Dimensionality Reduction Analysis Applied to Breast Cancer Data

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*—Breast Cancer is one of the most aggressive cancer types and has a great impact in cancer mortality, mainly in women. This work analyzes tumor cells characteristics in order to provide an effective method to preprocess and reduce the dimension of the data. The presence of several predictors may provide meaningless or redundant information, what increase the cost of a Machine Learning-based techniques. In this paper, we present a framework based in Principal Component Analysis (PCA) and data normalization to clean the data and extract only the more relevant parameters.**

***Index Terms*—Breast Cancer, Dimensionality Reduction, Linear regression, Machine learning, Principal Component Analysis.**

I. Introduction

[1], [2], [3], [4].

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), 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 [5]. 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, wh ich 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, to be confirmed throughout the development of the study.

Feature Extraction and Dimensionality reduction [3], Clustering and Class Separation [6], [7]. All the works aim to increase accuracy in breast cancer diagnosis. [8]. ORGANIZAR ISSO NA INTRO.

II. Methods

A. Data set overview

The breast cancer is characterized when the cells in the women's breast start growing uncontrollaby, forming 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 569 rows, *i.e*, 32 features for 569 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.

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 by performing a Principal Component Analysis (PCA), 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 their definition, with order of appearance.

The diagnosis predictor is the categorical variable, which it can be either **M** for Malignant, or **B** for Benign. ACHAR UM LOCAL NO TEXTO PRA ENCAIXAR ESSE TRECHO

TABLE I: Summary of Predictors Information.

Tag	Description	Units
id	Patient identification	-
Diagnosis	Sample class	-
Radius Mean	Mean value of lobes' radius	
Texture Mean	Mean value of surface texture	
Perimeter Mean	Mean value of lobes' outer perimeter	
Area Mean	Mean value of lobes' area	
Smoothness Mean	Mean value of smoothness level	
Compactness Mean	Mean value of tumor cell compactness	
Concavity Mean	Mean value of tumor cell concavity	
Concave Points Mean	Mean value of tumor cell concave points	
Symmetry Mean	Mean value of tumor cell symmetry	
Fractal Dimension Mean	Mean value of tumor cell fractal dimension	
Radius SE	Error of radius	
Texture SE	Error of texture	
Perimeter SE	Error of perimeter.	
Area SE	Error of area	
Smoothness SE	Error of smoothness	
Compactness SE	Error of compactness	
Concavity SE	Error of concavity	
Concave Points SE	Error of concave points	
Symmetry SE	Error of symmetry	
Fractal Dimension SE	Error of fractal dimension	
Radius Worst	Worst tumor cell radius value	
Texture Worst	Worst tumor cell texture value	
Perimeter Worst	Worst tumor cell perimeter value	
Area Worst	Worst tumor cell area value	
Smoothness Worst	Worst tumor cell smoothness value	
Compactness Worst	Worst tumor cell compactness value	
Concavity Worst	Worst tumor cell compactness value	
Concave Points Worst	Worst concave points value	
Symmetry Worst	Worst tumor cell symmetry value	
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. Thus, in a non-normalized data set, if there are large differences between the range of some variables, those with the highest range would have much more relevance to the model, since the variance is larger, that is,

it carries more information.

B. Data cleaning

It was checked if there were any missing values, but it was found that all data is fully completed, which eliminates the need of any data compensation technique.

However, the first column containing the *id* of the patients is not relevant for the analysis, so it's discarded during the analysis.

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 II: 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.

The following image shows the conditional histograms of the data:

As it's possible to see, the majority of variables present normal and exponential distribution. For the most part of the

