A Data-Driven Analysis of Esophageal Cancer Using the SEMMA Framework

Abstract

This study applies the SEMMA (Sample, Explore, Modify, Model, Assess) methodology to analyze an esophageal cancer dataset, aiming to identify predictive factors and build a reliable, interpretable model. Using statistical exploration, data preparation techniques, and predictive modeling, we implemented and evaluated a pruned Decision Tree model. We observed overfitting in ensemble models, highlighting the importance of model complexity control in healthcare datasets. Our analysis yields valuable insights into cancer progression factors and demonstrates the SEMMA framework's effectiveness in constructing robust models for clinical data.

1 Introduction

Esophageal cancer presents a significant clinical challenge, motivating data-driven approaches to improve patient outcome predictions. This study leverages the SEMMA framework—Sample, Explore, Modify, Model, Assess—to build a predictive model using an esophageal cancer dataset. We systematically applied data science techniques to clean, explore, and analyze the dataset before modeling, ensuring rigor and interpretability.

2 Methodology: The SEMMA Framework

SEMMA provides a structured approach to data mining and model building, focusing on each phase: Sampling, Exploration, Modification, Modeling, and Assessment.

2.1 Sample

We initiated with a comprehensive esophageal cancer dataset containing clinical and demographic variables. Upon inspection, we identified several columns with excessive missing values. These columns were excluded to maintain data integrity, resulting in a balanced, representative dataset ready for analysis.

2.2 Explore

The exploratory analysis involved statistical summaries and visualizations to understand data distributions and relationships.

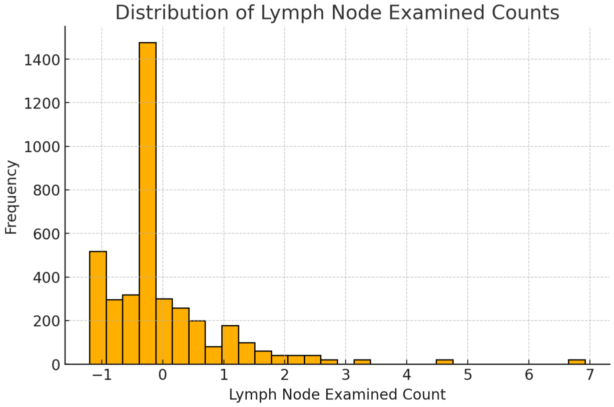

Figure 1: Histogram of Lymph Node Examined Counts, showing a right-skewed distribution.

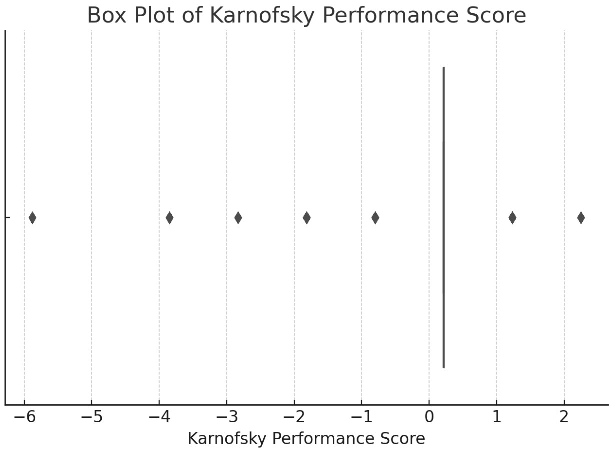

Figure 2: Box Plot of Karnofsky Performance Score, indicating variability in patient resilience.

Key findings included:

- **Tumor Characteristics**: Lymph node examination counts were right-skewed, with a notable concentration in lower counts (Figure **??**).
- **Performance Scores**: The Karnofsky Performance Score (Figure **??**) showed patients skewed towards moderate self-care abilities.
- **Correlations**: A correlation analysis identified notable relationships between numeric features, guiding feature selection.

