**THE BREAST CANCER DIAGNOSTIC WISCONSIN**

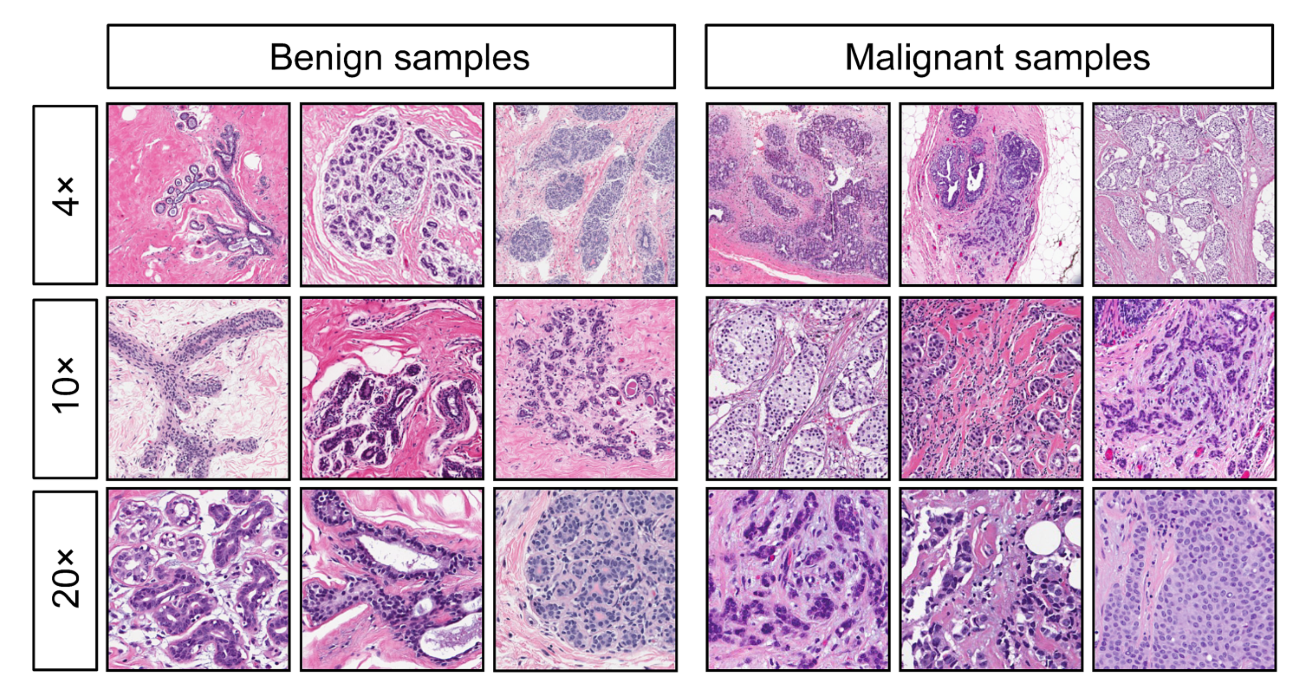
Manoj Kumar Surabhi

Sumanth Reddy Thandra

December 2024

**Abstract:**

Breast cancer, one of the most malignant types of cancers, has been seriously threatening both the physical and mental health of women in the world. However, this is curable if detected in an early-stage as a non-metastatic disease thus underlying the importance of early detection and need to conduct research in these lines. The Breast Cancer Diagnostic Wisconsin dataset provides a robust foundation for such investigations, comprising features derived from fine-needle aspiration (FNA) biopsies of breast masses. This paper will focus exclusively on predicting whether a tumour is benign or malignant. The relationship between predictors will be examined and variables will be excluded if not useful for prediction. Subsequently, both linear and non-linear models will be trained on the dataset using cross-validation methods. The top-performing models will be then tested on a separate test set, and the overall best model will be selected based on predictive Kappa and other relevant performance metrics.