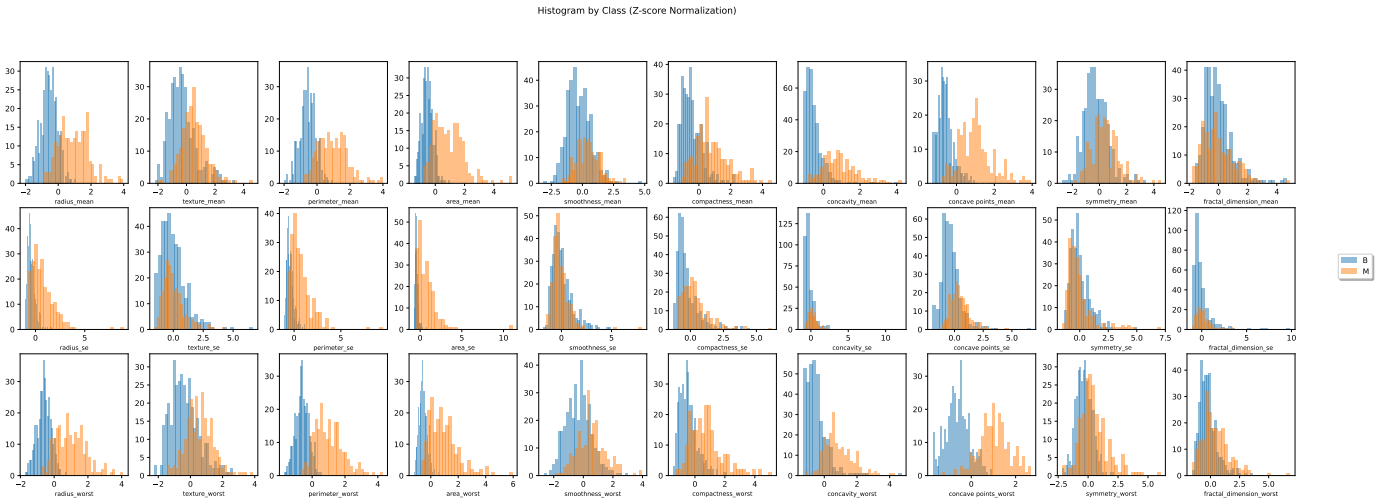

Fig. 1: Conditional Histogram.

TABLE III: 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

worst values for all the variables, it's not possible to infer a distribution density difference between then, since their values are really close to one another. Therefore, these variables are probably not informative for the final prediction. However, for values of radius mean, perimeter mean, area mean and smoothness mean, the difference is clear, indicating that those variables may have more relevance.

Besides, some of the predictors, such as concave points mean and symmetry mean have similar distributions, which indicates a linear relationship between then. The same is found analysing the following columns of errors: radius, texture, perimeter, smoothness and compactness.

The linear relantionship between these predictors will be analysed using the pair plot soon.

The figure 1 shows that some predictors show very distinct distribution. Cite which predictors has very similar distributions, and how many have very different histogram.

D. Predictors Relationship and Dimensionality reduction

- 21 pairs presents a correlation coefficients $|p| > 0.9$. - 15 pairs presents a correlation coefficients $|p| > 0.95$. What

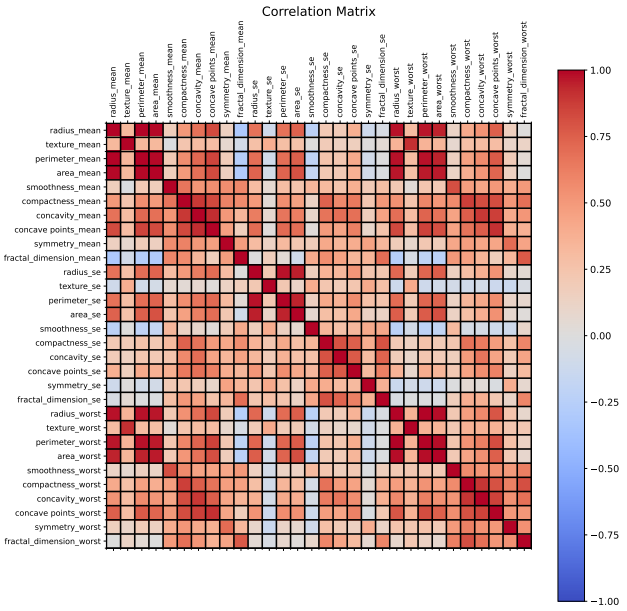

Fig. 2: Correlation Matrix presented as a Heatmap.

