Assignment2_BTC1877H

Leen Madani

2023-11-02

Assignment Objective:

The objective of this assignment is to train and use a number of models for both regression and classification, as well as to perform survival analysis. You will use a data set from the University of Wisconsin where each record represents follow-up data for one breast cancer case after surgery. The data set contains features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Information about the outcome of the patient is also included, such as time to recurrence or time to last seen, for those who have not experienced recurrence yet.

Q1 Regression (7 points)

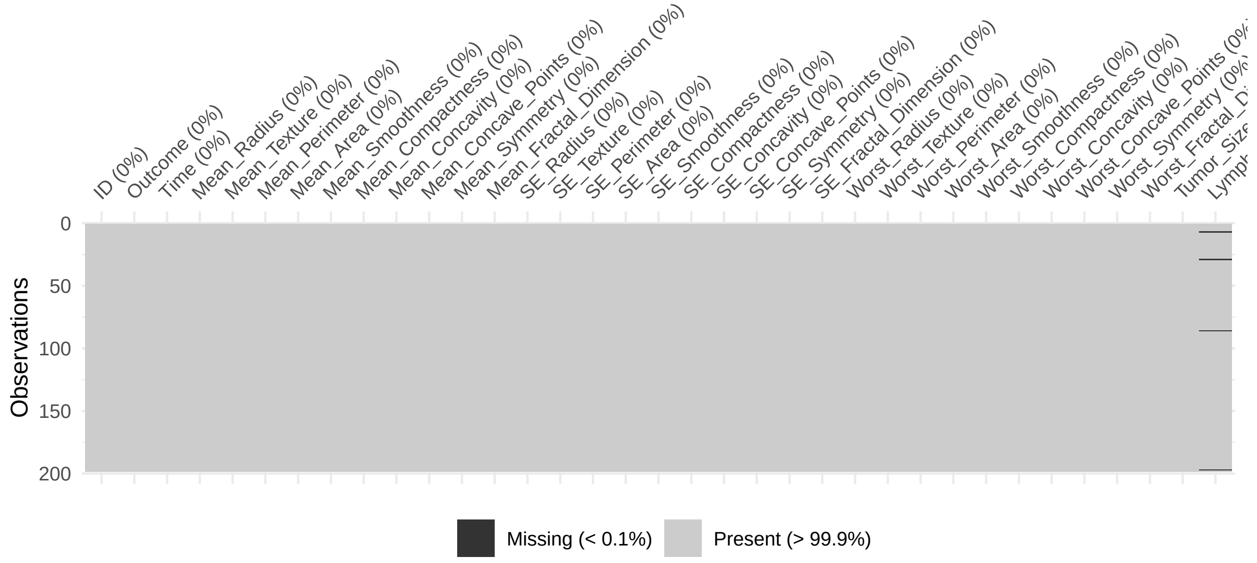

Figure 1: Visualization of the Amount of Missing Data.

