Table of Contents

1. Introduction

Breast cancer, a complex and heterogeneous disease, remains a significant global health concern with profound implications for individuals and societies alike. Amidst the evolving landscape of oncological research, datasets play a pivotal role in unraveling the intricacies of this malignancy. The Wisconsin Breast Cancer dataset (WBCD) stands as a valuable repository, offering a nuanced glimpse into the cellular characteristics of breast tumors. As postgraduate scholars dedicated to advancing our understanding of cancer biology, we delve into the depths of this dataset to extract meaningful insights that may pave the way for improved diagnostic precision and treatment strategies.

A Brief Overview

The WBCD originates from fine-needle aspirates of breast masses, capturing the microscopic details of cell nuclei in biopsied tissues. Comprising features computed from digitized images, this dataset facilitates a multidimensional exploration of cell morphology and structure. The primary objective is to distinguish between benign and malignant tumors based on these quantitative attributes.

Features of Interest

1. Radius Features: Encompassing mean, standard error, and worst (largest) values, the radius features describe the average size of the tumor cells.

2. Texture Features: Reflecting variations in grayscale intensity, texture features provide insights into the spatial arrangement of cell nuclei.

3. Perimeter, Area, and Smoothness: These features quantify the geometric properties of cell boundaries and the homogeneity of cell sizes.

4. Compactness and Concavity: Offering measures of the compactness and concavity of the cell nuclei, these features contribute to the characterization of irregularities in cellular structure.

5. Symmetry and Fractal Dimension: Symmetry features highlight the balance in cell shape, while fractal dimension captures the complexity of the tumor boundary.

2. Business Understanding

n the realm of breast cancer diagnosis, leveraging the Wisconsin Breast Cancer dataset (WBCD) is a strategic move towards precision medicine. The goal is crystal clear: enhance diagnostic accuracy and treatment decisions. By analyzing the cellular intricacies encoded in the dataset's features, our mission is

to unearth patterns and biomarkers that can redefine the game in personalized breast cancer care. The business impact? More targeted treatments, improved patient outcomes, and a step closer to a future where breast cancer is not just understood but effectively managed.

3. Data Understanding

3.1 Data Collection

This particular dataset is one of the most prominent dataset in the realm of the machine learning we were able to gather a csv file for this data needed for this project. It is readily available on the archives of UC Irvine (https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic)

3.2 Describe the Data

This data set had a lot of attributes equating to almost 32 attributes. And we soon find out that this is not complete, we can easily remove the data from this. This data was very informative since the null rate was 0 in all the attributes of the dataset.

Attribute Name	**Description**
Id	The Id of the patient in the dataset.
Unnamed	It was a null value in the dataset
Radius Mean	Standard deviation of gray-scale value
Perimeter Mean	Size of the core tumor
Area Mean	Area of the core tumor
Smoothness Mean	Local variation in radius lengths
Compactness Mean	Compactness Mean
Concavity Mean	Severity of concave portions of the contour
Concave Points Mean	Number of concave portions of the contour
Symmetry Mean	Symmetry of cell nuclei
Fractal Dimension Mean	| Coastline approximation - 1
Radius SE	Standard error of the mean of distances
Texture SE	Standard error of standard deviation of

	gray-scale
Perimeter SE	Standard error of perimeter
Area SE	Standard error of area
Smoothness SE	Standard error of local variation in radius lengths
Compactness SE	Standard error of compactness
Concavity SE	Standard error of severity of concave portions
Concave Points SE	Standard error of number of concave portions
Symmetry SE	Standard error of symmetry
Fractal Dimension SE	Standard error of fractal dimension
Radius Worst	Worst (largest) value of distances
Texture Worst	Worst (largest) value of standard deviation of gray-scale
Perimeter Worst	Worst (largest) value of perimeter
Area Worst	Worst (largest) value of area
Smoothness Worst	Worst (largest) value of local variation in radius lengths
Compactness Worst	Worst (largest) value of compactness
Concavity Worst	Worst (largest) value of severity of concave portions
Concave Points Worst	Worst (largest) value of number of concave portions
Symmetry Worst	Worst (largest) value of symmetry
Fractal Dimension Worst	Worst (largest) value of fractal dimension

Table 1. Description of data

3.3 Explore Data

The first review gives us that most of the features are derived from each other for Concavity Mean, Concavity Worst, Concave Points Worse these are almost derived from each other and it makes sense.

