**Group Project**

**Breast Cancer**

Alejandra Barrios– 301314502

Paula Carrillo – 301314500

Sebastian Gonzalez – 301336125

Saksham Tiwari- 301334725

Centennial College

Business Analytics Capstone

David Parent

July 15, 2024

**Contents**

[Introduction 3](#_Toc171517848)

[Data Setup and Exploration 3](#_Toc171517849)

[Variable Analysis 5](#_Toc171517850)

[Descriptive Statistics 6](#_Toc171517851)

[Over sampling 8](#_Toc171517852)

[Decision Tree Model 9](#_Toc171517853)

[Maximal Tree 9](#_Toc171517854)

[Misclassification Tree 9](#_Toc171517855)

[Average Square Error Tree 9](#_Toc171517856)

[Optimal Tree 9](#_Toc171517857)

[Impute 10](#_Toc171517858)

[Adjust Outliers 10](#_Toc171517859)

[Skewness 11](#_Toc171517860)

[Replace Dummies 11](#_Toc171517861)

[Logistic Regression Model 11](#_Toc171517862)

[Full Regression 12](#_Toc171517863)

[Forward Regression 12](#_Toc171517864)

[Backward Regression 12](#_Toc171517865)

[Stepwise Regression 12](#_Toc171517866)

[Optimal Regression 12](#_Toc171517867)

[Neural Network Model 14](#_Toc171517868)

[Neural Network: Cap & Floor 14](#_Toc171517869)

[Neural Network: Transform Variables 15](#_Toc171517870)

[Neural Network: Best Regression (3H) 15](#_Toc171517871)

[Neural Network: 2H 15](#_Toc171517872)

[Neural Network: 4H 15](#_Toc171517873)

[Neural Network: 5H 15](#_Toc171517874)

[Neural Network: 6H 15](#_Toc171517875)

[Neural Network: 7H 15](#_Toc171517876)

[Neural Network: 8H 15](#_Toc171517877)

[Optimal Neural Network 15](#_Toc171517878)

[Assessment 16](#_Toc171517879)

[Conclusion and Recommendations 18](#_Toc171517880)

[Appendix 20](#_Toc171517881)

# Introduction

The objective of this project is to determine the characteristics of the patients who are more likely to die of breast cancer with a dataset extracted from Kaggle Website[[1]](#footnote-2).

The project has several steps including an initial exploratory analysis to understand a general panorama of the records; cleaning and managing outliers of the dataset; running different models such as decision trees, regressions, and Neural Networks to predict which patients are more likely to die according to different attributes of the dataset; comparing each model considering Average Squared Error (ASE) and ROC index; finally, choosing the best model using ASE as a validation assessment rating and interpret its results to make recommendations to the health industry.

# Data Setup and Exploration

The dataset chosen is from patients with breast cancer, it has a total of 4024 records with the following variables:

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Description** | **Role** | **Level** |
| Age | The age in years of each patient | Input | Interval |
| Race | The ethnicity or skin color of the patient (White, Black, Other) | Input | Nominal |
| Marital Status | Status of relationship of the patient (Divorced, Married, Separated, Single, and Widowed) | Input | Nominal |
| T Stage | Indicates the size of the main tumor. The higher the number after T, the larger the tumor  **T1**: Tumor is 2 cm (3/4 of an inch) or less across.  **T2**: Tumor is more than 2 cm but not more than 5 cm (2 inches) across.  **T3**: Tumor is more than 5 cm across.  **T4**: Tumor of any size growing into the chest wall or skin. This includes inflammatory breast cancer. | Input | Nominal |
| N Stage | Refer to the number and location of lymph nodes that contain cancer. The higher the number after N, the more lymph nodes that contain cancer.  **N1**: Cancer has spread to 1 to 3 lymph nodes under the arm with at least one area of cancer spread greater than 2 mm across.  **N2**: Cancer has spread to 4 to 9 lymph nodes under the arm, or cancer has enlarged the internal mammary lymph nodes.  **N3**: Cancer has spread to 10 or more axillary lymph nodes, with at least one area of cancer spread greater than 2 mm. | Input | Nominal |
| 6th Stage | **IIA:** The tumor is less than 2 centimeters and less than four axillary lymph nodes have cancer cells present.  **IIB:** The tumor is between 2 and 5 centimeters and has spread to less than four axillary lymph nodes.  **IIIA:** The tumor is larger than the approximate size of a small lime (over 5 centimeters), AND the cancer has spread to 1, 2, or 3 lymph nodes under the arm or near the breastbone.  **IIIB:** the tumor has grown into the muscles of the chest wall or skin.  **IIIC:** The cancer has spread to 10 or more axillary lymph nodes | Input | Nominal |
| Differentiate | How the cells look like. Going from well differentiated to undifferentiated. | Rejected | Nominal |
| Grade | **1**: The cancer cells are well differentiated. They look almost like normal cells.  **2**: The cancer cells are moderately differentiated. They are between grades 1 and 3.  **3**: The cancer cells are poorly differentiated or undifferentiated. They look less normal, or more abnormal, than healthy cells.  **anaplastic; Grade IV:** The cells look undifferentiated or abnormal. | Input | Nominal |
| A Stage | **Regional:** The cancer has spread outside the breast to nearby lymph nodes.  **Distant:** The cancer has spread to distant parts of the body (Such as lungs, liver, bones, etc.) | Input | Nominal |
| Tumor Size | Size of tumor in mm | Input | Interval |
| Estrogen Status | **Positive:** Cancer cells have receptors for estrogen.  **Negative:** Otherwise | Input | Nominal |
| Progesterone Status | **Positive:** Cancer cells have receptors for progesterone.  **Negative:** Otherwise | Input | Nominal |
| Regional Node Examined | Total number of regional lymph nodes that were removed and examined by the pathologist. | Input | Interval |
| Reginol Node Positive | Number of regional lymph nodes examined that were positive for cancer. | Input | Interval |
| Survival Months | Number of months that the person is being alive after the diagnosis. | Rejected | Interval |
| Status | If the patient survived (1 dead - 0 alive) | Target | Binary |

## Variable Analysis

In this section, a Heatmap and Chi-Square Test of independence was created to discover any relationship between the variables for numerical and categorical variables, respectively.

**Heatmap**

The Heatmap identifies that the variable ‘Survival Months’ has a strong relationship negative correlation (-0.6) with ‘Status’ (Target label). This correlation suggests that as the number of survival months increases, the likelihood of Status being 1 (Dead) decreases, which aligns with the expected outcome that longer survival time correlates with being live.

**A diagram of a number of patients

Description automatically generated with medium confidence**

For this reason, Survival Months is rejected in the model to address the curse of dimensionality and ensure that the model remains focused and free from redundancy.

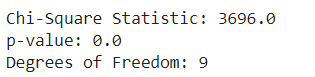
**Chi-Square Test of independence**

A chi-square test was conducted between the categorical variables ‘Differentiate’ and ‘Grade since it appears to have a relationship. The Contingency table shows the frequency of each category of ‘Differentiate’ compared across the categories of ‘Grade’. The null hypothesis for this test is that there is no relationship between the two variables; whereas the alternative hypothesis is that there is a relationship between the two variables.

**A white background with black text

Description automatically generated**

The Chi-square test, with a significance level of 0.05, shows that the p-value is 0.0, meaning that the null hypothesis is rejected and concluding that there is an association between the two variables. A p-value of 0.0 indicates a extremely strong evidence against the null hypothesis.

****

For this reason, the variable differentiate is rejected in the model.

## Descriptive Statistics

The following table presents the measures of central tendency for each variable of the dataset. The dataset has no missing values for any variable which means it is not necessary to apply techniques to manage these values; also, the numerical variables present a range (minimun and maximun) and mean. For instance, the age of patients is between 30 years and 69 years with a mean of 54.52. The regional Node Positive vary between 1 tumor and 46 tumors and mean of 5.60, which it could indicate that it migh have outliers as well as tumor size with a range between 1 tumor and 140 tumors and mean of 33.17.