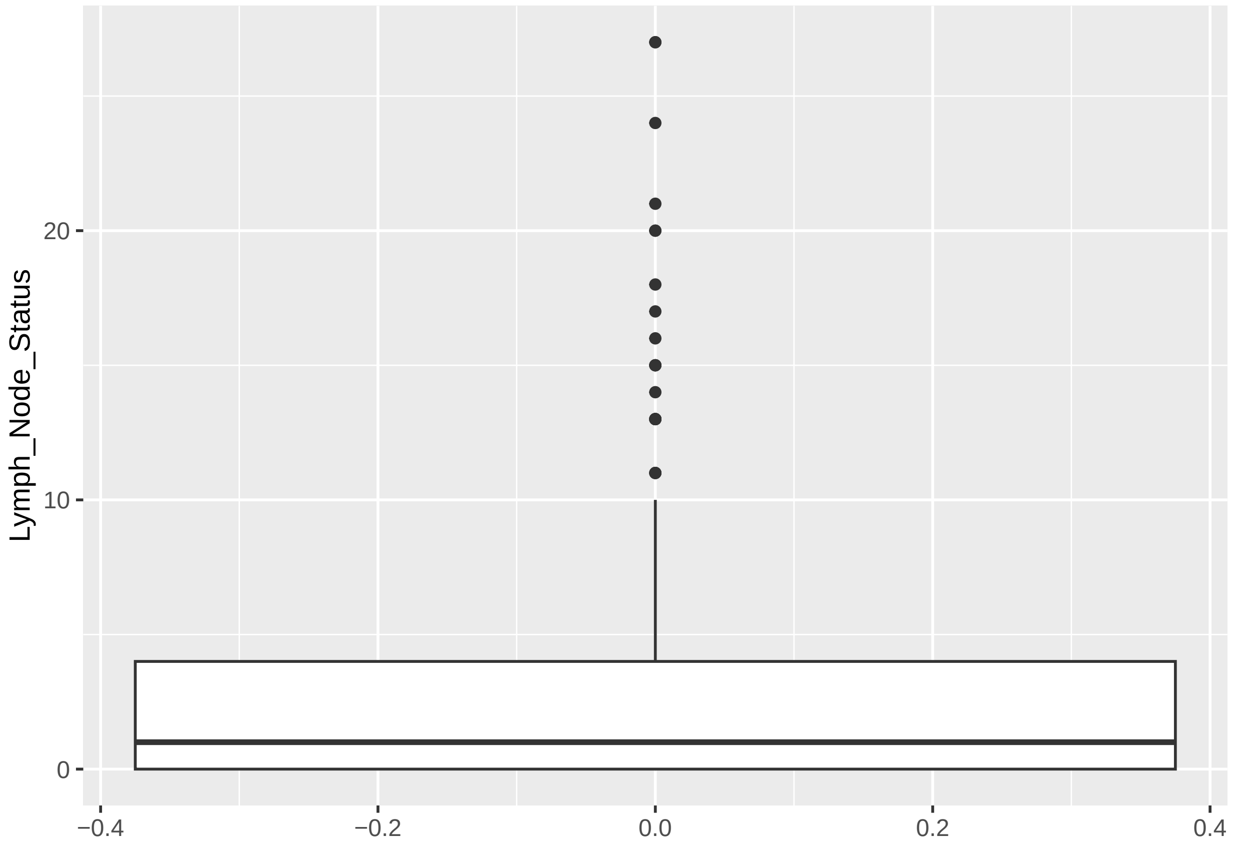

Figure 2: Boxplot to check for outliers in Lymph Node Status variable before data cleaning or encoding into categorical variable.

