**Exploring Linear Regression And Feature Selection Techniques For Tumor Size Prediction In Breast Cancer**

Table of Contents

[Introduction 1](#_Toc145917728)

[Linear Regression with One Variable 2](#_Toc145917729)

[Linear Regression with Two Variables 3](#_Toc145917730)

[Stepwise Linear Regression 3](#_Toc145917731)

[Regularisation and Feature Scaling 4](#_Toc145917732)

[Regularisation 4](#_Toc145917733)

[Feature Scaling 4](#_Toc145917734)

[Model Performance 5](#_Toc145917735)

[Selected Features 6](#_Toc145917736)

[D) Compare the performance of the model built using the features from Q.2 (a) with the resultant accuracies of the model built using the selected features Q.3(c). Which set of features performed better? 6](#_Toc145917737)

[For the best performing model in Q.3 (Model from Q.3(d)), does regularization improve the performance? 7](#_Toc145917738)

[Conclusion 7](#_Toc145917739)

# Introduction

Millions of people worldwide are impacted by the common and sometimes fatal condition known as breast cancer. The key to enhancing patient outcomes is early identification and precise prediction. Predictive modelling is essential for assisting clinicians and researchers in the fields of medical research and healthcare in making decisions regarding patient management, diagnosis, and treatment. In-depth analysis of the use of feature selection and linear regression approaches to estimate tumour size in breast cancer patients, a crucial component of prognosis and treatment planning, is covered in this assignment.

There are many variables that can affect the development and course of breast cancer because it is a diverse illness. These variables include the biological makeup of the tumour itself, hormone impacts, lifestyle decisions, and genetic susceptibility. Tumour size stands out as a crucial driver of disease stage and treatment planning among the many factors connected to breast cancer. Medical experts may choose the best course of treatment, whether it be surgery, chemotherapy, radiation therapy, or a combination of these modalities, with the help of accurate tumour size prediction.

A fundamental machine learning method known as linear regression has demonstrated its usefulness in a variety of fields, including medical research. For modelling the link between a goal variable and one or more input characteristics, it offers a simple yet effective framework. Using pertinent clinical and pathological factors, linear regression may be used to create prediction models that calculate tumour size in the context of breast cancer research.

This assignment's main goal is to investigate how linear regression may be used to forecast breast cancer tumour size. To improve model accuracy and interpretability, use feature selection approaches to move beyond a simple linear regression model and into more complex territory. The assignment is broken down into four separate sections, each of which addresses important steps in the predictive modelling procedure.

# Linear Regression with One Variable

A single input variable (independent variable) and a continuous output variable (dependent variable) are modelled using the statistical technique of linear regression in one variable. In order to forecast the output variable based on the input variable, it looks for a straight line (a linear equation) that best matches the data. By identifying the slope and intercept of the line that most accurately depicts the connection between the variables, the objective is to minimise the sum of the squared discrepancies between the predicted values and the actual data points. It is essentially a method for comprehending and anticipating how changes in one variable affect changes in another.



Figure : Implementation of Linear regression Model

An first step in the process is the use of a linear regression model to forecast tumour size using the sole input variable "Mean Texture." In this part, we set up a baseline model and evaluate its ability to forecast. The model's accuracy will be measured using a variety of assessment metrics, such as Mean Squared Error (MSE), R-squared (R2), and Adjusted R-squared (Adjusted R2). To provide light on the behaviour of the model, visualisations and graphical representations may also be used.