Turns out this data is best for showcasing PDA and LCA since these would reduce our dimension drastically.

3.4 Verify Data Quality

This data is quite uniform since if we take a look at all the distributions of the values, it is very obvious they are uniformly distributed and looks clean. And since these values are derived from each other they are quite related to each other.

4.0 Data Preparation

4.1 Select Data

I looked at the data and decided to keep most of the features and not skip anything except of a few things and data. For eg: I decided to remove ID and Unnamed attributes since they did not deem necessary to me. All the other attributes like Concave Mean, Compactness worst weer all retained.

4.2 Clean Data

As Stated Earlier, I remove the data attributes that didn't feel necessary and did not need to be in the data. Furthermore the diagnosis attribute was object, I converted that to a Boolean or INt 64 attribute just for the sake of uniformity in the data.

```
diagnosis                 object
```

```
diagnosis                  int64
```

Image 1: The converted class

4.3 Construct Data

I looked into the data to create a distribution of the attributes to better understand the data visually which was a great step.

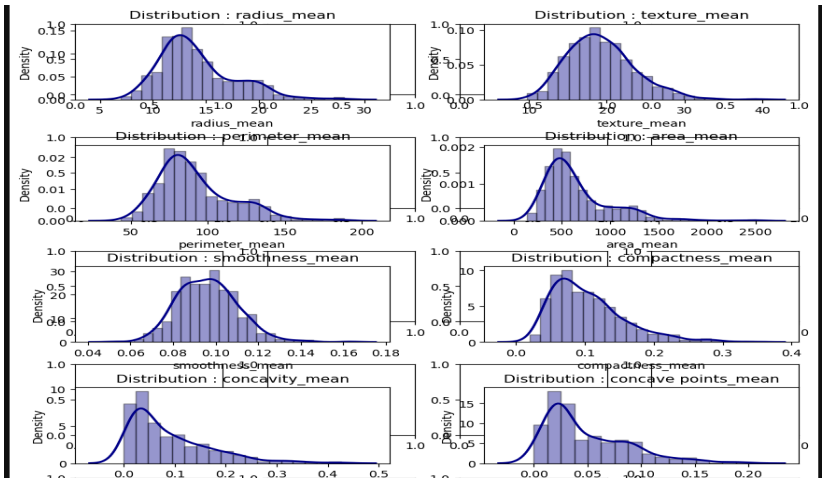

Image 2 Data Distribution of Attributes

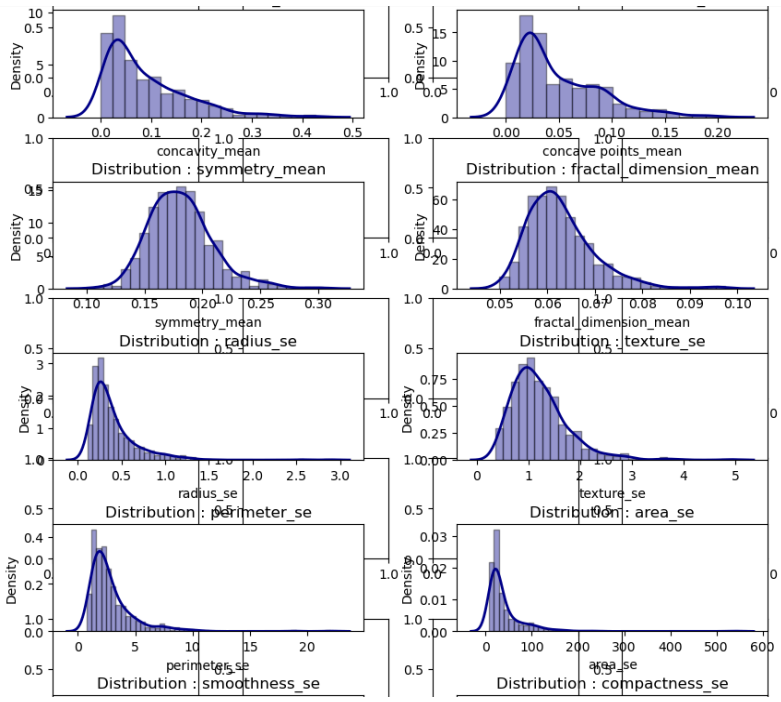

Image 3 Data Distribution of Attributes

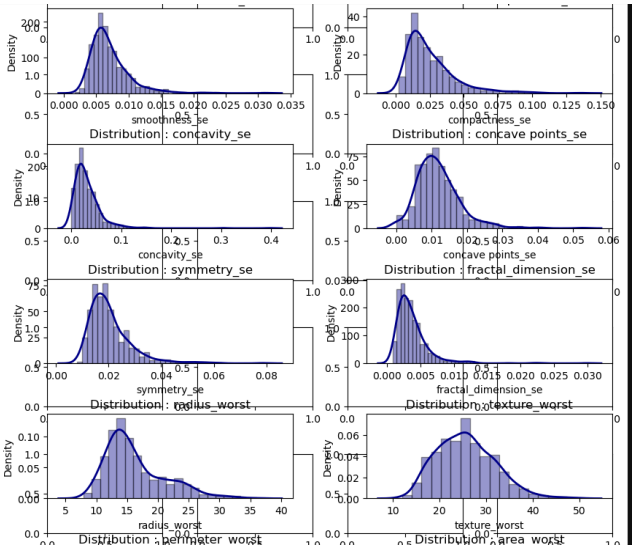

Image 4 Data Distributions

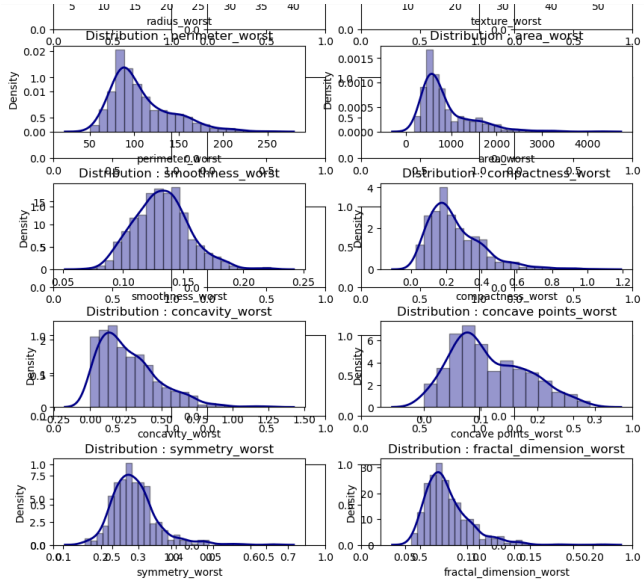

Image 5 Data Distribution

The Distribution gives us the idea the data is uniformly distributed and is a very through data conversions and once this was understood we applied dimension reduction to our attributes.
Namely LDA and PCA

Principal Component Analysis:
I finalized the principal components to be 5 based on the screen plot that was generated from the dataset.

