*Practical Data Analysis*

*fs349 & clt39*

## **Part A: Technical Analysis**

Within our analysis, we selected the Decision Tree Classifier for its capability to model complex decision boundaries through hierarchical attribute splits. This choice is backed by its demonstrated versatility in previous high-grade analyses, highlighting its ability to derive meaningful insights from data. The model's transparent decision-making facilitates an intuitive understanding of how features impact classification, aligning with our goals to achieve high classification accuracy while understanding the data structure deeply.  
To optimize our Decision Tree Classifier's performance, we adopted a systematic approach to parameter tuning, utilizing GridSearchCV to explore a range of parameter values, with a focus on max\_depth, min\_samples\_split, and min\_samples\_leaf. This search is vital for identifying the optimal balance between the model's complexity and its generalization capability, which is directly linked to the classifier's accuracy.  
A critical element of our analysis was understanding the bias-variance trade-off, crucial for optimal model performance. By plotting a validation curve across different max\_depth values, we observed the impact of tree depth on model accuracy. Lower depths result in high bias and underfitting, while excessive depth causes high variance and overfitting. This analysis is key to determining an appropriate model complexity that balances bias and variance, ensuring the model's robustness and reliability. Our examination of the bias-variance trade-off provides a detailed understanding of how various parameter settings influence the model's generalization to unseen data.

## **Part B: Model Insights and Real-Life Applicability**

Upon analysing the learned Decision Tree model, we noted that features such as Uniformity of Cell Size and Bare Nuclei are critical in classification decisions, consistent with their early emergence as crucial nodes within the decision trees. This observation aligns with medical research findings, emphasizing their significance in diagnosing breast cancer[[1]](#endnote-1). The prominence of these features in our model underscores their potential as diagnostic markers, supporting the Decision Tree's ability to reveal clinically relevant insights.  
Our evaluation reveals a strong correlation between specific features and likelihood of malignancy in breast cancer diagnoses, resonating with established medical knowledge. For instance, Uniformity of Cell Size is extensively documented as a significant indicator of malignancy, where higher uniformity suggests increased cancer risk.[[2]](#endnote-2) This corroboration with medical literature not only validates our model's predictive accuracy but also enhances its credibility as a diagnostic tool.  
While the model precisely identifies significant attributes, cautious interpretation is necessary. Due to the inherent nature of Decision Trees, there is a potential for overemphasizing certain features influenced by dataset specifics or inherent biases. Therefore, while the insights are invaluable, they should complement, not replace, thorough clinical analysis.  
Comparing the model's insights with real-world applications underscores that despite accurate feature identification, the intricacy of medical diagnostics demands a multifaceted approach. The significance of various interacting factors in cancer prognosis highlights the necessity for an integrated diagnostic process that combines algorithmic predictions with clinical expertise to optimize patient outcomes[[3]](#endnote-3)

# Works Cited

* Dua, D., and Graff, C. UCI Machine Learning Repository. Available at <http://archive.ics.uci.edu/ml>.
* Street, W.N., et al. (1993). Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology. Available at <https://www.researchgate.net/publication/2512520_Nuclear_Feature_Extraction_For_Breast_Tumor_Diagnosis>.
* Wolberg, W.H., et al. (1995). Breast cancer diagnosis and prognosis via linear programming. Available at <https://www.jstor.org/stable/171686>.

## Appendix

A graph of a line graph

Description automatically generated

Figure 1: Error vs Model Complexity Graph with Regions of Interest

A diagram of a company structure

Description automatically generated

Figure 2: Best Decision Tree Classifier Diagram

A graph with text on it

Description automatically generated

Figure 3: Feature Importance per Class

1. Dua, D., and Graff, C. UCI Machine Learning Repository. [↑](#endnote-ref-1)
2. Street, W.N., et al. (1993). Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology. [↑](#endnote-ref-2)
3. Wolberg, W.H., et al. (1995). Breast cancer diagnosis and prognosis via linear programming. [↑](#endnote-ref-3)