

**SEMESTER ONE 2024/2025 ACADEMIC YEAR 1**

**SCHOOL COMPUTING AND IMFORMATICS TECHNOLOGY**

**DEPARTMENT OF COMPUTER SCIENCE**

**MASTER OF SCIENCE IN COMPUTER SCIENCE**

**MCS 7103**

**MACHINE LEARNING**

**ASSIGNMENT ONE**

**Eugene Munyaneza**

**2024/HD05/21936U**

**2400721936**

Table of Contents

[EXPLORATORY DATA ANALYSIS FOR ANALYSING BREAST CANCER RISK ESTIMATION FOR CUBAN WOMEN 3](#_Toc177568810)

[INTRODUCTION 3](#_Toc177568811)

[EXPLORATORY DATA ANALYSIS 3](#_Toc177568812)

[DATASET 3](#_Toc177568813)

[QUESTIONS 4](#_Toc177568814)

[PREPROCESSING 4](#_Toc177568815)

[*Duplicate values* 4](#_Toc177568816)

[*Missing values* 4](#_Toc177568817)

[Normalization 5](#_Toc177568818)

[ANALYSIS 5](#_Toc177568819)

[UNIVARIET ANALYSIS 6](#_Toc177568820)

[MULTIVARIET ANALYSIS 6](#_Toc177568821)

[CONCLUSION 6](#_Toc177568822)

[Works Cited 7](#_Toc177568823)

# EXPLORATORY DATA ANALYSIS FOR ANALYSING BREAST CANCER RISK ESTIMATION FOR CUBAN WOMEN

## INTRODUCTION

Breast cancer is one of the leading causes of cancer-related deaths worldwide (Siegel , Giaquinto, & Jemal, 2024). Detecting cancer early before it has spread gives us the best chance for its treatment to succeed (Cancer Research UK, 2023). Breast cancer detection can therefore help to save so many lives.

Detecting breast cancer is however, a very challenging process, especially as we don’t clearly know the cause of cancer (American Cancer Society, 2024). But with the improvements in Machine Learning, we have a chance at analyzing the vast amount of data collected to trace patterns.

In this report, we are going to explore sample data on breast cancer diagnosis collected in Cuba and identify trends, patterns and correlations.

## EXPLORATORY DATA ANALYSIS

### DATASET

* Our data is from BCSC Hispanic dataset.
* We shall be examining medical data of 1,697 cases of breast cancer diagnosis in Cuban women.
* Our data has 22 features.
* Our target variable is cancer diagnosis: Yes, has cancer and No doesn’t have cancer.

data columns

**Risk factor**: Description - Data type (Values)

1. **age**: Patient's age - Quantitative ( 20 - 90)
2. **menarche**: Age of menarche - Quantitative ( 8 - 17)
3. **menopause**: Age of menopause - Quantitative ( 0, 30 - 60)
4. **agefirst**: Age at first successful delivery - Quantitative ( 0, 9 - 46)
5. **children**: Number of children born alive - ( 0 - 6)
6. **breastfeeding**: Time breastfed in months - Quantitative ( 0 - 72(months))
7. **nrelbc**: Number of first-degree relatives with breast cancer - Quantitative ( 0 - 2)
8. **biopsies**: Number of breast biopsies - Quantitative ( 0 - 5)
9. **hyperplasia**: Atypical hyperplasia - Qualitative ( no, yes)
10. **race**: Race - Quantitative ( white, mixed, black)
11. **year**:
12. **imc**:
13. **weight**: Patients weight at screening - Quantitative - Kgs
14. **exercise**: Weekly physical activity - Quantitative ( 0 - 7)
15. **alcohol**: Alcohol consumption - Qualitative ( no, yes)
16. **tobacco**: Tobacco consumption - Qualitative ( no, yes)
17. **allergies**: Number of allergies suffered - Qualitative ( no, dermatitis, laryngitis, medications, other)
18. **emotional**:
19. **depressive**: Whether a patient is depressed or not - Qaulitative (yes, no)
20. **histologicalclass**:
21. **birads**: Breast Imaging-Reporting and Data System (BI-RADS) (0,1,2,3,4,5,6 broken into A,B and C)
22. **cancer**: Breast cancer diagnosis - Qualitative ( no, yes)

### QUESTIONS

1. Do we have sufficient data?
2. Is our data trustworthy and clean?
3. Does our data meet our requirements?
4. What are we targeting?

We have **1,697** samples collected and I think that this can be sufficient for our model.

The data was collected by trusted sources (Breast cancer risk factors in Cuban women, 2024). As we explore, we shall find out if the data is clean or if we can clean it.

We have **22** features in our data, and I think this is sufficient to meet our requirements.

We aim to predict whether a patient has cancer or not. In our sample date of **1,697** patient data, **1,160** have cancer and **537** do not have cancer.

### PREPROCESSING

#### Duplicate values

None

#### Missing values

We have 7 missing values ,namely.

1. Biopsies: 1
2. year: 537
3. imc: 7
4. weight: 10
5. allergies: 276
6. histologicalclass: 537
7. birads: 537

We have 537 patients without cancer, these are the same patients who are missing the following features year of diagnosis, histological class and birads. We shall therefore ignore these are they do not affect our classification.