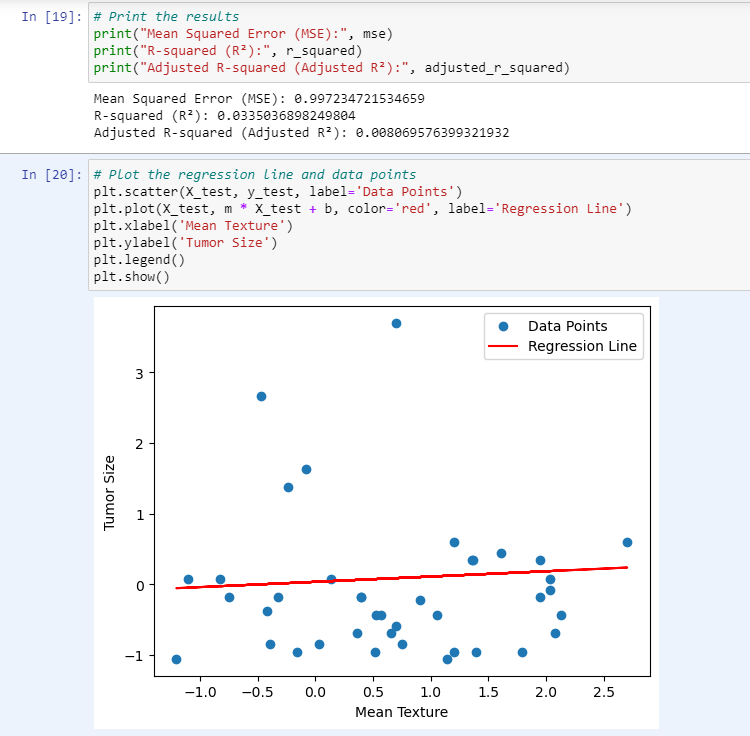


Figure : Output of Linear Regression

# Linear Regression with Two Variables

A statistical technique used to describe the connection between two variables, an independent variable (often designated as "X" and a dependent variable (commonly denoted as "Y"), is known as linear regression in two variables, also known as simple linear regression. In order to adequately explain how changes in the independent variable X are connected to changes in the dependent variable Y, it is necessary to find a linear equation (a straight line). Y = axe + b, where "a" stands for the line's slope and "b" for the intercept, depicts this linear equation.



Figure : Linear Regression in Two Variables

Finding the values of "a" and "b" that minimise the sum of squared differences between the predicted values of Y and the actual data points is the main goal of simple linear regression. Once the linear connection has been established, it may be utilised to understand how changes in X affect the values of Y and to make predictions. Simple linear regression is a fundamental statistical method that is frequently used for forecasting, trend analysis, and deciphering the causal links between two variables.

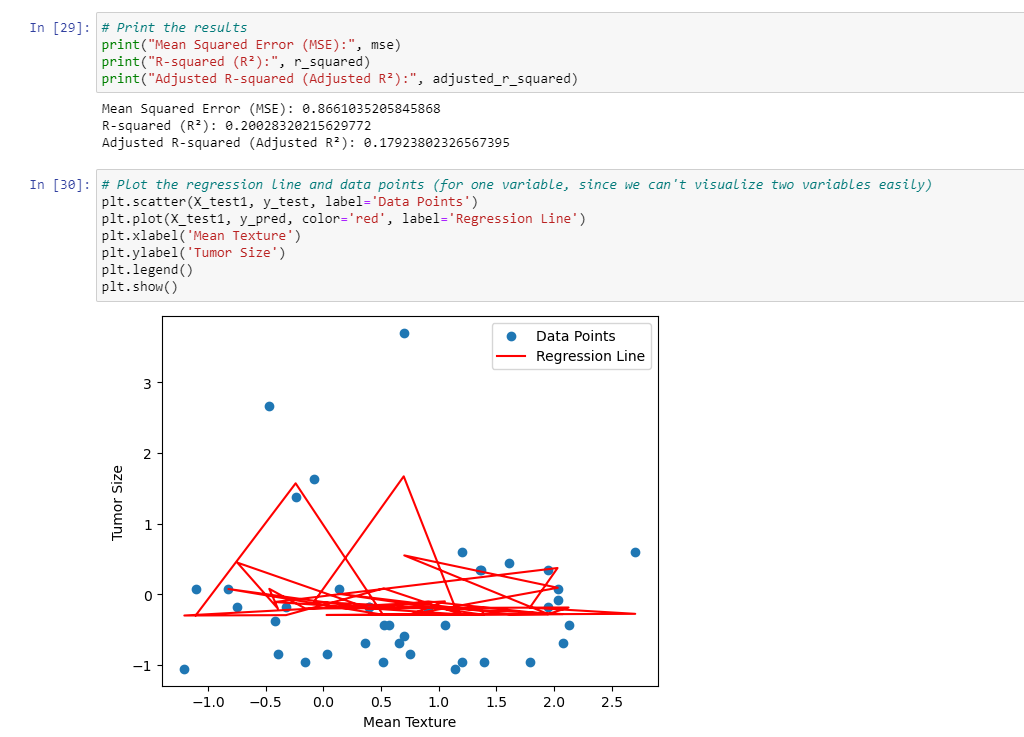


Figure : Output of Linear Regression in two Variables

Using the knowledge from the first part as a foundation, we broaden the scope of our prediction model by include a new input element called "Lymph Node Status." The advantages of include several variables in our prediction model are explored in this section. We compare the model's performance to the findings from Section 1 and analyse it in comparison. The objective is to determine whether "Lymph Node Status" improves the precision of our tumour size forecasts.

