**DECISION TREE FOR BREAST CANCER ANALYSIS - TASK 1**

|  |  |
| --- | --- |
| **NAME** | ABDUL MUIZZ, MUHAMMAD ABDULLAH |
| **REGISTRATION NUMBER** | FA23-BCS-117, FA23-BCS-108 |
| **DATE OF SUBMISSION** | 26TH SEPTEMBER 2025 |
| **MODERATOR** | SIR UMAR NAUMAN |

**Table of Contents:**

[**Abstract:** 3](#_Toc209687165)

[**Problem Statement:** 4](#_Toc209687166)

[**Dataset Description:** 5](#_Toc209687167)

[**Feature Categories:** 5](#_Toc209687168)

[**Data Split:** 5](#_Toc209687169)

[**Methodology:** 6](#_Toc209687170)

[**1. Data Preprocessing** 6](#_Toc209687171)

[**2. Exploratory Data Analysis** 6](#_Toc209687172)

[**3. Feature Selection Strategies** 6](#_Toc209687173)

[**4. Model Configuration** 6](#_Toc209687174)

[**Data Analysis and Visualization:** 7](#_Toc209687175)

[**Hypothesis Validation:** 7](#_Toc209687176)

[**Feature Correlation Analysis:** 7](#_Toc209687177)

[**Feature Selection Approaches:** 9](#_Toc209687178)

[**Approach 1: All Original Features (30 features)** 9](#_Toc209687179)

[**Approach 2: Reduced Correlation Set (3 features)** 9](#_Toc209687180)

[**Approach 3: PCA Transformation (5 components)** 9](#_Toc209687181)

[**Model Implementation:** 10](#_Toc209687182)

[**Decision Tree Configuration:** 10](#_Toc209687183)

[**Tree Structure Analysis:** 10](#_Toc209687184)

[**Results and Analysis:** 13](#_Toc209687185)

[**Performance Comparison:** 13](#_Toc209687186)

[**Key Findings:** 13](#_Toc209687187)

[**Feature Importance Analysis:** 13](#_Toc209687188)

[**Model Interpretability:** 13](#_Toc209687189)

[**Random Forest Comparison:** 15](#_Toc209687190)

[**Implementation:** 15](#_Toc209687191)

[**Results:** 15](#_Toc209687192)

[**Analysis:** 15](#_Toc209687193)

[**Random Forest Advantages:** 15](#_Toc209687194)

[**List of Figures:** 17](#_Toc209687195)

[**Conclusion:** 18](#_Toc209687196)

[**Key Findings:** 18](#_Toc209687197)

[**Medical Relevance:** 18](#_Toc209687198)

[**Limitations:** 18](#_Toc209687199)

[**Future Work:** 18](#_Toc209687200)

[**Practical Applications:** 19](#_Toc209687201)

[**References:** 20](#_Toc209687202)

## **Abstract:**

This project implements Decision Tree classification algorithms to predict breast cancer diagnosis using the Wisconsin Breast Cancer Dataset. The study compares three different feature selection approaches: using all original features, reducing correlated features, and applying Principal Component Analysis (PCA). A Random Forest classifier is also implemented for performance comparison. The original feature set achieved the highest test accuracy of 94% with Decision Trees, while Random Forest achieved 97% accuracy.

## **Problem Statement:**

Breast cancer is one of the most common cancers affecting women worldwide. Early and accurate diagnosis is crucial for effective treatment. This project aims to:

* Build a Decision Tree classifier to predict breast cancer diagnosis (malignant vs benign).
* Compare different feature selection techniques and their impact on model performance.
* Analyze which features are most important for cancer prediction.
* Evaluate Decision Trees against Random Forest for this classification task.
* Provide interpretable models that can assist in medical decision making.

## **Dataset Description:**

**Source:** sklearn.datasets - Breast Cancer Wisconsin Dataset  
**Total Samples:** 569 cases  
**Classes:** 2 (Malignant, Benign)  
**Features:** 30 numerical features

### **Feature Categories:**

All features are computed from digitized images of cell nuclei and include measurements for:

* **Radius:** Mean distances from center to points on perimeter
* **Texture:** Standard deviation of gray scale values
* **Perimeter:** Perimeter of cell nuclei
* **Area:** Area of cell nuclei
* **Smoothness:** Local variation in radius lengths
* **Compactness:** (perimeter²/area - 1.0)
* **Concavity:** Severity of concave portions of contour
* **Concave Points:** Number of concave portions of contour
* **Symmetry:** Symmetry of cell nuclei
* **Fractal Dimension:** "Coastline approximation" - 1

Each feature has three measurements: mean, standard error, and worst (largest) value.

