Gabriela Aleksandrova

Fontys University of Applied Sciences

Project proposal

A breast cancer screening tool/biomarker

Contents

[Section 1: Project introduction 0](#_Toc85704371)

[Project overview 0](#_Toc85704372)

[Added value 0](#_Toc85704373)

[Domain understanding 0](#_Toc85704374)

[Data collection 1](#_Toc85704375)

[Exploratory Data Analysis 2](#_Toc85704376)

[Raw data overview 2](#_Toc85704377)

[Data info 2](#_Toc85704378)

[Descriptive statistics 3](#_Toc85704379)

[Overview of BMI status of patients per classification group 3](#_Toc85704380)

[Classification variable overview 3](#_Toc85704381)

[Comparing univariate summary statistics per Classification category 4](#_Toc85704382)

[Exploring correlations between independent variables 5](#_Toc85704383)

[Conclusion from EDA 6](#_Toc85704384)

[Overview of target variable 6](#_Toc85704385)

[Modelling 6](#_Toc85704386)

[Evaluation and deployment 7](#_Toc85704387)

[Section 2. Societal impact assessment 9](#_Toc85704388)

[Technology impact 9](#_Toc85704389)

[Stakeholders 9](#_Toc85704390)

[Privacy 9](#_Toc85704391)

[Data 10](#_Toc85704392)

[Transparency 10](#_Toc85704393)

[Inclusivity 10](#_Toc85704394)

[Hateful and criminal actors 11](#_Toc85704395)

[Future 11](#_Toc85704396)

[Conclusion 11](#_Toc85704397)

# Section 1: Project introduction

## Project overview

The project goal is to present a breast cancer predicting model (biomarker) for early detection of the disease in the patients. The collected data itself is a combination of anthropometric measurements and blood analysis indicators. It includes 166 observations (per patient) obtained from an open access journal website for cancer research.

## Added value

An early detection of breast cancer is essential when it comes to effectively planning the treatment of the diagnosed which is far more affordable when the tumor is detected in its early phase. The biomarker’s usage can be applicable along with a routine blood test analysis.

## Domain understanding

Various contributive robust models are developed based on different data collected from specialized routine consultation. Specialists in this field noted that several hormones, apart from the age of the patients and their BMI are good candidates that can be used as predictors of the breast cancer presence. Thirteen years ago, a breast cancer specific antigen CA15-3 and insulin growth protein were classified as logistic regression predictors. In 2013, leptin, it’s ratio with Adiponectin, BMI and CA15-3 were used together as biomarker for this type of diagnosis. Resistin, from the other hand, was assessed as a predictor of postmenopausal breast cancer.

## Data collection

The data is collected from blood samples from women after an overnight fasting. Before the start of the experiment, anthropometric and demographic data - age, weight, and height of the patients was collected as well and the status of menopause (at least 12 months after menopause was a requirement) was checked. The fasting blood indicators were extracted in a laboratory and analyzed using commercial kits. A tumorectomy and mastectomy were used to obtain a tumor tissue from diagnosed patients which resulted in classifying the participants into two categories – diseased and healthy. A description of each of the indicators is included below:

* **Glucose**- an important energy source and a component of many carbohydrates.

**Measure unit**: **mmol/L (millimoles per litre)**

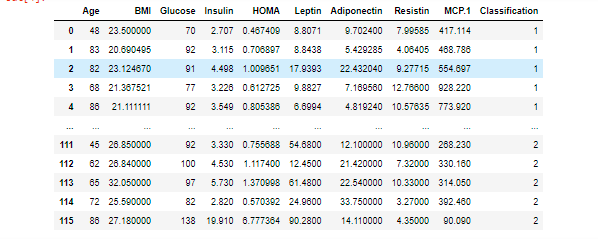
* **Insulin**- a pancreas produced hormone which regulates the amount of glucose in the blood.

**Measure unit : Insulin Units – 100 IU – 1 ml.**

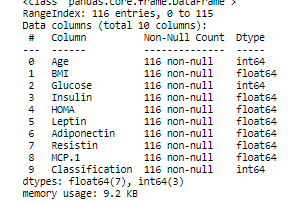
* **Resistin –**  known asAdipose tissue-specific secretory factor (ADSF) – specific cysteine-rich peptide hormone derived from adipose tissue.
* **Adiponectin** – a protein hormone which regulates the glucose levels.
* **MCP** – small cytokine that recruits monocytes, dendritic cells to the inflammation sites that are produced by tissue injury/infection.
* **HOMA** – indicator of insulin resistance
* **BMI** – body weight divided by the squared height of the person.

