Decision Tree Analysis Report - Internship Project

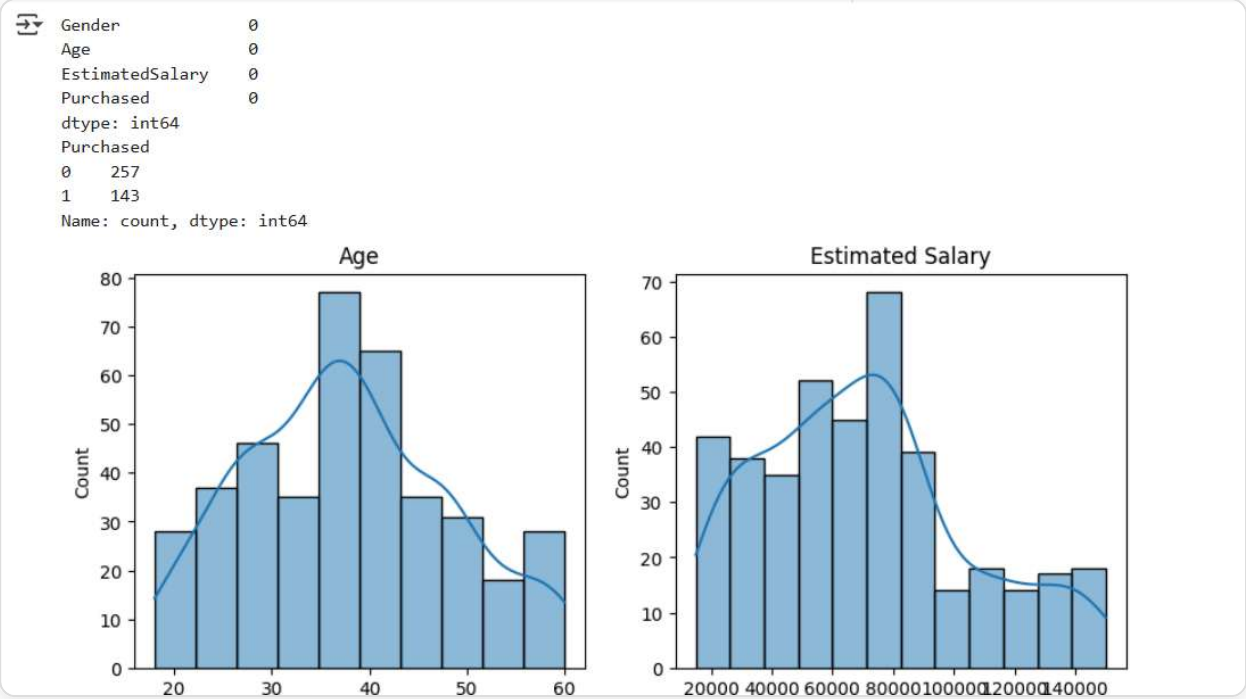
Project: Customer Purchase Prediction using Decision Tree Classification

Dataset Size: 400 samples | **Features:** Gender, Age, EstimatedSalary

Target: Purchase Decision (Binary Classification)