****

**Table of contents:**

**1.Background:**

Breast cancer is a type of cancer that develops in the cells of the breast, commonly forming in the milk ducts. It can spread to other parts of the body if untreated. The disease arises from abnormal growth of breast cells, influenced by genetic mutations, hormonal changes, lifestyle factors, or family history. Breast cancer is one of the most common cancers globally, primarily affecting women, though men can also develop it.

The Wisconsin Breast Cancer Dataset (WBCD) provides valuable insights into the diagnosis of breast cancer. Created by Dr. William H. Wolberg at the University of Wisconsin Hospitals in 1992, this dataset contains nuclear features derived from fine needle aspiration cytology (FNAC) biopsy tests of patients' breasts. Early detection and appropriate treatment, such as surgery, radiation, chemotherapy, hormonal therapy, or targeted therapy, are crucial in improving patient outcomes.

To this end, the early identification of BC through screening and detection methods is important so that the disease is identified during its initial stages when malignant cells are local to the breast(s) only. If not caught early, the malignant BC cells spread to other parts of the body and patients are often subjected to much more complex, invasive treatments. Patients with late diagnosis often have lower survival rates and may die soon after being diagnosed.

**2.Variable Introduction and Definitions**

The data was provided by the University of Wisconsin Hospitals and recorded in 1992 by Dr. William H. The target variable is the diagnosis, which has two classes: Benign and Malignant. Benign implies that a tumour is non-cancerous, while Malignant implies that it is cancerous.

Ten features are computed from a digitized image obtained from a fine needle aspirate (FNA) of a breast mass. These features describe the characteristics of the cell nuclei present in the image. For each image, the mean, standard error, and "worst (the largest, calculated as the mean of the three largest values) of these features were computed, resulting in a total of **32** features and 569 observations.



|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| diagnosis | Indicates whether a tumour is benign or malignant |
| radius\_mean | mean of distances from center to points on the perimeter |
| texture\_mean | Mean of gray-scale values |
| Perimeter\_mean | Mean of perimeter |
| area\_mean | Mean of the area |
| Smoothness\_mean | Mean of local variation in radius length |
| Compactness\_mean | Perimeter\_mean^2/area\_mean-1.0 |
| concavity\_mean | Severity of concave portions of the contour |
| concave\_points\_mean | Number of concave portions of the contour |
| symmetry\_mean | Mean of the symmetry |
| fractal\_dimension\_mean | “Coastline approximation – 1” |
| radius\_se | Standard error of distances from center to points on the perimeter |
| texture\_se | Standard error of gray-scale values |
| Perimeter\_se | Standard error of perimeter |
| area\_ se | Standard error of the area |
| Smoothness\_ se | Standard error of local variation in radius length |
| Compactness\_ se | Perimeter\_se ^2/area\_se-1.0 |
| concavity\_ se | Severity of concave portions of the contour |
| concave\_points\_ se | Number of concave portions of the contour |
| symmetry\_ se | Standard error of the symmetry |
| fractal\_dimension\_ se | “Coastline approximation – 1” |
| radius\_worst | worst of distances from center to points on the perimeter |
| texture\_ worst | worst of gray-scale values |
| Perimeter\_ worst | worst of perimeter |
| area\_ worst | worst of the area |
| Smoothness\_ worst | worst of local variation in radius length |
| Compactness\_ worst | Perimeter\_ worst ^2/area\_ worst -1.0 |
| concavity\_ worst | Severity of concave portions of the contour |
| concave\_points\_ worst | Number of concave portions of the contour |
| symmetry\_ worst | Mean of the symmetry |
| fractal\_dimension\_ worst | “Coastline approximation – 1” |

Based on the relationships between the predictors, highly correlated variables may be removed, or PCA (Principal Component Analysis) may be applied, depending on the model requirements. Additionally, the data will be examined for skewness, outliers, near-zero variance, and missing values. After these steps, both linear and nonlinear models will be developed and tested on the dataset.

