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, since the presence of several predictors may provide redundant information, what increases 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, that can preserve original data characteristics and feed a predicting model to provide a final diagnosis.**

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

I. Introduction

Breast Cancer is the most incident among women in the world and the leading cause of cancer death. In 2018, there were 2.1 million new cases, equivalent to 11.6% of all cancers estimated. Regardless of the socioeconomic development of the country, the incidence of this cancer ranks among the top positions of female malignant neoplasms, even though the most frequent type of diagnosed cancer substantially vary across countries [1].

After skin cancer, breast cancer is the most common cancer diagnosed in women in Brazil. Approximately 59,700 women were diagnosed in 2019, with a mortality rate of 13.68 per 100,000 habitants. There is not only one risk factor for breast cancer, however age over 50 is considered the most important [2]. Other factors that contribute to the increased risk of developing the disease are genetic factors and hereditary factors (cancer of ovary in the family), obesity, physical inactivity and exposure frequent exposure to ionizing radiation (environmental and behavioral factors) [1], [2].

The funding support for breast cancer has helped to emerge diagnosis and treatment advances. Survival rates have increased, and the number of deaths associated with the disease is declining, mainly due to factors such as earlier detection, better understanding of the disease and personalized approach [3]. 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 [4], [5].

The main characteristics of malignant tumors, that causes various types of cancer, may be used as machine learning models input to assist in clinical diagnosis or even to provide an automated diagnosis. Retrieving tumor features such as texture, smoothness, radius size, etc. The stage of preprocessing and analysis of these parameters is crucial to the early-stage breast cancer detection [6]. Various approaches exploiting the same dataset are available in literature, e.g, linear programming, machine learning, data mining, and optimization techniques [7]–[9].

Throughtout this paper, the relevance of such cell caracteristics will be analyzed by statitics metrics and data visualization tools aiming to find correlation with the diagnosis. The main goal is build framework capable to provide preprocessed data to use in a probabilistic model able to predict the malignancy of tumor cells. These steps precede more complex analysis such as feature extraction, dimensionality reduction, clustering and class separation [10]–[12].

Futhermore, the Principal Component Analysis (PCA), is applied on the search for relationship between the parameters and their class, that may be categorized as a binary classification problem [13]. The PCA is a celebrated method to compute predictors relationship and reduce the data dimension, and in this work we present the most relevant characteristics related to the malignant, based on the provided set.

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 diagnosis predictor is the categorical variable, which it can be either **M** for Malignant, or **B** for Benign [6].

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 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.

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.

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.

TABLE I: Summary of Predictors Information.

Tag	**Description**
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

C. Statistical Analysis

Aiming to present a general perspective of the dataset, using the method *describe* from the Pandas library, the table IV is a cutout of the entire table involving all parameters, here not shown in order to not difficult the comprehension.

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

As it's possible to see in Fig. 6, the majority of variables present normal and exponential distribution. For the most part of the 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.

D. Covariance Matrix and Predictors Relationship

The degree of the linear relationship between two variables may be measured using the covariance coefficients (p). It ranges from -1 to +1, where large positive and negative values indicates positively and negatively correlated data,respectively. Its absolute magnitude measures the degree of redundancy. If the covariance is close to zero, the data is uncorrelated [11].

In order to provide data covariance, we compute the covariance matrix and present it is a heatmap for visualization simplicity. Fig. 1 shows that various predictors are strongly correlated, mainly positive.

TABLE III: Covariance threshold analysis.

$\|p\| \geq$	Related Pairs
0.9	21
0.95	15

For deepen the analysis, we use threshold two values for $|p|$ to assess how correlated data values, as shown in table III. 21 predictors pairs present a correlation coefficients $|p| > 0.9$, and 15 present a correlation coefficients $|p| > 0.95$.