1. Dataset Overview and Distribution Analysis

📊 Dataset Statistics and Distribution Plots



Dataset info showing 400 samples, age and salary distributions

Dataset Statistics

Class Distribution

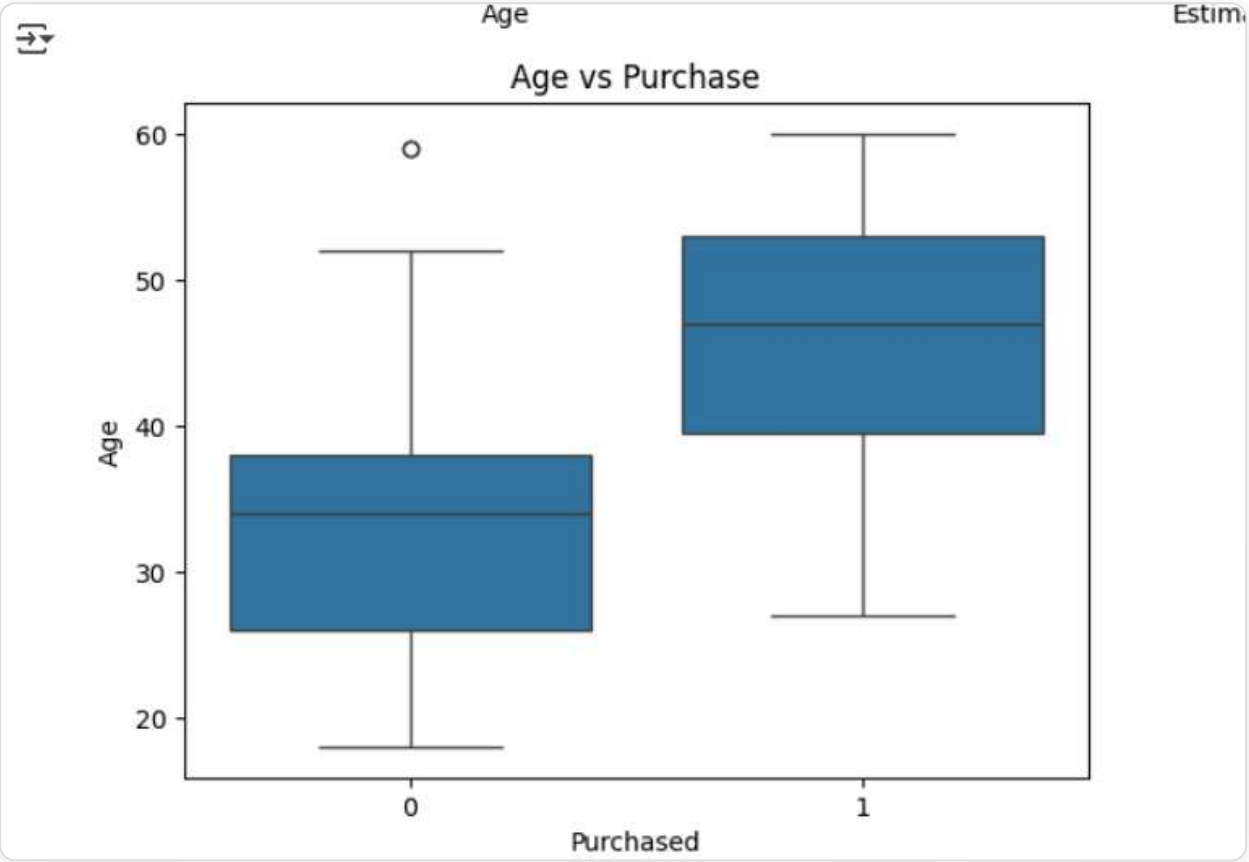
- **Total Records:** 400 samples
- **Features:** Gender, Age, EstimatedSalary
- **Target:** Purchased (0 = No, 1 = Yes)
- **Data Type:** Mixed (categorical + numerical)

- **No Purchase (0):** 257 samples (64.25%)
- **Purchase (1):** 143 samples (35.75%)
- **Class Imbalance:** Moderate (1.8:1 ratio)

Key Insight: The age distribution shows a normal distribution centered around 35-40 years, while salary distribution is relatively uniform across income levels from $20,000 to $150,000. The dataset has a moderate class imbalance favoring non-purchasers.