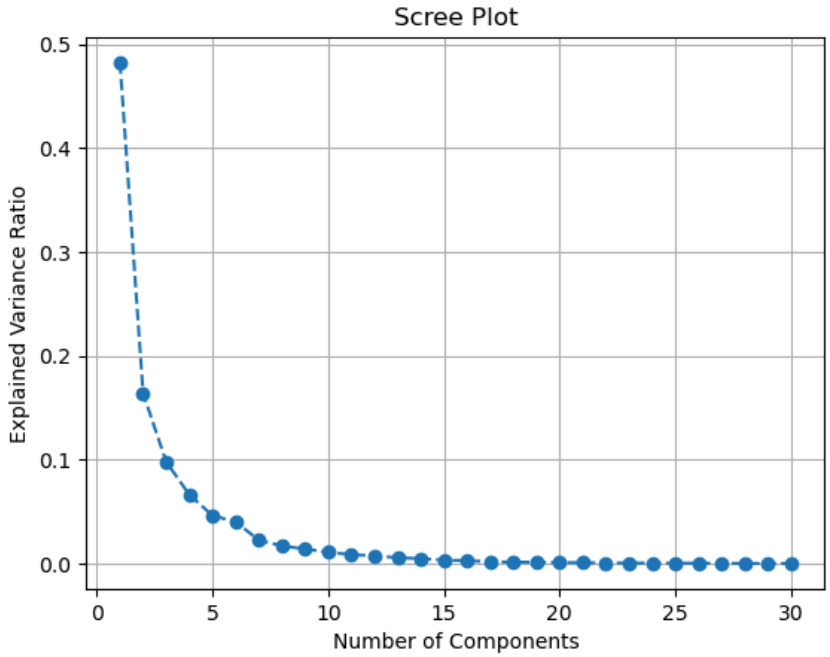

Image 6 Scree plot for identifying best N value

```
array([[ 4.60265441, -0.41075183,  3.58673363, -0.83191135, -0.03212493,
         0.39877632],
       [ 1.38906519, -1.17695852, -1.28921732, -0.70808726, -0.89083797,
        -0.5964347 ],
       [-2.07453086,  1.46991742,  1.82242056, -2.96106312, -0.24405341,
         1.72068785],
       ...,
       [-3.0473129 , -1.45644194, -1.17135382,  0.35041136,  1.01672826,
        -0.36142598],
       [-1.62385412,  0.85135294, -1.06476527,  1.36632565, -0.10819286,
         0.25755375],
       [-2.10599918,  0.46944449, -0.99214498,  0.26847057, -0.73017165,
        -0.85068029]])
```

Image 7 PCA converted values

Converted the entire attributes to look like something like this.

Linear Discriminant Analysis converted the dataset for us which is readily available on the sklearn modules

LDA values were very closely identified and turns out they are very minute since interclass relation is very high.

```
[[ 2.85785121e+00],
 [ 1.17253731e+00],
 [-1.37826623e+00],
 [-1.52635997e+00],
 [ 3.56084740e+00],
 [-1.42856836e+00],
 [ 4.42636560e+00],
 [-8.80581069e-01],
 [-3.49252081e-01],
 [-4.85834440e-01],
 [ 3.27937495e+00],
 [-1.04712431e+00],
 [-1.81829347e+00],
 [-1.96414726e+00],
 [-2.62599415e+00],
 [ 1.24356939e+00],
 [-1.65518980e+00],
 [-1.83187464e-01],
```

Image 8 LDA converted Data

4.4 Integrate Data

This dataset did not need any integration.

4.5 Format Data

This dataset did not need to be formatted since the initial file was csv which was sufficient for our data process and modeling.

5.0 Modeling and Evaluation

5.1 Select the Modeling Technique

Since we had predetermined modeling techniques, we used Support Vector Machines, Random Forest, MultiLayer Perceptron.
The processing for the respective modeling techniques is based on the datasets we had to perform some conversions in the dataset that standardized the features and applying PCA and LDA on the attributes.

5.2 Generate Test Design

I created a function that could generate all the models within the requirement and used it on the dataset. Every attribute in the training set was converted by fit_transform and the test was converted by transform.

5.3 Build Model

The Standardized dataset, the pca converted dataset and lda dataset were all fed into the modeling pipeline and created the required things needed for the project.

5.4 Asses Model

The confusion matrix, the accuracy and classification report was generated for the project provided below:

1. For Standardized Dataset where all the values were standardized for this project.

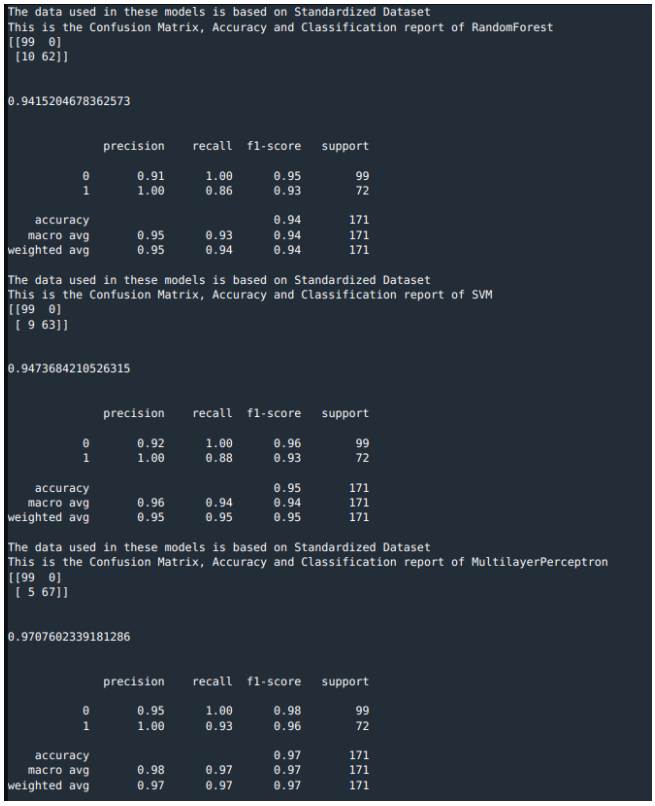

Image 9 Standardized Data Evaluation

2. PCA based values running through all the models with their respective accuracies

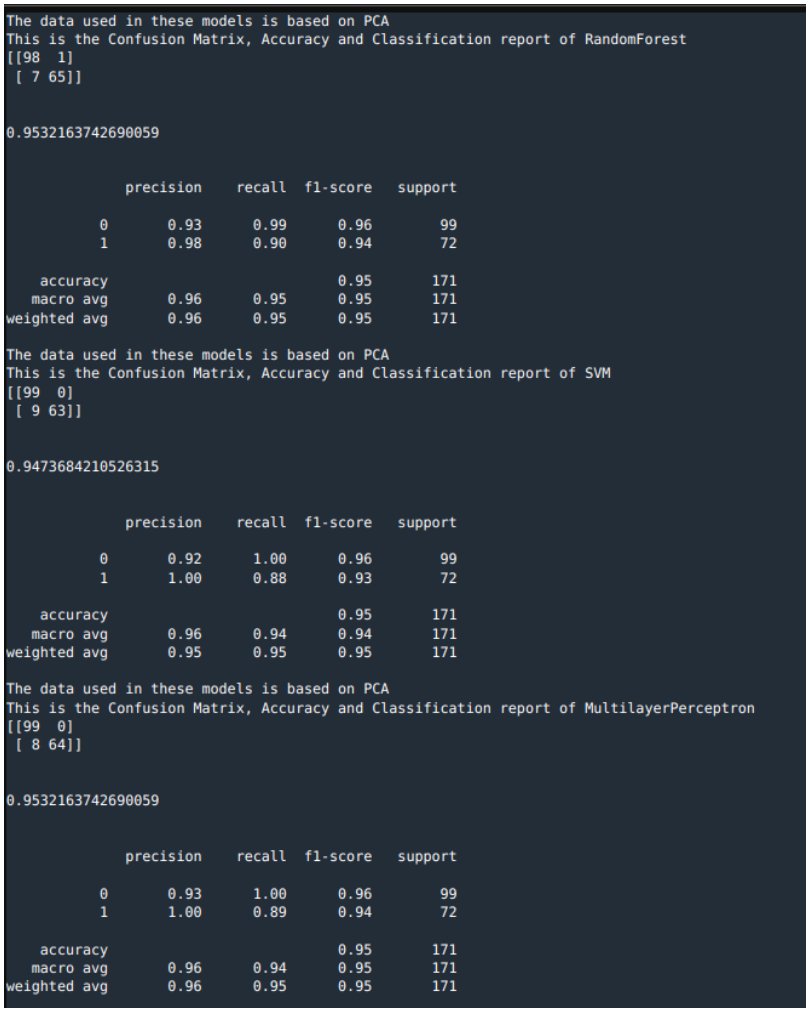

Image 10 PCA dataset Evaluation on all the models

3. LDA based values running through all the models with their respective accuracies

```
The data used in these models is based on LDA
This is the Confusion Matrix, Accuracy and Classification report of RandomForest
[[99  0]
 [ 9 63]]


0.9473684210526315


              precision    recall  f1-score   support

           0       0.92      1.00      0.96        99
           1       1.00      0.88      0.93        72

    accuracy                           0.95       171
   macro avg       0.96      0.94      0.94       171
weighted avg       0.95      0.95      0.95       171

The data used in these models is based on LDA
This is the Confusion Matrix, Accuracy and Classification report of SVM
[[99  0]
 [10 62]]


0.9415204678362573


              precision    recall  f1-score   support

           0       0.91      1.00      0.95        99
           1       1.00      0.86      0.93        72

    accuracy                           0.94       171
   macro avg       0.95      0.93      0.94       171
weighted avg       0.95      0.94      0.94       171

The data used in these models is based on LDA
This is the Confusion Matrix, Accuracy and Classification report of MultilayerPerceptron
[[99  0]
 [ 9 63]]


0.9473684210526315


              precision    recall  f1-score   support

           0       0.92      1.00      0.96        99
           1       1.00      0.88      0.93        72

    accuracy                           0.95       171
   macro avg       0.96      0.94      0.94       171
weighted avg       0.95      0.95      0.95       171
```