The covariance evidence combined with scatter plot analysis in Figure 2 provide more information to corroborate that our data carries a lot of redundancy. We can observe that the pair plots: Radius Mean vs. Perimeter Mean, Radius Mean vs. Area

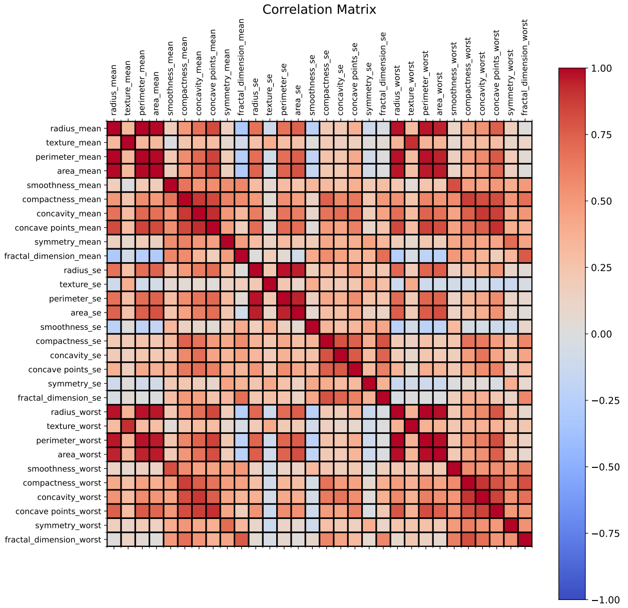

Fig. 1: Correlation Matrix presented as a Heatmap.

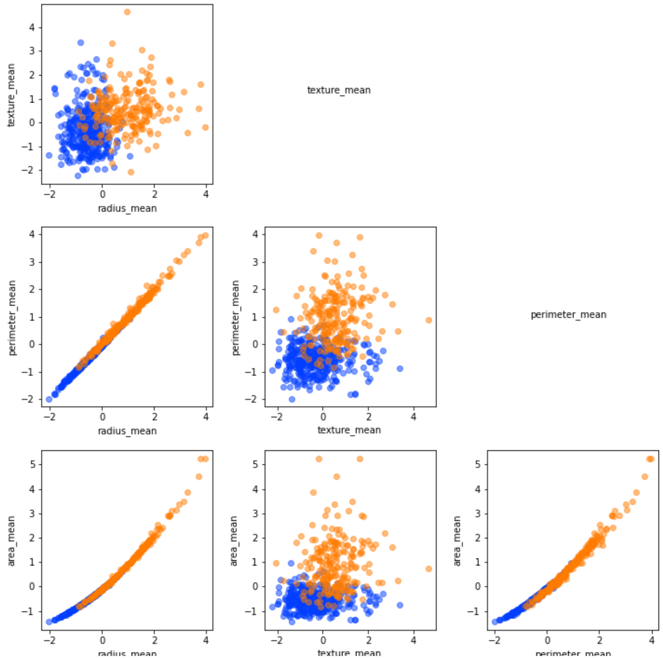

Fig. 2: Scatter Plot.

Mean and Perimeter Mean vs. Area Mean, present a linear behavior or very close to it.

E. Principal Component Analysis (PCA)

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.

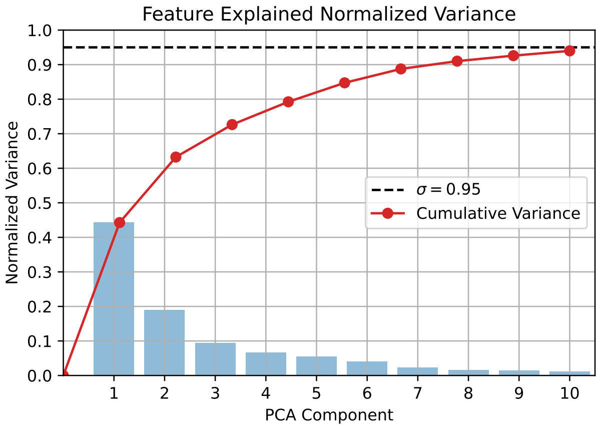

Fig. 3: Normalized Bar plot.

We can see in Fig. 3, the normalized variance, i.e, relevance on our context vs. the first 10 principal components (PC). We also present a cumulative curve, which shows that the first 2 PC carries more than 60% of the variance, leading to 95% with 10 PC, what we can interprete as only 10 columns of this dataset preserve 95% from the original information. It means dividing the complexity by 3 and improving the storage and processing cost.

