Project 2

Decision Tree

1 Description

In this assignment, you are going to build decision trees on real-world datasets using `scikit-learn`.

The datasets you will be working on include:

- **Binary class dataset**: The UCI Heart Disease dataset is used for classifying whether a patient has a heart disease or not based on age, blood pressure, cholesterol level, and other medical indicators. This dataset includes 303 samples, with labels indicating presence (1) or absence (0) of heart disease.
- **Multi-class dataset**: The Palmer Penguins dataset is used for classifying penguin species based on physical characteristics. The dataset includes 344 samples of three penguin species: `Adelie`, `Chinstrap`, and `Gentoo`, with features such as bill length, flipper length, body mass, and sex.
- **Additional dataset**: You have to find another dataset and build the decision tree for it. Please provide a detailed description of the dataset information in your report.

  Your dataset must:

  - Contain both features and labels for supervised learning.
  - Include at least 300 samples for meaningful analysis.
  - Contain multiple classes or at least two binary classes.

2 Specifications

You are required to write **Python Notebooks** (.ipynb) and use the `scikit-learn` library to complete the following tasks described for the **Heart Disease dataset**.

For the remaining datasets (**Penguins dataset** and your **additional dataset**), perform similar tasks as with the Heart Disease dataset.

While there are no strict guidelines for code organization, each task must be clearly documented and fully comply with all specified requirements.

2.1 Preparing the datasets

This task sets up the training and test datasets for the upcoming experiments.

Using the features and labels above, please prepare the following four subsets:

- `feature_train`: a set of training samples.
- `label_train`: a set of labels corresponding to the samples in `feature_train`.
- `feature_test`: a set of test samples with a structure to `feature_train`.
- `label_test`: a set of labels corresponding to the samples in `feature_test`.

You need to shuffle the dataset before splitting and ensure it is split in a stratified fashion. Other parameters (if there are any) should remain at their default settings.

There will be experiments on training and test sets with different proportions, including 40/60, 60/40, 80/20, and 90/10 (train/test); therefore, you will need 16 subsets in total.

Visualize the class distributions in all datasets (the original set, training sets, and test sets) across all proportions to demonstrate that they have been appropriately prepared.

2.2 Building the decision tree classifiers

This task involves conducting experiments on the designated train/test proportions listed above. You need to fit an instance of `sklearn.tree.DecisionTreeClassifier` (using information gain) to each training set and visualize the resulting decision tree with Graphviz.

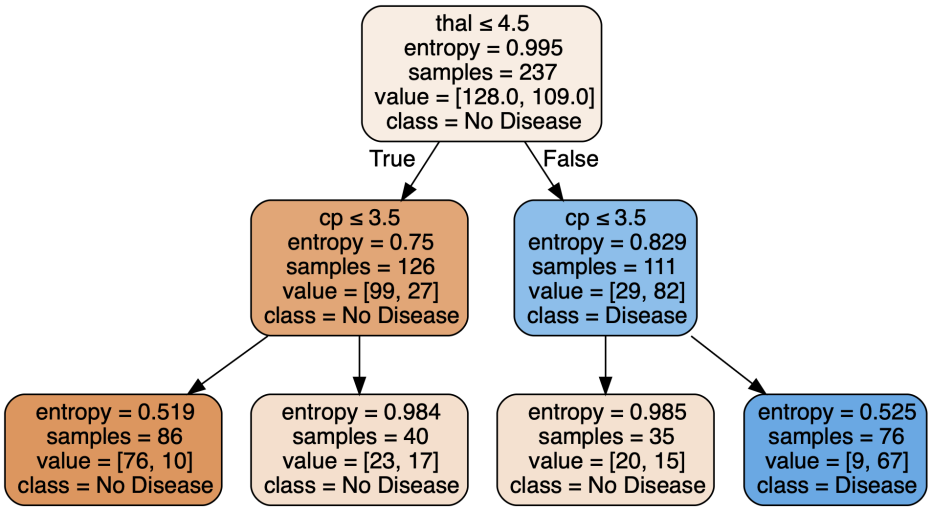

Figure 1: Example for a decision tree classifier (with depth = 2).

