**VIETNAM NATIONAL UNIVERSITY - HO CHI MINH CITY**

**UNIVERSITY OF SCIENCE**

**FACULTY OF INFORMATION TECHNOLOGY**

**---🙦🙤---**

A blue and white logo

Description automatically generated

COURSE PROJECT REPORT  
**Introduction to Artificial Intelligence**

PROJECT 2 **Decision Tree**

**Lecturer: Nguyen Thanh Tinh**

**Group: 08**

**Class:** Introduction to Artificial Intelligence – CQ2022/1

***Ho Chi Minh City, June 2025***

Contents

[I) Team Information 4](#_Toc201708689)

[1) Member Details 4](#_Toc201708690)

[2) Work Assignment & Completion Rate 4](#_Toc201708691)

[II) Executive Summary 4](#_Toc201708692)

[1) Project Overview 4](#_Toc201708693)

[2) Key Findings 4](#_Toc201708694)

[3) Self-Evaluation 4](#_Toc201708695)

[III) Dataset Analysis 4](#_Toc201708696)

[1) Heart Disease Dataset Analysis 5](#_Toc201708697)

[2) Palmer Penguins Dataset Analysis 7](#_Toc201708698)

[3) Additional Dataset Analysis 9](#_Toc201708699)

[IV) Comparative Analysis 11](#_Toc201708700)

[1) Cross-Dataset Performance Comparison 11](#_Toc201708701)

[2) Impact of Dataset Characteristics 11](#_Toc201708702)

[3) Decision Tree Behavior Analysis 11](#_Toc201708703)

[4) Practical Implications 12](#_Toc201708704)

[V) Technical Implementation Details 12](#_Toc201708705)

[1) Code Structure and Organization 12](#_Toc201708706)

[2) Libraries and Dependencies 12](#_Toc201708707)

[3) Reproducibility 13](#_Toc201708708)

[VI) Conclusions and Future Work 13](#_Toc201708709)

[1) Key Findings Summary 13](#_Toc201708710)

[2) Limitations 13](#_Toc201708711)

[3) Future Research Directions 13](#_Toc201708712)

[VII) References 13](#_Toc201708713)

[VIII) Appendices 13](#_Toc201708714)

[1) Appendix A: AI Tools Declaration 14](#_Toc201708715)

[2) Appendix B: Additional Visualizations 14](#_Toc201708716)

[3) Appendix C: Code Snippets 14](#_Toc201708717)

[4) Appendix D: Data Dictionary 14](#_Toc201708718)

# Team Information

## Member Details

|  |  |  |
| --- | --- | --- |
| **Student ID** | **Full Name** | **Role** |
| 22120037 | Nguyen Van Chien | Heart Disease Dataset Analysis |
| 22120144 | Ma Cat Huynh | Palmer Penguins Dataset Analysis |
| 22120149 | Nguyen Phan Duc Khai | Additional Dataset Analysis |
| 22120158 | Nguyen Van Khanh | Project Coordinator & Comparative Analysis |

## Work Assignment & Completion Rate

|  |  |  |
| --- | --- | --- |
| **Member** | **Assigned Tasks** | **Completion Rate** |
| Nguyen Van Chien | Heart Disease dataset (Tasks 2.1-2.4) | 100% |
| Nguyen Van Khanh | Palmer Penguins dataset (Tasks 2.1-2.4) | 100% |
| Ma Cat Huynh | Additional dataset selection & analysis (Tasks 2.1-2.4) | 100% |
| Nguyen Phan Duc Khai | Coordination, quality control, comparative analysis (Task 2.5) | 100% |

# Executive Summary

## Project Overview

[Brief description of the project objectives and scope]

## Key Findings

[High-level summary of main results across all three datasets]

## Self-Evaluation

* Overall Project Completion Rate: [XX%]
* Major Achievements: [List key accomplishments]
* Challenges Faced: [Describe main difficulties encountered]
* Areas for Improvement: [Self-reflection on project execution]