As for the categorical varibles, the table shows the mode which indicates the most frequent category of the variable. The most popular characteristics in patients are the marital status married, race white, grade 2 which means the cancer cells are moderately differentiated, N1 stage which means the cancer has spread to 1 to 3 lymph nodes, T2 stage which means Tumor is more than 2 cm but not more than 5 cm.

A screenshot of a computer

Description automatically generated

The graphs represent the distribution of N stage and T stage according to the status, showing that the most frequent patients dead are in N1 stage with 270 patients, and in T2 stage with 303 patients who have died.

A screenshot of a computer screen

Description automatically generated A screenshot of a graph

Description automatically generated

The histograms show the distribution of the variables Tumor size, Reginol Node Positive, Age and Regional Node Examined. The tumor size and Reginol Node positive is skewed in the right tail.

A screen shot of a graph

Description automatically generated A graph showing a number of bars

Description automatically generated

A graph of age distribution

Description automatically generated A graph of a number of bars

Description automatically generated

The box plot shows how data are distributed and any outliers. In this case, as mentioned earlier, tumor size has outliers; the median of the tumor size for patients that are alive is less than the median of the patients that died.

The next graph shows the that the tumor size of the patients alive (Status is 0) has more variability than the patients dead (Status is 1) although the patients dead have greater outliers.

A screenshot of a computer

Description automatically generated

# Over sampling

The original dataset had 4024 records, 3408 of them had the status of “alive”, and 616 “Dead”. A random sampling was performed to balance the dataset and avoid bias in the model; 616 from the 3408 “alive” records were randomly choosing in Microsoft Excel.

For the target variable **Status**, the original dataset was modified from Nominal to binary, replacing “Dead” to 1 and “Alive” to 0.

# Decision Tree Model

## Maximal Tree

The Average -Squared error for the Maximal Tree is 0.236363. Figure 1 and Figure 2 in the Appendix section exhibit the statistics results and the diagram for the Maximal Tree.

## Misclassification Tree

The Average -Squared error for the Misclassification Tree is 0.229525. Figure 3 and Figure 4 in the Appendix section exhibit the statistics results and the diagram for the Misclassification Tree.

## Average Square Error Tree

The Average-Squared error for the Average Square Error Tree is 0.228703. Figure 5 and Figure 6 in the Appendix section exhibit the statistics results and the diagram for the Average Square Error Tree.

## Optimal Tree

The ASE decision tree with the average square error of 0.228703 is the optimal choice of tree for predicting patients who are more likely to die of breast cancer. Having the lowest average square error indicated the model predictions are closer to the true values on average. The decision tree is better suited for identifying high-risk individuals, contributing to more effective and potentially life-saving interventions.

Based upon the average square error tree, patients with positive or missing progesterone status are less likely to die at 44.10% compared to negative status with 67.3%, indicating **negative progesterone status** might be a positive prognosis indicator.

With **N stages** N3 and N2 the survival rate lowers as potential dead percentage holds to be 60.1%, with N1 stage having 32.45% and this aligns with expected relationship between advance stages and reduced survival.

Under **Age**, younger patients have less death rate of 28% compared to older patients at 41.11%. This suggests age might play a role in deaths within this specific group.

**Progesterone status and N stage seems to be the important factors influencing deaths in this model.**

A screenshot of a computer

Description automatically generated

# Impute

The dataset did not contain any missing values, for that reason the imputation was not needed. The following image shows there were not any missing values for non of the variables.

A screenshot of a computer

Description automatically generated

# Adjust Outliers

In the previous image, there are 2 variables with a skewness higher than 1: Reginol Node Positive and Tumor Size. To address this, capping and flooring was employed to reduce the skewness of those two variables. Although this technique resulted in a reduction in skewness, both variables still presented skewness values exceeding 1.

A screenshot of a computer

Description automatically generatedA screenshot of a computer screen

Description automatically generated

# Skewness

To continue reducing the skewness, a log transformation method was used in both variables mentioned above, to make the data more closely to a normal distribution. The method reduced the skewness in less than 1 for both variables with its respective levels.

