HW9

612415013 蕭宥羽

1. How to execute codes.
   - 原始程式碼（不設定最大深度）：

```
####################
# 題目: 更改 tree 的深度
model = DecisionTreeClassifier(random_state=42)
####################
```

   - Max_depth = 3

```
####################
# 題目: 更改 tree 的深度
model = DecisionTreeClassifier(random_state=42, max_depth=3)
####################
```

   - Max_depth = 5

```
####################
# 題目: 更改 tree 的深度
model = DecisionTreeClassifier(random_state=42, max_depth=5)
####################
```

   - Max_depth = 7

```
####################
# 題目: 更改 tree 的深度
model = DecisionTreeClassifier(random_state=42, max_depth=7)
####################
```

2. Experimental results
   - 原始程式碼（不設定最大深度）：
     - 準確率

```
Accuracy: 0.77
```

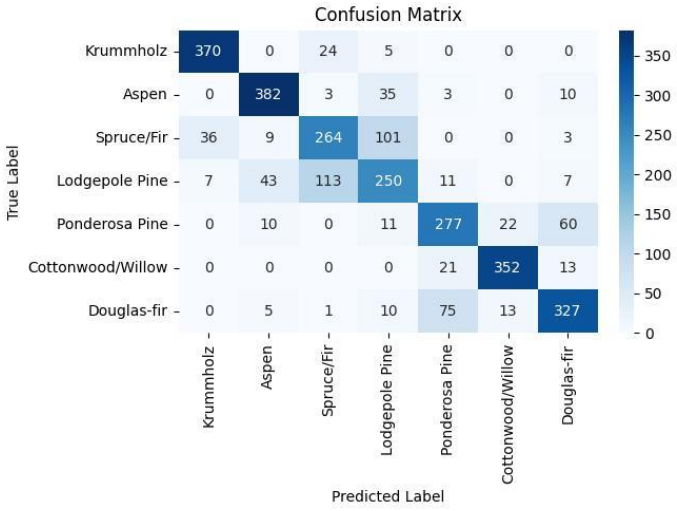
     - Feature

Feature	Importance
Elevation	0.413249
Horizontal_Distance_To_Fire_Points	0.090462
Horizontal_Distance_To_Roadways	0.089369
Horizontal_Distance_To_Hydrology	0.068992
Hillshade_9am	0.061754
Aspect	0.038547
Hillshade_Noon	0.034490
Vertical_Distance_To_Hydrology	0.032935
Hillshade_3pm	0.028591
Slope	0.019119
Soil_Type3	0.014799
Wilderness_Area1	0.013588
Soil_Type4	0.013381
Soil_Type10	0.010529
Soil_Type39	0.007644
Wilderness_Area4	0.005005
Soil_Type17	0.004707
Soil_Type32	0.004478
Wilderness_Area3	0.004469
Soil_Type12	0.004207
Soil_Type38	0.003960
Soil_Type29	0.003599
Soil_Type33	0.003194
Soil_Type6	0.003138
Soil_Type2	0.002999
Soil_Type22	0.002645
Soil_Type11	0.002557
Soil_Type30	0.002471
Soil_Type35	0.002217
Soil_Type24	0.001890
Soil_Type31	0.001668
Soil_Type20	0.001513
Soil_Type1	0.001443
Soil_Type23	0.001430
Soil_Type40	0.001092
Soil_Type16	0.000994
Soil_Type13	0.000677
Wilderness_Area2	0.000540
Soil_Type19	0.000536
Soil_Type5	0.000433
Soil_Type36	0.000185
Soil_Type26	0.000181
Soil_Type34	0.000174
Soil_Type27	0.000148

- 前 10 筆資料分類的結果

Id	Actual	Predicted
6ayWrW52q9tNzbW	6	6
0GEpYEmMMaz1Kqs	2	2
IljBNXffVG0Em4q	6	4
b86tOO0UtnhVqRw	1	1
NaZLqGwSyMJBGEl	2	2
rwMh0RKC3sdtD7P	2	2
HvdLlTPQ0xQJmcL	3	3
v5DvSlMYMOlTzwi	3	6
Rn94CHKoAlyDp1K	6	6
to8D8C1zfTBFEoY	6	6