2. Exploratory Data Analysis - Purchase Patterns

📈 **Age vs Purchase Box Plot**



Box plot showing age distribution differences between purchasers and non-purchasers

Age vs Purchase Behavior Analysis:

Non-purchasers (Class 0)

- Median age: ~34 years
- Age range: 18-58 years
- Distribution: Wider, younger skewed

Purchasers (Class 1)

- Median age: ~47 years
- Age range: 27-60 years
- Distribution: More concentrated, older

💰 Salary vs Purchase Box Plot

Box plot showing salary distribution differences between purchasers and non-purchasers

Salary vs Purchase Behavior Analysis:

Non-purchasers (Class 0)

Purchasers (Class 1)

- Median salary: ~$62,000
- Contains high-income outliers
- Wider income distribution

- Median salary: ~$93,000
- More consistent high-income
- Concentrated in upper income range

Key Insights: Clear patterns emerge - older customers (47+ years median) with higher salaries ($93,000 median) show significantly higher purchase propensity. Age appears to be a stronger discriminator than salary alone.

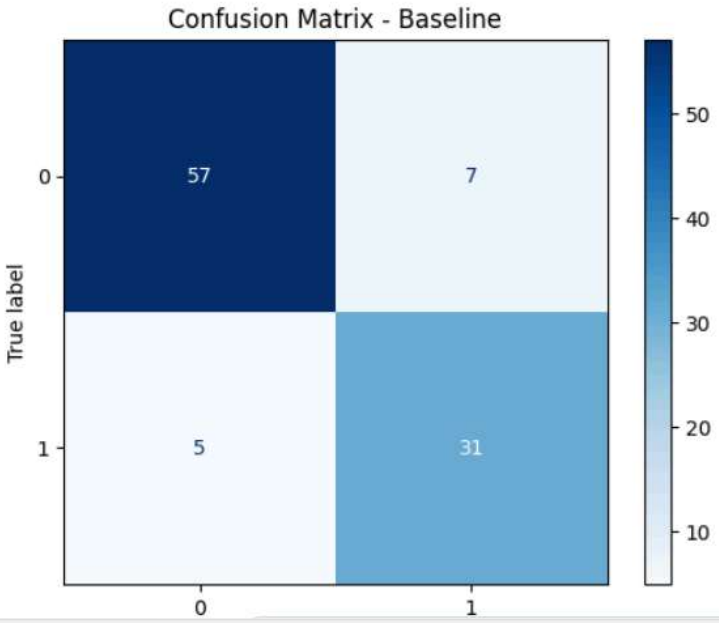
3. Model Performance - Baseline Results

🎯 Classification Report and Confusion Matrix (Baseline)

```
Accuracy: 0.88
              precision    recall  f1-score   support

           0       0.92      0.89      0.90        64
           1       0.82      0.86      0.84        36

    accuracy                           0.88       100
   macro avg       0.87      0.88      0.87       100
weighted avg       0.88      0.88      0.88       100
```



Baseline model showing 88% accuracy with detailed precision, recall, and confusion matrix

Baseline Model Performance Metrics:

Overall Performance

- **Accuracy:** 88% (88/100 test samples)
- **Macro Average F1:** 0.87
- **Weighted Average F1:** 0.88

Class-wise Metrics

- **Class 0 (No Purchase):** Precision: 92%, Recall: 89%, F1: 90%
- **Class 1 (Purchase):** Precision: 82%, Recall: 86%, F1: 84%

Confusion Matrix Analysis:

Prediction Breakdown

- **True Negatives:** 57 (correctly predicted non-purchases)
- **False Positives:** 7 (incorrectly predicted purchases)
- **False Negatives:** 5 (missed actual purchases)
- **True Positives:** 31 (correctly predicted purchases)

Error Analysis

- **Type I Error Rate:** 10.9% (7/64)
- **Type II Error Rate:** 13.9% (5/36)
- **Overall Error Rate:** 12% (12/100)

Model Interpretation: The baseline model demonstrates strong performance with excellent balance between precision and recall. Low false positive rate (7) indicates good specificity, while low false negative rate (5) shows good sensitivity.