A screenshot of a medical report

Description automatically generated

# Replace Dummies

For the study case, no categorical variable was replaced; instead, the categorical variables were kept as the original variables without any substitution to not loss any relevant information. The categorical variables in the dataset represent different patterns or conditions associated with breast cancer that are crucial to have a comprehensive understanding within the data.

# Logistic Regression Model

Four regression models were applied: Full Regression, Forward Regression, Backward Regression, and Stepwise Regression. Average Square Error was used as the metric to decide which is the optimal regression model.

## Full Regression

The Average Square Error for the Full Regression is 0.211012 **(Appendix – Figure 7).**

## Forward Regression

The Average Square Error for the Forward Regression is 0.211064 (Appendix – Figure 8).

## Backward Regression

The Average Square Error for the Backward Regression is 0.20826 (Appendix – Figure 9).

## Stepwise Regression

The Average Square Error for the Stepwise Regression is 0.211064 (Appendix – Figure 10).

## Optimal Regression

After comparing the Average Square Error (ASE) of the four regressions, the optimal regression with the lower ASE is Backward Regression.

A screenshot of a computer

Description automatically generated

The variables that are included in the final model are:

* Grade.
* LOG\_REP\_Reginol\_Node\_Positive.
* Progesterone\_Status.
* REP\_Age.
* REP\_Regional\_Node\_Examined.
* T\_Stage.

A screenshot of a computer

Description automatically generated

However, the variables with the higher Chi-Square are the most important variables for the model. For this model the variables that are more important are:

* Progesterone\_Status with a chi-square of 28.55.
* LOG\_REP\_Reginol\_Node\_Positive with a chi-square of 22.03.
* T\_Stage T1 with a chi-square of 8.13.
* REP\_Age with a chi-square of 7.50.

The next image indicates the odds ratio estimates for each input, following with the interpretation:

A screenshot of a computer screen

Description automatically generated

* For **Grade**, the odds ratio **(1 vs anaplastic; Grade IV)** estimate equals 0.001. This means that cases with Grade 1 are 99.9% less likely to die of breast cancer than cases with anaplastic Grade 4.
* For **Grade**, the odds ratio **(2 vs anaplastic; Grade IV)** estimate equals 0.001. This means that cases with Grade 2 are 99.9% less likely to die of breast cancer than cases with anaplastic Grade 4.
* For **Grade**, the odds ratio **(3 vs anaplastic; Grade IV)** estimate equals 0.001. This means that cases with Grade 3 are 99.9% less likely to die of breast cancer than cases with anaplastic Grade 4.
* For **LOG\_REP\_Reginol\_Node\_Positive**, the odds ratio estimate equals 1.871. This means that every time **LOG\_REP\_Reginol\_Node\_Positive** goes up by the factor of 2.74 the probability to die of breast cancer increases by 87.1%.
* For **Progesterone\_Status,** the odds ratio estimate equals 3.725. This means that cases with Negative Progesterone are 3.725 times more likely to die of breast cancer that cases with Positive Progesterone.
* For **REP\_Age,** the odds ratio estimate equals 1.027. This means that for each additional year, the probability of die of breast cancer increases by 2.7%.
* For **REP\_Regional\_Node\_Examined,** the odds ration estimate equals 0.974. This means that for each additional node examined the probability of die of breast cancer goes down by 97.4%.
* For **T\_Stage**, the odds ratio **(T1 vs T4)** estimate equals 0.269. This means that cases with T1 are 73.1% less likely to die of breast cancer than cases with anaplastic T4.
* For **T\_Stage**, the odds ratio **(T2 vs T4)** estimate equals 0.414. This means that cases with T2 are 58.6% less likely to die of breast cancer than cases with anaplastic T4.
* For **T\_Stage**, the odds ratio **(T3 vs T4)** estimates equals 0.392. This means that cases with T3 are 60.8% less likely to die of breast cancer than cases with anaplastic T4.

