4, n2: 1, Gini: 0.32)
└Node (Feature: p_bmi, Threshold: -0.2195, Gini: 0.486111)
├Node (Feature: personal_Hypertension_Year, Threshold: -0.276, Gini: 0.490393)
├Node (Feature: eGFR_pre, Threshold: -0.265, Gini: 0.430604)
├Node (Feature: SBP_pre, Threshold: 0.892, Gini: 0.4928)
├Node (Feature: p_bmi, Threshold: -0.9435, Gini: 0.432133)
├Node (Feature: p_bmi, Threshold: -1.3105, Gini: 0.5)
├Leaf node (Label: 0, n1: 1, n2: 4, Gini: 0.32)
└Leaf node (Label: 1, n1: 4, n2: 1, Gini: 0.32)
└Leaf node (Label: 0, n1: 1, n2: 8, Gini: 0.197531)
└Leaf node (Label: 1, n1: 5, n2: 1, Gini: 0.277778)
└Node (Feature: eGFR_pre, Threshold: 1.194, Gini: 0.358239)
├Node (Feature: eGFR_pre, Threshold: -0.035, Gini: 0.406111)
├Leaf node (Label: 1, n1: 13, n2: 1, Gini: 0.132653)
└Node (Feature: eGFR_pre, Threshold: 0.344, Gini: 0.453686)
├Node (Feature: eGFR_pre, Threshold: 0.197, Gini: 0.48)
├Node (Feature: personal_Hypertension_Year, Threshold: -0.7485, Gini: 0.48)
├Leaf node (Label: 1, n1: 4, n2: 1, Gini: 0.32)
└Leaf node (Label: 0, n1: 2, n2: 3, Gini: 0.48)
└Leaf node (Label: 0, n1: 0, n2: 5, Gini: 0)
└Node (Feature: eGFR_pre, Threshold: 0.786, Gini: 0.349636)
├Leaf node (Label: 1, n1: 11, n2: 0, Gini: 0)
└Node (Feature: SBP_pre, Threshold: 0.361, Gini: 0.455)
├Leaf node (Label: 1, n1: 11, n2: 4, Gini: 0.391111)
└Leaf node (Label: 0, n1: 2, n2: 3, Gini: 0.48)
└Leaf node (Label: 1, n1: 16, n2: 1, Gini: 0.110727)
└Node (Feature: p_bmi, Threshold: -1.143, Gini: 0.495)
├Node (Feature: eGFR_pre, Threshold: 0.148, Gini: 0.466939)
├Node (Feature: p_bmi, Threshold: -1.569, Gini: 0.385633)
├Leaf node (Label: 0, n1: 2, n2: 3, Gini: 0.48)
└Leaf node (Label: 1, n1: 15, n2: 3, Gini: 0.277778)
└Node (Feature: p_bmi, Threshold: -1.431, Gini: 0.486111)
├Leaf node (Label: 0, n1: 1, n2: 4, Gini: 0.32)
└Leaf node (Label: 1, n1: 4, n2: 3, Gini: 0.489796)
└Node (Feature: eGFR_pre, Threshold: 0.8135, Gini: 0.457278)
├Node (Feature: eGFR_pre, Threshold: 0.6255, Gini: 0.422832)
├Node (Feature: eGFR_pre, Threshold: 0.1045, Gini: 0.453145)
├Node (Feature: SBP_pre, Threshold: -0.401, Gini: 0.387812)
├Leaf node (Label: 1, n1: 4, n2: 3, Gini: 0.489796)
└Node (Feature: SBP_pre, Threshold: 0.0145, Gini: 0.312175)
├Leaf node (Label: 0, n1: 0, n2: 13, Gini: 0)
└Node (Feature: eGFR_pre, Threshold: -0.53, Gini: 0.444444)
├Node (Feature: eGFR_pre, Threshold: -0.9715, Gini: 0.48)
├Leaf node (Label: 0, n1: 2, n2: 3, Gini: 0.48)
└Leaf node (Label: 1, n1: 4, n2: 1, Gini: 0.32)
└Leaf node (Label: 0, n1: 0, n2: 8, Gini: 0)
└Node (Feature: p_bmi, Threshold: -0.4785, Gini: 0.46281)
├Leaf node (Label: 0, n1: 2, n2: 4, Gini: 0.444444)