### **Data Split:**

* **Training Set:** 80% (455 samples)
* **Test Set:** 20% (114 samples)

## **Methodology:**

### **1. Data Preprocessing**

* Loaded breast cancer dataset from sklearn
* Split data into training and testing sets (80-20 split)
* No missing values to handle
* Features were already normalized in the dataset

### **2. Exploratory Data Analysis**

* Generated hypothesis: Size and shape of cell nuclei are good predictors
* Created visualization to validate hypothesis
* Analyzed feature correlations

### **3. Feature Selection Strategies**

Three approaches were implemented:

1. **All Features:** Using complete 30-feature set
2. **Correlation-Based Reduction:** Manual selection of low-correlation features
3. **PCA Transformation:** Dimensionality reduction to 5 principal components

### **4. Model Configuration**

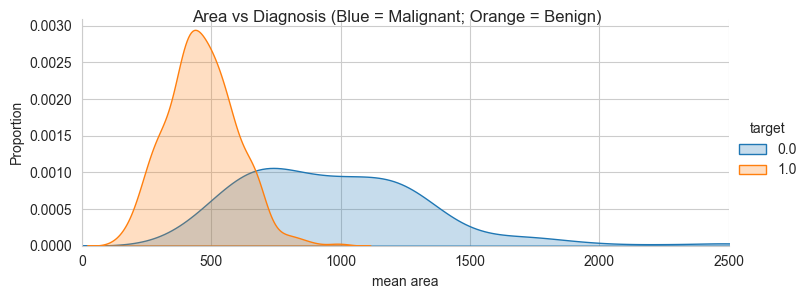
* **Algorithm:** DecisionTreeClassifier
* **Parameters:** max\_depth=3, min\_samples\_leaf=12
* **Evaluation:** Training and test accuracy comparison

## **Data Analysis and Visualization:**

### **Hypothesis Validation:**

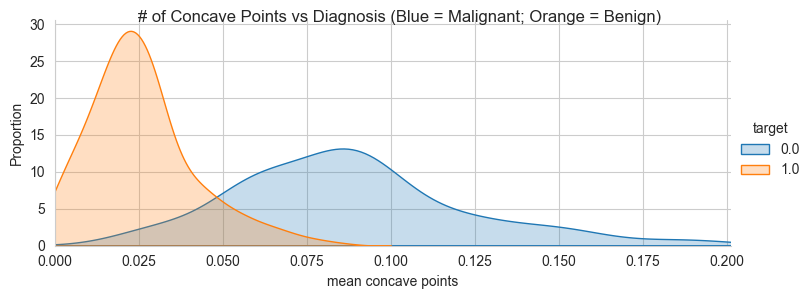
The initial hypothesis stated that cancerous cell nuclei are typically larger and more misshapen than benign ones. This was validated through:

1. **Mean Area Analysis:** KDE plots showed clear separation between malignant and benign tumors based on mean area measurements



**Figure 1:** Kernel Density Estimate plot of 'mean area' distribution for malignant vs benign tumors

1. **Concave Points Analysis:** Distribution plots revealed distinct patterns in mean concave points between the two classes.

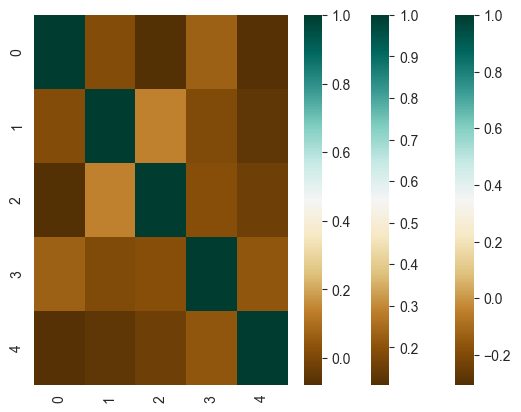


**Figure 2:** KDE plot of 'mean concave points' showing class separation

### **Feature Correlation Analysis:**

A correlation heatmap revealed high correlation among many features, particularly between:

* Area, perimeter, and radius measurements
* Different statistical measures (mean, error, worst) of the same feature
* Compactness and concavity measurements



