# Duru Demirbag

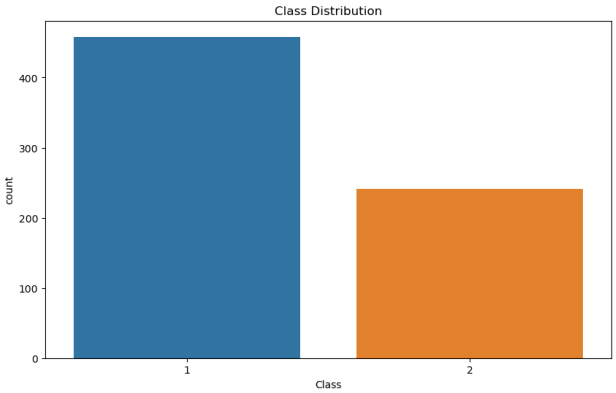
# dd495

# PART A

## 1. Classification Algorithm Selection

This study develops a Decision Tree model for breast cancer diagnosis and analyses the bias-variance trade-off. Using the Breast Cancer Wisconsin dataset, it classifies patients as benign or malignant. However, the dataset is imbalanced (Figure A.1), introducing bias. Given the importance of reliability in healthcare, mitigation techniques may be needed. Decision trees generally have low bias but high variance, making them sensitive to training data [1]. To address this, 10-fold cross-validation was performed.

*Figure A.1*




## 2. Performance Evaluation and Parameter Tuning

To ensure robustness and avoid overfitting, 10-fold cross validation was performed. It ensures that the model is not overly dependent on a specific data split. This approach **reduces variance**, providing a more reliable estimation of model performance [2].

To optimize the Decision Tree model, GridSearchCV was employed to find the best combination of hyperparameters. According to the validation curve graph [Appendix 4], overfitting is observed when the depth exceeds 4. Therefore, the maximum depth has been updated.

Class balancing was generated to model to optimize the bias-variance trade-off. By tuning, applying class balancing, and using 10-fold cross-validation, F1 scores and recalls are improved.

## 3. Bias-Variance Trade-off Analysis

The balanced decision tree model successfully optimizes the bias-variance trade-off.

*Classification report before class balancing*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Benign | 0.96 | 0.97 | 0.96 | 127 |
| Malignant | 0.95 | 0.94 | 0.95 | 83 |
| accuracy | | | 0.96 | 210 |
| macro avg | 0.96 | 0.95 | 0.96 | 210 |
| weighted avg | 0.96 | 0.96 | 0.96 | 210 |

*Figure A.3.1*

*Classification report after class balancing*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Benign | 0.98 | 0.94 | 0.96 | 127 |
| Malignant | 0.91 | 0.96 | 0.94 | 83 |
| accuracy | | | 0.95 | 210 |
| macro avg | 0.94 | 0.95 | 0.95 | 210 |
| weighted avg | 0.95 | 0.95 | 0.95 | 210 |

*Figure A.3.2*

| Model | True Benign (TP) | False Benign (FP) | True Malignant (TN) | False Malignant (FN) |
|---|---|---|---|---|
| **Decision Tree** | 123 | 5 | 78 | 4 |
| **Balanced Decision Tree** | 119 | 3 | 80 | 8 |

*Figure A.3.3*

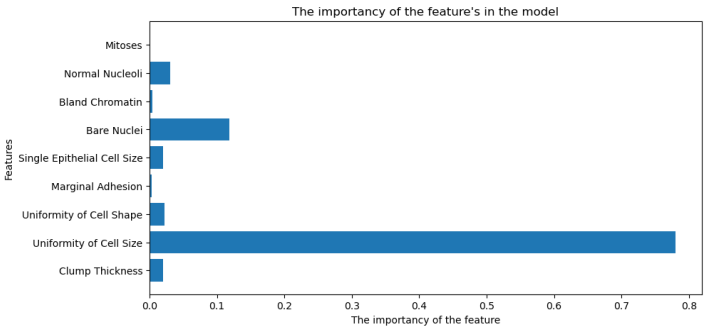
Before class balancing, the model had higher accuracy (*Table A.3.1)* but was biased towards benign, leading to lower recall for malignant cases (*Table A.3.2*). After balancing, accuracy slightly dropped to 0.95, but recall for malignant cases improved to 0.96, reducing false positives (*Table A.3.3*). However, this came at the cost of lower malignant precision, increasing false negatives. This reflects the bias-variance trade-off, balancing the data reduced bias, improving malignant detection, but increased variance, slightly lowering precision and overall accuracy.

