NAMIBIA UNIVERSITY OF SCIENCE AND TECHNOLOGY

Faculty of Computing and Informatics

School of Computing

Department of Software Engineering

DTA621 Data Analytics, SEMESTER 4, 2023



# Group member list:

|  |  |  |
| --- | --- | --- |
| **SN** | **Name** | **Student Number** |
| 1 | Orilio Naobeb | 223077593 |
| 2 | Jerome Nanuseb |  |
| 3 | Mwilima kamwi | 223070068 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Contents

[Group member list: 2](#_Toc179534049)

[Problem Statement: 4](#_Toc179534050)

[Introduction 6](#_Toc179534051)

[The Pie Chart Below Shows Distribution of Breast Cancer Class (Recurrence vs No Recurrence) 6](#_Toc179534052)

[Age Distribution of Patients 8](#_Toc179534053)

[Correlation Between Tumor Size and Age 9](#_Toc179534054)

[Distribution of Malignancy Degrees 10](#_Toc179534055)

[Relationship Between Involved Lymph Nodes and Recurrence 11](#_Toc179534056)

# Problem Statement:

The Breast Cancer Dataset is collected to help develop machine learning models that can predict the likelihood of breast cancer recurrence in patients. Specifically, the dataset aims to address the following problem:

**Purpose**: The primary goal is to build a predictive model that can determine whether a patient is likely to experience a recurrence of breast cancer after treatment. By analysing patient characteristics such as age, tumor size, and lymph node involvement, the model can assist physicians in making informed decisions about further treatments, follow-up strategies, and patient care.

**Problem to Solve**: The dataset is intended to help classify patients into two groups: those who will experience no recurrence of breast cancer and those who will experience a recurrence. Early and accurate predictions of recurrence are crucial for improving treatment outcomes and optimizing patient care.

**Why Use This Data**: The data contains key medical information about patients, such as age, tumor size, and lymph node status, which are known factors influencing cancer recurrence. Machine learning models can be trained on this data to identify patterns that are not immediately obvious to human experts.

**Dataset Columns and Explanation**:

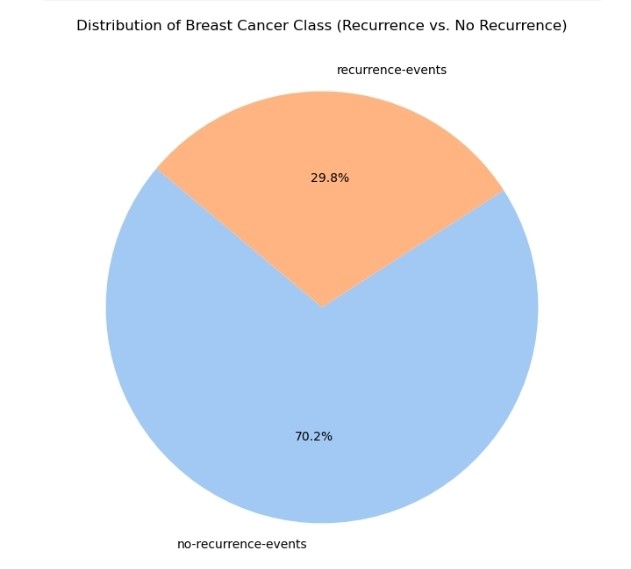
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **Type** | **Values / Ranges** | **Description** | **Missing Values** |
| Class (Target) | Categorical (nominal) | no-recurrence-events, recurrence-events | Indicates whether the patient has experienced a recurrence of breast cancer or not. This is the target variable for the predictive model. |  |
| Age | Categorical (ordinal) | Ranges: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99 | The age of the patient in 10-year intervals. Age is often a significant factor in cancer prognosis. |  |
| Menopause | Categorical (nominal) | lt40, ge40, premeno | Indicates the patient’s menopausal status. Hormonal changes related to menopause can impact cancer risk and recurrence. |  |
| Tumor Size | Categorical (ordinal) | Ranges: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59 | The size of the tumor in millimeters (mm). Larger tumors may have a higher risk of recurrence. |  |
| Inv-nodes | Categorical (ordinal) | Ranges: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39 | The number of involved lymph nodes. The higher the number, the greater the chance of cancer spreading. |  |
| Node-caps | Categorical (nominal) | yes, no | Indicates whether there is a capsular invasion (spread of cancer cells beyond the lymph node boundary). | 8 missing values |
| Deg-malig | Categorical (ordinal) | 1 (low), 2 (medium), 3 (high) | The degree of malignancy or severity of the cancer cells. A higher degree of malignancy indicates more aggressive cancer. |  |
| Breast | Categorical (nominal) | left, right | Indicates which breast was affected by cancer. This helps in analyzing if the recurrence is related to the location. |  |
| Breast-quad | Categorical (nominal) | left-up, left-low, right-up, right-low, central | Identifies the quadrant of the breast where the tumor was located. Different quadrants may have varying recurrence risks. | 1 missing value |
| Irradiat | Categorical (nominal) | yes, no | Indicates whether the patient received radiation therapy. Radiation is often used to reduce recurrence risk. |  |

# Introduction

The analysis of breast cancer recurrence provides valuable insights into the distribution of patient outcomes, tumor characteristics, and demographic factors. This report aims to explore the patterns of recurrence events in breast cancer patients, highlighting critical data points such as the imbalance between recurrence and no-recurrence cases, the age distribution of patients, the correlation between tumor size and age, and the degree of malignancy across different breast quadrants. Additionally, it examines the relationship between recurrence events and involved lymph nodes, offering a comprehensive look at how these factors might influence predictive modeling and treatment decisions.

