Gabriela Aleksandrova

Fontys University of Applied Sciences

End user report

A breast cancer screening tool/biomarker

Contents

[Section 1:Data exploration results 1](#_Toc87535192)

[EDA conclusion and insights 1](#_Toc87535193)

[Section 2:Model deployment,usage and reccomendations 2](#_Toc87535194)

[Model performance 2](#_Toc87535195)

[Model sensitivity and specificity 3](#_Toc87535196)

[Model FP rate 3](#_Toc87535197)

[System deployment info 4](#_Toc87535198)

[Example use case 4](#_Toc87535199)

[Future improvement 4](#_Toc87535200)

[Conclusion 6](#_Toc87535201)

# Section 1: Data exploration results

This section outlines the results from the data exploration and modelling as well as the respective conclusions.

## EDA conclusion and insights

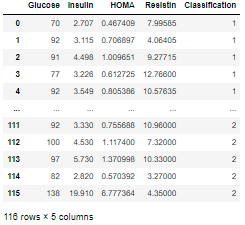
The initial data consisted of demographic and anthropometric data about each patient along with the blood indicator levels. Each observation corresponded to one patient. The goal of the EDA was to initially explore the data and its components and find the features that are least important in predicting whether a patient has breast cancer. Part of the EDA was exploring the univariate statistics per category, e.g. what are the median, mean, max and min values to check if the data is suitable for modelling. The illustration below shows the average distribution for healthy (1) and diseased (2) patients.

A screenshot of a computer

Description automatically generated with low confidence

Illustration: Average levels of blood analysis indicators per category

After the initial data investigation, the next step was to overview the relationship between the features and the target variable. Statistical methods (such as Kendall rank correlation and Variance Inflation Factor) were applied to determine which variables should be eliminated (these processes are described further in the EDA notebook) and afterwards, the data was prepared for modelling. This resulted in the following dataset:



*Illustration 1.Results from EDA.Final dataset.*

The resulting dataset has 4 features and 1 target variable.The conclusion from the EDA analysis is that the demographic and anthropometric data about the patients don’t add any value to the prediction as well as some blood analysis indicators such as Leptin and Adiponectin.

# Section 2:Model deployment,usage and reccomendations

## Model performance

It is important to have certain metrics (KPIs) to keep track of while building the model. The output of the model is a category and thus it is important that each patient is classified correctly towards the respective category. In the following sections, several metrics will be overviewed and specific KPIs will be presented, respectively.

### Model sensitivity and specificity

There are two measures that are commonly used to evaluate the performance of screening tests: the **sensitivity** and **specificity** of the test. The sensitivity of the test reflects the probability that the screening test will be positive among those who are diseased. In contrast, the specificity of the test reflects the probability that the screening test will be negative among those who, in fact, do not have the disease.

|  |  |  |
| --- | --- | --- |
|  | *Weighted average* | |
| Model type | Sensitivity (Precision) | Specificity (Recall) |
| SVM | 0.84 | 0.82 |

KPI:

* It is advisable both sensitivity and recall to be as high as possible and it is important that there is not a big mismatch between them.

### Model FP rate

A False Positive is the incorrect identification of anomalous data as such, i.e. classifying as “abnormal” data which is in fact normal.

|  |  |
| --- | --- |
| Model type | FP rate |
| SVM | 0.07 |

KPI:

* The FP should be as low as possible.

## System deployment info

|  |  |
| --- | --- |
| Deployment recommendation | The screening tool should be part of the existing system where the blood analysis results are stored in a shared database and uploaded to the web platform through APIs. |
| Input data | Blood analysis results |
| Input data format | .csv files |
| Output | 0 (no breast cancer) ,1 (breast cancer) |
| Output type | Binary |

## Example use case

***Stakeholders: Laboratory assistant and patient:***

*A woman ( age : >45) is referred from its GP for blood test. She comes to the laboratory with the required documents, the assistant there takes the blood sample. At the end of this procedure, the employee informs the woman about the possibility of examining further the blood indicator results via the new breast cancer screening tool. If she is interested, the assistant gives her the login credentials to the platform where the blood analysis results will be published and explains where to find the additional section of the breast cancer screening result.*

## Future improvement

The illustration below shows a comparison of the training and validation accuracy on a different training size. The validation itself is an evaluation metric of the actual skill of the model. This is essential, since during the validation procedure, the model is trained on smaller data splits in contrast with training it on a single split, which is not fully reliable, because the model will ‘learn too much’ from such data distribution, memorize the noise in the data and therefore, will not perform well on unseen data. With that being mentioned, the goal is to have bigger amount of representative data which will improve the validation accuracy to the point there is not significant mismatch between the validation and the training accuracy. Currently, the model doesn’t use enough data for training and validation and therefore, it needs re-training on more representative sample, for example a collection of blood analysis results formatted in the same way as the .csv file used for training. Although labelling each observation manually will take time, it is guaranteed that the model performance will significantly improve after more data is added.

Chart, line chart

Description automatically generated

*Illustration 2. Comparison between training and validation accuracy.*

## Conclusion

There are several main points that can be concluded:

* The tool should be properly integrated in the system where the records of the blood analysis results are stored
* The model should be re-trained. Twice more data should be used for training (looking at the learning curve graph) to achieve better alignment between the training and validation accuracy.