final report

2023-12-19

Abstract

Introduction

Data and Methods

Descriptive Data

Data Cleaning and Preparation

The dataset utilized in this study was derived from a comprehensive breast cancer database. The initial step in data preparation involved the standardization of variable names to ensure consistency. In addition, we converted several categorical variables into factors with defined levels. Specifically, we recoded the survival status variable into a binary format with `Dead` as 1 and `Alive` as 0. Finally, a new variable `node_positive_prop` was created and calculated based on the ratio of `reginol_node_positive` to `regional_node_examined`. This variable represents the proportion of examined nodes that were found to be positive.

Variable and Model Selection Procedures

After preprocessing the dataset, we subdivided our dataset into categorical and numerical variables in order to have a general outline data patterns. In **Table 1**, we summarized the essential statistics of all the categorical variables, which includes variable names, number of missing values, unique and top counts.

For numerical variables, we employed boxplot visualizations to effectively represent their distribution patterns. As illustrated in **Figure 1**, these boxplots serve as a comprehensive visualization, which include potential outliers, important quartiles, and medians of all the numerical variables. # Assumption Checking

Result

Variable Selection

Model Selection

Conclusion

Appendix

Table

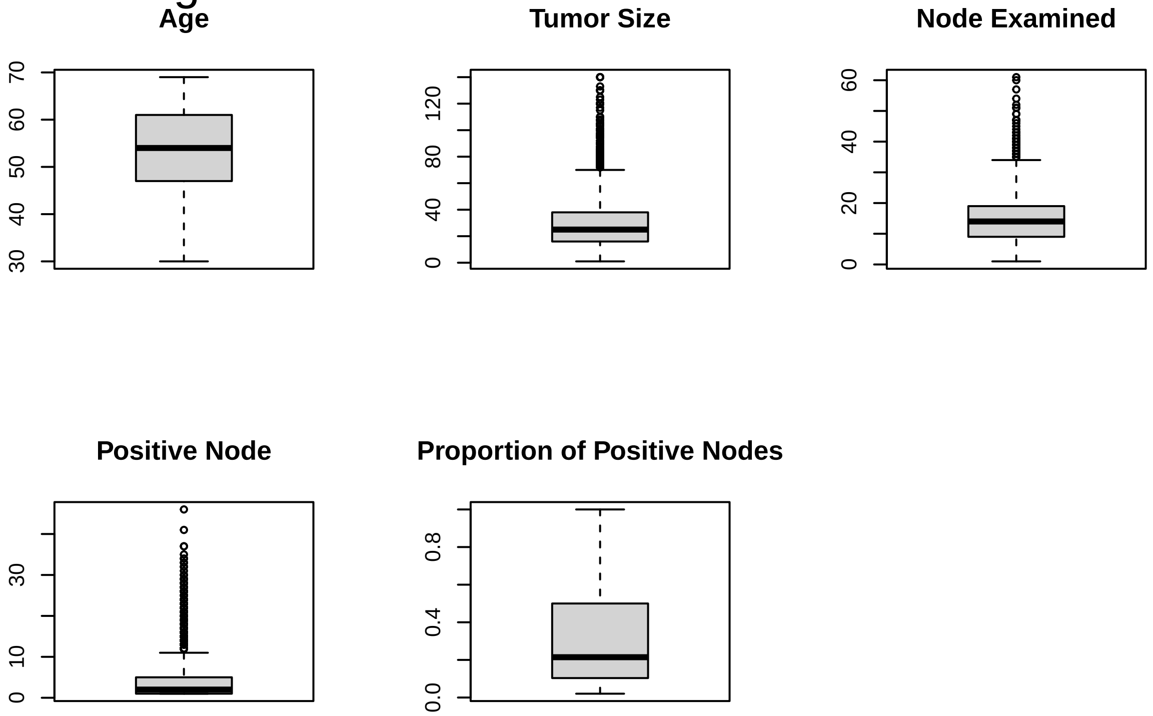
Table 1: Summary Statistics of Categorical Variables

Variable	Missing	Unique Counts	Top Counts
race	0	3	Whi: 3413, Oth: 320, Bla: 291
marital_status	0	5	Mar: 2643, Sin: 615, Div: 486, Wid: 235

Variable	Missing	Unique Counts	Top Counts
t_stage	0	4	T2: 1786, T1: 1603, T3: 533, T4: 102
n_stage	0	3	N1: 2732, N2: 820, N3: 472
x6th_stage	0	5	IIA: 1305, IIB: 1130, III: 1050, III: 472
differentiate	0	4	Mod: 2351, Poo: 1111, Wel: 543, Und: 19
grade	0	4	2: 2351, 3: 1111, 1: 543, ana: 19
a_stage	0	2	Reg: 3932, Dis: 92
estrogen_status	0	2	Pos: 3755, Neg: 269
progesterone_status	0	2	Pos: 3326, Neg: 698
status	0	2	0: 3408, 1: 616
node_positive_prop	0	NA	NA

Figure

Figure 1: Numerical Variables Distribution



Age

Tumor Size

Node Examined

Positive Node

Proportion of Positive Nodes

Contribution