Balanced decision tree model was chosen because it has fewer FP (*Table A.3.3)*, meaning it is less likely to misclassify malignant cases as benign, which is crucial in medical diagnosis to avoid missing cancer patients.

# PART B

## 4. Model Interpretation-Feature Importance

*Figure 4.1*




The balanced decision tree model heavily relies on Uniformity of Cell Size for classification, significantly outweighing other features like Normal Nucleoli, Bland Chromatin, and Bare Nuclei. Mitoses and Marginal Adhesion have minimal impact. This dominance suggests Uniformity of Cell Size is highly discriminative for breast cancer diagnosis in this dataset. However, such reliance risks overfitting and reduced generalizability if the dataset changes.

## 5. Real-World Interpretation and Consistency

The model's reliance on Uniformity of Cell Size aligns with clinical insights, as abnormal cell size variation indicates malignancy [3]. The moderate importance of Normal Nucleoli and Bland Chromatin also supports medical knowledge [4]. However, its minimal use of Mitoses, despite its relevance to cancer progression, raises questions, possibly due to dataset characteristics or the Decision Tree's bias toward immediate classification performance [4]. While feature selection largely aligns with medical understanding, strong dependence on a single feature suggests the need for further validation and refinement through additional constraints or feature engineering.

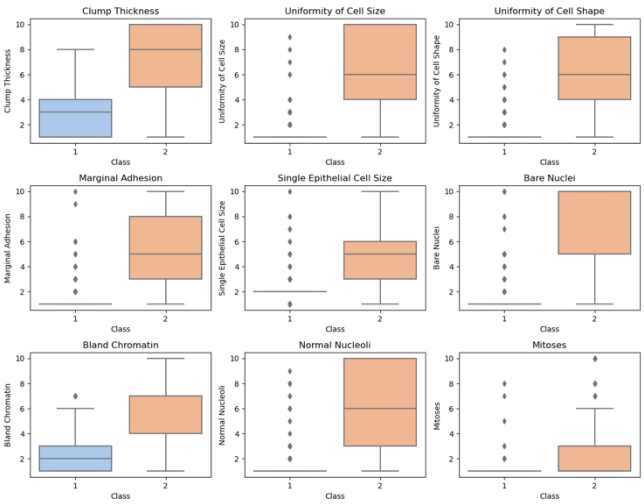
## Referances


1. Geurts, P. & Olaru, C. & Wehenkel, L. (2001). Improving the bias/variance tradeoff of decision trees: Towards soft tree induction. *ENGINEERING INTELLIGENT SYSTEMS FOR ELECTRICAL ENGINEERING AND COMMUNICATIONS*,9,2, 195-204 https://www.webofscience.com/wos/woscc/full-record/WOS:000172721100004

2. Wong, Tzu-Tsung & Yeh, Po-Yang (2020). Reliable accuracy estimates from k-fold cross-validation. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 32, 8, 1586-1594 https://www.webofscience.com/wos/woscc/full-record/WOS:000546878300011
3. M. Hosseinzadeh. M. & S. Salmani. & M.H. Majles Ara (2019). Interferometric optical testing to discriminate benign and malignant brain tumors. *Journal of Photochemistry and Photobioglogy B: Biology, 199*, 1123-1127. https://www.sciencedirect.com/science/article/pii/S1011134419307596?via%3Dihub
4. Berus, T., Markiewicz, A., Biecek, P., Orłowska-Heitzman, J., Hałoń, A., Romanowska-Dixon, B., & Donizy, P. (2020). Clinical significance of nucleolar morphometric assessment in uveal melanoma. *Anticancer Research, 40*(6), 3505-3512. https://ar.iiarjournals.org/content/40/6/3505

