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

**UNIVERSITY OF SCIENCE**

**FACULTY OF INFORMATION TECHNOLOGY**

**---ab---**

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***

[I) Team Information 4](#_Toc201868726)

[1) Member Details 4](#_Toc201868727)

[2) Work Assignment & Completion Rate 4](#_Toc201868728)

[II) Executive Summary 4](#_Toc201868729)

[1) Project Overview 4](#_Toc201868730)

[2) Key Findings 4](#_Toc201868731)

[3) Self-Evaluation 4](#_Toc201868732)

[III) Dataset Analysis 4](#_Toc201868733)

[1) Heart Disease Dataset Analysis 5](#_Toc201868734)

[2) Palmer Penguins Dataset Analysis 11](#_Toc201868735)

[3) Additional Dataset Analysis 18](#_Toc201868736)

[IV) Comparative Analysis 25](#_Toc201868737)

[1) Cross-Dataset Performance Comparison 25](#_Toc201868738)

[2) Impact of Dataset Characteristics 25](#_Toc201868739)

[3) Decision Tree Behavior Analysis 26](#_Toc201868740)

[4) Practical Implications 27](#_Toc201868741)

[V) Technical Implementation Details 28](#_Toc201868742)

[1) Code Structure and Organization 28](#_Toc201868743)

[2) Libraries and Dependencies 30](#_Toc201868744)

[3) Reproducibility 31](#_Toc201868745)

[VI) References 32](#_Toc201868746)

[VII) Appendices 32](#_Toc201868747)

[1) Appendix A: AI Tools Declaration 32](#_Toc201868748)

[2) Appendix B: Additional Visualizations 33](#_Toc201868749)

[3) Appendix C: Code Snippets 33](#_Toc201868750)

[4) Appendix D: Data Dictionary 33](#_Toc201868751)

Contents

# Team Information

## Member Details

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

### 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]]

Insights:

**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]]

Insights:

**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]]

Insights:

**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]]

Insights:

### 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 |

**Depth Analysis Insights:**

* 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

## 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: UCI Dermatology Dataset (Erythemato-Squamous Diseases)
* Source: UCI Machine Learning Repository
* Samples: 366 patients (after preprocessing: 366 samples retained)
* Features: 34 medical indicators + 1 target variable (after one-hot encoding: 109 features)
* Target: Multi-class classification (6 disease types)
* Domain: Medical diagnosis and dermatological disease classification
* Original Class Distribution:
  + Class 1 (Psoriasis): 112 samples (30.6%)
  + Class 2 (Seborrheic dermatitis): 61 samples (16.7%)
  + Class 3 (Lichen planus): 72 samples (19.7%)
  + Class 4 (Pityriasis rosea): 49 samples (13.4%)
  + Class 5 (Chronic dermatitis): 52 samples (14.2%)
  + Class 6 (Pityriasis rubra pilaris): 20 samples (5.5%)
  + Class Imbalance Ratio: 5.6:1 (Psoriasis to Pityriasis rubra pilaris - significant imbalance)
* Feature Categories:
  + Clinical Attributes (12 features): erythema, scaling, definite borders, itching, koebner phenomenon, polygonal papules, follicular papules, oral mucosal involvement, knee and elbow involvement, scalp involvement, family history, age
  + Histopathological Attributes (22 features): melanin incontinence, eosinophils infiltrate, PNL infiltrate, fibrosis papillary dermis, exocytosis, acanthosis, hyperkeratosis, parakeratosis, clubbing rete ridges, elongation rete ridges, thinning suprapapillary epidermis, spongiform pustule, munro microabcess, focal hypergranulosis, disappearance granular layer, vacuolisation damage, spongiosis, saw tooth appearance, follicular horn plug, perifollicular parakeratosis, inflammatory monoluclear infiltrate, band like infiltrate
* Medical Context:
  + Erythemato-squamous diseases share overlapping clinical and histopathological features
  + Accurate differential diagnosis is challenging due to feature similarity
  + Both clinical examination and histopathological analysis are crucial for diagnosis