# Dataset Analysis

## Heart Disease Dataset Analysis

### Dataset Description

* **Dataset**: UCI Heart Disease Dataset
* **Source**: UCI Machine Learning Repository
* **Samples**: 303 patients
* **Features**: 13 medical indicators + 1 target variable
* **Target**: Binary classification (0: No Disease, 1: Disease)
* **Domain**: Medical diagnosis and cardiovascular health assessment
* **Original Class Distribution:**
  + Class 0 (No Disease): 160 samples (53.9%)
  + Class 1 (Disease): 137 samples (46.1%)
  + Class Balance Ratio: 1.17:1 (relatively balanced dataset)
* **Feature Description:** The dataset contains 13 medical features that are critical indicators for heart disease diagnosis:
  + age: Patient age in years (29-77 years)
  + sex: Gender (1=Male, 0=Female)
  + cp: Chest pain type (1=Typical angina, 2=Atypical angina, 3=Non-anginal pain, 4=Asymptomatic)
  + trestbps: Resting blood pressure in mm Hg (94-200 mm Hg)
  + chol: Serum cholesterol level in mg/dl (126-564 mg/dl)
  + fbs: Fasting blood sugar (1=>120 mg/dl, 0=≤120 mg/dl)
  + restecg: Resting electrocardiographic results (0=Normal, 1=ST-T abnormality, 2=Left ventricular hypertrophy)
  + thalach: Maximum heart rate achieved (71-202 bpm)
  + exang: Exercise induced angina (1=Yes, 0=No)
  + oldpeak: ST depression induced by exercise relative to rest (0-6.2)
  + slope: Peak exercise ST segment slope (1=Upsloping, 2=Flat, 3=Downsloping)
  + ca: Number of major vessels colored by fluoroscopy (0-3)
  + thal: Thallium stress test result (3=Normal, 6=Fixed defect, 7=Reversible defect)

### Data Preparation

**Preprocessing Steps Performed:**

* Missing Value Handling: Replaced '?' characters with NaN and removed 6 samples with missing data
* Data Type Conversion: Converted categorical features (ca, thal) to numeric format
* Target Variable Processing: Converted multi-class target (0-4) to binary classification (0=No Disease, 1=Disease)
* Final Dataset: 297 samples after cleaning (removed 6 samples with missing values)

**Stratified Train/Test Split Analysis:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Split Ratio** | **Training Samples** | **Test Samples** | **Class 0 (Train)** | **Class 1 (Train)** | **Class 0 (Test)** | **Class 1 (Test)** |
| 40/60 | 118 (39.7%) | 179 (60.3%) | 64 (54.2%) | 54 (45.8%) | 96 (53.6%) | 83 (46.4%) |
| 60/40 | 178 (59.9%) | 119 (40.1%) | 96 (53.9%) | 82 (46.1%) | 64 (53.8%) | 55 (46.2%) |
| 80/20 | 237 (79.8%) | 60 (20.2%) | 128 (54.0%) | 109 (46.0%) | 32 (53.3%) | 28 (46.7%) |
| 90/10 | 267 (89.9%) | 30 (10.1%) | 144 (53.9%) | 123 (46.1%) | 16 (53.3%) | 14 (46.7%) |
| Split Ratio | Training Samples | Test Samples | Class 0 (Train) | Class 1 (Train) | Class 0 (Test) | Class 1 (Test) |

**Class Distribution Analysis:**

* All splits maintain excellent stratification with class proportions consistently around 54% (No Disease) and 46% (Disease)
* The stratified splitting ensures representative samples in both training and testing sets
* Smaller test sets (90/10) may reduce statistical reliability of performance estimates

### Decision Tree Implementation

**Model Configuration:**

* Algorithm: DecisionTreeClassifier (scikit-learn)
* Criterion: Entropy (Information Gain)
* Random State: 42 (for reproducibility)
* Default Parameters: No initial constraints on depth, min\_samples\_split, etc.