**Figure 3:** Correlation heatmap of all 30 original features

## **Feature Selection Approaches:**

### **Approach 1: All Original Features (30 features)**

* **Advantage:** Complete information retention
* **Disadvantage:** High correlation between features
* **Result:** Potential overfitting due to redundant information

### **Approach 2: Reduced Correlation Set (3 features)**

Selected features with low inter-correlation:

* **mean texture**
* **mean area**
* **mean symmetry**

**Rationale:** These features represent different aspects of cell nuclei characteristics

### **Approach 3: PCA Transformation (5 components)**

* Applied Principal Component Analysis to reduce dimensionality
* Retained 5 principal components
* **Advantage:** Uncorrelated features while preserving most variance
* **Disadvantage:** Loss of feature interpretability

## **Model Implementation:**

### **Decision Tree Configuration:**

python

DecisionTreeClassifier(max\_depth=3, min\_samples\_leaf=12)

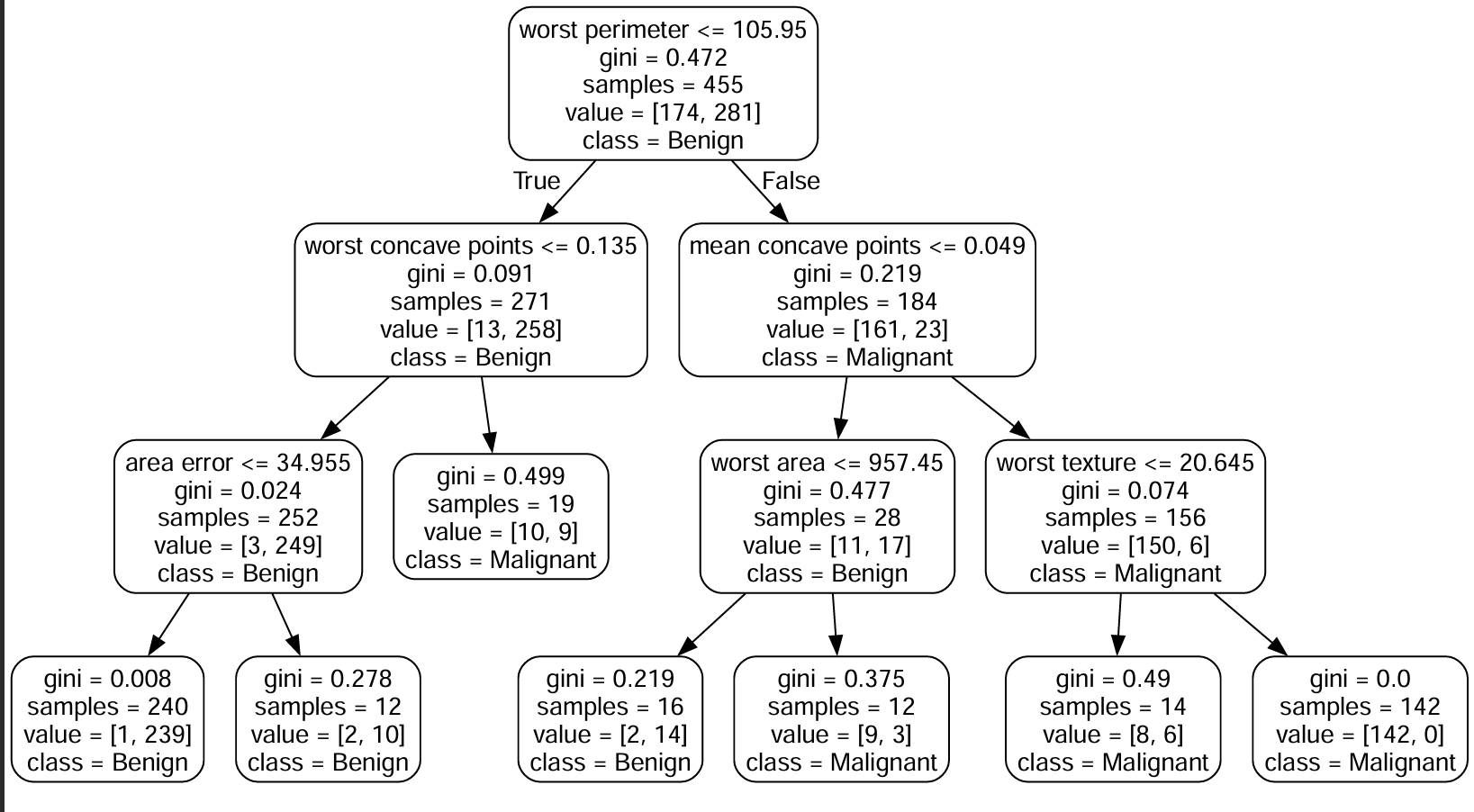
**Parameter Justification:**

* **max\_depth=3:** Prevents overfitting while maintaining interpretability
* **min\_samples\_leaf=12:** Ensures each leaf has sufficient samples for reliable predictions

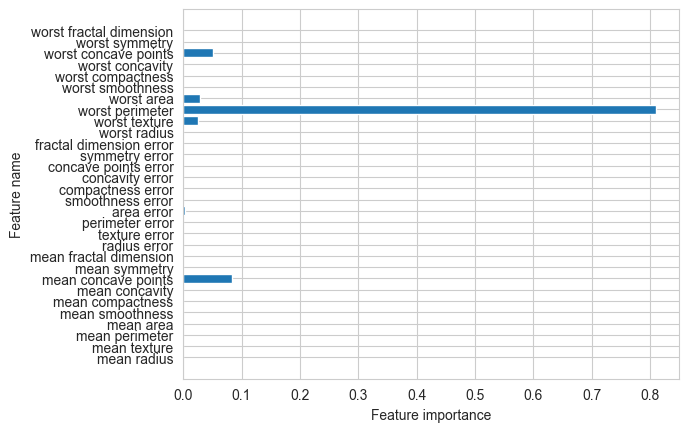
### **Tree Structure Analysis:**

Each decision tree was visualized to understand the decision-making process:

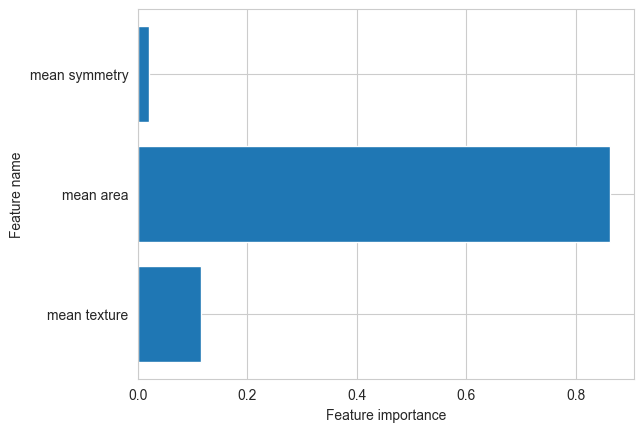
* **Tree 1 (All Features):** Focused on 'worst perimeter' and 'worst concave points'.
* **Tree 2 (Reduced Features):** Used 'mean area' and 'mean texture' as primary decision nodes.
* **Tree 3 (PCA Features):** Based on principal components (less interpretable).