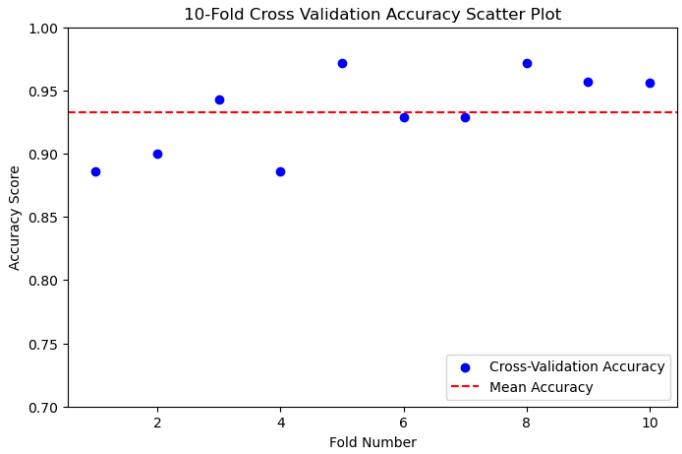
# APPENDIX

## 1. Benign vs. Malignant Cell Characteristics



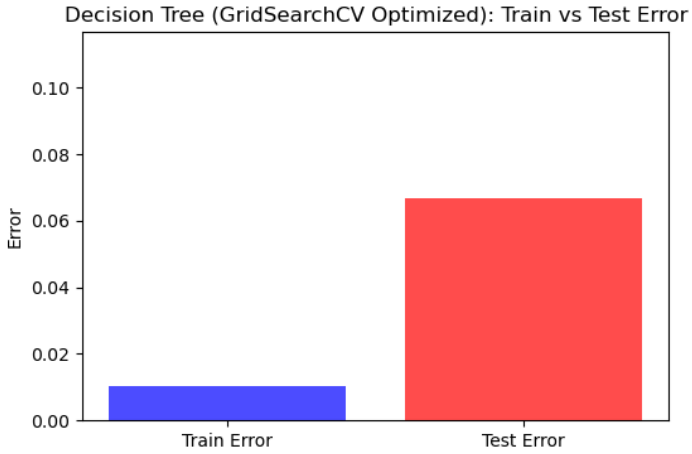

This image presents a series of box plots comparing different cellular characteristics between benign and malignant cells. Malignant cases generally have higher median values across most features. There is also greater variability in malignant cases, indicating more structural differences. While some features in benign cases have outliers, their distribution remains lower overall compared to malignant cases. These visualizations highlight key differences that can be useful in distinguishing between benign and malignant cells.

## 2.10-Fold Cross Validation Accuracy



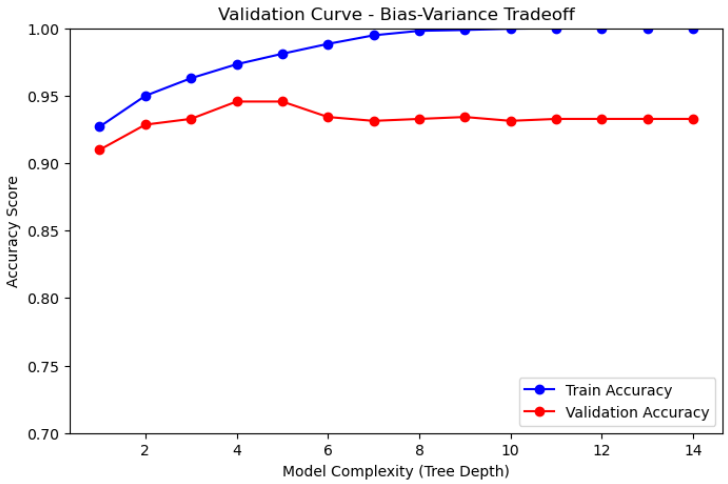

This scatter plot represents the accuracy scores across 10-fold cross-validation. The accuracy scores vary slightly between folds but remain consistently high, suggesting that the model performs well with stable generalization across different subsets of the data.