**Decision Tree Characteristics by Split:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Split Ratio** | **Tree Depth** | **Number of Nodes** | **Training Accuracy** | **Interpretability** |
| 40/60 | 8 | 45 | ~95% | Moderate complexity |
| 60/40 | 9 | 57 | ~96% | Higher complexity |
| 80/20 | 10 | 67 | ~97% | Complex structure |
| 90/10 | 11 | 73 | ~98% | Very complex |

**Key Observations:**

* Trees become deeper and more complex with larger training sets
* Unrestricted trees show signs of potential overfitting with very high training accuracy
* Most important features identified: cp (chest pain type), thalach (max heart rate), ca (major vessels)

### Performance Evaluation

**Classification Reports and Confusion Matrices**

**40/60 Split Results:**

Classification Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| No Disease | 0.81 | 0.83 | 0.82 | 96 |
| Disease | 0.80 | 0.78 | 0.79 | 83 |
| accuracy |  |  | 0.81 | 179 |
| macro avg | 0.81 | 0.80 | 0.80 | 179 |
| weighted avg | 0.81 | 0.81 | 0.81 | 179 |

Confusion Matrix:

[[80 16]

[18 65]]

**60/40 Split Results:**

Classification Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| No Disease | 0.84 | 0.88 | 0.86 | 64 |
| Disease | 0.85 | 0.80 | 0.82 | 55 |
| accuracy |  |  | 0.84 | 119 |
| macro avg | 0.84 | 0.84 | 0.84 | 119 |
| weighted avg | 0.84 | 0.84 | 0.84 | 119 |

Confusion Matrix:

[[56 8]

[11 44]]

**80/20 Split Results:**

Classification Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| No Disease | 0.88 | 0.91 | 0.89 | 32 |
| Disease | 0.89 | 0.86 | 0.87 | 28 |
| accuracy |  |  | 0.88 | 60 |
| macro avg | 0.88 | 0.88 | 0.88 | 60 |
| weighted avg | 0.88 | 0.88 | 0.88 | 60 |

Confusion Matrix:

[[29 3]

[ 4 24]]

**90/10 Split Results:**

Classification Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| No Disease | 0.94 | 0.94 | 0.94 | 16 |
| Disease | 0.93 | 0.93 | 0.93 | 14 |
| accuracy |  |  | 0.93 | 30 |
| macro avg | 0.93 | 0.93 | 0.93 | 30 |
| weighted avg | 0.93 | 0.93 | 0.93 | 30 |

Confusion Matrix:

[15 1]

[ 1 13]]

**Performance Insights:**

* Accuracy Progression: Performance improves with larger training sets (81% → 84% → 88% → 93%)
* Precision Analysis:
  + No Disease class: Consistently high precision (81-94%), indicating low false positive rate
  + Disease class: Strong precision (80-93%), showing reliable positive predictions
  + Both classes maintain balanced precision across splits
* Recall Analysis:
  + No Disease class: Excellent recall (83-94%), effectively identifying healthy patients
  + Disease class: Good recall (78-93%), critical for not missing disease cases
  + Medical Significance: Higher recall for disease class is crucial to minimize false negatives
* F1-Score Balance:
  + Consistent F1-scores between classes indicate well-balanced model performance
  + No significant bias toward either class across different training sizes
* Training Set Size Impact:
  + Larger training sets (80/20, 90/10) show superior performance
  + However, 90/10 split may have inflated performance due to small test set (only 30 samples)
  + 80/20 split appears optimal for reliable performance estimation
* Medical Context Considerations:
  + False Negatives (FN): Missing disease cases - most critical error type
  + False Positives (FP): Incorrectly diagnosing healthy patients - less critical but still important
  + The model shows acceptable balance between sensitivity and specificity

### Depth Analysis

**Accuracy vs. Depth Analysis (80/20 Split)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **max\_depth** | **None** | **2** | **3** | **4** | **5** | **6** | **7** |
| Accuracy | 0.8833 | 0.8000 | 0.8167 | 0.8500 | 0.8667 | 0.8833 | 0.8833 |
| Tree Nodes | 67 | 7 | 11 | 17 | 23 | 35 | 51 |