- 混淆矩陣



- Max_depth = 3
  - 準確率

```
Accuracy: 0.60
```

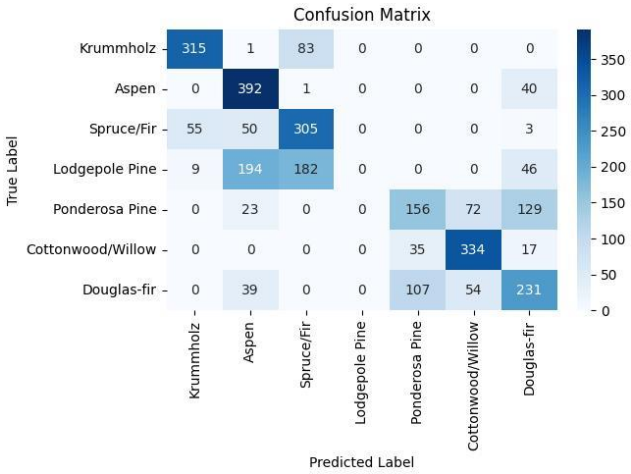
  - Feature

Feature	Importance
Elevation	0.892229
Hillshade_9am	0.075465
Soil_Type4	0.032306

  - 前 10 筆資料分類的結果

Id	Actual	Predicted
6ayWrW52q9tNzbW	6	6
0GEpYEmMMaz1Kqs	2	2
IljBNXffVG0Em4q	6	5
b86tOO0UtnhVqRw	1	1
NaZLqGwSyMJBGEl	2	2
rwMh0RKC3sdtD7P	2	2
HvdLlTPQ0xQJmcL	3	1
v5DvSlMYMOlTzwi	3	1
Rn94CHKoAlyDp1K	6	6
to8D8C1zfTBFEoY	6	6

- 混淆矩陣



- Max_depth = 5
  - 準確率

```
Accuracy: 0.66
```

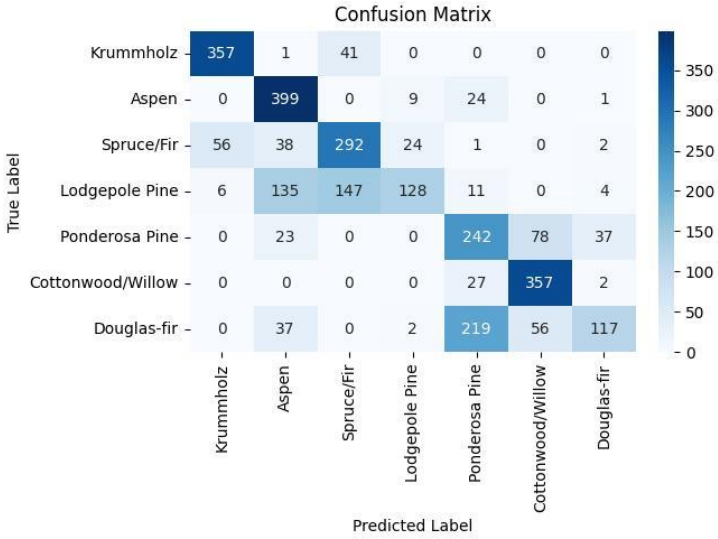
  - Feature

Feature	Importance
Elevation	0.715720
Hillshade_9am	0.059571
Horizontal_Distance_To_Hydrology	0.041703
Horizontal_Distance_To_Fire_Points	0.034755
Soil_Type4	0.025501
Soil_Type3	0.025215
Wilderness_Area1	0.023996
Horizontal_Distance_To_Roadways	0.019267
Soil_Type10	0.013068
Soil_Type39	0.008775
Soil_Type38	0.007666
Soil_Type17	0.006248
Soil_Type32	0.004887
Soil_Type22	0.003915
Wilderness_Area4	0.003808
Aspect	0.003222
Slope	0.002683

  - 前 10 筆資料分類的結果

Id	Actual	Predicted
6ayWrW52q9tNzbW	6	4
0GEpYEmMMaz1Kqs	2	2
IljBNXffVG0Em4q	6	5
b86tOO0UtnhVqRw	1	1
NaZLqGwSyMJBGEl	2	2
rwMh0RKC3sdtD7P	2	2
HvdLlTPQ0xQJmcL	3	3
v5DvSlMYMOlTzwi	3	1
Rn94CHKoAlyDp1K	6	6
to8D8C1zfTBFEoY	6	6