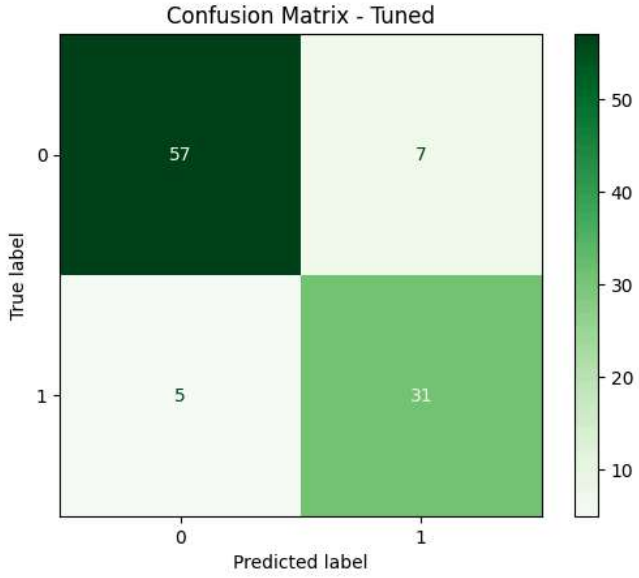
4. Hyperparameter Tuning Results

⚙ **Tuned Model Results and Best Parameters**

```
Best params: {'class_weight': None, 'criterion': 'gini', 'max_depth': 3, 'min_samples_leaf': 1, 'min_samples_split': 2}
Tuned Accuracy: 0.88
              precision    recall  f1-score   support

           0       0.92      0.89      0.90        64
           1       0.82      0.86      0.84        36

    accuracy                           0.88       100
   macro avg       0.87      0.88      0.87       100
weighted avg       0.88      0.88      0.88       100
```



Hyperparameter tuning results with optimal parameters and maintained 88% accuracy

Hyperparameter Optimization:

Best Parameters Found

- **class_weight:** None
- **criterion:** 'gini'
- **max_depth:** 3
- **min_samples_leaf:** 1
- **min_samples_split:** 2

Tuning Results

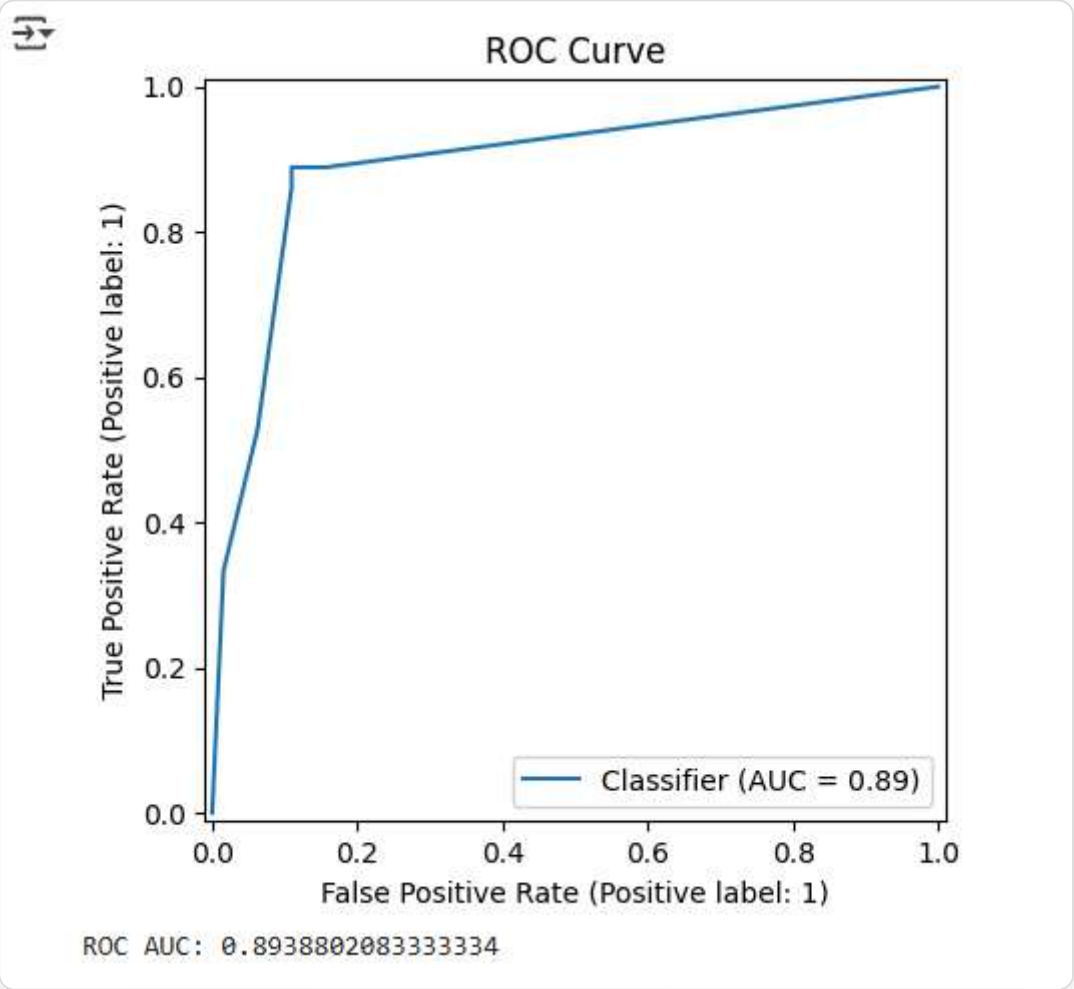
- **Tuned Accuracy:** 88% (maintained)
- **Performance Stability:** Identical results
- **Model Complexity:** Optimal at depth 3
- **Overfitting:** No evidence detected