**Visualizations**

* **Tree Complexity Analysis:**
  + **max\_depth=2:** Simple 7-node tree with basic decision rules
  + **max\_depth=3:** 11-node tree capturing primary diagnostic patterns
  + **max\_depth=4-5:** Moderate complexity (17-23 nodes) with good interpretability
  + **max\_depth=6-7:** Higher complexity (35-51 nodes) approaching full tree performance
  + **max\_depth=None:** Full complexity (67 nodes) with potential overfitting
* **Accuracy vs. Depth Trend:**
  + Initial performance at depth 2: 80.0% accuracy
  + Steady improvement from depth 2 to 6
  + Plateau effect at depth 6-7 and unlimited depth (88.33%)
  + Optimal performance achieved at moderate depths

**Depth Analysis Insights:**

**Key Findings:**

* Optimal Depth Identification:
  + Best Performance: max\_depth = 6, 7, or None (all achieve 88.33% accuracy)
  + Recommended Depth: max\_depth = 6 for optimal balance
  + Reasoning: Achieves maximum accuracy with 35 nodes vs. 67 nodes for unlimited depth
* Overfitting vs. Underfitting Analysis:
  + Underfitting (depth ≤ 3): Simplified models miss important diagnostic patterns
  + Optimal Range (depth 4-6): Good balance between complexity and generalization
  + Potential Overfitting (depth > 6): Diminishing returns with increased complexity
* Trade-off Between Complexity and Performance:
  + Interpretability: Depth 2-3 trees are easily interpretable by medical professionals
  + Performance: Depth 6 provides best accuracy while maintaining reasonable interpretability
  + Clinical Usability: Moderate depth trees (4-6) offer good compromise for medical decision support
* Generalization Capability:
  + Plateau in performance suggests the dataset's diagnostic patterns are captured by depth 6
  + Further increases in depth add noise rather than meaningful patterns
  + Model stability achieved at moderate depths indicates good generalization potential
* Medical Decision Support Implications:
  + Depth 3-4: Suitable for basic screening tools with high interpretability
  + Depth 5-6: Optimal for comprehensive diagnostic support systems
  + Depth > 6: May be too complex for clinical interpretation without sacrificing performance

**Recommended Configuration:** Based on the analysis, **max\_depth = 6** provides the optimal configuration for heart disease prediction, offering maximum accuracy (88.33%) while maintaining clinical interpretability with 35 decision nodes

## Palmer Penguins Dataset Analysis

### Dataset Description

* **Dataset**: Palmer Penguins Dataset
* **Source**: Palmer Station Long Term Ecological Research (LTER) Program
* **Samples**: 344 penguins
* **Features**: 7 biological and environmental variables + 1 target variable
* **Target**: Multi-class classification (3 penguin species)
* **Domain**: Ecological research and species classification
* **Original Class Distribution**:
  + Adelie: 152 samples (44.2%)
  + Gentoo: 124 samples (36.0%)
  + Chinstrap: 68 samples (19.8%)
  + Class Imbalance Ratio: 2.24:1 (Adelie to Chinstrap - moderate imbalance)
* **Feature Description:** The dataset contains 7 key features for penguin species classification:
  + **bill\_length\_mm**: Length of penguin's bill in millimeters (32.1-59.6 mm)
  + **bill\_depth\_mm**: Depth/height of penguin's bill in millimeters (13.1-21.5 mm)
  + **flipper\_length\_mm**: Length of penguin's flipper in millimeters (172-231 mm)
  + **body\_mass\_g**: Body mass in grams (2,700-6,300 g)
  + **sex**: Gender (Male/Female) - one-hot encoded to sex\_male
  + **island**: Breeding island location - one-hot encoded to:
  + **island\_Dream**: Dream Island indicator
  + **island\_Torgersen**: Torgersen Island indicator
  + **year**: Year of observation (2007-2009)
* **Ecological Context:**
  + **Adelie penguins**: Most widespread, found on all three islands
  + **Gentoo penguins**: Larger species, primarily on Biscoe Island
  + **Chinstrap penguins**: Smallest group, mainly on Dream Island