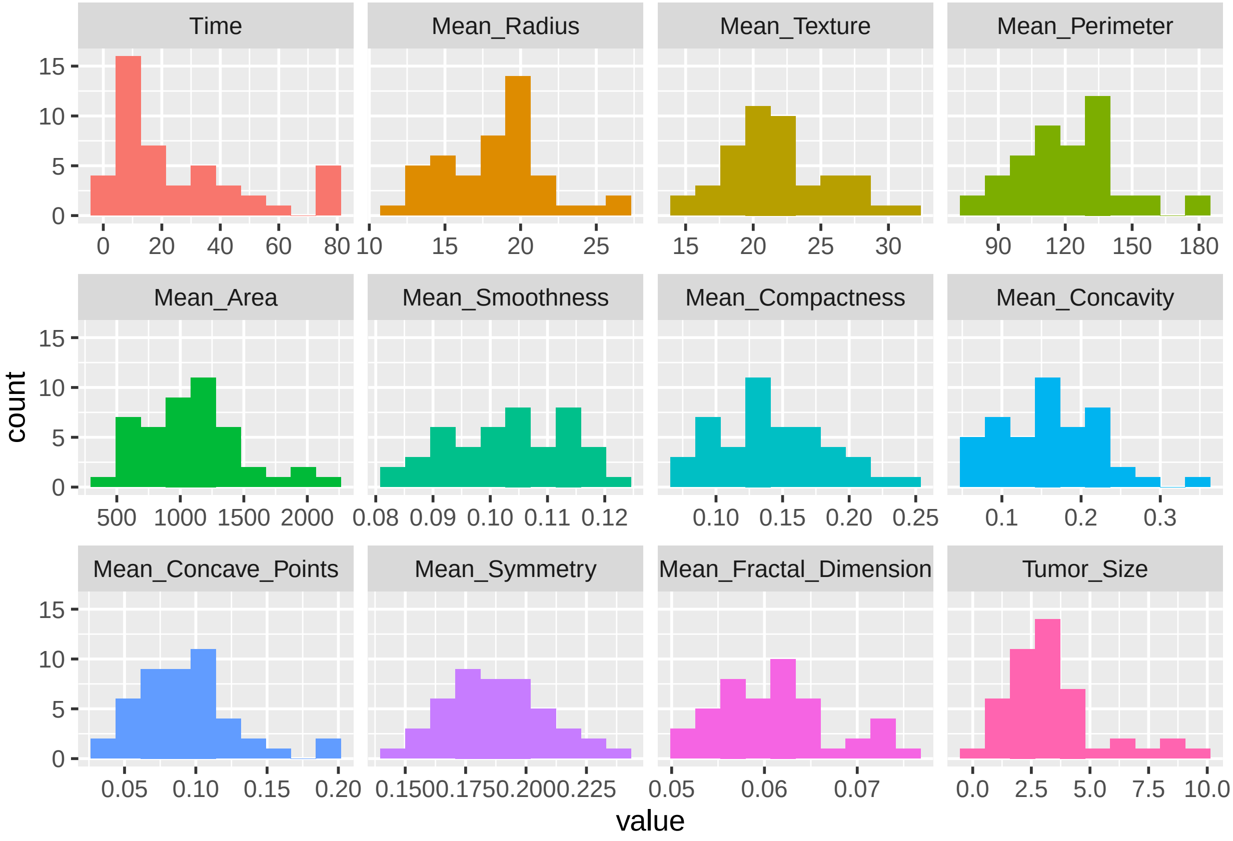

Figure 3: Visualization of the Continuous Variables using Histograms.