### Data Preparation

**Preprocessing Steps Performed:**

* **Missing Value Handling**:
  + Age feature: Class-wise mode imputation for missing values (diseases often affect specific age groups)
  + No samples were removed; all 366 samples retained
* **One-Hot Encoding**: Applied to all categorical features (33 features) except age
  + Original 34 features expanded to 109 features after encoding
  + Used drop='first' to avoid multicollinearity
* **Feature Engineering**:
  + Age treated as continuous numerical variable
  + Categorical scales (0-3) properly encoded as binary indicators
* **Data Quality**: High-quality dataset with minimal missing data

**Stratified Train/Test Split Analysis:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Split Ratio** | **Training Samples** | **Test Samples** | **Class 1 (Train)** | **Class 2 (Train)** | **Class 3 (Train)** | **Class 4 (Train)** | **Class 5 (Train)** | **Class 6 (Train)** |
| 40/60 | 146 (39.9%) | 220 (60.1%) | 45 (30.8%) | 24 (16.4%) | 29 (19.9%) | 19 (13.0%) | 21 (14.4%) | 8 (5.5%) |
| 60/40 | 219 (59.8%) | 147 (40.2%) | 67 (30.6%) | 37 (16.9%) | 43 (19.6%) | 29 (13.2%) | 31 (14.2%) | 12 (5.5%) |
| 80/20 | 292 (79.8%) | 74 (20.2%) | 89 (30.5%) | 49 (16.8%) | 57 (19.5%) | 39 (13.4%) | 42 (14.4%) | 16 (5.5%) |
| 90/10 | 329 (89.9%) | 37 (10.1%) | 100 (30.4%) | 55 (16.7%) | 65 (19.8%) | 44 (13.4%) | 47 (14.3%) | 18 (5.5%) |

**Class Distribution Analysis**:

* Excellent stratification maintained across all splits
* Class proportions remain consistent: ~31% Psoriasis, ~17% Seborrheic dermatitis, ~20% Lichen planus, ~13% Pityriasis rosea, ~14% Chronic dermatitis, ~5% Pityriasis rubra pilaris
* Significant class imbalance with Class 6 being the minority class
* 90/10 split has very small test set (37 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 Depth** | **Number of Nodes** | **Training Accuracy** | **Test Accuracy** | **Model Complexity** |
| 40/60 | Complex | ~150+ nodes | ~95% | 80% | High complexity |
| 60/40 | Complex | ~120+ nodes | ~96% | 81% | High complexity |
| 80/20 | Complex | ~100+ nodes | ~97% | 92% | Moderate complexity |
| 90/10 | Complex | ~80+ nodes | ~98% | 89% | Moderate complexity |

**Key Decision Features Identified**:

* **elongation\_rete\_ridges**: Primary discriminator at root node
* **vacuolisation\_damage**: Major secondary split
* **fibrosis\_papillary\_dermis**: Important for Class 1 (Psoriasis) identification
* **koebner\_phenomenon**: Distinguishes between multiple classes
* **age**: Helpful for age-related disease patterns
* **oral\_mucosal\_involvement**: Specific diagnostic indicator

### Performance Evaluation

**Classification Reports and Confusion Matrices**

**40/60 Split Results:**

Classification Report for (40/60 split):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 1 | 0.89 | 0.82 | 0.85 | 67 |
| 2 | 0.56 | 0.89 | 0.69 | 37 |
| 3 | 1.00 | 0.91 | 0.95 | 43 |
| 4 | 0.92 | 0.80 | 0.86 | 30 |
| 5 | 0.68 | 0.61 | 0.64 | 31 |
| 6 | 1.00 | 0.50 | 0.67 | 12 |
| accuracy |  |  | 0.80 | 220 |
| macro avg | 0.84 | 0.76 | 0.78 | 220 |
| weighted avg | 0.84 | 0.80 | 0.81 | 220 |