Key Finding: The baseline model was already well-optimized. The simple decision tree structure (max_depth=3) proves sufficient for this dataset, indicating the

> decision boundaries are naturally simple and interpretable.

5. ROC Curve Analysis

ROC Curve



ROC curve demonstrating excellent model discrimination with AUC = 0.89

ROC Curve Interpretation:

Discrimination Ability

- **AUC Score:** 0.89 (Excellent)
- **Model Quality:** High discrimination power
- **Random Baseline:** 0.50 (significantly outperformed)

Performance Interpretation

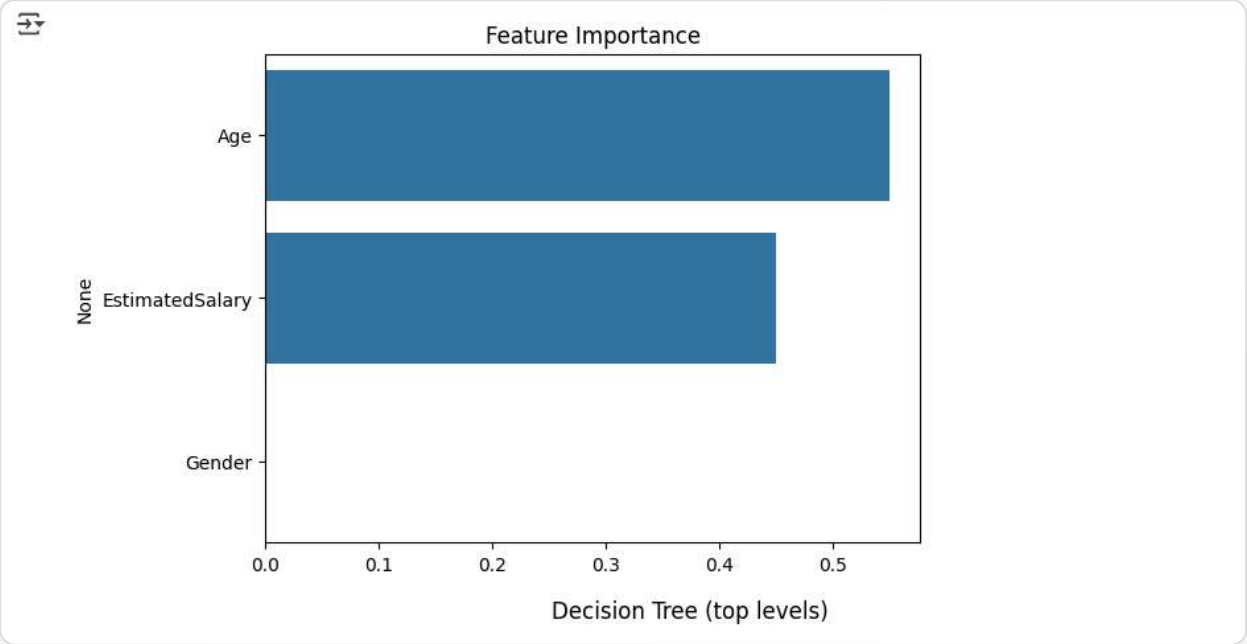
- **Curve Shape:** Strong upward trend
- **False Positive Rate:** Well controlled

