

**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 2](#_Toc176974803)

[INTRODUCTION 2](#_Toc176974804)

[EXPLORATORY DATA ANALYSIS 2](#_Toc176974805)

[CONCLUSION 7](#_Toc176974806)

[Works Cited 7](#_Toc176974807)

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

## INTRODUCTION

Detecting breast cancer is 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 analyze sample data on breast cancer diagnosis collected in Cuba.

## 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.

#### QUESTIONS

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

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.

Our data

A white rectangular grid with many small squares

Description automatically generated with medium confidence

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)

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

A screenshot of a computer code

Description automatically generated

Total Cases: 1697

Total Positives: 1160

Total Negatives: 537

Missing value analysis.

A screenshot of a computer program

Description automatically generated

We have 537 cases where patients didn't have cancer. Which corresponds with the value for year of diagnosis, histological class and birads.

These columns therefore won't be needed in our data analysis.

Id is also not needed in our predictions.

Emotional, Allergies and Depressive is also irrelevant in our analysis.

Biopsies will not help in our analysis as they are not done for those who didn't get diagnosed with breast cancer.



#### INSIGHTS

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



From the above results we can see that some fields actually have a negative correlation with getting breast cancer.

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