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

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 [14]: 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 (P) 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.

$$Y = PX \tag{1}$$

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 (X) points in order to perform a reorientation.

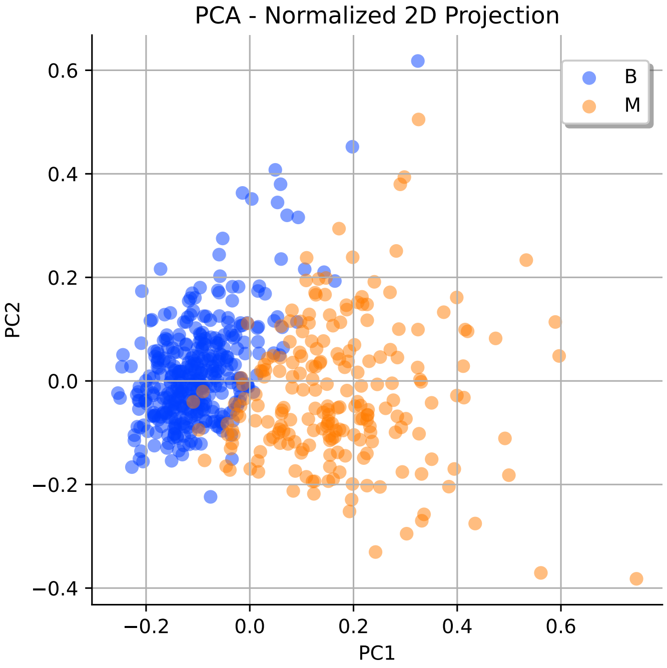

Fig. 4: Data scatter plot on the two principal components domain.

After compute the projection, we can plot as shown in Fig. 4, and observe that this component allow to split the data almost perfectly in two clusters.

Fig. 5 shows the loading, i.e, each predictor relevance on the projection variances of this 2 PC.

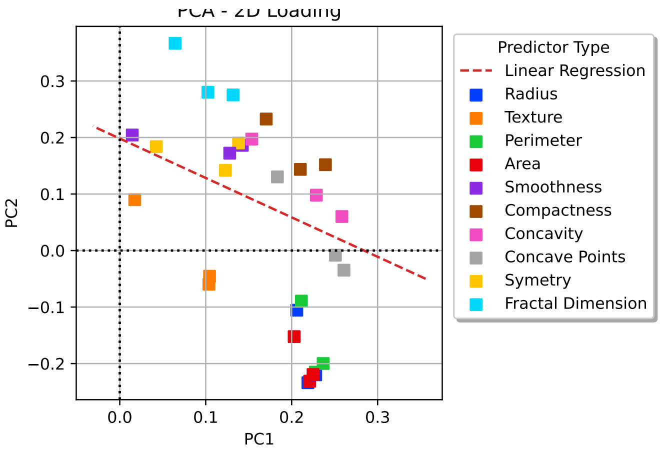

Fig. 5: PCA loading with two principal components.

IV. Conclusion

This work explored tatistical analyzes and the PCA application to visualize with clarity the behavioral patterns of the predictors in each class, in addition to the scatter trends, both being in a compact graphical representation.

It is interesting to note that the dataset obtained at the end of this extraction of predictors represents the general idea (95%) of the original dataset with only 10 PC. However, it is up to the modeler to decide whether this loss is offset by reduced computational complexity. It is worth mentioning that there are methods to further decrease the impact of predictor extraction of a dataset, although such methods have not been developed for the problem addressed here.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre *et al.*, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J Clin.*, vol. 68, pp. 394–424, 11 2018.

[2] C. Dieguez, Ed., *Estimativa 2020 : Incidência de Câncer no Brasil*, 1st ed. Instituto Nacional de Câncer José Alencar Gomes da Silva, 2019.

[3] M. M. Rivera-Franco and L.-R. Eucario, "Delays in breast cancer detection and treatment in developing countries," *Breast cancer : basic and clinical research*, vol. 12, pp. 1–5, 01 2018.