# Neural Network Model

The Neural Models in the Breast Cancer Model are nine. Seven of these Neural Models are using just the most important variables of the best regression model to mitigate the curse of dimensionality. In this case, the backward regression. The average error was selected as the model selection criterion. Also, the maximum iterations were changed to 100 and the maximum time to 10 minutes.

The other two models were included after the Cap and Floor (NN Cap and Floor) and the Transform Variables (NN Transform Variables) modifications. These models were included to validate if they have lower (ASE) than the Neural Models connected to the backward regression that is the best regression model.

## Neural Network: Cap & Floor

The Cap & Floor Neural Network has the model selection criterion Profit/Loss and three hidden units. The Average Square Error is 0.20948 (Appendix – Figure 11).

## Neural Network: Transform Variables

The Transform Neural Network has the model selection criterion Profit/Loss and three hidden units. The Average Square Error is 0.211469 (Appendix – Figure 12).

## Neural Network: Best Regression (3H)

The Neural Network: Best Regression has the model selection criterion Average Error and three hidden units. The Average Square Error is 0.208826 (Appendix – Figure 13).

## Neural Network: 2H

The Neural Network: Best Regression has the model selection criterion Average Error and two hidden units. The Average Square Error is 0.210514 (Appendix – Figure 14).

## Neural Network: 4H

The Neural Network: Best Regression has the model selection criterion Average Error and four hidden units. The Average Square Error is 0.210002 (Appendix – Figure 15).

## Neural Network: 5H

The Neural Network: Best Regression has the model selection criterion Average Error and five hidden units. The Average Square Error is 0.20793 (Appendix – Figure 16).

## Neural Network: 6H

The Neural Network: Best Regression has the model selection criterion Average Error and six hidden units. The Average Square Error is 0.207433 (Appendix – Figure 17).

## Neural Network: 7H

The Neural Network: Best Regression has the model selection criterion Average Error and seven hidden units. The Average Square Error is 0.208127 (Appendix – Figure 18).

## Neural Network: 8H

The Neural Network: Best Regression has the model selection criterion Average Error and eight hidden units. The Average Square Error is 0.208862 (Appendix – Figure 19).

## Optimal Neural Network

After running all the Neural Networks models, the best model is the Neural Network using 6 hidden units. The ASE is 0.207433 which is the lowest comparing the other models. Also, this model has a ROC index of 0.737 which is the highest as well. These two parameters indicate that the 6 hidden units neural network model is the best among the neural networks.

A screenshot of a computer

Description automatically generated

The following graph shows that the iteration where with the lowest ASE for the Valid dataset is in the third iteration, that means that the model has converged.

A screen shot of a graph

Description automatically generated

# Assessment

The results shows that the best model is the Neural Network after regression with 6 hidden units with an ASE of 0.207433.

Using ROC as a validation assessment rating the best model is the Neural Network with 5 hidden units with a ROC of 0.741

A screenshot of a computer

Description automatically generated A screenshot of a graph

Description automatically generated

A graph showing a curve

Description automatically generated with medium confidence

The cumulative response chart show that the Neural Network with 3 hidden units is the most effective model based on the response rate; choosing the best 5% of the records the response rate of this Neural Network is 100%.

A screen shot of a computer

Description automatically generated

As for the lift, choosing the best 5% of the records the best lift is from the Neural Network with 3 hidden units with a lift value of 1.996764

A screen shot of a graph

Description automatically generated

# Conclusion and Recommendations

This project identifies the different characteristics of patients who are more likely to die from breast cancer; to have a consist and reliable model some changes had to be done to the dataset; first, balance the records between dead and alive patients to avoid bias; second, using cap and floor technique to manage the outliers; Finally, transforming the variables to manage properly the skewness of the variables.

After running the different predictive models including decision tree models, logistic regression models, and neural network models,

the backward regression model stands out as the most effective model to identifying individuals that have a higher risk of dying of breast cancer. Although the best model with the ASE criteria is the Neural Network with 6 hidden units, the backward regression was chosen as the best model due to Neural Network cannot be interpretated.