**3. Preprocessing of the predictors**

Before building any model, it is crucial to preprocess the data. Some preprocessing steps include identifying correlations, applying transformations, creating dummy variables, and imputing missing values. For this dataset, there is no need to create dummy variables since all the variables are continuous. Additionally, there are no missing values in the observations, and no near-zero variance variables were found. Two redundant variables, named id and x, were removed, leaving 30 predictors.

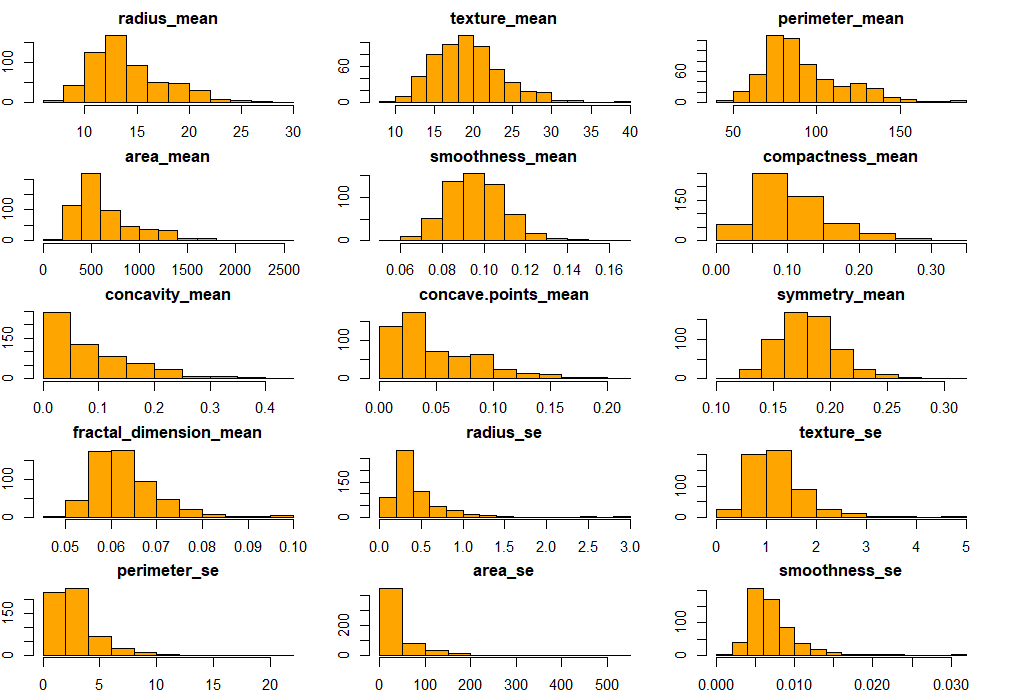
1. **Missing values and near- zero variances**