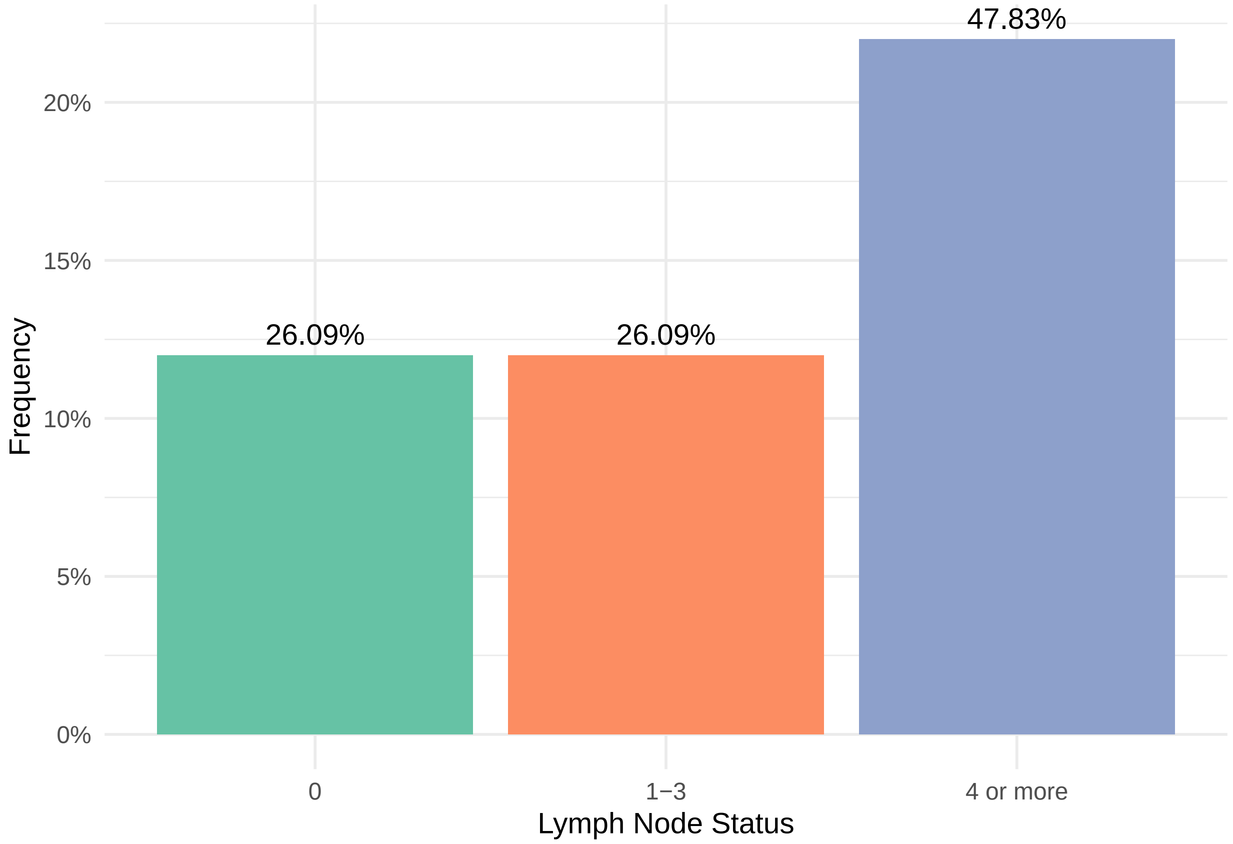

Figure 4: Visualization of the categorical variable 'Lymph Node Status' as frequency distribution bar plot. '0' indicates no positive axillary lymph nodes observed, '1-3' indicates 1 to 3 positive axillary lymph nodes observed, and '4 or more' indicates 4 or more positive axillary lymph nodes observed.

Table 1: Descriptive Statistics for Continuous Predictors

Variable	Mean	Standard Deviation	Variation Coef.	1st Quartile	Median	3rd Quartile	Skewness
Time	25.57	22.73	0.89	9.00	16.50	36.75	1.10
Mean_Radius	18.33	3.37	0.18	15.66	18.82	20.26	0.31
Mean_Texture	21.76	3.70	0.17	19.07	21.36	24.16	0.34
Mean_Perimeter	121.10	22.91	0.19	103.90	123.55	133.42	0.38
Mean_Area	1081.98	397.26	0.37	799.87	1090.00	1278.50	0.79
Mean_Smoothness	0.10	0.01	0.10	0.09	0.10	0.11	-0.17
Mean_Compactness	0.14	0.04	0.29	0.11	0.13	0.17	0.27
Mean_Concavity	0.16	0.06	0.38	0.11	0.16	0.21	0.38
Mean_Concave_Points	0.09	0.03	0.37	0.07	0.09	0.11	0.85
Mean_Symmetry	0.19	0.02	0.11	0.17	0.19	0.20	0.32
Mean_Fractal_Dimension	0.06	0.01	0.10	0.06	0.06	0.07	0.41
Tumor_Size	3.47	2.03	0.58	2.35	3.00	4.00	1.49