### Data Preparation

**Preprocessing Steps Performed:**

* **Missing Value Handling:**
  + Numerical features: 2 missing values in bill measurements imputed using IterativeImputer
  + Categorical features: 11 missing sex values filled with mode (most frequent value)
  + Final dataset: 344 samples retained (no samples removed)
* **One-Hot Encoding**: Applied to categorical variables (sex, island)
* **Feature Engineering**: Converted categorical features to binary indicators
* **Data Quality**: High-quality dataset with minimal missing data (3.2% overall)

**Stratified Train/Test Split Analysis:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Split Ratio** | **Training Samples** | **Test Samples** | **Adelie (Train)** | **Gentoo (Train)** | **Chinstrap (Train)** | **Adelie (Test)** | **Gentoo (Test)** | **Chinstrap (Test)** |
| 40/60 | 137 (39.8%) | 207 (60.2%) | 61 (44.5%) | 49 (35.8%) | 27 (19.7%) | 91 (44.0%) | 75 (36.2%) | 41 (19.8%) |
| 60/40 | 206 (59.9%) | 138 (40.1%) | 91 (44.2%) | 74 (35.9%) | 41 (19.9%) | 61 (44.2%) | 50 (36.2%) | 27 (19.6%) |
| 80/20 | 275 (79.9%) | 69 (20.1%) | 122 (44.4%) | 99 (36.0%) | 54 (19.6%) | 30 (43.5%) | 25 (36.2%) | 14 (20.3%) |
| 90/10 | 309 (89.9%) | 35 (10.1%) | 137 (44.3%) | 111 (35.9%) | 61 (19.7%) | 15 (42.9%) | 13 (37.1%) | 1. (20.0%) |

**Class Distribution Analysis:**

* Excellent stratification maintained across all splits
* Class proportions remain consistent: ~44% Adelie, ~36% Gentoo, ~20% Chinstrap
* Balanced representation ensures reliable model evaluation
* 90/10 split has very small test set (35 samples) which may limit evaluation reliability

### Decision Tree Implementation

**Model Configuration:**

* **Algorithm:** DecisionTreeClassifier (scikit-learn)
* **Criterion:** Entropy (Information Gain)
* **Random State:** 42 (for reproducibility)
* **Default Parameters:** No initial depth constraints

**Decision Tree Characteristics by Split:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Split Ratio** | **Tree Nodes** | **Training Accuracy** | **Test Accuracy** | **Model Complexity** |
| 40/60 | 13 | ~92% | 95.17% | Low complexity |
| 60/40 | 17 | ~94% | 97.83% | Moderate complexity |
| 80/20 | 25 | ~96% | 98.55% | Higher complexity |
| 90/10 | 25 | ~97% | 100.00% | High complexity |

**Key Decision Features Identified:**

* **flipper\_length\_mm**: Primary discriminator at root node (≤ 206.5 mm)
* **bill\_length\_mm**: Secondary split for species differentiation
* **bill\_depth\_mm**: Important for distinguishing Adelie from others
* **body\_mass\_g**: Helps separate Gentoo (larger) from other species
* **island\_Dream** and **island\_Torgersen**: Geographic indicators

### Performance Evaluation

**Classification Reports and Confusion Matrices**

**40/60 Split Results:**

Classification Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| Adelie | 0.97 | 0.96 | 0.96 | 91 |
| Chinstrap | 0.91 | 1.00 | 0.95 | 41 |
| Gentoo | 0.96 | 0.92 | 0.94 | 75 |
| accuracy |  |  | 0.95 | 0.95 |
| macro avg | 0.95 | 0.96 | 0.95 | 207 |
| weighted avg | 0.95 | 0.95 | 0.95 | 207 |

Confusion Matrix:

[[87 1 3]

[ 0 41 0]

[ 3 3 69]]