Furthermore, the following features do not have a relationship with cancer, and we also exclude them; emotional, allergies and depression.

The Id feature is used as a counter only, and therefore we also exclude it.

We are left with the following features of interest.

1. Biopsies: 1
2. imc: 7
3. weight: 10

#### Handling missing data

##### Biopsies.

On analyzing biopsies, it is discovered that the mean is 1.31, minimum value is 0, maximum value is 5. The mean is skewed to the minimum value, so we examine the percentiles, 25% is 0, 50% is 1 and 75% is 2. The 50% percentile is therefore chosen to replace the missing value.

##### Imc

On analyzing imc, it is discovered that the mean is 25.53, the minimum is 5 and the maximum is 88.8. Again, the mean is far from the maximum, therefore we examine the percentiles. The 25% is 22.8, 50% is 25.1 and 75% is 27.3. It is clear that the mean is a more common value and so it was chosen to replace the missing data.

##### weight

On analyzing weight, it is discovered that the mean is 70.45, the minimum is 13 and the maximum is 240. Again, the mean is far from the maximum, therefore we examine the percentiles. The 25% is 61.9, 50% is 69.2 and 75% is 79.2. The 50th percentile is a more common value and so it was chosen to replace the missing data.

Normalization

We shall normalize for none-numeric values, the following fields are the none-numeric values.

1. Cancer (No and Yes)
2. Hyperplasia (No and Yes)
3. Alcohol (No and Yes)
4. Tobacco (No and Yes)
5. Race (Black, Mixed and White).

We shall utilize label encoding to encode the data as follows.

1. cancer mapping: {'No': 0, 'Yes': 1}
2. hyperplasia mapping: {'No': 0, 'Yes': 1}
3. alcohol mapping: {'No': 0, 'Yes': 1}
4. tobacco mapping: {'No': 0, 'Yes': 1}
5. race mapping: {'Black': 0, 'Mixed': 1, 'White': 2}

## ANALYSIS

Analyze the data to find how many cases were positive and negative.

Total Cases: 1697

Total Positives: 1160

Total Negatives: 537

Missing value analysis.

### UNIVARIET ANALYSIS

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **age** | **menarche** | **nrelbc** | **biopsies** | **hyperplasia** | **race** | **imc** | **weight** | **alcohol** | **tobacco** | **cancer** |
| **count** | 1697 | 1697 | 1697 | 1697 | 1697 | 1697 | 1697 | 1697 | 1697 | 1697 | 1697 |
| **mean** | 51.48 | 11.73 | 1.14 | 1.31 | 0.46 | 1.17 | 25.53 | 70.49 | 0.68 | 0.65 | 0.68 |
| **std** | 11.93 | 1.84 | 0.36 | 1.18 | 0.499 | 0.80 | 4.95 | 12.77 | 0.466 | 0.48 | 0.47 |
| **min** | 20 | 8 | 1 | 0 | 0 | 0 | 5 | 13 | 0 | 0 | 0 |
| **25%** | 45 | 10 | 1 | 0 | 0 | 0 | 22.8 | 62 | 0 | 0 | 0 |
| **50%** | 53 | 12 | 1 | 1 | 0 | 1 | 25.1 | 69.2 | 1 | 1 | 1 |
| **75%** | 61 | 13 | 1 | 2 | 1 | 2 | 27.3 | 79.2 | 1 | 1 | 1 |
| **max** | 90 | 17 | 4 | 5 | 1 | 2 | 88.8 | 240 | 1 | 1 | 1 |

### MULTIVARIET ANALYSIS

Using inbuilt tools in panda, I was able to analyze the correlation between the features and our outcome.

A summary of the correlations ranked from highest to lowest.

|  |  |
| --- | --- |
| Factor | Correlation |
| alcohol | 0.57 |
| tobacco | 0.54 |
| hyperplasia | 0.5 |
| age | 0.35 |
| menopause | 0.29 |
| nrelbc | 0.2 |
| weight | 0.08 |
| race | -0.08 |
| menarche | -0.12 |
| imc | -0.15 |
| agefirst | -0.17 |
| children | -0.43 |
| breastfeeding | -0.46 |

# CONCLUSION

From our table we can see that drinking alcohol has the highest impact, followed by smoking, hyperplasia, patients age, menopause, nrelbc, and weight.

Race, menarche, imc, agefirst, number of children and breastfeeding don't have a positive correlation.

# Works Cited

American Cancer Society. (2024, 1 1). *Understanding the Causes of Cancer*. Retrieved from Cancer.org: https://www.cancer.org/cancer/risk-prevention/understanding-cancer-risk.html

*Breast cancer risk factors in Cuban women*. (2024, 08 30). Retrieved from Mendeley Data: https://data.mendeley.com/datasets/7jhddnpz2p/1

Cancer Research UK. (2023, 03 30). *Why is early cancer diagnosis important?* Retrieved from Cancer Research UK: https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important/1000

Siegel , R. L., Giaquinto, A. N., & Jemal, A. (2024). Cancer statistics, 2024. *CA: A Cancer Journal for Clinicians*, 1–114.