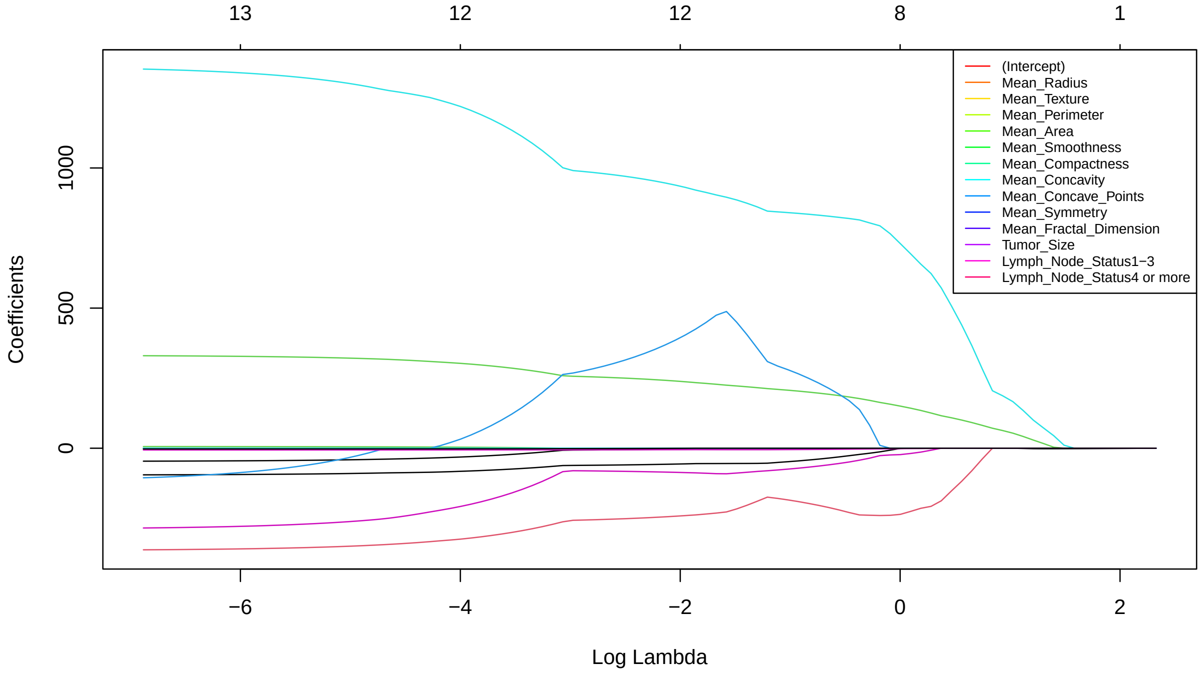

Figure 5: LASSO Coefficient Paths for Predicting Time to Recurrence.This figure illustrates the paths of the coefficients for 12 predictors as the regularization parameter lambda increases in a LASSO model. Predictors include various tumor characteristics such as radius, texture, and size, as well as lymph node status. Each line represents a predictor, with the value of the coefficient shrinking towards zero as the penalty for model complexity increases. The plot aids in identifying which predictors have the most significant impact on the time to recurrence as the model becomes more parsimonious.

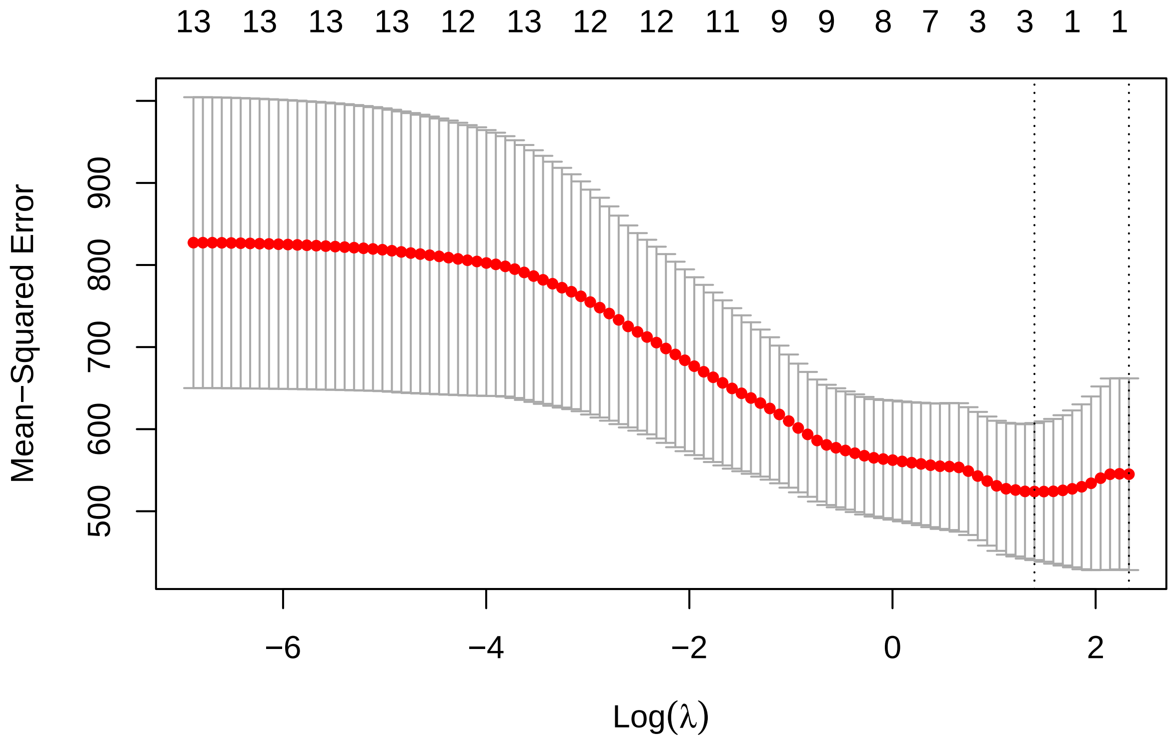

Figure 6: Cross-validation to Determine Optimal Regularization Parameter in LASSO Model. This figure represents the cross-validation process used to identify the optimal value of lambda that minimizes the mean squared error in a LASSO regression model. The plot illustrates the variation of the mean squared error as the regularization parameter lambda changes.