**60/40 Split Results:**

Classification Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| Adelie | 1.00 | 0.95 | 0.97 | 61 |
| Chinstrap | 0.93 | 1.00 | 0.96 | 27 |
| Gentoo | 0.98 | 1.00 | 0.99 | 50 |
| accuracy |  |  | 0.98 | 138 |
| macro avg | 0.97 | 0.98 | 0.98 | 138 |
| weighted avg | 0.98 | 0.98 | 0.98 | 138 |

Confusion Matrix:

[[58 2 1]

[ 0 27 0]

[ 0 0 50]]

**80/20 Split Results:**

Classification Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| Adelie | 1.00 | 0.97 | 0.98 | 30 |
| Chinstrap | 1.00 | 1.00 | 1.00 | 14 |
| Gentoo | 0.96 | 1.00 | 0.98 | 25 |
| accuracy |  |  | 0.99 | 69 |
| macro avg | 0.99 | 0.99 | 0.99 | 69 |
| weighted avg | 0.99 | 0.99 | 0.99 | 69 |

Confusion Matrix:

[[29 0 1]

[ 0 14 0]

[ 0 0 25]]

**90/10 Split:**

Classification Report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| Adelie | 1.00 | 1.00 | 1.00 | 15 |
| Chinstrap | 1.00 | 1.00 | 1.00 | 7 |
| Gentoo | 1.00 | 1.00 | 1.00 | 13 |
| accuracy |  |  | 1.00 | 35 |
| macro avg | 1.00 | 1.00 | 1.00 | 35 |
| weighted avg | 1.00 | 1.00 | 1.00 | 35 |

Confusion Matrix:

[[15 0 0]

[ 0 7 0]

[ 0 0 13]]

**Performance Insights:**

* Exceptional Overall Performance: Accuracy ranges from 95.17% to 100% across all splits
* Species-Specific Analysis:
  + Adelie penguins: Consistently high precision (97-100%) and recall (95-100%)
  + Chinstrap penguins: Perfect recall (100%) across all splits, excellent precision (91-100%)
  + Gentoo penguins: Strong performance with precision (96-100%) and recall (92-100%)
* Multi-Class Classification Excellence:
  + Balanced performance across all three species
  + No significant bias toward majority class (Adelie)
  + Minority class (Chinstrap) performs exceptionally well
* Training Set Size Impact:
  + Clear performance improvement with larger training sets: 95.17% → 97.83% → 98.55% → 100%
  + 90/10 split shows perfect performance (potentially due to small test set)
  + 80/20 split provides optimal balance of performance and evaluation reliability
* Confusion Matrix Analysis:
  + Main confusion occurs between Adelie and Gentoo species
  + Chinstrap penguins are most easily distinguishable (perfect recall)
  + Geographic and morphological features provide strong discriminative power
* Ecological Implications:
  + Physical measurements (bill, flipper, body mass) are highly effective for species identification
  + Island location provides additional discriminative information
  + Results support known biological differences between penguin species

### Depth Analysis

**Accuracy vs. Depth Analysis (80/20 Split)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **max\_depth** | **None** | **2** | **3** | **4** | **5** | **6** | **7** |
| Accuracy | 0.9855 | 0.9710 | 0.9710 | 0.9855 | 0.9855 | 0.9855 | 0.9855 |
| Tree Nodes | 25 | 7 | 11 | 17 | 23 | 25 | 25 |

**Visualizations**

* **Tree Complexity Progression:**
  + **max\_depth=2:** Simple 7-node tree with basic flipper/bill length splits
  + **max\_depth=3:** 11-node tree capturing primary species patterns
  + **max\_depth=4-7:** Optimal complexity (17-25 nodes) achieving maximum performance
  + **max\_depth=None:** Full tree with 25 nodes - same performance as constrained versions
* **Accuracy vs. Depth Pattern:**
  + Performance dip at depths 2-3 (97.10% accuracy)
  + Sharp improvement at depth 4, reaching optimal performance (98.55%)
  + Plateau effect from depth 4 onwards - no overfitting observed
  + Stable performance indicates well-separated classes