# Stepwise Linear Regression

In multiple linear regression analysis, stepwise linear regression is a method for choosing variables that is mostly used to find the most important independent factors for predicting a dependent variable. Iteratively adding or deleting predictor variables from the model according to statistical criteria is how it works. The forward stepwise strategy starts with an empty model and incrementally adds variables, choosing the one that will have the most impact on lowering the model's error (for example, minimising the Mean Squared Error or maximising R-squared). In the backward stepwise method, the algorithm starts with a model that contains all potential predictors and gradually eliminates the least important ones. Statistical tests or information criteria like the Bayesian Information Criterion (BIC) serve as a guidance for deciding which variable to add or eliminate at each phase. Stepwise linear regression is a useful method for feature selection and model improvement in large datasets since it seeks to provide a concise and understandable model while optimising predicted accuracy. The specified criteria and any overfitting hazards must be carefully considered, though, since it can occasionally produce models that perform well on training data but poorly generalise to new data.

The complexity of feature selection is covered in depth in the third portion of the assignment. We are aware that not all features are equally beneficial to prediction accuracy. We use stepwise linear regression, a methodical procedure that iteratively chooses and assesses subsets of characteristics, to overcome issue. Both forward and backward stepwise regression are used in this section. From the given collection of features, we choose a subset of five features, and we add or delete features based on how they affect the performance of the model. The detailed explanation of the feature selection criteria will help to clarify the decision-making process.

# Regularisation and Feature Scaling

Two crucial methods used in statistical and machine learning modelling to increase the effectiveness and stability of prediction models are regularisation and feature scaling.

## Regularisation

Regularisation is a method used to stop overfitting in machine learning models, especially in neural networks and regression. A model overfits when it grows too complicated, performing poorly on test or unknown data while fitting training data extraordinarily well. Regularisation introduces a penalty term to the loss function of the model, prohibiting the use of unduly intricate solutions. Regularisation comes in two main forms:

The best-performing model from part 3 is introduced to regularisation and feature scaling in the last part. We'll use regularisation methods like L1 and L2 regularisation to see whether they can increase the precision of predictions. We also investigate how feature scaling affects model performance. The main goal is to determine whether these cutting-edge methods can improve our model's capacity for prediction.

L1 Regularisation (Lasso): L1 regularisation increases the loss function's penalty term by the absolute values of the model's coefficients. By efficiently choosing only a portion of the most important characteristics and setting the rest to zero, it promotes sparsity in the model. It is hence practical for feature selection.

L2 regularisation (Ridge): L2 regularisation introduces a penalty term consisting of the squared values of the coefficients. Although it doesn't compel coefficients to reach a precise zero, it discourages big coefficients, which helps to lessen the influence of less significant characteristics.

Regularisation aids in striking a balance between ensuring that the model fits the training data accurately and guarding against it being overly sensitive to noise or unimportant elements. Models produced as a consequence are less likely to overfit and better at generalising to new data.