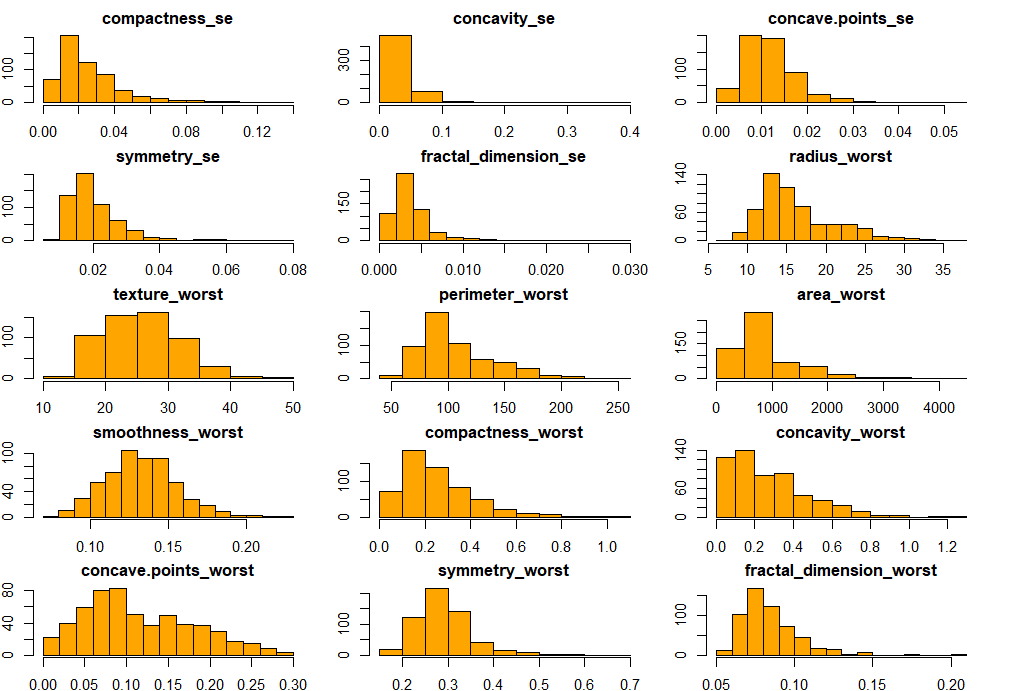
Before training a model on the data, it is essential to check for missing values. If any are found, they must be imputed. For this dataset, no missing values were found. Additionally, near-zero variance variables do not provide useful information and can lead to overfitting. Upon inspection, near-zero variance variables were identified.