**Depth Analysis Insights:**

* Optimal Depth Identification:
  + Best Performance: max\_depth ≥ 4 (all achieve 98.55% accuracy)
  + Recommended Depth: max\_depth = 4 for optimal simplicity
  + Reasoning: Achieves maximum accuracy with only 17 nodes vs. 25 for unlimited depth
* Overfitting vs. Underfitting Analysis:
  + Underfitting (depth ≤ 3): Simplified models miss important species distinctions
  + Optimal Range (depth ≥ 4): Captures all necessary biological patterns
  + No Overfitting Observed: Unlimited depth doesn't degrade performance
* Biological Pattern Capture:
  + Depth 2-3: Captures basic size differences (Gentoo vs. others)
  + Depth 4+: Successfully models complex morphological relationships
  + Species Separability: Well-defined species boundaries require minimal tree depth
* Ecological Classification Implications:
  + Simple Rules Suffice: 4-level decision tree captures species differences
  + Interpretability: Moderate depth maintains biological interpretability
  + Robustness: Consistent performance suggests stable morphological patterns
* Practical Applications:
  + Field Research: Depth-4 trees suitable for field identification guides
  + Automated Classification: Higher depths acceptable for automated systems
  + Educational Use: Shallow trees (depth 2-3) useful for teaching species differences
* **Recommended Configuration:** Based on the analysis, **max\_depth = 4** provides the optimal configuration for Palmer Penguins classification, achieving maximum accuracy (98.55%) while maintaining interpretability with only 17 decision nodes.
* **Decision Tree Structure (Unlimited Depth):** The final tree uses a hierarchical approach:
  + **Primary split:** flipper\_length\_mm ≤ 206.5 (separates Adelie from larger species)
  + **Secondary splits:** bill\_length\_mm and bill\_depth\_mm (refine Adelie classification)
  + **Tertiary splits:** island location and body\_mass\_g (distinguish Gentoo from Chinstrap)
  + **Final classification:** Combination of morphological and geographical features
* This structure reflects the natural biological hierarchy where body size (flipper length) provides the primary species distinction, followed by bill morphology and geographic distribution patterns.

## Additional Dataset Analysis

### Dataset Selection and Description

**Dataset:** [Selected dataset name]

* **Source**: [Dataset source/URL]
* **Justification**: [Why this dataset was chosen]
* **Samples**: [Number ≥ 300]
* **Features**: [Number and detailed description]
* **Target**: [Classification type and classes]
* **Class Distribution**: [Original distribution]
* **Domain**: [Application domain/context]

### Data Preparation

**Preprocessing Steps:**

* One-hot encoding for categorical features: [List categorical features]
* [Other preprocessing steps]

**Train/Test Split Analysis:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Split Ratio | Training Samples | Test Samples | Adelie (Train) | Chinstrap (Train) | Gentoo (Train) | Adelie (Test) | Chinstrap (Test) | Gentoo (Test) |
| 40/60 |  |  |  |  |  |  |  |  |
| 60/40 |  |  |  |  |  |  |  |  |
| 80/20 |  |  |  |  |  |  |  |  |
| 90/10 |  |  |  |  |  |  |  |  |

### Decision Tree Implementation

**Model Configuration:**

* Algorithm: DecisionTreeClassifier (scikit-learn)
* Criterion: Information Gain (Entropy)
* Default parameters used for initial implementation

**Decision Tree Visualizations:** [Placeholder for Graphviz decision tree visualizations for each split ratio]

### Performance Evaluation

**Accuracy vs. Depth Analysis (80/20 Split)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| max\_depth | **None** | **2** | **3** | **4** | **5** | **6** | **7** |
| **Accuracy** |  |  |  |  |  |  |  |

**Visualizations**

[Placeholder for decision tree visualizations at different depths] [Placeholder for accuracy vs. depth chart]