## Exploratory Data Analysis

### Raw data overview



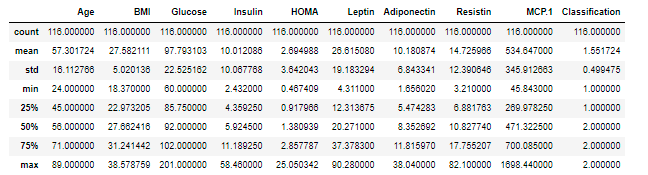
### Data info



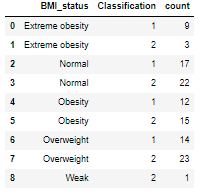
It is also noticeable there are no missing values (each column has 116 total entries as the size of the whole dataset).

### Descriptive statistics

This overview is important for exploring the values per variable, since they are not measured on the same scale.



### Overview of BMI status of patients per classification group



*Overview of Total number of classified patients per BMI\_status.*

### Classification variable overview

The last column from the dataset is ‘Classification’,which holds important information per patient. 1 stands for ‘diseased’ and 2 stands for ‘healthy’.



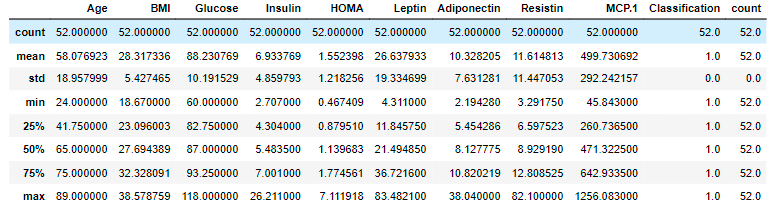
*Total number of diseased (Classification=1) and healthy (Classification=2) patients.*

### Comparing univariate summary statistics per Classification category

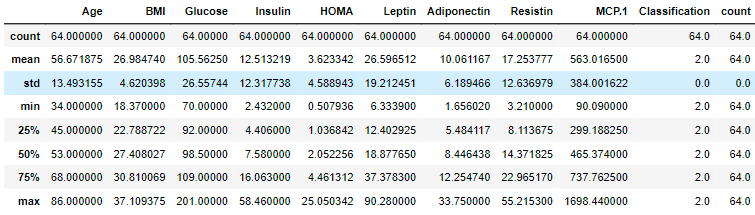
The first table is filtered to display summary statistics only about healthy people, the second one for the diseased ones.

Key findings:

* The glucose level for healthy people is ~20% lower than the one of the diseased.
* A nearly 50% increase in the average Insulin levels for diseased can be noticed
* Resistin level is also higher for diseased people

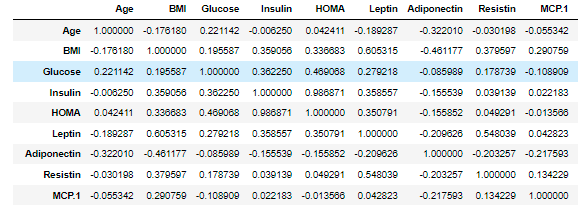


*Classification=1 (healthy patients)*

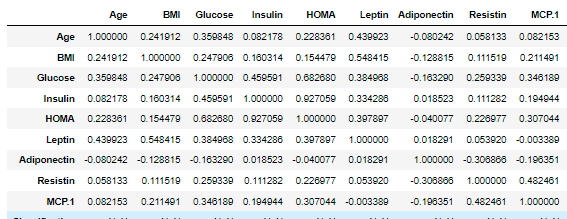
**

*Classification=2 (diseased patients)*

### Exploring correlations between independent variables



*Classification=1 (healthy patients)*



*Classification=2 (diseased patients)*

Additional insights can be derived from comparing the regression output for each category . Overall, it is noticeable that the correlation coefficients for some pairs in the disease table are higher than the ones of the healthy patients.

## Conclusion from EDA

Since all the variables in the dataset are important indicators for breast cancer, their combined effect on the target variable should be considered. The regression and univariate statistical outputs suggest that some values (for example Glucose and Insulin levels) are higher for the diseased people. Furthermore, the dataset contains almost the equal amount of information for diseased and healthy people, which is important information that must be checked before starting with data preprocessing.

## Overview of target variable

This section contains overview of the type of the target variable- the values it represents and the meaning of these values. This is important step towards choosing suitable modelling technique.

In the above included dataset, the variable that holds the prediction output is ‘Classification’ with values 1 for healthy patients and 2 for the diseased ones. Each row (observation) in the dataset is classified via one of these class labels.

## Modelling

After acknowledging the type of the target variable and the features that describe it (explained under the Data Collection section), the next question is:

* What kind of technique will be suitable to model this data?