- 混淆矩陣



- Max_depth = 7
  - 準確率

```
Accuracy: 0.71
```

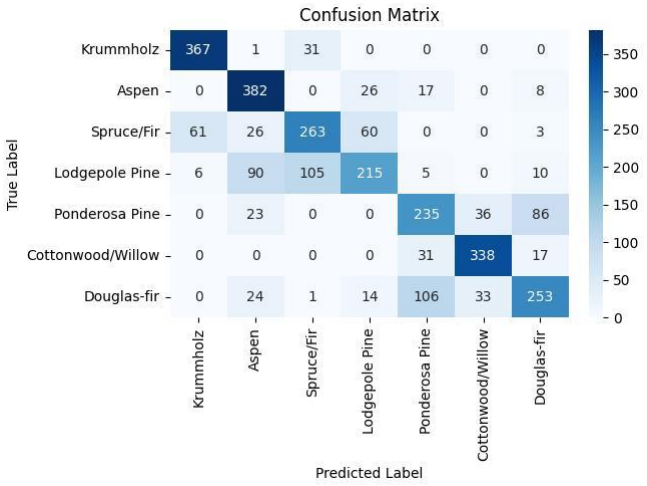
  - Feature

Feature	Importance
Elevation	0.632647
Hillshade_9am	0.062534
Horizontal_Distance_To_Hydrology	0.052989
Horizontal_Distance_To_Fire_Points	0.047484
Horizontal_Distance_To_Roadways	0.044374
Soil_Type4	0.022008
Soil_Type3	0.021761
Wilderness_Area1	0.020710
Soil_Type39	0.012045
Soil_Type10	0.011278
Wilderness_Area3	0.010904
Hillshade_Noon	0.010849
Aspect	0.009385
Soil_Type12	0.007061
Soil_Type38	0.006616
Vertical_Distance_To_Hydrology	0.005988
Soil_Type17	0.005392
Slope	0.004727
Soil_Type32	0.004218
Soil_Type22	0.003378
Soil_Type40	0.001761
Soil_Type13	0.000774
Soil_Type33	0.000585
Soil_Type31	0.000532

  - 前 10 筆資料分類的結果

Id	Actual	Predicted
6ayWrW52q9tNzbW	6	4
0GEpYEmMMaz1Kqs	2	2
IljBNXffVG0Em4q	6	4
b86tOO0UtnhVqRw	1	1
NaZLqGwSyMJBGEl	2	3
rwMh0RKC3sdtD7P	2	2
HvdLlTPQ0xQJmcL	3	3
v5DvSlMYMOlTzwi	3	1
Rn94CHKoAlyDp1K	6	6
to8D8C1zfTBFEoY	6	6

- 混淆矩陣



3. Conclusion

- 準確率比較
  - ✓ 原始程式碼（不設定最大深度）：達到最高的準確率 0.77，模型會選擇他認為的最佳深度來使用，而在本次作業中的正確率為最高。
  - ✓ Max_depth = 3:此模型的準確率為 0.60，反映了明顯的欠擬合問題。由於模型過於簡化，未能充分捕捉數據中的複雜特徵，從而導致預測能力不足。
  - ✓ Max_depth = 5:模型的準確率提高到 0.66，這表明模型已經開始捕捉到更多的數據特徵。雖然性能有所提升，但模型仍未能完全揭示數據的全部潛在模式。
  - ✓ Max_depth = 7:準確率進一步增至 0.71，顯示出在更深的樹結構下，模型的擬合能力得到加強，能夠辨識更細微的數據模式。
- 前 10 筆資料比較

  圖中的每一個表格都列出了數據點的識別碼（Id），實際類別（Actual）和預測類別（Predicted）。
  - ✓ 深度 3 的模型在預測方面顯示出較大的誤差，特別是在 Ids "ILjBNXfVG0EM4q" 和 "HvdLLTPQ0xQJMcl" 的預測上，其中預測的類別與實際類別相差甚遠。這可能表明模型因為深度不夠而無法充分學習數據的特徵，導致欠擬合。
  - ✓ 深度 5 的模型在準確性上有所提升，如 Id "6ayWrW52q9tNzbw" 的預測從深度 3 的錯誤預測中糾正過來。這表明增加深度有助於模型捕捉更多的數據特徵，從而改進預測性能。
  - ✓ 深度 7 的模型進一步改善了一些預測，例如 Id "ILjBNXfVG0EM4q" 和 "6ayWrW52q9tNzbw"，顯示模型在更深的深度下能夠更有效地學習複雜的數據模式。不過，某些預測如 "v5DvSLMYM01Tzwi" 仍有誤差，這可能意味著即使深度增加，也可能需要其他方法來避免過擬合或進一步提升模型的泛化能力。
- 混淆矩陣比較
  - ✓ 深度為 3 的模型：