Interpret 1.3 and 1.4 Results: Lasso regression model & cross-validation to predict the time to recurrence

The optimal value of lambda λ, which in this model is 4.0450303, was found through the five-fold cross-validation and is the amount of penalty applied during the Lasso process. The penalty's purpose is to prevent overfitting by discouraging overly complex models. A lambda value of 4.0450303 suggests that the model requires a moderate level of penalization to balance the trade-off between bias and variance, minimizing the prediction error on new data.

The coefficient matrix indicates the effect size of each predictor variable on the outcome when the model is regularized by the optimal lambda. In this case:

The intercept 54.189931 is a constant term representing the baseline level of the response variable (time to recurrence) when all predictors are at their reference levels are zero.

A negative coefficient for Mean_Radius -1.8471901 implies that an increase in the mean radius is associated with a decrease in the time to recurrence. This could be interpreted as larger tumors being more aggressive and likely to recur sooner.

A positive coefficient for Mean_Smoothness 44.4578869 suggests that smoother tumors are associated with a longer time to recurrence. This might indicate that tumors with a smoother surface are less aggressive.

Similarly, a positive coefficient for Mean_Symmetry 3.4811757 implies that more symmetrical tumors may recur later than less symmetrical ones. Again, this might indicate that tumors that are more symmetrical are less aggressive.

The rest of the predictor variables (which showed dots (.) in the R output) indicate that their coefficients have been reduced to zero, effectively removing them from the model and reducing the risk of overfitting. This is a result of the Lasso penalization, which has determined that they do not contribute significantly to the model after accounting for the penalty on complexity.

It is important to note that in the Lasso regression output of coefficients, the reference category for the categorical variable 'Lymph_Node_Status' is not displayed. This is because the model uses one level of the categorical variable as a baseline to which the other levels are compared, and this baseline is integrated into the model's intercept. This practice helps to prevent multicollinearity by avoiding the inclusion of overly correlated variables, known as the 'dummy variable trap'. In our model, the '0' category of 'Lymph_Node_Status' is the reference level. As a result, the coefficients for the levels '1-3' and '4 or more' represent the change in the response variable relative to this reference level.

In summary, the Lasso model has identified Mean_Radius, Mean_Smoothness, and Mean_Symmetry as significant predictors of the time to recurrence and mean that they may play a big role in affecting breast-patient outcomes.

Question 2 Classification (7 points)

Table 2: Area Under the Curve (AUC) Results for from Lasso and Tree Classification Models

Iteration	AUC Lasso	AUC Tree
1	0.7261905	0.4739975
2	0.6405844	0.5662338
3	0.5850000	0.5002778
4	0.6375758	0.5045455
5	0.6497890	0.5411392

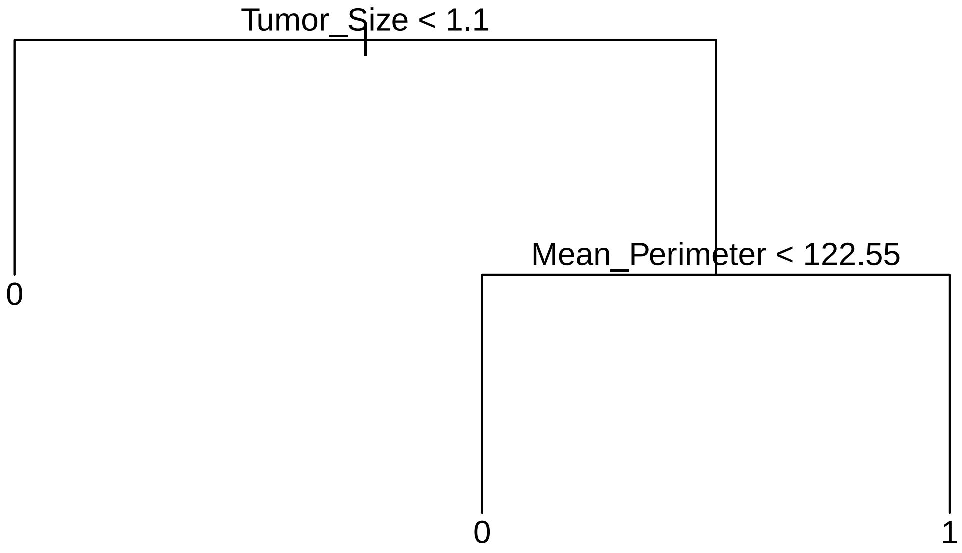

Figure 7: Visualization of full pruned classification tree. Nodes represent the conditions that split the data, and leaves represent the outcome classifications. The tree is pruned to avoid overfitting and to improve the model's generalization to unseen data.