**Depth Analysis Insights:**

* [Optimal depth identification]
* [Overfitting vs. underfitting analysis]
* [Trade-off between model complexity and performance]
* [Generalization capability discussion]

### Depth Analysis

**Accuracy vs. Depth Analysis (80/20 Split)**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| max\_depth | **None** | **2** | **3** | **4** | **5** | **6** | **7** |
| **Accuracy** |  |  |  |  |  |  |  |

**Visualizations**

[Placeholder for decision tree visualizations at different depths] [Placeholder for accuracy vs. depth chart]

**Depth Analysis Insights:**

* [Optimal depth identification]
* [Overfitting vs. underfitting analysis]
* [Trade-off between model complexity and performance]
* [Generalization capability discussion]

# Comparative Analysis

## Cross-Dataset Performance Comparison

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Samples** | **Features** | **Classes** | **Best Accuracy** | **Optimal Depth** | **Training Time** |
| Heart Disease |  |  |  |  |  |  |
| Palmer Penguins |  |  |  |  |  |  |
| [Additional Dataset] |  |  |  |  |  |  |

## Impact of Dataset Characteristics

**Sample Size Effect:**

* [Analysis of how sample size affects model performance]
* [Comparison of performance stability across different splits]

**Feature Complexity Effect:**

* [Impact of number of features on decision tree structure]
* [Categorical vs. numerical feature handling]

**Class Complexity Effect:**

* [Binary vs. multi-class classification performance]
* [Class imbalance impact across datasets]

## Decision Tree Behavior Analysis

**Depth Patterns:**

* [Common patterns in optimal depth across datasets]
* [Overfitting tendencies comparison]

**Feature Importance:**

* [Top features identified by each model]
* [Feature selection insights]

## Practical Implications

**Model Selection Recommendations:**

* [Guidelines for choosing optimal parameters based on dataset characteristics]
* [Trade-offs between interpretability and performance]

# Technical Implementation Details

## Code Structure and Organization

**Repository Structure:**

|  |
| --- |
| project/  ├── notebooks/  │ ├── heart\_disease\_analysis.ipynb  │ ├── palmer\_penguins\_analysis.ipynb  │ ├── additional\_dataset\_analysis.ipynb  │ └── comparative\_analysis.ipynb  ├── data/  │ ├── heart\_disease.csv  │ ├── palmer\_penguins.csv  │ └── [additional\_dataset].csv  ├── src/  │ ├── data\_preprocessing.py  │ ├── model\_training.py  │ └── visualization.py  └── reports/  └── final\_report.pdf |

## Libraries and Dependencies

* pandas: [version] - Data manipulation
* scikit-learn: [version] - Machine learning implementation
* matplotlib/seaborn: [version] - Data visualization
* graphviz: [version] - Decision tree visualization
* numpy: [version] - Numerical computations

## Reproducibility

**Random State Settings:** [Seed values used for reproducible results]  
**Environment:** [Python version and key library versions]

# Conclusions and Future Work

## Key Findings Summary

* [Main insights from decision tree analysis across all datasets]
* [Performance patterns identified]
* [Optimal configuration recommendations]

## Limitations

* [Dataset limitations encountered]
* [Model limitations observed]
* [Analysis constraints]

## Future Research Directions

* [Potential improvements to the analysis]
* [Additional algorithms to compare]
* [Extended dataset analysis possibilities]

# References

[1] [UCI Heart Disease Dataset reference]

[2] [Palmer Penguins Dataset reference]

[3] [Additional dataset reference]

[4] [Scikit-learn documentation]

[5] [Other relevant references]

# Appendices

## Appendix A: AI Tools Declaration

Tools Used: [List any AI tools used (ChatGPT, etc.)]

Prompts Used:

[Specific prompts used for each AI tool interaction]

[Context and purpose of AI assistance]

## Appendix B: Additional Visualizations

[Any additional charts, graphs, or visualizations not included in main sections]

## Appendix C: Code Snippets

[Key code snippets for critical functions if needed]

## Appendix D: Data Dictionary

[Detailed feature descriptions for all datasets]