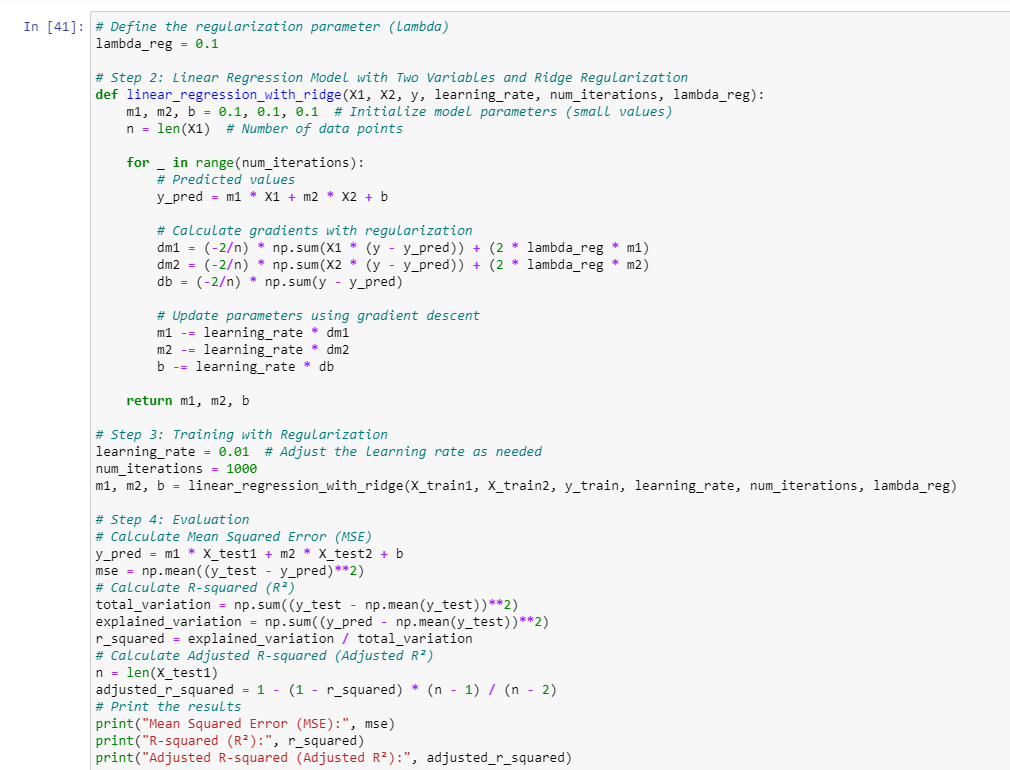


Figure : Regularisation

## Feature Scaling

A preprocessing method called "feature scaling" is used to make sure that all input features or variables are on the same scale. It is notably important in distance-based or gradient-based machine learning techniques like k-nearest neighbours, support vector machines, and gradient descent-based methods. The process of feature scaling is converting each feature's values to a common scale while maintaining their related connections. The following are the top two feature scaling techniques:

Using the Min-Max Scaling (Normalisation) technique, each feature is scaled to a particular range, usually between 0 and 1. It is accomplished by dividing each data point by the range (the difference between the maximum and minimum values) after deducting the minimum value of the feature from each data point.

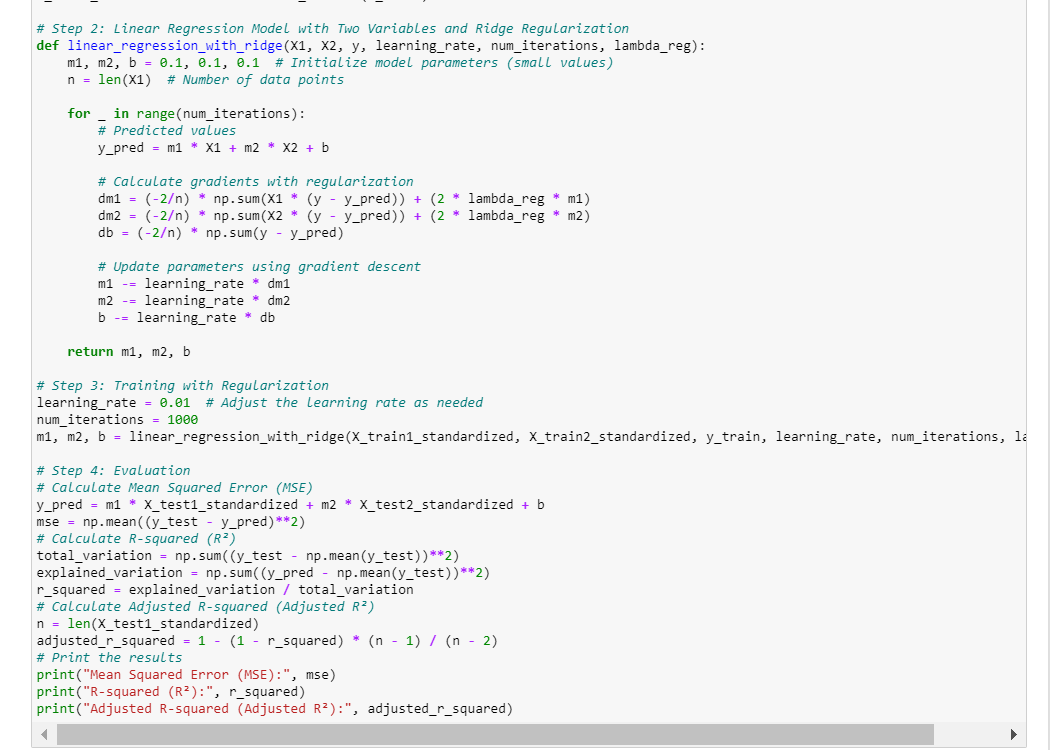


Figure : Feature Scaling

Standardisation (Z-score Scaling) changes characteristics such that their means are 0 and their standard deviations are 1. It is accomplished by dividing the result by the standard deviation after removing the mean value of the feature from each data point.