2.3 Evaluating the decision tree classifiers

For each of the above decision tree classifiers, predict the samples in the corresponding test set and generate a report using `classification_report` and `confusion_matrix`.

	precision	recall	f1-score	support
No Disease	0.66	0.97	0.78	32
Disease	0.92	0.43	0.59	28
accuracy			0.72	60
macro avg	0.79	0.70	0.69	60
weighted avg	0.78	0.72	0.69	60

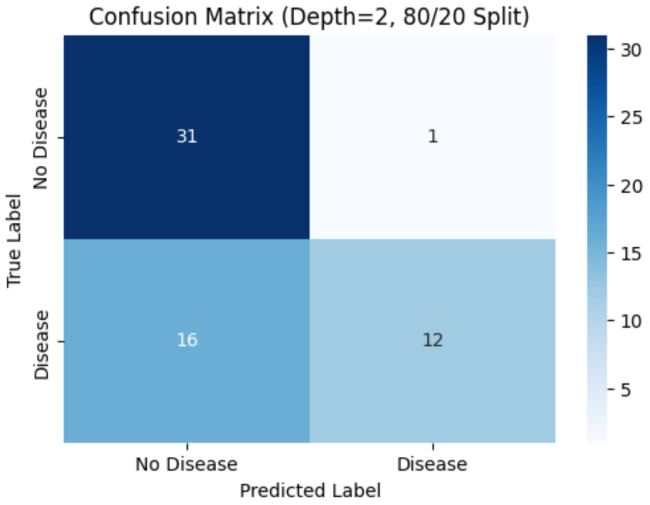

Figure 2: Example for Classification Report and Confusion Matrix.

How do you interpret the classification report and the confusion matrix? Based on the results, provide your insights into the performance of these decision tree classifiers.

2.4 The depth and accuracy of a decision tree

This task focuses on the 80/20 training and test sets. You need to consider that how the depth of the decision tree affects classification accuracy.

You can specify the maximum depth of a decision tree by adjusting the `max_depth` parameter. Try the following values for parameter `max_depth`: None, 2, 3, 4, 5, 6, 7. Then:

- Provide the decision trees, visualized using Graphviz, for each `max_depth` value.
- Report the `accuracy_score` (on the test set) of the decision tree classifier for each value of the `max_depth` parameter in the following table.

`max_depth`	None	2	3	4	5	6	7
Accuracy							

- Provide charts and your insights on the statistics reported above.

2.5 Repeat for Other Datasets

You are required to repeat the same workflow described above for both the **Penguins dataset** and your chosen **Additional dataset**. For categorical features, please use **one-hot encoding**.

After completing the experiments for all datasets, write a comparative analysis in your report. Discuss how characteristics of each dataset — including the number of classes, number of features, and sample size — affect the decision tree's performance. Use tables or plots to summarize your findings and support your conclusions.

3 Requirements

3.1 Report

The report must include the following sections:

- Member information (Student ID, full name, etc.).
- Work assignment table, which includes information on each task assigned to team members, along with the completion rate of each member compared to the assigned tasks. For example, student A has a percentage of completion 90% and the group work has a total score of 9.0, then A receives a score of $9.0 * 90\% = 8.1$.
- A self-evaluation of the completion rate of the project and other requirements.
- All visualizations must be presented in the .ipynb file, while statistical results and insights must be presented in the report.
- The report needs to be well-formatted and exported to PDF. If there are figures cut off by the page break, etc., points will be deducted.
- References (if any).

3.2 Submission

- All reports, code, etc., must be contributed in the form of a compressed file (.zip, .rar, .7z) and named according to the format: `StudentID1_StudentID2_etc.zip/.rar/.7z`.
- If the compressed file is larger than 25MB, prioritize compressing the report and source code. Images and other large files may be uploaded to the Google Drive and shared via a link.

4 Assessment

The detailed assessment criteria for this project are outlined as follows:

No.	Criteria	Score
1	Analysis of the Heart Disease dataset.	30%
2	Analysis of the Palmer Penguins dataset.	30%
3	Analysis of an additional dataset.	30%
4	Comparative analysis of all three datasets.	5%
5	Well-structured and formatted notebooks.	5%
	Total	**100%**

The detailed assessment criteria for each dataset are outlined as follows:

No.	Criteria	Score
1	Data preparation.	30%
2	Implement decision tree classifiers.	20%
3	Performance evaluation of decision tree.	
	- Classification report and confusion matrix.	10%
	- Insights.	10%
4	Depth and accuracy of decision trees.	
	- Visualization (trees, tables, charts).	20%
	- Insights.	10%
	Total	**100%**

5 Notices

Please pay attention to the following notices:

- This is a **GROUP** assignment. Each group has 4 members.
- Duration: about 3 weeks.
- If you use AI tools (e.g., ChatGPT), you must clearly declare the prompts used in appendix; otherwise, it will be considered plagiarism.
- Any plagiarism, any tricks, or any lie will have a 0 point for the course grade.