[4] G. W. Sledge, E. P. Mamounas, G. N. Hortobagyi, H. J. Burstein, P. J. Goodwin, and A. C. Wolff, "Past, present, and future challenges in breast cancer treatment," *J Clin Oncol.*, vol. 32, pp. 1979–1986, 07 2014.

[5] F. Cardoso *et al.*, "ESO-ESMO 2nd international consensus guidelines for advanced breast cancer (ABC2)," *Breast*, vol. 23, pp. 489–502, 07 2014.

[6] 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.

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

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

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

[12] 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/

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

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

TABLE IV: Unconditional Data Statistics for the Mean Predictors

	radius	***texture***	***perimeter***	***area***	***smoothness***	***compactness***	***concavity***	***concave points***	***symmetry***	***fractal dimension***
count	569	569	569	569	569	569	569	569	569	569
mean	14.12	19.28	91.96	654.88	0.09	0.10	0.08	0.04	0.18	0.06
std	3.52	4.30	24.29	351.91	0.01	0.05	0.07	0.03	0.02	0.00
skewness	0.94	0.65	0.99	1.64	0.45	1.19	1.40	1.17	0.72	1.30
min	6.98	9.71	43.79	143.50	0.05	0.01	0.00	0.00	0.10	0.04
25%	11.70	16.17	75.17	420.30	0.08	0.06	0.02	0.02	0.16	0.05
50%	13.37	18.84	86.24	551.10	0.09	0.09	0.06	0.03	0.17	0.06
75%	15.78	21.80	104.10	782.70	0.10	0.13	0.13	0.07	0.19	0.06
max	28.11	39.28	188.50	2501.00	0.16	0.34	0.42	0.20	0.30	0.09

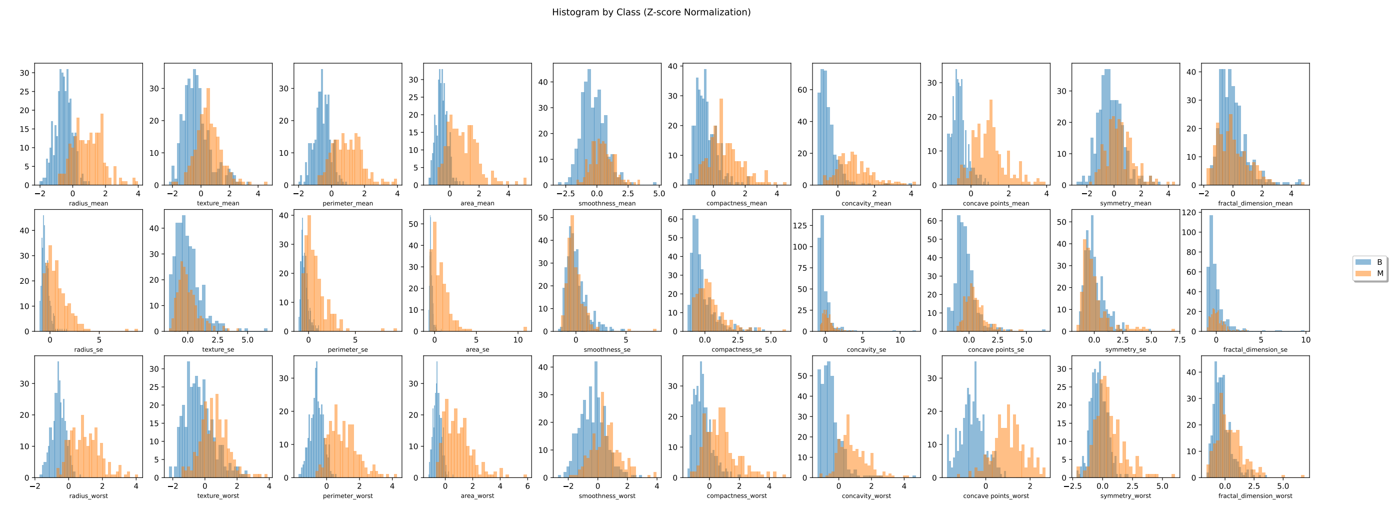

Fig. 6: Conditional Histogram.