The features‘ values are numerical, and the target variable is categorical (the categories are represented as numbers). In this case, it would be suitable to model this data via Logistic regression, a model that classifies the instances towards certain class based on their impact on the odds ratio of the class. An example visualization is included below:

Timeline

Description automatically generated

Visualization of the relationship between Glucose (independent variable) and the target variable (in a binary format; 1=healthy and 0=diseased)

## Evaluation and deployment

Even if the model evaluation results are accurate, a further verification is needed from the expert that understands how the process of detecting breast cancer works. More specifically, this person should have an expertise in evaluating if each observation/row is correctly labelled by observing the biomarkers’ level of each individual patient and if the combined effect of the variables in the dataset should be considered. Moreover, if the model is not suitable, this person will give recommendations about the data- what information is significant and what is lacking. To conclude, there will be a feasibility deduction and the model will be deployed when there is more clarification on how the dataset should be extended and if the model is suitable.

Without being an expert in this field, I am setting the following model evaluation goal:

* Accuracy, precision, recall-~80% (more data should be added and when this happens, the accuracy should be >90% since this is an important medicine technology)
* Sensitivity -avoiding false negatives and identifying true positives-people correctly identified as having certain condition not considering and classifying the ones that do not have the condition; this is an important measure, since it serves as a confirmation of the usefulness of the test. Although, it doesn’t consider the people that do not have the condition when in fact they do have it. It is significant that the patients are correctly classified as ‘diseased’.
* Specificity- identifying true negatives, avoiding false positives-identifying people who do not have the condition. It is significant that the patients are correctly classified as ‘healthy’.

When it comes to Sensitivity and Specificity, I think it is important that they both have high result since they relate to the effectiveness of the test.

# Section 2. Societal impact assessment

The goal of this section is to outline the impact of the technology on the society -the benefits (added value) from introducing the project. The assessment consists of different subjects which are elaborated with the goal of defining if there are misalignments with the overall project mission and how they can be tackled.

## Technology impact

The biomarker assessment tool will provide the chance to many women to take immediate action on planning the treatment of breast cancer. It can be offered every time a woman undergoes a routine blood test, and the result is immediate.

## Stakeholders

* Healthy women (in menopause) who want a breast cancer biomarker assessment in addition to a blood test routine analysis
* Oncologist who understands and interprets the biomarker prediction

## Privacy

The data is collected in lawful manner according to the health data protection rules and regulations stated in the GDPR. Each observation can’t be tracked back to the patient it belongs to.

The assessment tool doesn’t store any personal information such as name or address which is explained in advance to everyone that are interested to undergo the assessment. Furthermore, there is a limited access to the tool – only those healthcare specialists having the domain knowledge can work with it.

## Data

The collected data can be easily extended in future to contain observations for different ages or simply more observations for ages that are already present in the data. There is no variable/observation in the data that can be considered as subjective or exposing any details about particular patient.

## Transparency

The patients will be thoroughly informed about the way of working of the biomarker and how the information is being processed by it. One of the obligations of the specialist providing this service will be to answer any type of questions related to the usage of the tool so that the patient gets to know more about its beneficence and make informed decision.

## Inclusivity

The technology is mainly recommended for woman over 50 who are in menopause as an addition to their routine blood analysis test in exchange to a small fee. The reason for that recommendation is because certain age groups are prone to the disease more than others. Nevertheless, the technology is accessible to woman on every age, but it should be noted that the accuracy of the tool prediction might not be the same since the data observations are mainly collected and labelled for >50 years woman.

## Hateful and criminal actors

There is always a chance that a data breach occurs. This can happen if for example a hacker gains access to the medical system where the observations are stored or due to unauthorized access. In this case, it is advisable that there is incident response plan to stop the leakage/repair the system that is executed immediately after suspicion of data breach. Therefore, to mitigate this risk, an extensive security strategy must be developed or if it exists already, they it be integrated with the tool.

## Future

There is a significant chance that the technology will continue to have the same meaningful effect and positive impact on the woman’s health. The information that is being used by the tool can be extended with a lot of new features that will provide even better classification. A perfect future scenario will be the biomarker usage to be extended for further classification of the stage of the tumor and complete guidance of every step that has to be taken during this phase ,the type of treatment as well screening of cancer recurrence.

# Conclusion

Contemplating on the different subjects from the Potential Impact Assessment was extremely insightful when it comes to additional aspects that must be considered. For instance, security concerns that might occur and type of information that can be added to the model. The social impact assessment serves as blueprint of the improvements that can be made in future to the model.