Interpret 2.1 and 2.2

The study set out to develop predictive models for a binary outcome indicating patient recurrence, using a dataset with 194 records and 13 predictors after the exclusion of non-predictive attributes such as ID and Time. The predictors encompassed various tumor characteristics and lymph node statuses, with the outcome recoded to binary format, where ‘1’ signified recurrence and ‘0’ indicated no recurrence. An initial examination of the class distribution of “Outcome” showed 148 cases without recurrence and 46 cases with recurrence.

Dataset and Preliminary Analysis: The distribution of lymph node status was found to be balanced, with no need for collapsing levels. A 50/50 split of the dataset was implemented, ensuring equal and balanced distribution in both training and test sets.

LASSO for Classification:: The LASSO model was employed for classification first. The optimal lambda value, determined via cross-validation, was 0.00033. This value was pivotal in generating predictions on the test set. Model performance was asssessed using Area under the curve (AUC) The higher AUC values for the lasso indicates a better performance compared to the tree and suggests that the lasso model can better differentiate between patients with recurrence and those without.

Classification Tree (CART): The classification tree model was pruned to avoid overfitting, with the optimal size determined through cross-validation to be 10 terminal nodes. The performance of the pruned tree on the test set yielded lower AUC values as shown in table 2, indicating that the model’s ability to distinguish between the two outcomes is less effective than the lasso classification model.

Pruned Tree (Figure 7): The pruned tree consists of two primary predictors, “Tumor_Size” and “Mean_Perimeter,” which play a significant role in classifying patients into recurrence and non-recurrence categories. By considering the splits in the tree, healthcare providers can make informed decisions on treatments. For example, patients with smaller tumor sizes (less than 1.1) may be at a lower risk of recurrence, while additional considerations, such as “Mean_Perimeter,” are taken into account for patients with larger tumor sizes.

Model Comparison and Conclusion:

The performance of the LASSO and CART models was compared using the validation set method. We performed five different iterations, each with a different random split of the data. Lasso classifcation emerged as the superior model, with a higher AUC compared to the classification tree model. This suggests that the lasso model has a better predictive performance and could be considered a more reliable tool for clinical decision-making in this context. The consistent superiority of the lasso regression model across the splits would further validate its robustness as the preferred method for this predictive task.

Question 3 Survival Analysis (6 points)

3.1

Censoring occurs when the information about an event of interest is incomplete. In the context of time-to-event data, such as time to recurrence of a disease, censoring is present when the event (recurrence) has not occurred for some subjects during the study period. There are several reasons for censoring. The most common type, which is right censoring, happens when the study ends before the event occurs, or the subject leaves the study early. Left censoring occurs when a subject has already experienced the event before the study begins. Finally, interval censoring is when the event occurs in a time interval between two observation points. **For our data, the censored observations are those where the patient did not have a recurrence during the study period. The variable “Outcome” indicates whether recurrence has occurred (R) or not (N). When Outcome is N (or 0), it represents a censored observation because the event (recurrence) has not been observed.**

3.2

a)

The Kaplan-Meier survival analysis provides insights into the distribution of time to recurrence among breast cancer patients post-surgery. The median time to recurrence was calculated at 16.5 months. This value represents the point at which half of the patients experienced a recurrence, indicating a substantial risk within the first year and a half post-treatment.