    這個模型對於大多數樹種的分類有較大的誤差。對於 "Krummholz" 和 "Douglas-fir" 的預測相對較準，但對於 "Spruce/Fir" 的混淆明顯，預測為其他類別的情況較多。特別是 "Ponderosa

Pine" 被經常誤分類為 "Lodgepole Pine"。

- ✓ 深度為 5 的模型：

  準確率提高，尤其是對 "Spruce/Fir" 和 "Ponderosa Pine" 的分類更為準確。"Krummholz" 的分類準確度略有下降，但其他樹種的分類準確度提高。"Lodgepole Pine" 和 "Ponderosa Pine" 之間的混淆減少，但仍存在一定程度的誤分。

- ✓ 深度為 7 的模型：

  這個模型進一步提高了分類的準確率，尤其是對 "Douglas-fir" 和 "Cottonwood/Willow"。"Krummholz" 和 "Spruce/Fir" 的分類準確度仍然很高，顯示模型在更深的深度下能夠更好地捕捉這些類別的特徵。然而，對於 "Aspen" 和 "Douglas-fir" 的分類準確度雖有提升，但依然有部分誤判

- 混淆矩陣比較
  - ✓ 深度為 3 的模型：
    - 在這個淺層模型中，「Elevation」(海拔) 是最重要的特徵，其重要性達到 0.412369。這顯示在較簡單的模型中，海拔高度是影響分類決策的主要因素。
    - 「Horizontal Distance to Fire Points」(與火源點的水平距離) 和「Hillshade 9am」(上午 9 點的陰影) 也是相對重要的特徵，但其重要性遠低於海拔。
    - 土壤類型 (Soil Type) 的重要性普遍較低。
  - ✓ 深度為 5 的模型：
    - 「Elevation」的重要性略有下降，變為 0.715720，但仍然是最重要的特徵。
    - 「Hillshade 9am」的重要性提升，表明在更複雜的模型中，環境光線條件對分類決策的影響增大。
    - 特定的土壤類型，如「Soil Type4」的重要性有所提升，顯示隨著模型深度的增加，模型開始考慮更多細節特徵。
  - ✓ 深度為 7 的模型：
    - 「Elevation」和「Hillshade 9am」的重要性略有下降，這可能是因為模型在更深的層次中學會了從更多的特徵中提取信息。
    - 「Horizontal Distance to Hydrology」(與水文站的水平距離) 的重要性在這個深度下顯著提升，顯示水源對於模型在此深度的決策中變得更加關鍵。
    - 更多的土壤類型變得顯著，如「Soil Type29」和「Soil Type32」，這表明更深的模型能夠捕捉到更細微的地質差異對分類決策的影響。

隨著決策樹模型深度的增加，從深度 3 到深度 7，我們可以觀察到模型的特徵重要性分佈發生了顯著變化，這反映了模型在解釋和利用數據方面的演進。在深度較低的模型中，海拔 (Elevation) 作為單一特徵對分類決策的影響極為重要，但隨著模型深度的增加，其他特徵如「Hillshade 9am」和不同的土壤類型開始顯示出更高的重要性。這表明更深的模型能夠更好地捕捉和整合多種特徵，從而進行更精細

的數據分析。此外，更深的模型在特徵利用上展現了更大的多樣性和平衡，這有助於提高模型對複雜數據模式的理解。然而，這也伴隨著過擬合的風險，因為模型可能過度適應訓練數據中的細節，而在未見數據上的表現可能不佳。因此，選擇適當的模型深度是一個關鍵的步驟，需要在學習能力和泛化能力之間找到平衡。這些洞察促使我們在模型開發過程中更加注重特徵選擇和模型調整，以確保既能捕捉數據的內在規律，也能適應新的、未知的情境。

4. Discussion

在進行這次的作業過程中，我初步對題目的要求感到有些困惑，導致我花費了大量的時間去仔細分析和比較不同的程式碼段落，以確保我能夠完全理解每一行程式的作用和它們如何影響最終的輸出結果。這個過程雖然漫長且充滿挑戰，但它極大地豐富了我的學習經驗，讓我對決策樹模型的深度和特徵重要性有了更深入的了解。此外，這次作業也提升了我的分析能力和解決問題的技巧，使我在面對複雜的數據分析問題時更加自信。通過助教提示如何調整模型參數，我學會了如何調整模型的深度以達到更好的預測準確率，同時也理解了過擬合與欠擬合之間的微妙平衡。每一次的調整都讓我對模型的運作機制有了更透徹的理解，這不僅僅是技術上的提升，更是對數據科學核心概念的深化認識。