└─Leaf node (Label: 1, n1: 5, n2: 0, Gini: 0)
└─Leaf node (Label: 0, n1: 0, n2: 7, Gini: 0)
└─Leaf node (Label: 1, n1: 6, n2: 3, Gini: 0.444444)
└─Node (Feature: eGFR_pre, Threshold: -0.259, Gini: 0.4206)
├─Node (Feature: eGFR_pre, Threshold: -0.716, Gini: 0.232988)
├─Node (Feature: eGFR_pre, Threshold: -1.442, Gini: 0.301065)
├─Leaf node (Label: 0, n1: 1, n2: 20, Gini: 0.0907029)
└─Node (Feature: SBP_pre, Threshold: 1.077, Gini: 0.375)
├─Node (Feature: p_bmi, Threshold: 0.7335, Gini: 0.444444)
├─Node (Feature: p_bmi, Threshold: 0.3425, Gini: 0.498615)
├─Leaf node (Label: 0, n1: 4, n2: 9, Gini: 0.426036)
└─Leaf node (Label: 1, n1: 5, n2: 1, Gini: 0.277778)
└─Leaf node (Label: 0, n1: 1, n2: 10, Gini: 0.165289)
└─Leaf node (Label: 0, n1: 1, n2: 13, Gini: 0.132653)
└─Leaf node (Label: 0, n1: 2, n2: 37, Gini: 0.0973044)
└─Node (Feature: personal_Hypertension_Year, Threshold: -0.1925, Gini: 0.482853)
├─Node (Feature: personal_Hypertension, Threshold: -2.502, Gini: 0.499764)
├─Leaf node (Label: 1, n1: 5, n2: 1, Gini: 0.277778)
└─Node (Feature: personal_Hypertension_Year, Threshold: -0.9915, Gini: 0.497566)
├─Leaf node (Label: 0, n1: 1, n2: 6, Gini: 0.244898)
└─Node (Feature: p_bmi, Threshold: 0.57, Gini: 0.49992)
├─Node (Feature: p_bmi, Threshold: 0.4065, Gini: 0.48)
├─Node (Feature: SBP_pre, Threshold: 0.246, Gini: 0.499635)
├─Node (Feature: personal_Hypertension_Year, Threshold: -0.7135, Gini: 0.471655)
├─Leaf node (Label: 1, n1: 4, n2: 1, Gini: 0.32)
└─Node (Feature: personal_Hypertension_Year, Threshold: -0.535, Gini: 0.492188)
├─Leaf node (Label: 0, n1: 3, n2: 4, Gini: 0.489796)
└─Leaf node (Label: 1, n1: 6, n2: 3, Gini: 0.444444)
└─Node (Feature: SBP_pre, Threshold: 1.169, Gini: 0.429688)
├─Leaf node (Label: 0, n1: 1, n2: 10, Gini: 0.165289)
└─Leaf node (Label: 1, n1: 4, n2: 1, Gini: 0.32)
└─Leaf node (Label: 0, n1: 0, n2: 8, Gini: 0)
└─Node (Feature: p_bmi, Threshold: 1.0585, Gini: 0.472318)
├─Leaf node (Label: 1, n1: 14, n2: 2, Gini: 0.21875)
└─Node (Feature: eGFR_pre, Threshold: 0.023, Gini: 0.475309)
├─Leaf node (Label: 0, n1: 1, n2: 8, Gini: 0.197531)
└─Leaf node (Label: 1, n1: 6, n2: 3, Gini: 0.444444)
└─Node (Feature: p_bmi, Threshold: 0.038, Gini: 0.42)
├─Node (Feature: personal_Hypertension_Year, Threshold: 0.6415, Gini: 0.497778)
├─Leaf node (Label: 1, n1: 6, n2: 2, Gini: 0.375)
└─Leaf node (Label: 0, n1: 2, n2: 5, Gini: 0.408163)
└─Node (Feature: eGFR_pre, Threshold: 0.6955, Gini: 0.360992)
├─Node (Feature: personal_Hypertension_Year, Threshold: 0.3635, Gini: 0.433076)
├─Node (Feature: p_bmi, Threshold: 0.362, Gini: 0.489796)
├─Leaf node (Label: 0, n1: 1, n2: 5, Gini: 0.277778)
└─Node (Feature: eGFR_pre, Threshold: 0.053, Gini: 0.497778)
├─Leaf node (Label: 1, n1: 4, n2: 1, Gini: 0.32)
└─Node (Feature: p_bmi, Threshold: 0.867, Gini: 0.48)
├─Leaf node (Label: 0, n1: 1, n2: 4, Gini: 0.32)
└─Leaf node (Label: 1, n1: 3, n2: 2, Gini: 0.48)
└─Node (Feature: p_bmi, Threshold: 0.643, Gini: 0.32)
├─Node (Feature: personal_Hypertension_Year, Threshold: 1.3365, Gini: 0.46281)

```
│   │ ├─Leaf node (Label: 1, n1: 3, n2: 2, Gini: 0.48)
│   │ └─Leaf node (Label: 0, n1: 1, n2: 5, Gini: 0.277778)
│   └─Leaf node (Label: 0, n1: 0, n2: 9, Gini: 0)
└─Leaf node (Label: 0, n1: 0, n2: 14, Gini: 0)
Accuracy: 84.375%
```