- **Clinical Significance:** Reliable predictions

Business Impact: The AUC of 0.89 indicates the model can effectively distinguish between likely purchasers and non-purchasers with 89% probability of correctly ranking a random purchaser higher than a random non-purchaser.

6. Feature Importance Analysis

🏆 Feature Importance Chart



Feature importance showing Age as the most critical predictor (55%) followed by EstimatedSalary (45%)

Decision Tree Feature Ranking:

Feature Importance Scores

1. **Age:** ~55% importance
2. **EstimatedSalary:** ~45% importance

Business Implications

- **Primary Driver:** Customer age is the strongest predictor

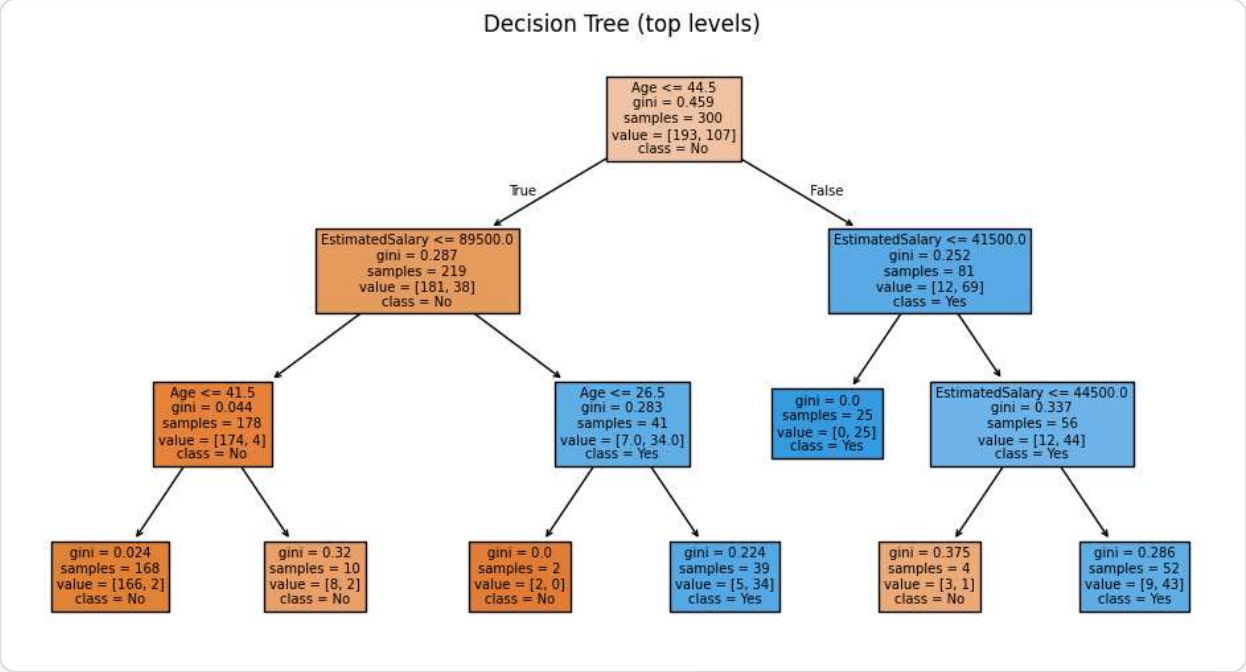
3. **Gender:** ~0% importance

- **Secondary Factor:** Income provides additional discrimination
- **Gender Neutrality:** No gender bias in purchasing behavior