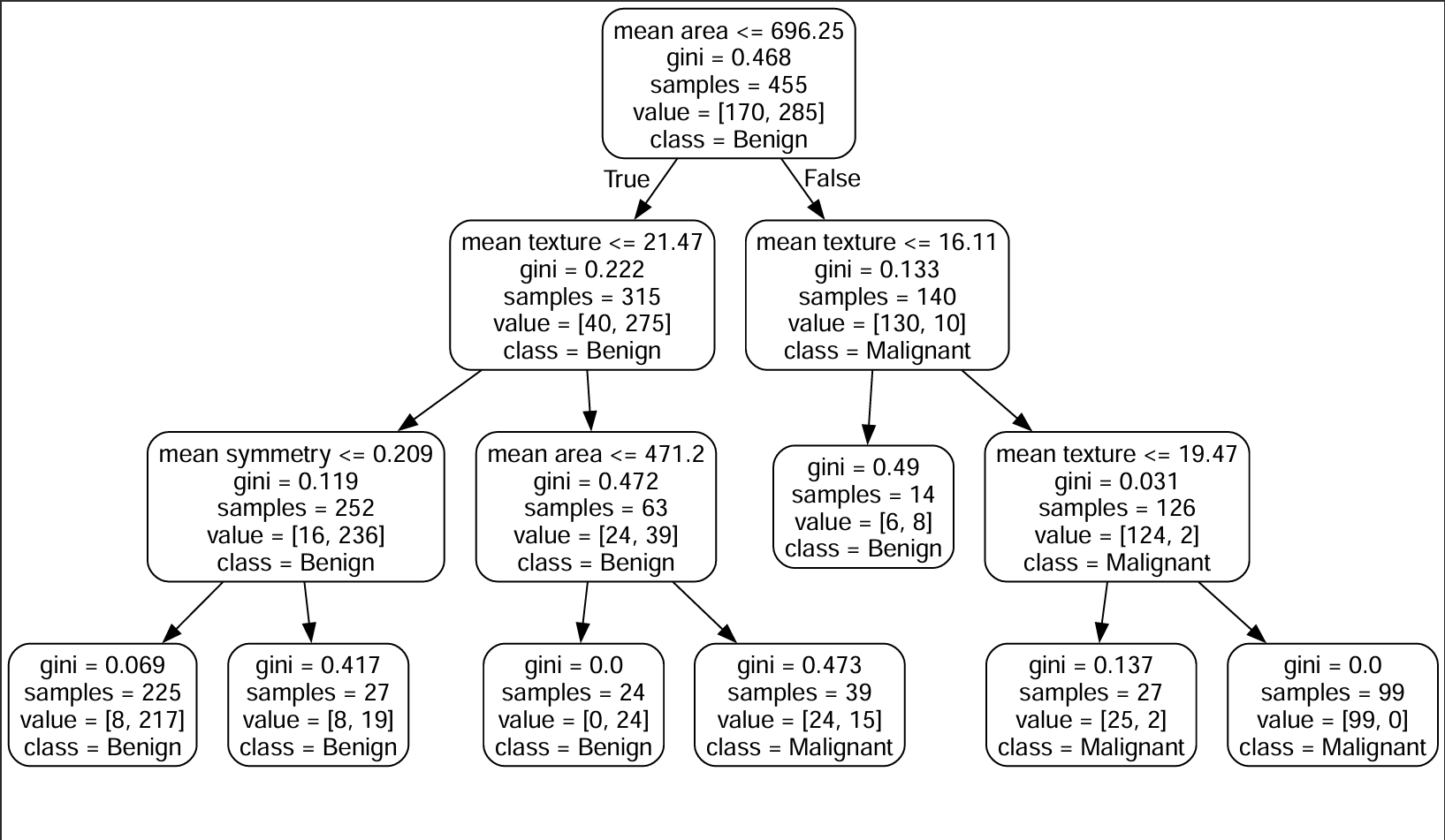
**Figure 4:** Decision Tree visualization using all original features



**Figure 5:** Feature importance plot for Decision Tree with all features.



**Figure 6:** Retain only features that were produced through PCA transformation.



**Figure 7:** Decision Tree visualization using reduced feature set (3 features)

## **Results and Analysis:**

### **Performance Comparison:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature Set** | **Training Accuracy** | **Test Accuracy** | **Number of Features** |
| **All Original Features** | 95% | **94%** | 30 |
| **Reduced Correlation** | 91% | 84% | 3 |
| **PCA Transformed** | 96% | 91% | 5 |

### **Key Findings:**

1. **Best Performance:** All original features achieved highest test accuracy (94%)
2. **Feature Importance:** 'worst perimeter' was identified as the most important feature
3. **Reduced Features:** Significant accuracy drop (84%) when using only 3 features
4. **PCA Results:** Good balance between dimensionality reduction and performance (91%)

### **Feature Importance Analysis:**

From the complete feature set model:

1. **worst perimeter**: Most important predictor
2. **worst concave points:** Second most important
3. **mean area**: Strong predictor in reduced feature set
4. **mean texture**: Consistent across different models

### **Model Interpretability:**

Decision trees provided clear, interpretable rules:

* If worst perimeter > threshold 🡪 Check worst concave points
* If mean area > threshold 🡪 Likely malignant
* Combination of size and shape features provides best discrimination

## **Random Forest Comparison:**

### **Implementation:**

* **Algorithm:** RandomForestClassifier
* **Feature Set:** All original 30 features
* **Evaluation:** Learning curves and confusion matrix analysis