Confusion Matrix (40/60):

[[55 7 0 0 5 0]

[ 2 33 0 0 2 0]

[ 0 2 39 1 1 0]

[ 0 5 0 24 1 0]

[ 5 6 0 1 19 0]

[ 0 6 0 0 0 6]]

**60/40 Split Results:**

Classification Report for (60/40 split):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 1 | 0.98 | 0.96 | 0.97 | 45 |
| 2 | 0.67 | 0.75 | 0.71 | 24 |
| 3 | 0.85 | 0.79 | 0.82 | 29 |
| 4 | 0.68 | 0.65 | 0.67 | 20 |
| 5 | 0.64 | 0.67 | 0.65 | 21 |
| 6 | 1.00 | 1.00 | 1.00 | 8 |
| accuracy |  |  | 0.81 | 147 |
| macro avg | 0.80 | 0.80 | 0.80 | 147 |
| weighted avg | 0.81 | 0.81 | 0.81 | 147 |

**80/20 Split Results:**

Classification Report for (80/20 split):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 1 | 0.96 | 0.96 | 0.96 | 23 |
| 2 | 0.75 | 1.00 | 0.86 | 12 |
| 3 | 1.00 | 0.80 | 0.89 | 15 |
| 4 | 0.90 | 0.90 | 0.90 | 10 |
| 5 | 1.00 | 0.90 | 0.95 | 10 |
| 6 | 1.00 | 1.00 | 1.00 | 4 |
| accuracy |  |  | 0.92 | 74 |
| macro avg | 0.93 | 0.93 | 0.92 | 74 |
| weighted avg | 0.93 | 0.92 | 0.92 | 74 |

**90/10 Split Results:**

Classification Report for (90/10 split):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 1 | 1.00 | 0.92 | 0.96 | 12 |
| 2 | 0.83 | 0.83 | 0.83 | 6 |
| 3 | 1.00 | 1.00 | 1.00 | 7 |
| 4 | 0.80 | 0.80 | 0.80 | 5 |
| 5 | 0.67 | 0.80 | 0.73 | 5 |
| 6 | 1.00 | 1.00 | 1.00 | 2 |
| accuracy |  |  | 0.89 | 37 |
| macro avg | 0.88 | 0.89 | 0.89 | 37 |
| weighted avg | 0.90 | 0.89 | 0.89 | 37 |

**Performance Insights:**

* **Accuracy Progression**: Clear improvement with larger training sets (80% → 81% → 92% → 89%)
* **Class-Specific Analysis**:
  + **Class 1 (Psoriasis)**: Consistently excellent performance (89-100% precision, 82-96% recall)
  + **Class 2 (Seborrheic dermatitis)**: Problematic class with frequent misclassifications
  + **Class 3 (Lichen planus)**: Strong performance (85-100% precision, 79-100% recall)
  + **Class 4 (Pityriasis rosea)**: Moderate performance (68-92% precision, 65-90% recall)
  + **Class 5 (Chronic dermatitis)**: Variable performance (64-100% precision, 61-90% recall)
  + **Class 6 (Pityriasis rubra pilaris)**: Perfect performance in larger splits despite being minority class
* **Training Set Size Impact**:
  + 80/20 split shows optimal balance of performance and reliability (92% accuracy)
  + 90/10 split may have inflated performance due to very small test set (37 samples)
  + Larger training sets significantly improve model generalization
* **Multi-Class Classification Challenges**:
  + Class 2 acts as a "confusion hub" - frequently receives misclassifications from other classes
  + Class 6 (minority class) performs surprisingly well, indicating distinct feature patterns
  + Feature overlap between Classes 2, 4, and 5 causes most misclassifications

### Depth Analysis

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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **max\_depth** | **None** | **2** | **3** | **4** | **5** | **6** | **7** |
| **Accuracy** | 0.9189 | 0.4730 | 0.4730 | 0.6216 | 0.6216 | 0.8108 | 0.8649 |