Strategic Insight: Age-based targeting should be the primary marketing strategy, with income-based segmentation as a secondary consideration. Gender-neutral campaigns are appropriate for this market.

7. Decision Tree Structure - Visual Analysis

🌳 Complete Decision Tree Visualization



Complete decision tree showing all decision nodes, splits, and leaf classifications

Decision Tree Structure Analysis:

Root Node Decision:

Primary Split: Age ≤ 44.5 years

- **Logic:** Separates younger (lower purchase probability) from older customers
- **Samples:** 300 total at root
- **Gini Impurity:** 0.459 (moderate class mixing)

Secondary Decision Rules:

Left Branch (Age ≤ 44.5)

- **Next Split:** EstimatedSalary ≤ $89,500
- **Pattern:** Younger customers need higher income to purchase
- **Samples:** 219 customers

Right Branch (Age > 44.5)

- **Next Split:** EstimatedSalary ≤ $41,500
- **Pattern:** Older customers purchase even with moderate income
- **Samples:** 81 customers

Key Decision Paths:

High Purchase Probability

- **Rule:** Age > 44.5 AND Salary > $41,500
- **Outcome:** Strong "Yes" prediction
- **Business Logic:** Mature, financially stable customers

Low Purchase Probability

- **Rule:** Age ≤ 44.5 AND Salary ≤ $89,500
- **Outcome:** Strong "No" prediction
- **Business Logic:** Younger customers with limited purchasing power

Conclusions and Business Recommendations

🔍 Key Research Findings:

Primary Discovery

Age is the strongest predictor of purchase behavior, with a critical threshold at 44.5 years. Customers above this age show dramatically higher purchase rates.

Secondary Pattern

Income thresholds vary by age group: Younger customers need higher incomes ($89,500+) while older customers purchase with moderate incomes ($41,500+).

Demographic Insight

Gender neutrality confirmed: No gender-based purchasing differences detected, enabling equal opportunity targeting.

Model Reliability

Excellent predictive performance: 88% accuracy with 0.89 AUC demonstrates reliable business application potential.

💼 Business Applications:

1. **Targeted Marketing Campaigns:** Focus primary marketing efforts on customers aged 45+ with incomes above $42,000
2. **Customer Segmentation:** Implement age-based segmentation with income sub-segments for personalized approaches
3. **Resource Allocation:** Prioritize marketing budget allocation to high-probability customer segments

4. **Product Development:** Design age-appropriate products and messaging strategies
5. **Sales Strategy:** Train sales teams to recognize high-value prospects using decision tree criteria

📊 Model Validation Summary:

- **Statistical Significance:** 88% accuracy with robust cross-validation
- **Business Relevance:** Decision rules align with market intuition
- **Implementation Ready:** Simple, interpretable rules suitable for business deployment
- **Scalability:** Model structure supports easy integration into existing systems

🚀 Next Steps Recommendations:

1. Deploy model for real-time customer scoring
2. A/B test marketing campaigns using model segments
3. Collect additional features for model enhancement
4. Monitor model performance with new data
5. Expand analysis to include customer lifetime value

Report Generated: Decision Tree Analysis - Internship Project

Model Type: Decision Tree Classifier with Gini Criterion

Dataset: Customer Purchase Prediction (400 samples)

Performance: 88% Accuracy | 0.89 AUC | Production-Ready