Through visual representations such as pie charts, bar graphs, and scatter plots, this analysis delves into the complexity of breast cancer recurrence and its potential impact on both clinical outcomes and data-driven models. Understanding these patterns is crucial for identifying high-risk patient groups, refining recurrence prediction models, and developing more effective treatment strategies.

# The Pie Chart Below Shows Distribution of Breast Cancer Class (Recurrence vs No Recurrence)



The pie chart above represents the distribution of breast cancer classes in terms of recurrence events versus no-recurrence events. The two segments of the pie chart provide the following information:

• No-recurrence-events: 70.2% of patients did not experience a recurrence of breast cancer after treatment.

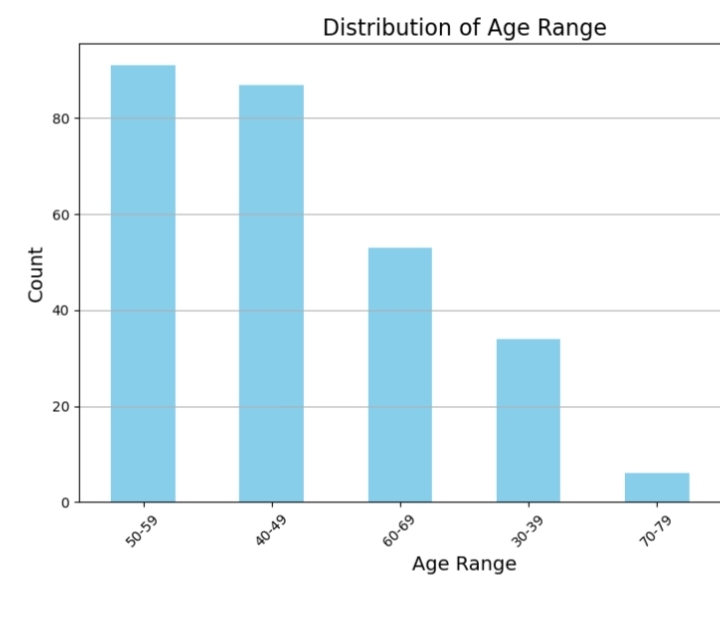
• Recurrence-events: 29.8% of patients experienced a recurrence.

Interpretation:

• Skewness: The data is imbalanced, with a majority of patients (70.2%) having no recurrence, compared to 29.8% who experienced recurrence.

Impact on Model Performance: Since the data is not balanced (more no-recurrence cases than recurrence cases), this could affect the performance of a model, particularly if you're using classification algorithms. Models may be biased toward predicting "no recurrence" more often due to the higher proportion of such cases in the data. You may need to consider techniques like oversampling the minority class (recurrence-events) or undersampling the majority class, or using evaluation metrics that handle imbalanced data, such as precision, recall, and F1-score.

# Age Distribution of Patients

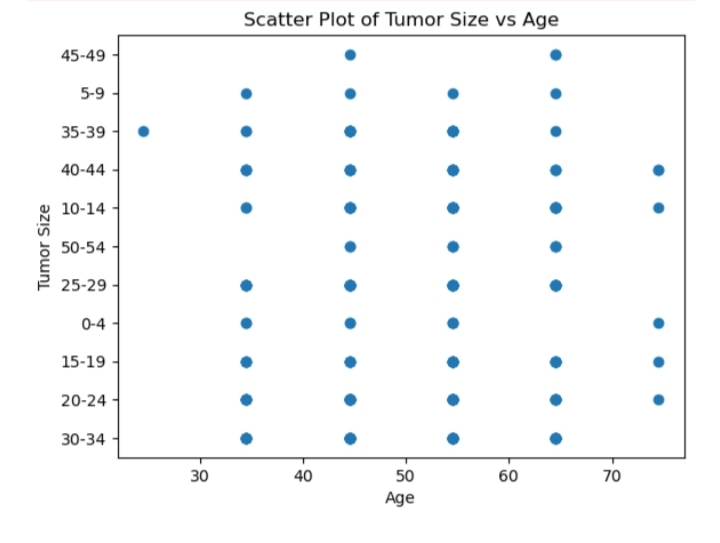


The age distribution across different ranges is depicted in the last image (Distribution of Age Range). The bar chart shows that the majority of patients fall within the age ranges of 50-59 and 40-49, followed by the 60-69 range. The 30-39 and 70-79 age ranges represent smaller portions of the dataset.