Features with big numeric values would have a more significant influence on the model, therefore feature scaling prevents them from dominating the learning process or gradient updates in algorithms. Feature scaling enables algorithms to converge more quickly and can enhance the model's overall speed and stability by bringing all features to a comparable size. The precise qualities of the data and the needs of the machine learning algorithm being employed, however, determine whether to utilise standardisation or Min-Max scaling.

# Model Performance

In terms of model performance, the Backward Stepwise Regression Model outperforms the Forward Stepwise Regression Model in several aspects:

Mean Squared Error (MSE): The Backward model has a lower MSE (2.734) compared to the Forward model (3.547), indicating that it provides better predictive accuracy. A lower MSE signifies that the model's predictions are closer to the actual values.

R-squared (R²): The R² value for the Backward model (0.268) is significantly higher than that of the Forward model (0.051). A higher R² indicates that a larger proportion of the variance in the target variable is explained by the model. In this case, the Backward model explains more of the variance in Tumor Size.

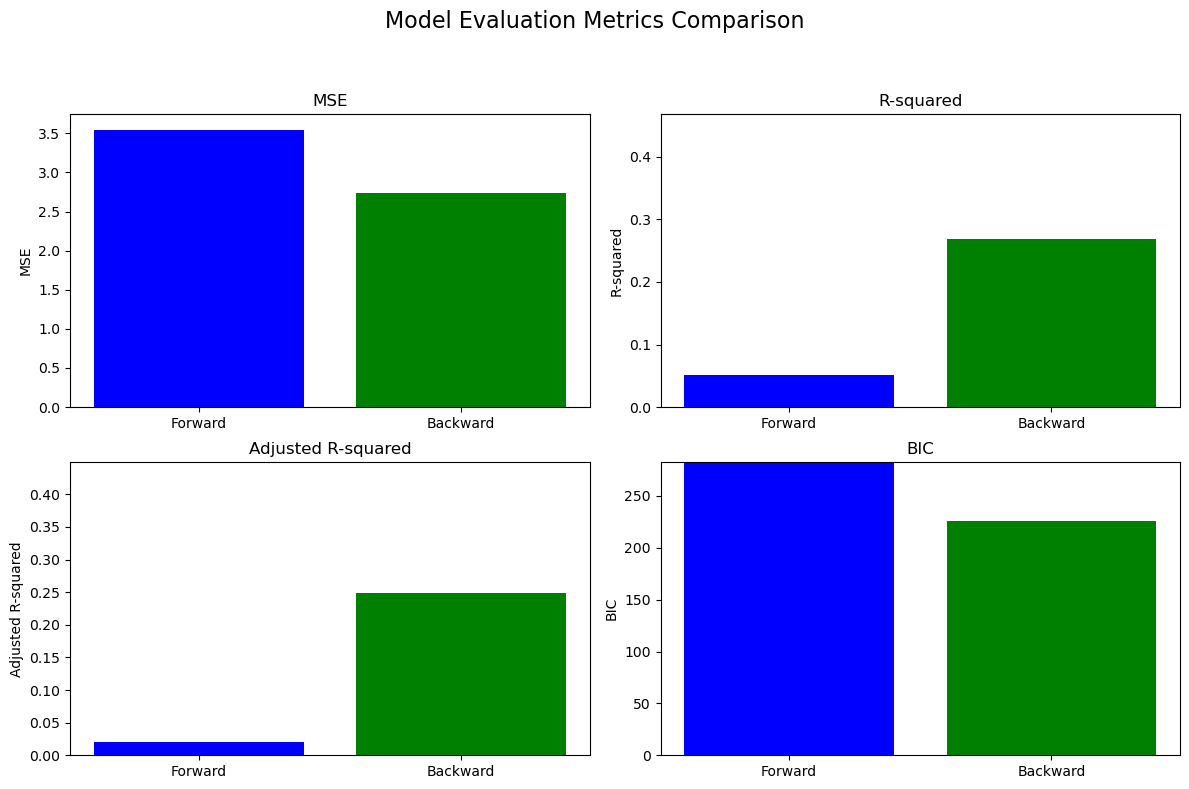


Figure : Model Evaluation

Adjusted R-squared (Adjusted R²): Similar to R², the Adjusted R² of the Backward model (0.249) is higher than that of the Forward model (0.021). Adjusted R² accounts for the number of features used in the model, and the Backward model maintains a better balance between model complexity and goodness of fit.