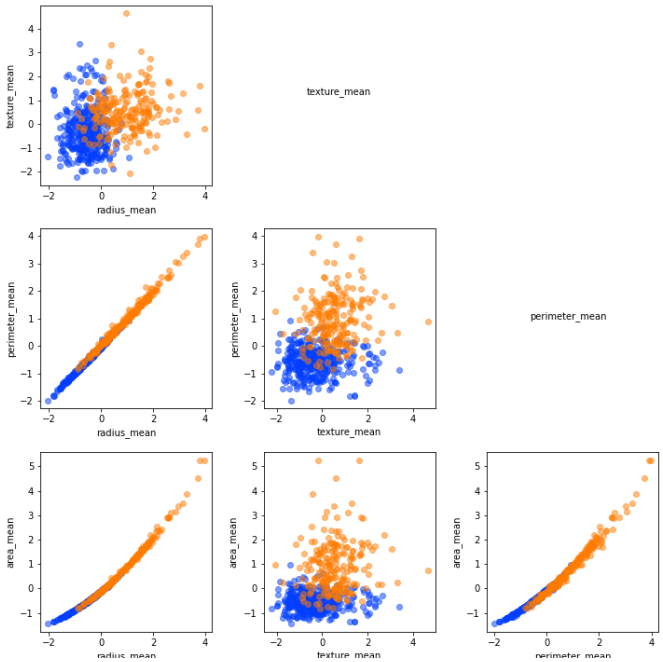

Fig. 3: Scatter Plot.

E. Principal Component Analysis (PCA)

[1], [2], [9], [10]

Principal Component Analysis (PCA) is a dimensionality reduction techninque that works by transforming a large set of variables into a smaller one that aims to represent as much as possible it's done by all the data variables. The justification for its implementation is that, at most part of the cases, it is worth losing a little accuracy in order to make a smaller data set. Besides, with a more compact data frame, it's less expansive to construct a machine learning model, because it'll work faster and the analysis will be easier. Thus, the goal is to preserve as much information as possible even after having performed the dimensionality-reduction.

The PCA is a variance sensitive method, which won't be an issue, since all the columns were standardized with the *Z-Score* application.

Then, the next step is to compute the Covariance Matrix, which is a $d \times d$ symmetric matrix with its entries being the covariance associated with all possible pairs of variables. Therefore, the covariance matrix of the breast cancer data set is a 31×31 matrix:

$$\begin{bmatrix} Cov(1,1) & Cov(1,2) & Cov(1,3) & \dots & Cov(1,4) \\ Cov(2,1) & Cov(2,2) & Cov(2,3) & \dots & Cov(2,4) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov(31,1) & Cov(31,2) & Cov(31,3) & \dots & Cov(31,31) \end{bmatrix}$$

The use of the covariance matrix is to identify where it's possible to find redundant information, according to the correlation between a pair or variables. If the covariance between a pair of variables is highly positive, it means that they are correlated and the increase of one implies the increase of the other. For covariance between pairs being negative, they called inversely correlated, the increase of one implies the decrease of the other.

The next stage is to compute the eigenvectors and its eigenvalues of the covariance matrix in order to identify the principal components. But first, the definition of Principal Components [10]: principal components are directions along which the variance of the data reaches its maximal value. They are linear combinations of the initial variables, related in such a way that the new variables are uncorrelated

and the most amount of information is present mainly within the initial components.

Since they are vectors, the principal components represent the directions of the data that explains a maximal variance, The fact that high variance indicates more information comes from the concept of entropy.

Finally, the mathematics behind this algorithm for the first principal component consists in finding a line that maximizes the average of square distances from the points to the origin. Then, for the second component, it's done the same, but with the condition of being orthogonal to the first line found, since they must be uncorrelated. The process is the same for the remaining components. That's where the importance of eigenvectors and eigenvalues lies, the first represents the direction of the the axes where the variance is maximum and the latter the coefficients attached to it.

As previously said, the first components always have more relevance, *i.e*, contain more information. Mathematically, once the eigenvectors are ordered according to their eigenvalues, the rank of principal components in significance is found as well.

Finally, from the eigenvectors a feature vector is created containing all of these eigenvectors or just some of them, after judging if some component is necessary or not, when they have less significance. So far, no data transformation was done apart from the normalization, so it's now necessary to recast the data along the principal component axes, which is done so by using the feature vector on the standardized data points in order to perform a reorientation.

$$FinalDataset = FeatureVector^T * StandardizedDataset^T$$

IV. Experimental Results

Describe our statistic results and figures.