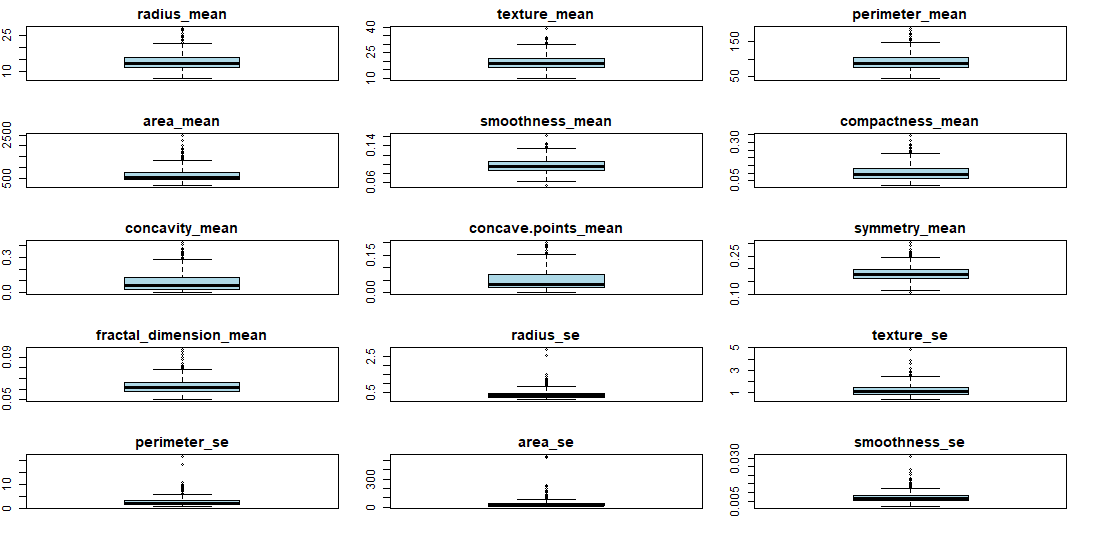
1. **Transformations**

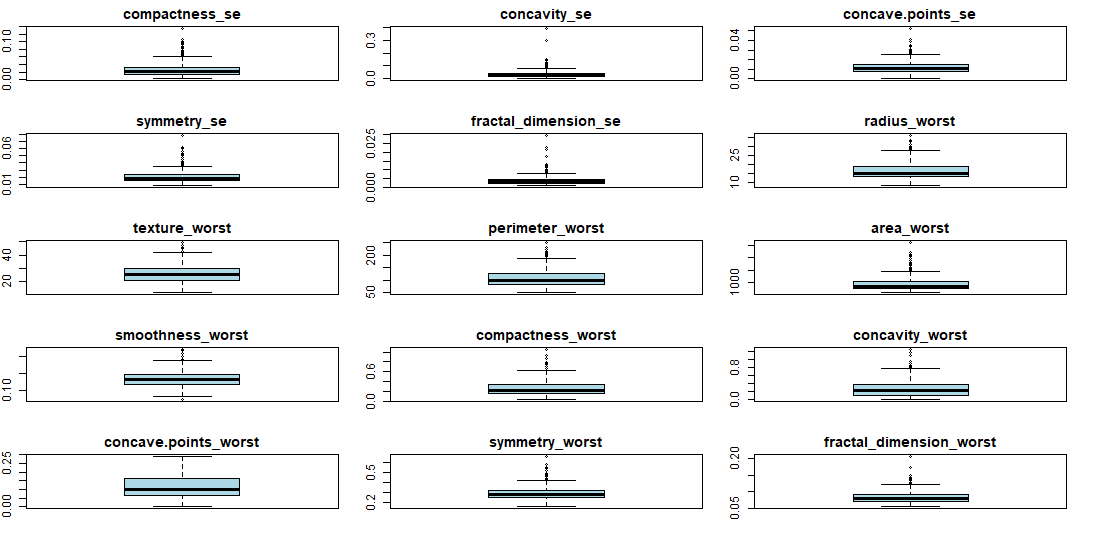
Skewness and outliers in the data lead to reduced predictive performance and bias. Applying transformations to the data can improve performance. Histograms and boxplots are the best tools for visualizing skewness and outliers in the predictors. The table, along with the histograms and boxplots below, displays the skewness and outliers of each continuous predictor before any transformations.