決策樹從根節點開始，依據各節點的最佳特徵與門檻分割資料。每個非葉節點顯示所依據的特徵名稱、門檻值以及該節點的 Gini 指數。透過遞迴分割，最終形成許多葉節點，這些葉節點代表了分類的最終決策。

3.2 節點特徵與門檻值

- 在每個內部節點，程式遍歷所有 7 個臨床特徵，並利用相鄰數值的平均作為候選門檻，選擇使分割後加權 Gini 指數最小的門檻。這代表該特徵在該門檻值下能夠有效區分正向與負向病例。

3.3 Gini 指數的角色

- Gini 指數作為不純度的度量，用以評估資料分割後的純淨度。數值越低表示該節點中資料越單一。各節點在決策樹構建過程中都會計算 Gini 指數，以尋找最佳分割點，從而逐步提升分類效果。

3.4 葉節點的分類與統計

- 當無法繼續分割或滿足最小樣本數要求時，節點被標記為葉節點。在葉節點中，計算正向（n1）與負向（n2）樣本數，依據 n1 與 n2 的比較結果來決定葉節點的最終分類：
    - 若 n1 ≥ n2，則該葉節點分類為正向。
    - 否則，分類為負向。
- 這種策略確保每個葉節點皆依據多數原則作出最合理的分類決定。
- 子葉最後分出來會相同的問題也解決。

3.5 決策樹的準確率

- 程式分別在兩個情境下評估決策樹的準確率：
    1. 訓練/測試分割情境：利用 450 筆訓練資料建立樹模型，再以 190 筆測試資料進行預測，其準確率為58.9474%，這是按照約，找到的相對能使準確率較高的訓練集與測試集切分方法。
    2. 全部資料訓練情境：全部 640 筆資料參與樹的建構，進而計算葉節點中的正確分類總數，並以此除以 640 得到準確率。
- 從結果中可以看出，當模型僅使用部分資料進行訓練時，測試集準確率可能較低，但能較真實地反映模型在實際應用中的表現；而全資料訓練則可能因過擬合而使得訓練準確率偏高，因此兩種情境各有其參考價值。

3.6 物件導向設計與記憶體管理：程式採用 C++ 的 unique_ptr 來管理動態配置的節點，確保記憶體不會泄露(Smart Pointer: can manage dynamic memory automatically)，同時利用 linked list（節點間的指標連結）來實現樹結構，達到模組化、易於擴展的目的。

3.7 分割準則與最小樣本數：採用 Gini 指數作為分割準則能夠有效衡量資料的不純度；同時通過設置最小樣本數限制，避免模型過度分割從而導致過擬合。這在臨床資料這類可能具有雜訊的情境中尤為重要。

3.8 視覺化與解釋性：程式提供 DOT 格式的導出功能，使得決策樹結構能夠以圖形方式呈現，方便後續分析與解釋。對於每個節點，不僅記錄了所用特徵與門檻值，還包含了該節點的 Gini 指數，這有助於使用者更直觀地理解模型的分割邏輯。

---

4. 未來改進

- 可考慮進一步引入交叉驗證來評估模型在不同資料分割下的穩定性。
- 探討其他分割準則（例如熵值）或後剪枝技術，以期望獲得更好的泛化能力。
- 根據不同臨床資料特性，或許可以針對特定特徵進行更細緻的預處理或權重調整，進一步提升模型的分類效能。
- 可使用隨機森林

結論

本作業成功根據匿名臨床資料構建了一棵二元決策樹，並利用 Gini 指數作為分割標準、linked list（透過 smart pointer）來實現樹狀結構，同時採用了最小樣本數作為防止過度擬合的策略。從整體樹結構、各節點的分割條件與不純度測量，到葉節點的最終分類與準確率計算，都展現了決策樹在分類問題中的有效性與解釋性。未來可針對模型的泛化能力和穩定性進行進一步優化與調整。