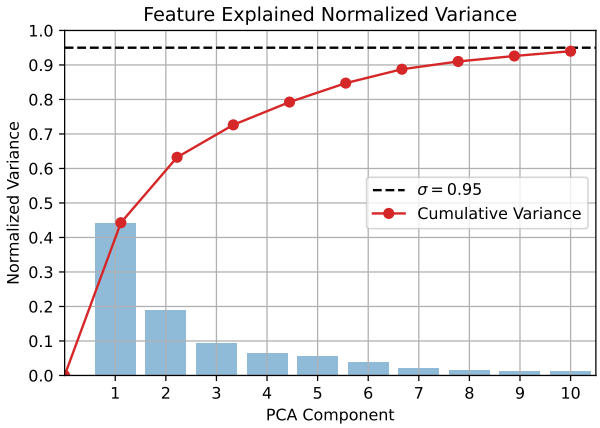

Fig. 4: Normalized Bar plot.

V. Discussion

Further description about our results and what it implies.

VI. Conclusion

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

VII. Further Work

This command is used for a strong suggestion.

This command is used for minor changes suggestion.

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

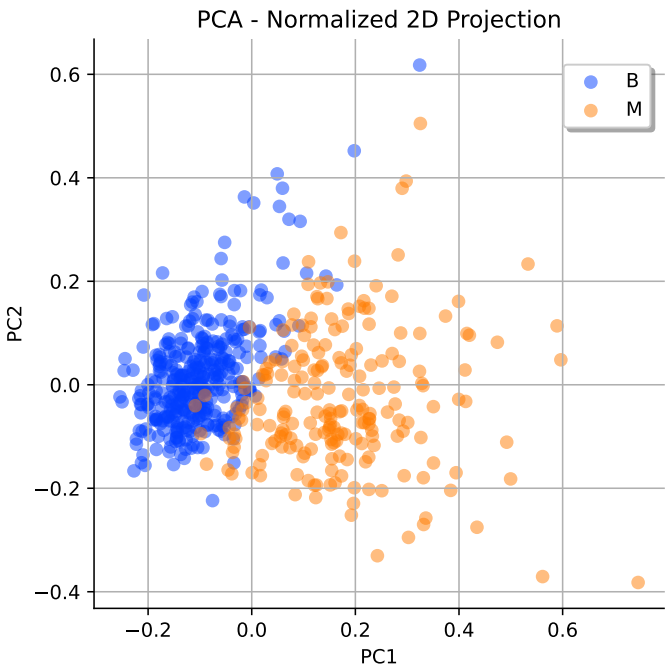

Fig. 5: Data scatter plot on the two principal components domain.

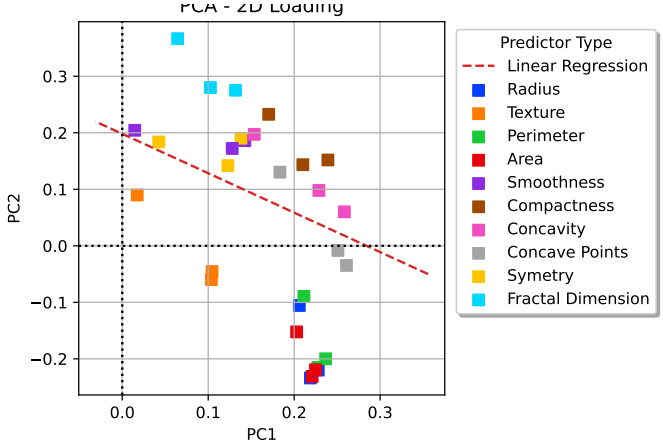

Fig. 6: PCA loading with two principal components.

REFERENCES

[1] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer, 2013.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, 2nd ed. Springer, 2009.

[3] N. Street, W. Wolberg, and O. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," *Proc. Soc. Photo-Opt. Inst. Eng.*, vol. 1905, pp. 861–870, 01 1993.

[4] H. Abdi and L. J. Williams, "Principal component analysis," *WIREs Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[5] Dummy, *Dummy Title*, 2nd ed. publisher, 2008.

[6] O. Mangasarian, N. Street, and W. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," *Operations Research*, vol. 43, pp. 570–577, 02 1995.

[7] K. P. Bennett and O. L. Mangasarian, "Robust linear programming discrimination of two linearly inseparable sets," *Optimization Methods and Software*, vol. 1, no. 1, pp. 23–34, 1992.

[8] W. H. Wolberg *et al.*, "Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates," *Cancer letters*, vol. 77, no. 2-3, pp. 163–171, 1994.

[9] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013. [Online]. Available: https://faculty.marshall.usc.edu/gareth-james/ISL/

[10] M. Ringnér, *What is principal component analysis*, 3rd ed. Nature Biotechnology, 2001.