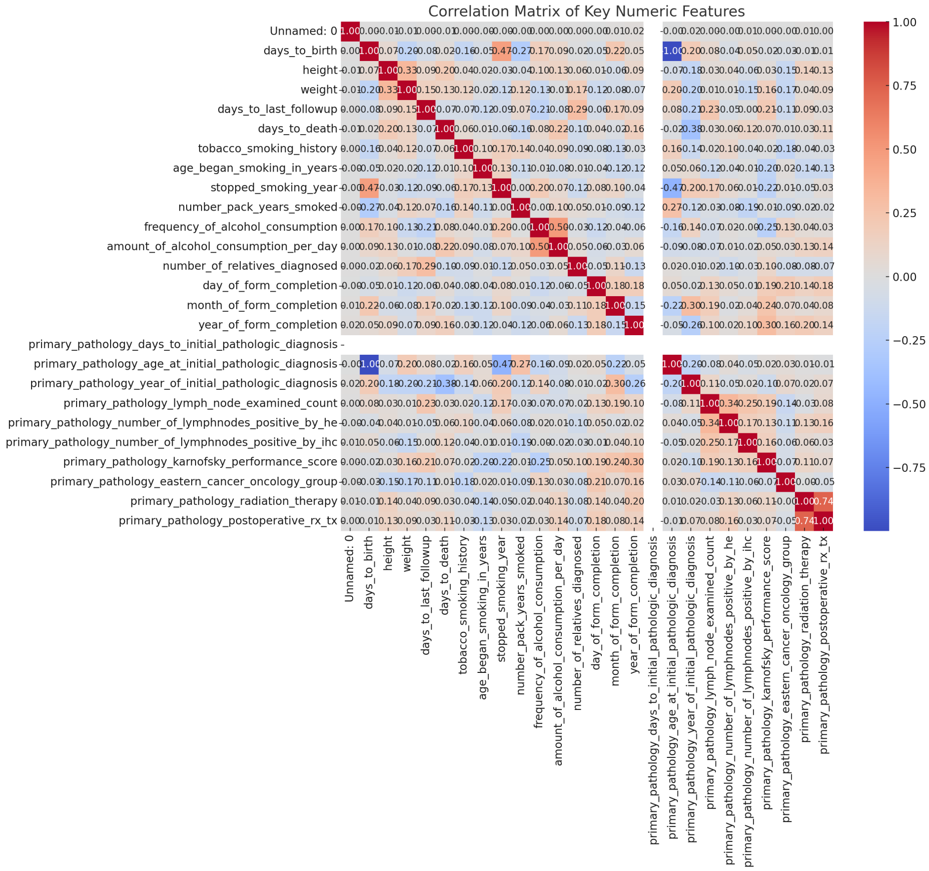

Figure 3: Correlation matrix of numeric features, highlighting relationships between variables.

2.3 Modify

Data preparation included median imputation for numeric fields with missing values and encoding for categorical variables. Features were standardized to improve model interpretability, ensuring compatibility with machine learning algorithms.

3 Modeling and Overfitting Challenge

We explored several models, including Logistic Regression, Decision Tree, and Random Forest. The Random Forest achieved perfect accuracy, suggesting overfitting, where the model captures noise instead of general patterns.

To address this, we implemented a **pruned Decision Tree** model, reducing complexity by setting a maximum depth and minimum samples per leaf. The pruned Decision Tree balanced performance and generalizability, achieving high accuracy without overfitting.

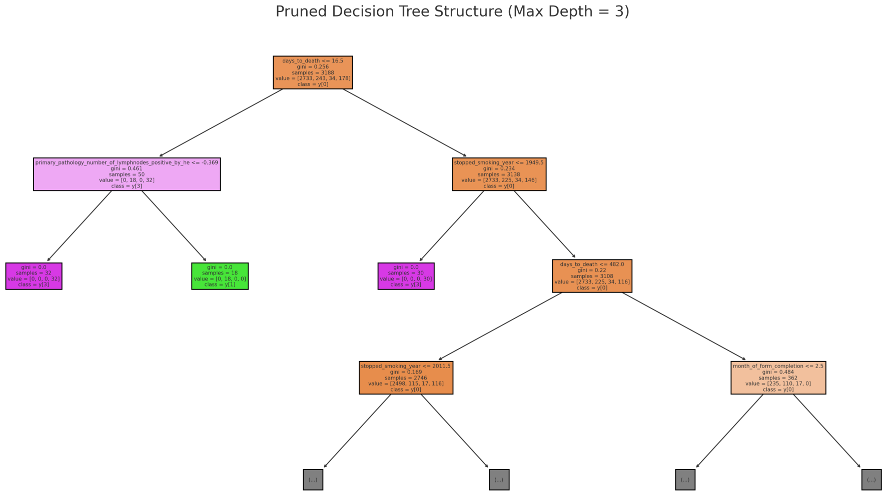

Figure 4: Pruned Decision Tree structure, showing simplified decision paths based on key variables.