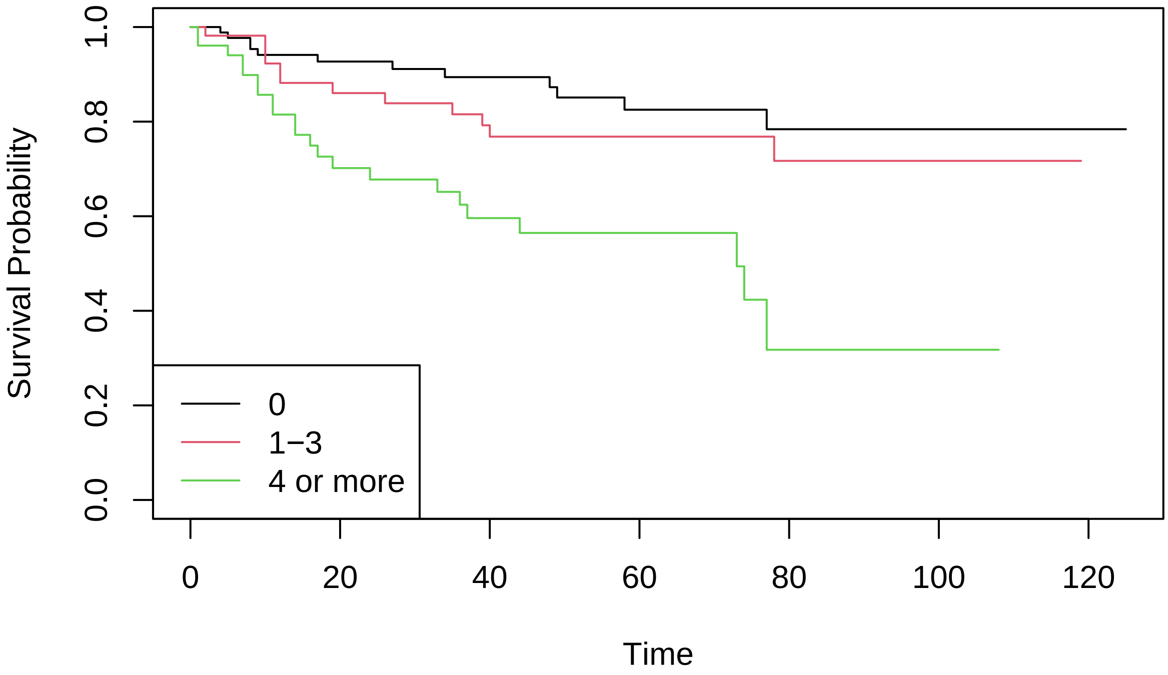

Figure 8: Kaplan-Meier Survival Curves for Different Lymph Node Status Categories .This figure displays Kaplan-Meier survival curves for different lymph node status categories (0, 1-3, 4 or more) in the context of breast cancer recurrence. The x-axis represents time, and the y-axis represents the estimated survival probability.

b)

Kaplan-Meier survival curves were generated to observe the differences in survival probabilities across different categories of axillary node status: ‘0’, ‘1-3’, and ‘4 or more’, whcih reflect the extent of cancer spread to axillary lymph nodes.

Based on the Kaplan-Meier survival curve graph, the curves revealed distinct patterns for each category. Patients with ‘0’ lymph node status showed the highest survival probability, suggesting a lower risk of recurrence. In contrast, those with ‘4 or more’ lymph nodes affected exhibited a significantly lower survival probability, underlining a higher risk of recurrence. The ‘1-3’ category showed intermediate survival probabilities.

c)

A Log-Rank test was performed to statistically compare the survival functions across the three levels of axillary node status. The test yielded a chi-square value of 18.3 on 2 degrees of freedom with a p-value of approximately 1×10ˆ-4, indicating a significant difference in survival distributions among the groups. This result emphasizes that axillary node status is a strong predictor of recurrence, with more affected nodes correlating with higher recurrence rates.

Table 3: Log-Rank Test Output

Lymph Node Status	N	Observed	Expected	(O-E)ˆ2/E	(O-E)ˆ2/V
0	87	12	21.4	4.125	7.762
1-3	56	12	14.4	0.391	0.574
4+	51	22	10.2	13.523	17.682

3.3

The Cox Proportional Hazards Model was utilized to identify predictors associated with the hazard of breast cancer recurrence, using the same variables as in the previous analyses. The findings are shown below:

Mean Radius and Mean Perimeter: A significant negative coefficient for Mean Radius and a positive one for Mean Perimeter suggest contrasting effects on recurrence risk. These features may reflect tumor characteristics influencing recurrence patterns. Mean Smoothness and Mean Fractal Dimension: Both showed significant coefficients, indicating their importance in predicting recurrence risk, potentially due to their association with tumor texture and complexity. Lymph Node Status: Patients with '4 or more' lymph nodes affected had a hazard ratio of 4.343, significantly higher than those with fewer affected nodes. This reaffirms the critical role of lymph node involvement in recurrence risk.

The Cox model showed a concordance index of 0.736, indicating a good predictive ability. The likelihood ratio, Wald, and Score tests all yielded significant p-values, confirming the model's robustness.