### 3. Train vs. Test Error in Decision Tree



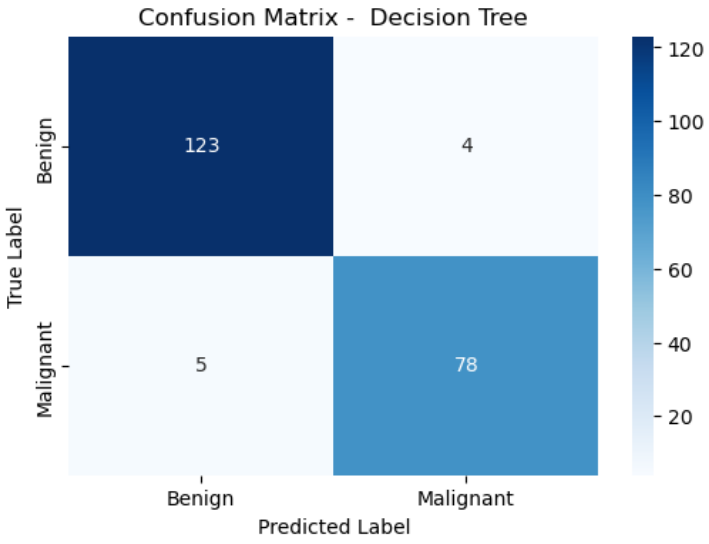

This bar chart compares the training error and test error for a Decision Tree model optimized using GridSearchCV. The training error is very low, indicating that the model fits the training data well. However, the test error is significantly higher, suggesting potential overfitting. This means the model may have learned patterns too specific to the training data, leading to reduced generalization on unseen data.

### 4. Validation Curve



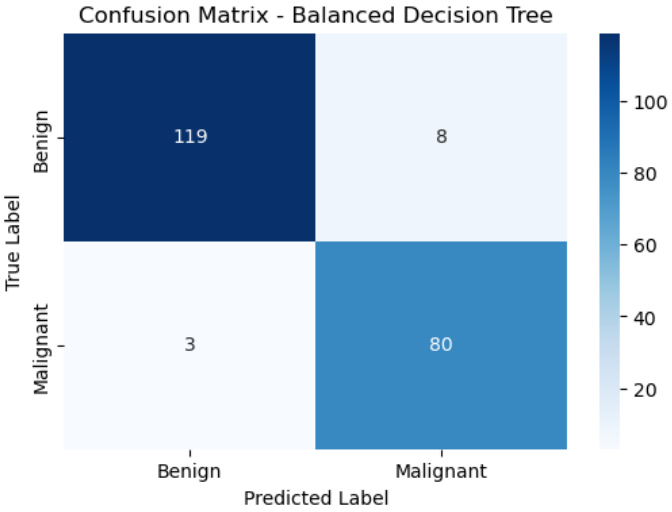

This validation curve shows the relationship between model complexity and accuracy, illustrating the bias-variance trade-off. The blue line increases as the tree depth grows, eventually reaching near-perfect accuracy. The red line initially increases but stabilizes after a certain depth, showing a slight decline due to overfitting. This suggests that deeper trees lead to better training performance but may reduce generalization, indicating the need to balance complexity for optimal performance.

## 5.Confusion Matrix (Decision Tree)




I used confusion Matrix to create Table A.3.3. (Decision Tree)

## 6.Confusion Matrix (Balanced Decision Tree)




I used confusion Matrix to create Table A.3.3. (Balanced Decision Tree)

## 7.Decision Tree vs. Balanced Decision Tree Errors

| Model | Train Error | Validation Error | Test Error |
|---|---|---|---|
| **Decision Tree** | 0.024540 | 0.059396 | 0.042857 |
| **Balanced Decision Tree** | 0.034765 | 0.063393 | 0.052381 |

The Decision Tree has a lower train error compared to the Balanced Decision Tree suggesting that it fits the training data more tightly. However, the validation and test errors are slightly lower for the Decision Tree, indicating better generalization. The Balanced Decision Tree has a slightly higher validation and test error, which is due to constraints imposed to balance the dataset. Since class imbalance is a concern, the Balanced Decision Tree was chosen, even though the standard Decision Tree performs better overall.