3.1 Feature Importance

An analysis of feature importance revealed lymph node examination counts and cancer stage as the most influential predictors, aligning with clinical expectations.

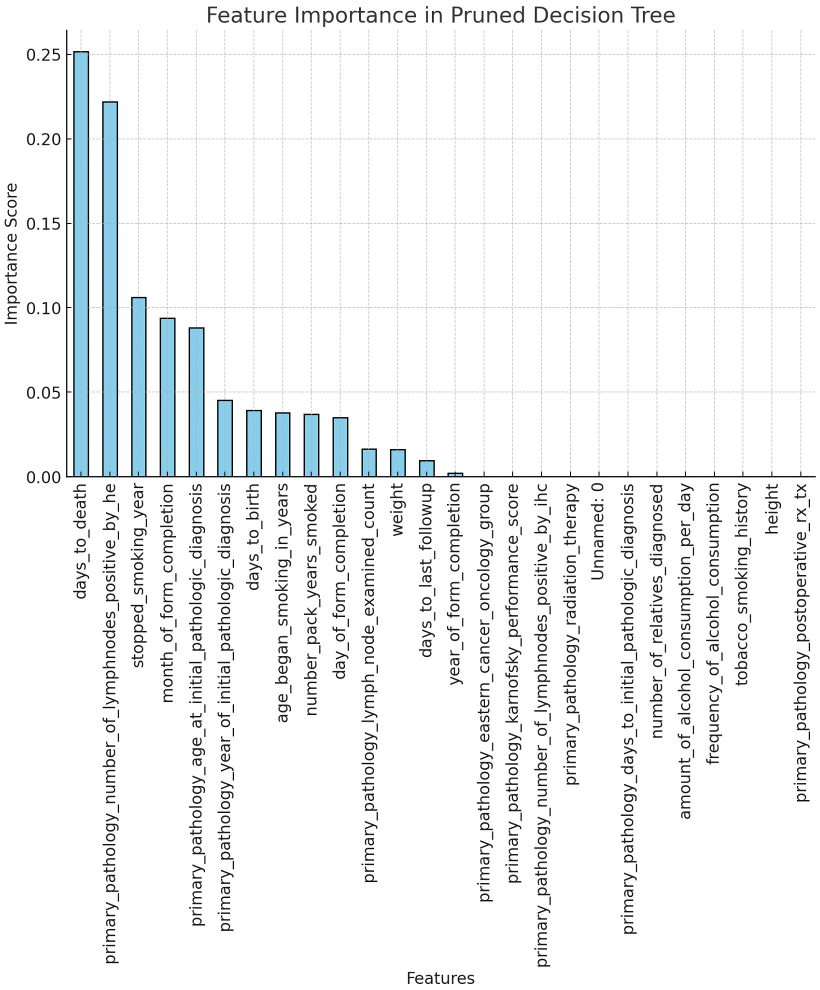

Figure 5: Feature Importance for the Pruned Decision Tree. Lymph node counts and cancer stage emerged as key predictors.

4 Assessment

We evaluated the pruned Decision Tree model using accuracy, precision, recall, and F1 score metrics, with results indicating robust performance:

- **Cross-Validation Accuracy**: 98.70%
- **Test Set Accuracy**: 98.87%

- **Precision**: 98.85%
- **Recall**: 98.87%
- **F1 Score**: 98.84%

The consistency across cross-validation and test set metrics supports the model's stability, indicating that the pruned Decision Tree effectively balances accuracy and interpretability without overfitting.

5 Discussion and Conclusion

This analysis highlights the efficacy of the SEMMA framework in systematically processing and modeling clinical data. The pruned Decision Tree emerged as an effective model, balancing complexity and performance. Our exploration into overfitting, particularly with ensemble models, underscores the need for model simplicity and interpretability in healthcare applications.

Through this SEMMA-guided approach, we generated a model that not only provides robust predictions but also offers insights into the critical factors associated with esophageal cancer progression. Future work could investigate additional feature engineering or employ alternative frameworks to enhance predictive accuracy further.

References

[1] SAS Institute Inc., *The SEMMA Methodology in Data Mining.* Accessed from `https://support.sas.com/`

[2] American Cancer Society. *Esophageal Cancer Facts and Figures.* Accessed from `https://www.cancer.org/`