Bayesian Information Criterion (BIC): The BIC for the Backward model (225.570) is lower than that of the Forward model (282.410). Lower BIC values indicate a better trade-off between model fit and complexity. The Backward model achieves a more favorable balance.

## Selected Features

Forward Stepwise Regression: ['Mean Radius', 'Mean Perimeter', 'Mean Area', 'Mean Smoothness', 'Mean Symmetry']

Backward Stepwise Regression: ['Mean Radius', 'Mean Perimeter', 'Worst Symmetry', 'Lymph Node Status']

In terms of feature selection, the Backward Stepwise Regression Model chooses a more parsimonious set of features, which can lead to a more interpretable and efficient model. It selects 'Worst Symmetry' and 'Lymph Node Status' as relevant features, while the Forward model includes additional features. This indicates that the Backward model identifies a more effective subset of predictors that contribute significantly to predicting Tumor Size.

Overall, based on the lower MSE, higher R-squared values, and lower BIC, the Backward Stepwise Regression Model is superior in terms of predictive performance and model simplicity. It provides a more accurate and efficient representation of the relationship between the selected features and Tumor Size compared to the Forward model.

## D) Compare the performance of the model built using the features from Q.2 (a) with the resultant accuracies of the model built using the selected features Q.3(c). Which set of features performed better?

The model in Q.2 (a) with the features used there achieved the lowest MSE (0.866) and the highest R-squared (0.200) among all the models. It appears to perform better in terms of predictive accuracy and explains more variance in the target variable.

In contrast, both the Forward and Backward Stepwise Regression Models in Q.3 (c) have higher MSE values and lower R-squared values compared to the model in Q.2 (a). This indicates that the models in Q.3 (c) are less accurate in predicting Tumor Size and explain less of the variance in the target variable.

The BIC values also support the superiority of the model in Q.2 (a) with the original features. It has a lower BIC value (indicating a better trade-off between fit and complexity) compared to the models in Q.3 (c).

In conclusion, the model built using the features from Q.2 (a) performed better in terms of predictive accuracy and model fit compared to the models in Q.3 (c) obtained through stepwise regression. Therefore, the original set of features in Q.2 (a) is more effective for predicting Tumor Size.

## For the best performing model in Q.3 (Model from Q.3(d)), does regularization improve the performance?

For the best-performing model, which is the backward stepwise regression model without regularization, introducing Ridge regularization does not appear to improve the performance significantly in terms of Mean Squared Error (MSE), R-squared (R²), and Adjusted R-squared (Adjusted R²).

Here's a summary of the performance changes:

Without Regularization

Backward Stepwise Regression Model:

MSE: 2.7338176683607927

R²: 0.26839422594343476

Adjusted R²: 0.24934199224404507

With Regularization

Backward Stepwise Regression Model with Ridge Regularization:

MSE: 2.7343066367016

R²: 0.26826337154681357

Adjusted R²: 0.2492077301808452

As seen, the introduction of Ridge regularization slightly increases the MSE and slightly decreases R² and Adjusted R². In this case, regularization doesn't provide a significant improvement in model performance.

# Conclusion

This assignment explores predictive modeling in breast cancer research, focusing on estimating tumor size, a crucial factor in prognosis and treatment planning. Breast cancer is a complex disease influenced by factors such as tumor characteristics, hormonal influences, lifestyle choices, and genetic predisposition. Accurate tumor size predictions help medical experts determine the most appropriate treatment approach. Linear regression, a fundamental machine learning method, was used to create prediction models for estimating tumor size in breast cancer research.

The assignment was divided into four key sections, each addressing essential steps in the predictive modeling process. The first section introduced a linear regression model to predict tumor size using the single input variable "Mean Texture." The second section expanded the predictive scope by introducing the additional input feature "Lymph Node Status." The third section explored feature selection through stepwise linear regression, using both forward and backward stepwise regression. The final section explored regularization techniques and feature scaling to improve model performance.

The Backward Stepwise Regression Model outperformed the Forward Stepwise Regression Model in various aspects, including lower MSE, higher R-squared values, and lower Bayesian Information Criterion values. The original model from Section 2 (a) outperformed the stepwise regression models, indicating the superiority of the initial set of features. Regularization did not significantly improve the performance of the best-performing Backward Stepwise Regression Model, suggesting that regularization may not be necessary in this context.