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 |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Contents

[Group member list: 2](#_Toc178998836)

[1. Problem Statement: 4](#_Toc178998837)

# 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. |  |