Interpretation:

• Most patients are concentrated between the ages of 40 and 59, indicating that breast cancer might be more common in this age group in the dataset. This insight is useful for identifying high-risk age brackets for recurrence analysis and treatment considerations.

# Correlation Between Tumor Size and Age

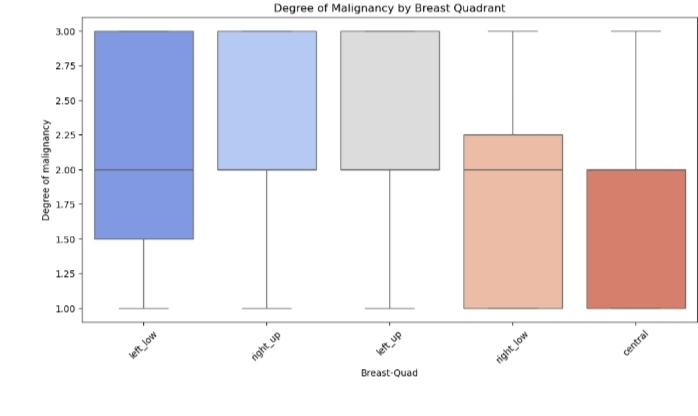


The scatter plot (Scatter Plot of Tumor Size vs. Age) shows the relationship between tumor size and patient age. Each dot represents a patient, with age on the x-axis and tumor size on the y-axis.

Interpretation:

• There doesn’t seem to be a strong visible correlation between age and tumor size. Tumor sizes are spread relatively evenly across different age groups, implying that age might not be a significant factor in determining tumor size. This means other factors such as genetics or tumor aggressiveness may play a more crucial role in recurrence.

# Distribution of Malignancy Degrees

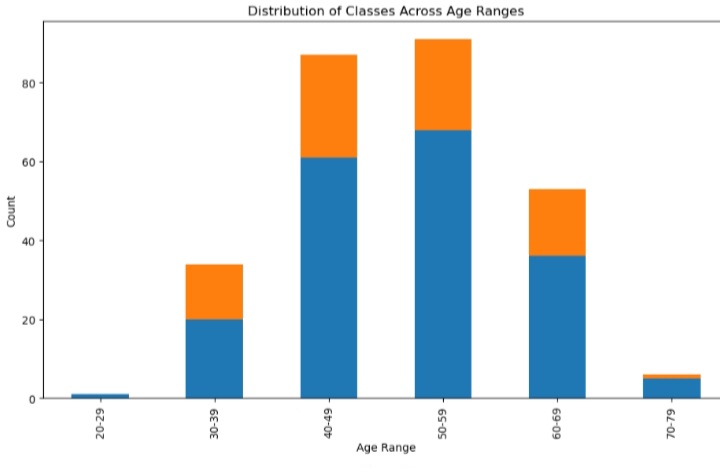


The first image (Degree of Malignancy by Breast Quadrant) shows a boxplot illustrating the distribution of malignancy degrees based on different breast quadrants.

Interpretation:

• The boxplots reveal a similar range of malignancy degrees across most quadrants, with the left low and right low quadrants showing slightly higher variance. Central breast quadrants show a lower median malignancy degree. This distribution can provide insights into whether certain areas of the breast are more prone to aggressive malignancies.

# Relationship Between Involved Lymph Nodes and Recurrence



The third image (Distribution of Classes Across Age Ranges) seems to show a stacked bar plot with two colors (blue and orange) representing different classes (likely recurrence vs. no recurrence) across age ranges.

Interpretation:

• The 40-49 and 50-59 age groups have the highest occurrences of both recurrence and no recurrence, as shown by the larger bars. The distribution across age groups suggests that these two age ranges may have a higher risk of recurrence. However, younger and older age groups have significantly lower counts, possibly due to either fewer cases or a lower likelihood of recurrence.

**Conclusion**

The dataset presented provides critical insights into various aspects of breast cancer analysis, including the distribution of recurrence events, patient demographics, tumor characteristics, and their potential implications for modeling and treatment. The visualizations, including pie charts, bar charts, scatter plots, and box plots, shed light on key factors such as recurrence rates, age distribution, tumor size, and malignancy degrees.

One prominent observation is the imbalance in recurrence events, where 70.2% of patients did not experience recurrence, while 29.8% did. This imbalance may significantly affect predictive models, requiring techniques to handle imbalanced data effectively. Furthermore, the age distribution suggests that breast cancer is more prevalent among patients aged 40-59, providing valuable information for risk analysis. However, there appears to be no strong correlation between age and tumor size, indicating that other factors, such as genetics, may play a more critical role in cancer aggressiveness. The degree of malignancy also varies slightly across breast quadrants, which may influence treatment decisions.

Finally, the analysis of recurrence events by age range highlights the increased likelihood of recurrence in patients aged 40-59, with fewer occurrences in younger and older groups. This comprehensive analysis provides a foundation for developing more targeted models and treatment plans that account for the complexities of breast cancer recurrence and patient characteristics.