**Depth Analysis Insights:**

* Optimal Depth Identification:
  + Best Performance: max\_depth = None (unlimited) achieves 91.89% accuracy
  + Recommended Depth: max\_depth = 7 for good balance (86.49% accuracy)
  + Critical Depth: Significant improvement occurs between depths 5-6 (62% → 81%)
* Overfitting vs. Underfitting Analysis:
  + Severe Underfitting (depth ≤ 3): Only ~47% accuracy - inadequate for medical diagnosis
  + Continued Underfitting (depth 4-5): 62% accuracy plateau - insufficient complexity
  + Optimal Range (depth 6-7): Major performance jump to 81-87%
  + Full Complexity (unlimited depth): Best performance but potential overfitting risk
* Medical Diagnostic Implications:
  + Shallow Trees (depth ≤ 5): Unsuitable for clinical use due to poor accuracy
  + Moderate Depth (6-7): Clinically acceptable performance while maintaining interpretability
  + Deep Trees (unlimited): Maximum accuracy but may be too complex for clinical interpretation
* Complex Multi-Class Pattern Recognition:
  + The dataset requires significant tree depth due to:
  + 6 distinct disease classes with overlapping symptoms
  + 34 original features (109 after encoding) with complex interactions
  + Subtle histopathological differences requiring detailed decision boundaries

**Recommended Configuration**: Based on the analysis, max\_depth = 7 provides an optimal balance between clinical interpretability and diagnostic accuracy (86.49%), while unlimited depth (91.89%) offers maximum performance for automated diagnostic systems.

**Decision Tree Structure Insights**:

* Primary splits focus on histopathological features (elongation\_rete\_ridges, vacuolisation\_damage)
* Secondary splits incorporate clinical features (age, koebner\_phenomenon)
* Deep branches necessary to distinguish between similar disease presentations
* Complex feature interactions require substantial tree depth for accurate classification

# Comparative Analysis

## Cross-Dataset Performance Comparison

**Performance Summary Across All Datasets:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Samples** | **Features** | **Classes** | **Best Accuracy** | **Optimal Split** | **Training Time** | **Complexity** |
| Heart Disease | 297 | 13 | 2 | 93.00% | 90/10 | Low | Medium |
| Palmer Penguins | 344 | 8 | 3 | 100.00% | 90/10 | Low | Low |
| Dermatology | 366 | 109 | 6 | 92.00% | 80/20 | Medium | High |

**Key Performance Insights:**

* **Palmer Penguins** achieved the highest peak performance (100%) with perfect classification on the 90/10 split
* **Heart Disease** showed consistent improvement across splits (81% → 84% → 88% → 93%)
* **Dermatology** demonstrated the most challenging classification task with significant performance variation (80% → 92% → 89%)

**Training Set Size Impact:**

* All datasets benefit from larger training sets, with performance generally improving from 40/60 to 80/20 splits
* The 90/10 split shows inflated performance due to very small test sets (30-37 samples)
* Optimal performance-reliability balance achieved at 80/20 split for most datasets

## Impact of Dataset Characteristics

* **Sample Size Effect:**
  + **Correlation with Performance**: Larger datasets (Dermatology: 366 samples) don't necessarily yield better performance due to increased complexity
  + **Stability**: Heart Disease (297 samples) shows most stable performance improvement with training size
  + **Reliability**: Palmer Penguins (344 samples) demonstrates most consistent high performance across splits
* **Feature Complexity Effect:**
  + **Low Complexity** (Palmer Penguins: 8 features): Highest performance, minimal overfitting
  + **Medium Complexity** (Heart Disease: 13 features): Balanced performance with interpretable results
  + **High Complexity** (Dermatology: 109 features after encoding): Most challenging, requires deeper trees
* **Feature Utilization Rates:**
  + Heart Disease: 85% (11/13 features typically used)
  + Palmer Penguins: 75% (6/8 features typically used)
  + Dermatology: 45% (~50/109 features typically used)
* **Class Complexity Effect:**
  + **Binary Classification** (Heart Disease): Consistent performance, balanced precision/recall
  + **3-Class Multi-class** (Palmer Penguins): Excellent separation between species
  + **6-Class Multi-class** (Dermatology): Significant confusion between similar disease classes
* **Class Imbalance Impact:**
  + Low imbalance (Heart Disease: 1.17:1) → Stable performance
  + Moderate imbalance (Palmer Penguins: 2.24:1) → High performance maintained
  + High imbalance (Dermatology: 5.6:1) → Class-specific performance variation