For the backward regression, the variables that are included in the model and that have a notable association with breast cancer mortality are: Grade, LOG\_REP\_Reginol\_Node\_Positive, Progesterone\_Status, REP\_Age, REP\_Regional\_Node\_Examined, and T\_Stage.

For the logistic regression, the odds ratio estimates provide a crucial information about the impact of individual variables on the likelihood of death from breast cancer. For the backward regression of the breast cancer model, the odds ratio estimates show:

* For the variable **Grade**, cases with Grades 1,2, and 3 are shown to be 99.9% less likely to die compared to cases with anaplastic Grade4.
* For **LOG\_REP\_Reginol\_Node\_Positive**, every time this variable goes up by the factor of 2.74 the probability to die of breast cancer increases by 87.1%.
* For **Progesterone\_Status,** cases with Negative Progesterone are 3.725 times more likely to die of breast cancer than cases with Positive Progesterone.
* For **REP\_Age,** for each additional year the probability of die of breast cancer increases by 2.7%.
* For **REP\_Regional\_Node\_Examined,** for each additional node examined the probability of die of breast cancer goes down by 97.4%.
* For **T\_Stage**, cases with T1 are 73.1% less likely to die of breast cancer than cases with anaplastic T4.
* For **T\_Stage**, cases with T2 are 58.6% less likely to die of breast cancer than cases with anaplastic T4.
* For **T\_Stage**, cases with T3 are 60.8% less likely to die of breast cancer than cases with anaplastic T4.

Some recommendations for the health industry include focusing on those patients who their progesterone is negative since they have a 272,5% more chance to die than patients with positive progesterone. Also, the number of positive nodes has a major impact on critical patients; the more positive node a patient has, the probability of dead increase in 87,1%.

The cases in anaplastic T4 stage are other important characteristic to consider since these patients have a 26,9% more probability to die. Finally, the age has also an impact with a 2,7% of more chances to die when the age increase.

The study also showed that anaplastic grade 4 has a greater impact in the target comparing with the other categories 1,2 and 3, with 0,1% chance not to die. For further research, it is recommended the use of more techniques for a deeper understanding of this categorical variable.

# Appendix

A screenshot of a computer

Description automatically generated

Figure 1. Maximal Tree Results

A computer screen shot of a computer

Description automatically generated

Figure 2. Maximal Tree Diagram

A screenshot of a computer

Description automatically generated

Figure 3. Misclassification Tree Results

A screenshot of a computer screen

Description automatically generated

Figure 4. Misclassification Tree Diagram

A screenshot of a computer

Description automatically generated

Figure 5. Average Squared Error Tree Results

A screenshot of a computer

Description automatically generated

Figure 6. Average Squared Error Tree Diagram

A screenshot of a computer

Description automatically generated

Figure 7. Full Regression Results

A screenshot of a computer

Description automatically generated

Figure 8. Forward Regression Results

A screenshot of a computer

Description automatically generated

Figure 9. Backward Regression Results

A screenshot of a computer

Description automatically generated

Figure 10. Stepwise Regression Results

A screenshot of a computer

Description automatically generated

Figure 11. Neural Network Cap and Floor Results

A screenshot of a computer

Description automatically generated

Figure 12. Neural Network Transform Variables Results

**A screenshot of a computer

Description automatically generated**

Figure 13. Neural Network 3 Hidden Units

A screenshot of a computer

Description automatically generated

Figure 14. Neural Network 2 Hidden Units

A screenshot of a computer

Description automatically generated

Figure 15. Neural Network 4 Hidden Units

**A screenshot of a computer

Description automatically generated**

Figure 16. Neural Network 5 Hidden Units

**A screenshot of a computer

Description automatically generated**

Figure 17. Neural Network 6 Hidden Units

A screenshot of a computer

Description automatically generated

Figure 18. Neural Network 7 Hidden Units

A screenshot of a computer

Description automatically generated

Figure 19. Neural Network 8 Hidden Units

1. Dataset link: <https://www.kaggle.com/datasets/reihanenamdari/breast-cancer>

   [↑](#footnote-ref-2)