Conference Paper Title

Brewton Morais
DETI
Universidade Federal do Ceará
Fortaleza, Brazil
brewtonlmorais@gmail.com

Lucas Abdalah
DETI
Universidade Federal do Ceará
Fortaleza, Brazil
lucasabdalah@alu.ufc.br

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—breast cancer, diagnosis, machine learning, linear model, linear regression, cell biopsy**

I. Introduction

This command is used for a strong suggestion.
This command is used for minor changes suggestion.

Breast Cancer is one of the most commom cancer type in women behind only of Skin Cancer. In Brazil, according to the Instituto Nacional de Câncer (INCA) translate to English?, approximately 59,700 women were diagnosed with Breast Cancer in 2019, with a mortality rate of 13.68 per 100,000 habitants. Despite the fact that there are a limited amount of data worlwide, in some regions such as Western Europe and North America, breast cancer is the one with the highest incidence.

Although the treatment for it is normally agressive, early-stage cancer detection may reduce the death rate in the long term [1]. Some exams such as mammography and contrast-enhaced (CE) digital mammography are commonly used to do the diagnosis, as well as the breast biopses aims to infer if the tumor is malignant or benign, requiring trained people.

With the purpose of improving Breast Cancer Diagnosis, the analysis of potential cancerous cells have been made, retrieving features such as texture, smoothness, radius size, etc. The study of these parameters can be crucial to the early-stage breast cancer detection, as it will be investigated in this work.

Throughtout this paper, the relevance of such cell caracteristics will be analyzed by statitics metrics and data visualization tools with the goal of fiding a correlation or a probability relationship to the final diagnosis. The final goal is to build a probabilistic model capable of predicting the malignancy of tumor cells.

Therefore this work focus on the search for linear relationship between the parameters and the class of data, which it's refered as M for malignant and B for benign cells, wich can be categorized as a binary classification problem. At first, it may be possible to use Linear Regression techniques and Maximum Likelihood Estimator to achieve this goal, which will be confirmed throughout the development of the study.

II. Breast Cancer Dataset

A. Context and parameters

The breast cancer is characterized when the cells in the women's breast start growing uncontrollaby, which forms tumors that can be either bening or malign.

The dataset contains cell parameters such as radius mean, texture mean, area mean, smoothness and so on, which are normally the parameters altered by a malignant tumor. It is composed by 32 columns and 539 rows, that is, 32 features for 539 samples. There are 2 columns between the 32 that do not represent any information concerning the cells, but the identification of the pacients and their final diagnosis as well. Therefore, the dataset contains two classes, Malignant (M) and Benign (B), characterizing a binary classification problem.

Although there are only two classes, the key challenge is to build a prediction model based on the weight of each feature on the final result, that is, the relevance on the final diagnosis. It's proposed a data investigation on the possible missing values as well as a dimensionality reduction since there are too many columns, what would make the model too complex and probably not able to generalize if all of these features were to be considered for the model construction.

Throughout this work, the programming language ***Python*** was employed using mainly the *Pandas* library, which deals with data frame, including the preprocessing and cleaning steps, as well as the plots and inferences.

B. Data set Variables

As it was previously stated, the data set contains 32 columns whose 30 of them correspond to the predictors. The following list corresponds to the set of all variables present in the data frame and its definition, with order of appearance.

1) ***id*** (primary key): identification of each patient, final target.
2) ***diagnosis*** (M/B): sample class, it can be either **M** for Malignant, or **B** for Benign.
3) ***radius mean***: mean value of lobes' radius.
4) ***texture mean***: mean value of surface texture.
5) ***perimeter mean***: mean value of lobes' outer perimeter.
6) ***area mean***: mean value of lobes' area.
7) ***smoothness mean***: mean value of smoothness level
8) ***compactness mean***: mean value of tumor cell compactness.

9) ***concavity mean***: mean value of tumor cell concavity.
10) ***concave points mean***: mean value of tumor cell concave points.
11) ***symmetry mean***: mean value of tumor cell symmetry.
12) ***fractal dimension mean***: mean value of tumor cell fractal dimension.
13) ***radius se***: error of radius.
14) ***texture se***: error of texture.
15) ***perimeter se***: error perimeter.
16) ***area se***: error of area.
17) ***smoothness se***: error of smoothness.
18) ***compactness se***: error of compactness.
19) ***concavity se***: error of concavity.
20) ***concave points se***: error of concave points.
21) ***symmetry se***: error of symmetry.
22) ***fractal dimension se***: error of fractal dimension.
23) ***radius worst***: worst tumor cell radius value.
24) ***texture worst***: worst tumor cell texture value.
25) ***perimeter worst***: worst tumor cell perimeter value.
26) ***area worst***: worst tumor cell area value.
27) ***smoothness worst***: worst tumor cell smoothness value.
28) ***compactness worst***: worst tumor cell compactness value.
29) ***concavity worst***: worst tumor cell compactness value.
30) ***concave points worst***: worst concave points value.
31) ***symmetry worst***: worst tumor cell symmetry value.
32) ***fractal dimension worst***: worst tumor cell fractal value.

III. Data Preprocessing

Before dealing with plots and inferences from the data, it is extremely important to perform a normalization in order to facilitate the visualization of the histograms and correlation plots. It is done by applying the z-score on the dataset.

A. Z-Score Normalization

Also called a standard score, the z-score is a normalization technique that gives the idea of how far from the mean a data point is. It can be placed on a *normal distribution* curve. In order to do so, it is necessary to know the mean μ and the standard deviation σ of the points.

Let $\bar{x}$ be the sample mean,

$$\bar{x} = \sum_i \frac{x_i}{N}$$

then

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$

Therefore, since now the data is located around a zero mean normal distribution with unitary variance, the identification of outliers has became easier.

The need to perform such normalization is that a model constructed from a non-normalized data set will probably have a biased result, since the model can be sensitive to variance. Therefore, if there are large differences between the range of some variables, those with the highest range will have more relevance to the model, since the variance is larger, that is, it carries more information.

B. Data review

It was check the existence of any missing values, but it was found that all data is fully completed, which eliminates the need of any data compesation technique.

C. Statistical Analysis

- To get a general perspective of the dataset, using the method *describe* from the Pandas library, the following table is a cutout of the entire table involving all parameters, here not shown in order to not difficult the comprehension:

TABLE I: Data Statistics 3 predictors

Stats	*radius mean*	*texture mean*	*perimeter mean*
count	569	569	569
mean	-1.256562	1.049736	1.272171
std	1.000880	1.000880	1.000880
min	-2.029648	-2.229249	-1.984504
25%	-6.893853	-7.259631	-6.919555
50%	-2.150816	-1.046362	-2.359800
75%	4.693926	5.841756	4.996769
max	3.971288	4.651889	3.976130

- The following table shows the mean statistics of 3 predictors grouped by Diagnosis: malignant or benign.

TABLE II: Data Statistics grouped by diagnosis

Diagnosis	*radius mean*	*texture mean*	*perimeter mean*
B	1.097064	-2.073335	1.269934
M	1.829821	-0.353632	1.685955

IV. Methods

Describe our methods and develloped work.

V. Experimental Results

Describe our statistic results and figures.

VI. Discussion

Further description about our results and what it implies.

VII. Conclusion

Quick recap about what we did, reinforce our results strengths and weakness.

VIII. Further Work

Quick recap about our work weakness and propose new approach to overcome its weakness.

References

[1] A. Author1 and C. CoAuthor1, *Title1*, 1st ed. publisher, 2001.