### **Results:**

* **Training Accuracy:** 100%
* **Test Accuracy:** 97%

### **Analysis:**

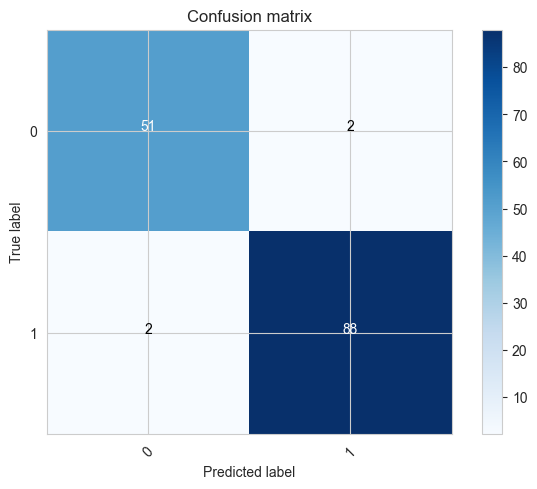
* **Learning Curve:** Small gap between training and cross-validation scores
* **Interpretation:** Low bias but potential variance (slight overfitting)
* **Confusion Matrix:** Minimal false positives and false negatives
* **Performance:** Superior to single Decision Tree (97% vs 94%)

### **Random Forest Advantages:**

1. **Higher Accuracy:** 3% improvement over single Decision Tree
2. **Robustness:** Better generalization through ensemble method
3. **Reduced Overfitting:** Averaging multiple trees reduces variance



**Figure 8:** Learning curve for Random Forest classifier



**Figure 9:** Confusion matrix for Random Forest test predictions

## **List of Figures:**

1. **Figure 1:** Kernel Density Estimate plot of 'mean area' distribution for malignant vs benign tumors. Page 8
2. **Figure 2:** KDE plot of 'mean concave points' showing class separation. Page 8
3. **Figure 3:** Correlation heatmap of all 30 original features. Page 9
4. **Figure 4:** Decision Tree visualization using all original features. Page 11
5. **Figure 5:** Feature importance plot for Decision Tree with all features. Page 12
6. **Figure 6:** Retain only features that were produced through PCA transformation. Page 12
7. **Figure 7:** Decision Tree visualization using reduced feature set (3 features). Page 13
8. **Figure 8:** Learning curve for Random Forest classifier. Page 17
9. **Figure 9:** Confusion matrix for Random Forest test predictions. Page 17

## **Conclusion:**

### **Key Findings:**

1. **Optimal Feature Set:** All 30 original features provided the best single Decision Tree performance (94%)
2. **Critical Features:** 'worst perimeter' and 'worst concave points' were most important for classification
3. **Feature Reduction Impact:** Aggressive feature reduction (30→3) caused 10% accuracy drop
4. **PCA Effectiveness:** Good compromise between dimensionality and performance
5. **Random Forest Superiority:** Ensemble method achieved 97% accuracy, outperforming single trees

### **Medical Relevance:**

* The model successfully validates the medical hypothesis that larger, more irregular cell nuclei indicate malignancy
* High accuracy (94-97%) makes it suitable for diagnostic assistance
* Interpretable features help medical professionals understand the reasoning

### **Limitations:**

1. **Small Dataset:** Only 569 samples may limit generalizability
2. **Feature Correlation:** High correlation among features may affect stability
3. **Binary Classification:** Only distinguishes malignant vs benign (no cancer staging)

### **Future Work:**

1. **Cross Validation:** Implement k fold cross validation for robust evaluation
2. **Feature Engineering:** Create new features combining existing measurements
3. **Advanced Algorithms:** Test Support Vector Machines and Neural Networks
4. **Clinical Validation:** Validate model performance on independent medical datasets
5. **Ensemble Methods:** Explore other ensemble techniques beyond Random Forest

### **Practical Applications:**

* **Diagnostic Support:** Assist pathologists in cancer diagnosis
* **Screening Programs:** Automated preliminary screening of cell samples
* **Research Tool:** Identify which cellular features are most predictive of malignancy
* **Educational Tool:** Help medical students understand cancer classification features

## **References:**

1. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine
2. Wolberg, W.H., Street, W.N., and Mangasarian, O.L. (1995). "Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates"
3. Breiman, L. (2001). "Random Forests". Machine Learning, 45(1), 5-32
4. Quinlan, J.R. (1986). "Induction of Decision Trees". Machine Learning, 1(1), 81-106
5. Scikit-learn Documentation: Decision Trees and Random Forest Classifiers