## Decision Tree Behavior Analysis

**Depth Patterns:**

* **Optimal Depth Requirements:**
  + Heart Disease: max\_depth = 6 (balanced performance-interpretability)
  + Palmer Penguins: max\_depth = 4 (biological patterns well-captured)
  + Dermatology: max\_depth = 7+ (complex medical relationships require depth)
* **Overfitting Tendencies:**
  + **Low Risk**: Palmer Penguins (stable performance across depths)
  + **Medium Risk**: Heart Disease (plateau effect at depth 6-7)
  + **High Risk**: Dermatology (dramatic performance drop at shallow depths)
* **Depth Sensitivity Analysis:**
  + Heart Disease: 0.083 accuracy range (low sensitivity)
  + Palmer Penguins: 0.015 accuracy range (very low sensitivity)
  + Dermatology: 0.446 accuracy range (high sensitivity)

**Feature Importance:**

**Primary Discriminative Features:**

* **Heart Disease:**
  + cp (chest pain type): Primary cardiac symptom indicator
  + thalach (max heart rate): Exercise tolerance measure
  + ca (major vessels): Anatomical severity marker
* **Palmer Penguins:**
  + flipper\_length\_mm: Primary species separator (≤206.5mm threshold)
  + bill\_length\_mm & bill\_depth\_mm: Secondary morphological features
  + Geographic features (island location): Supporting classification
* **Dermatology:**
  + elongation\_rete\_ridges: Root-level histopathological feature
  + vacuolisation\_damage: Major tissue damage indicator
  + fibrosis\_papillary\_dermis: Structural skin change marker

## Practical Implications

**Model Selection Recommendations:**

* **For Medical Diagnosis Applications:**
  + **Heart Disease Model**: Suitable for clinical decision support with 88% accuracy at optimal depth 6
  + **Dermatology Model**: Requires expert validation due to complexity and class confusion
* **For Ecological Research:**
  + **Palmer Penguins Model**: Excellent for automated species identification with 98.55% accuracy

**Performance vs. Interpretability Trade-offs:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Recommended Depth** | **Accuracy** | **Interpretability** | **Clinical Utility** |
| Heart Disease | 6 | 88.33% | High | Suitable for screening |
| Palmer Penguins | 4 | 98.55% | Very High | Field identification |
| Dermatology | 7 | 86.49% | Medium | Requires expert review |

**Domain-Specific Insights:**

* Medical Domains (Heart Disease, Dermatology):
  + Require balance between accuracy and interpretability
  + False negative minimization crucial for disease detection
  + Feature selection should align with clinical knowledge
* Ecological Domain (Palmer Penguins):
  + High accuracy achievable with simple models
  + Geographic features provide valuable supplementary information
  + Morphological measurements sufficient for species distinction

# Technical Implementation Details

## Code Structure and Organization

**Repository Structure:**

|  |
| --- |
| AI-Decision-Tree/  ├── datasets/  │ ├── additional\_dataset/  │ ├── heart\_disease/  │ └── palmer\_penguins/  ├── docs/  │ ├── reports/  │ ├── prompts/  │ ├── AI - Project 2.pdf  ├── notebooks/  │ ├── additional\_dataset/  │ ├── comparative\_analysis/  │ ├── heart\_disease/  │ ├── palmer\_penguins/  ├── results/  │ ├── additional\_dataset/  │ ├── comparative\_analysis/  │ ├── heart\_disease/  │ └── palmer\_penguins/  ├── src/  ├── venv/  ├── .gitignore  ├── README.md  └── requirements.txt |