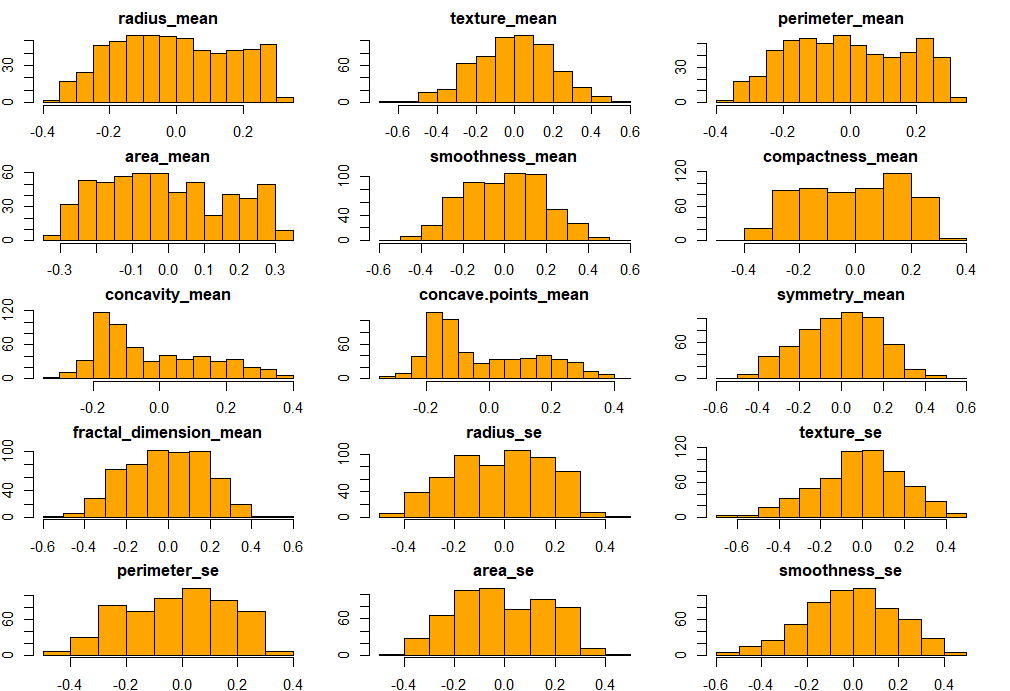


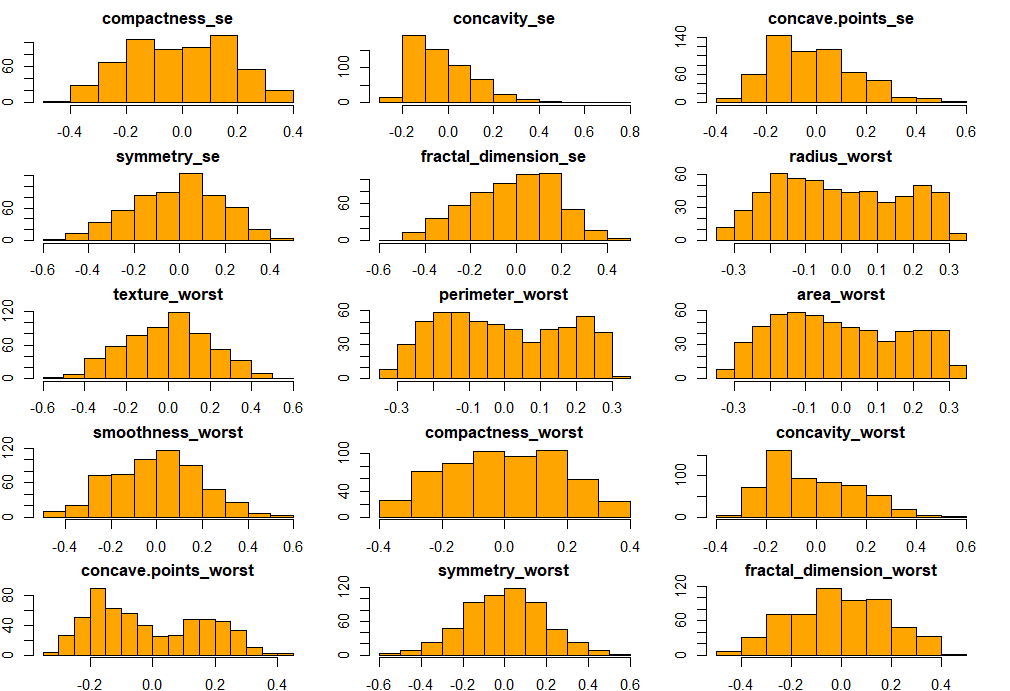


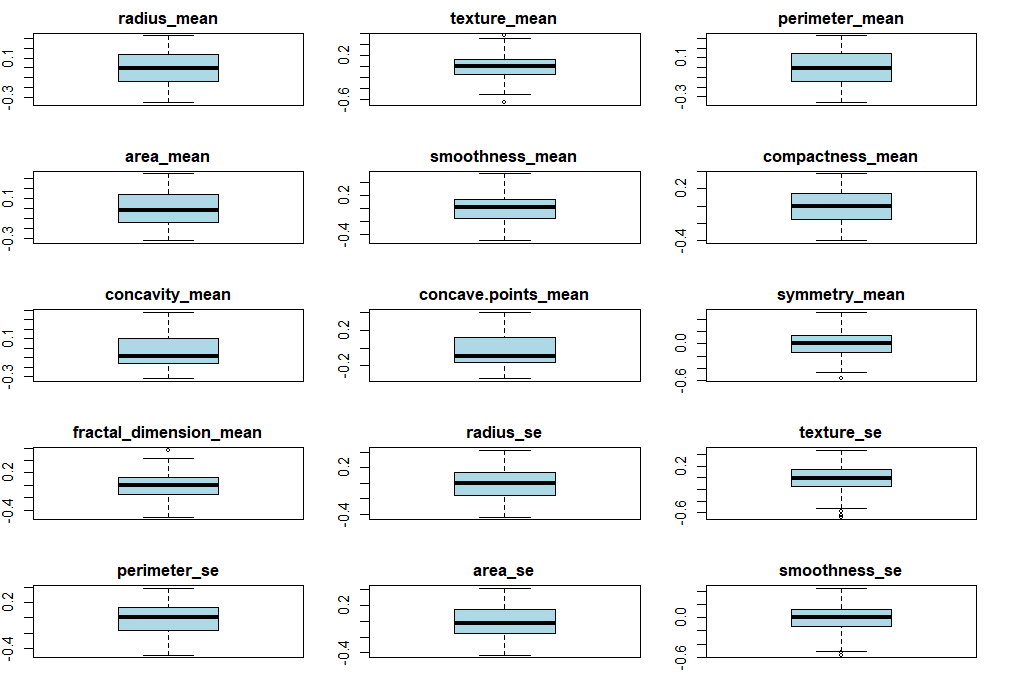


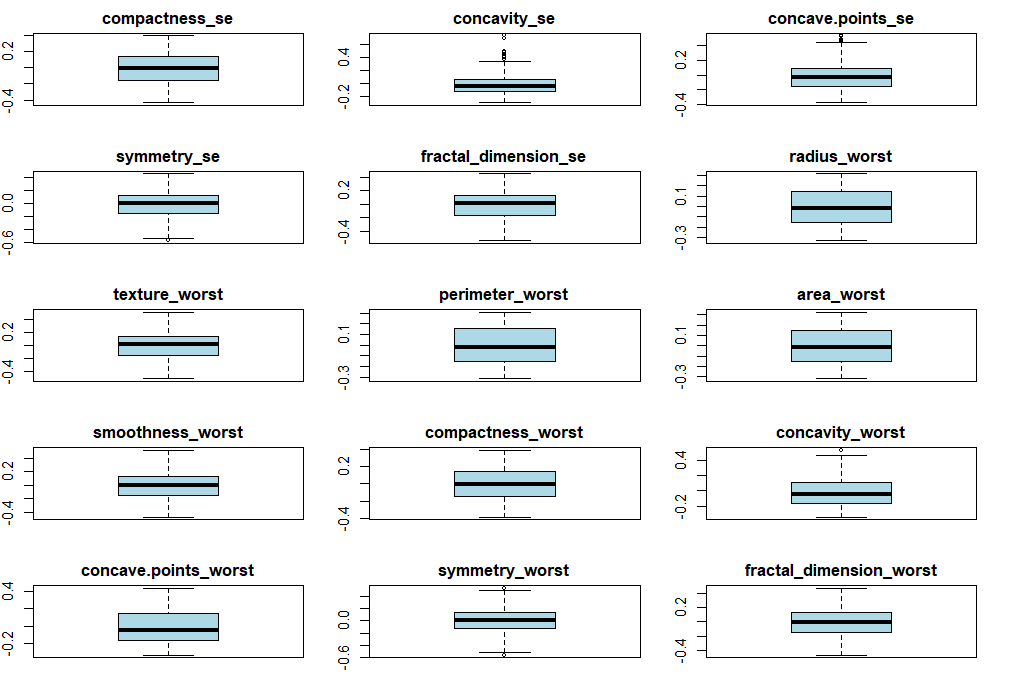


To account for the skewness and outliers of the predictors, Box-Cox transformation and spatial transformation will be applied. The histograms and boxplots below display the data after the transformations. It is evident that the skewness has been reduced, and fewer outliers are present.