**Implementation Architecture:**

* **Data Processing Pipeline:**
  + **Data Loading**: Individual dataset-specific loaders with validation
  + **Preprocessing**: Stratified preprocessing with one-hot encoding for categorical features
  + **Split Generation**: Consistent stratified splitting (40/60, 60/40, 80/20, 90/10)
  + **Model Training**: DecisionTreeClassifier with entropy criterion
  + **Evaluation**: Comprehensive metrics collection and visualization
* **Code Organization Principles:**
  + **Modular Design**: Separate notebooks for each dataset analysis
  + **Reproducibility**: Consistent random seeds (random\_state=42) across all experiments
  + **Comparative Structure**: Standardized analysis template for fair comparison
    - **Version Control**: Git-based collaboration with clear commit structure

## Libraries and Dependencies

**Core Dependencies:**

|  |
| --- |
| numpy==1.26.4 # Numerical computations and array operations  pandas==2.3.0 # Data manipulation and analysis  scikit-learn==1.4.0 # Machine learning algorithms and metrics  matplotlib==3.8.2 # Static visualizations and plots  seaborn==0.13.0 # Statistical data visualization  graphviz==0.20.1 # Decision tree visualization  jupyter==1.0.0 # Interactive notebook environment  plotly==5.18.0 # Interactive visualizations |

**Technical Stack Rationale:**

* **Scikit-learn Selection:**
  + **DecisionTreeClassifier**: Standard implementation with entropy criterion
  + **train\_test\_split**: Stratified splitting for consistent class distribution
  + **Classification metrics**: Comprehensive evaluation suite (precision, recall, F1, confusion matrix)
* **Visualization Strategy:**
  + **Matplotlib/Seaborn**: Static plots for report documentation
  + **Graphviz**: Tree structure visualization for interpretability analysis
  + **Plotly**: Interactive plots for detailed exploration (future enhancement)

**Data Processing Tools:**

* + **Pandas**: DataFrame operations, missing value handling, one-hot encoding
  + **NumPy**: Numerical operations, array manipulations, statistical calculations

## Reproducibility

**Random State Management:**

* **Consistent Seeds Applied:**
  + Data splitting: random\_state=42
  + Decision tree training: random\_state=42
  + Data shuffling: random\_state=42

**Environment Specification:**

|  |
| --- |
| # Python Version  Python 3.8+  # Virtual Environment Setup  python -m venv venv  venv\Scripts\activate # Windows  source venv/bin/activate # macOS/Linux  # Dependency Installation  pip install -r requirements.txt  # Notebook Execution  jupyter notebook |

* **Data Reproducibility Measures:**
  + **Fixed preprocessing steps**: Identical feature engineering across datasets
  + **Stratified splitting**: Maintains class distribution consistency
  + **Parameter standardization**: Same DecisionTreeClassifier configuration
  + **Evaluation consistency**: Identical metrics calculation across all models
* **Quality Assurance:**
  + **Code review process**: Team member cross-validation of implementations
  + **Result validation**: Multiple execution verification for consistency
  + **Documentation standards**: Comprehensive inline comments and markdown explanations
  + **Version control**: Git workflow with feature branches and merge reviews
* **Performance Benchmarking:**
  + **Execution timing**: Consistent hardware environment for fair comparison
  + **Memory usage**: Monitoring for large datasets (Dermatology with 109 features)
  + **Scalability testing**: Verification across different train/test split ratios
* **Collaboration Framework:**
  + **Team coordination**: Clear task assignment and progress tracking
  + **Code integration**: Standardized notebook structure for easy merging
  + **Result aggregation**: Centralized comparison analysis for fair evaluation
  + **Knowledge sharing**: Regular team meetings and documentation updates

# 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]