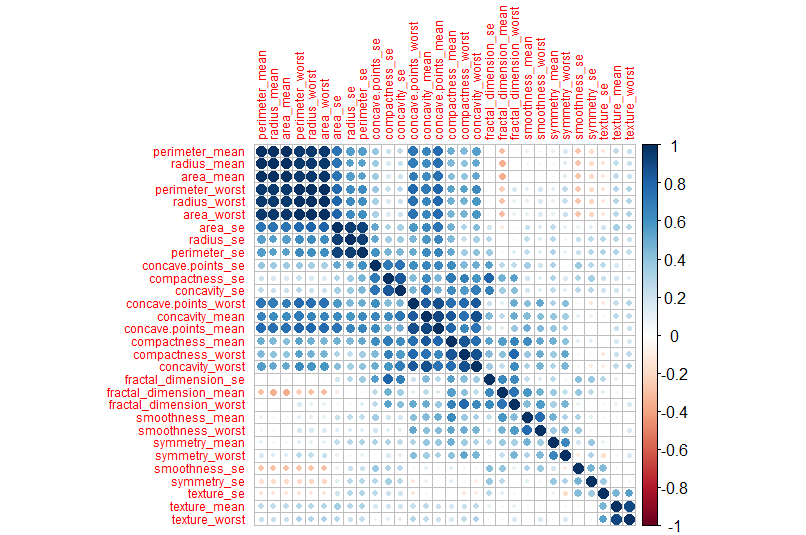






1. **Correlations**

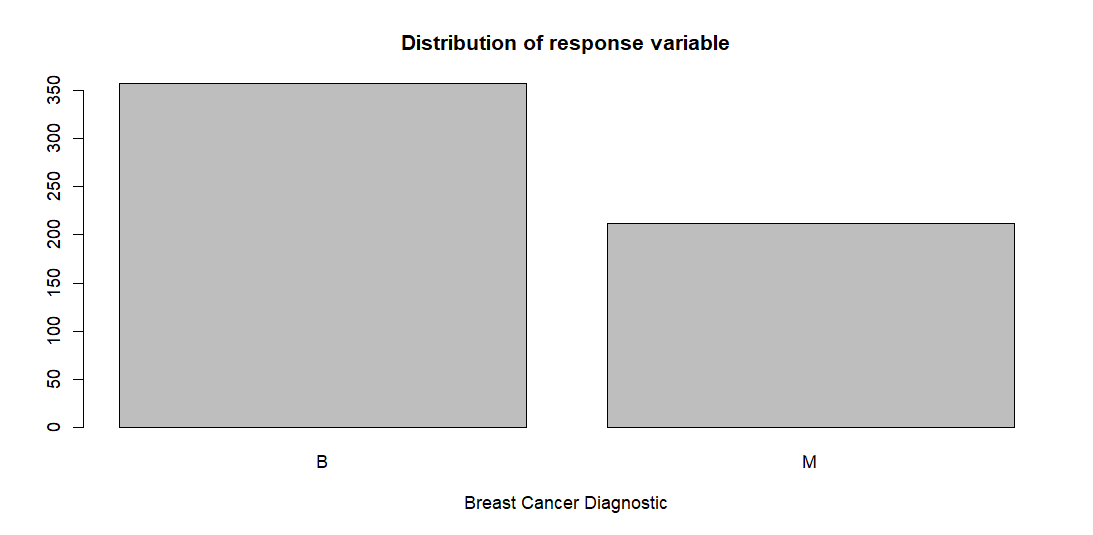
High collinearity between predictors can affect the performance of certain models. To address this, a correlation plot of the 30 remaining predictors was created. It was found that 12 variables have a high correlation, exceeding 0.9. The plot below illustrates the correlation between the predictors.



To address high correlation among predictors, PCA will be applied to certain models, including LDA, Logistic Regression, Naïve Bayes, KNN, and FDA. As part of the PCA process, the data will be automatically centered and scaled. For the remaining models, the data will be centered and scaled.

**4. Data Splitting**

The target variable is **diagnosis**, which has two classes: **Benign** and **Malignant**. The bar plot below illustrates the distribution of the diagnosis. It is evident that the two classes are imbalanced. To address this, stratified random sampling was applied. The data was split into 80% for training and 20% for testing, resulting in **456 samples for training** and **113 samples for testing**.



The two classes of diagnosis, **Benign** and **Malignant**, have **357** and **212** observations, respectively. To account for this class imbalance, stratified random sampling is employed for data splitting.