Image 11 LDA based evaluation

5.5 Evaluate results

Data	SVM	Random Forest	MLP
Standardized	94.15%	94.7%	97.7%
PCA	94.7%	95.32%	95%
LDA	94.7%	94.15%	94.7%

Table 2 Comparison of Results

5.6 Review Process

These values can indicate two things, one being that although the accuracy in the SVM model the accuracies have increased, they are not very significant. PCA based showed max accuracy on Random Forest and Multilayer perceptron. Upon closer look of the data, the confusion matrix seemed to be very identical. This is because the dataset is small for a complex model meaning overfitting is happening.

5.7 Determine next steps

In this case, since the model is not being deployed, our next steps involve further analysis of the results if deemed satisfactory. As that is the case, we have assessed the results of the models to be satisfactory, with further details given as to why in the subsequent chapters.

6.0 Discussion of Results

In this study, we explored the application of three different dimensionality reduction techniques—Standardization, Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA)—on a dataset focused on breast cancer classification. The performance of three machine learning models (Support Vector Machine - SVM, Random Forest, and Multilayer Perceptron - MLP) was evaluated on each dataset. The obtained accuracy values are presented in the table below:

Advantages of Standardization:
The Standardized dataset consistently demonstrated high accuracy across all three models, with MLP achieving an impressive 97.7% accuracy.
Standardization is effective in bringing features to a common scale, enabling models to converge faster and improving performance.

Advantages of PCA:

- PCA resulted in competitive accuracy values, particularly showcasing improvements in SVM and Random Forest models.
- PCA reduces the dataset's dimensionality while retaining as much variance as possible, contributing to a more efficient model training process.

Advantages of LDA:

- LDA maintained competitive performance, with SVM achieving the same accuracy as the Standardized dataset.
- LDA focuses on maximizing the separation between different classes, making it effective in scenarios where class discrimination is crucial.

Considerations:

- MLP performed exceptionally well on the Standardized dataset, indicating that the neural network model may benefit from standardized input features.
- PCA, by capturing the most informative features, proved beneficial for Random Forest, which is known for its sensitivity to feature importance.
- LDA's performance aligned closely with the original Standardized dataset, suggesting that in this specific context, LDA might not offer significant advantages over standardization.

Conclusion:

- Standardization consistently performed well, demonstrating the robustness of this preprocessing step.
- PCA showed promise, especially in improving the performance of SVM and Random Forest models.

- LDA, while maintaining competitive accuracy, did not outperform the Standardized dataset in this specific classification task.

Ultimately, the choice of dimensionality reduction technique should be guided by the specific characteristics of the dataset and the intricacies of the machine learning models employed.

7.0 Conclusion

In our exploration of breast cancer classification using machine learning, we delved into the dynamic interplay between dimensionality reduction techniques and model performance. The journey through Standardization, Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA) shed light on nuanced patterns within the Wisconsin Breast Cancer dataset.

The Standardized dataset emerged as a robust baseline, demonstrating consistent high accuracy across Support Vector Machine (SVM), Random Forest, and Multilayer Perceptron (MLP) models. The emphasis on feature scaling revealed its efficacy in enhancing convergence and boosting overall predictive capabilities.

PCA, a method adept at capturing essential information while reducing dimensionality, showcased its potential. Notably, SVM and Random Forest models reaped benefits from PCA, emphasizing its utility in scenarios where feature abundance might overshadow crucial patterns.

Linear Discriminant Analysis (LDA), designed to maximize class separability, maintained competitive accuracy but did not surpass the Standardized dataset's performance. This suggests that in this specific context, the discriminative power of LDA did not offer a significant advantage over straightforward standardization.

As we conclude this study, it is evident that the choice of preprocessing technique plays a pivotal role in shaping the success of breast cancer classification models. Standardization's reliability, PCA's dimensionality-taming prowess, and LDA's discriminative capabilities all contribute to the overarching goal of accurate and meaningful classification.
In the realm of machine learning, no one-size-fits-all solution prevails. Instead, our journey accentuates the need for thoughtful consideration, experimentation, and adaptation. Future investigations may unearth additional intricacies, refining the synergy between preprocessing methods and models for even more precise breast cancer diagnostics.

As we turn the page on this chapter, the quest for insights in breast cancer classification persists. The confluence of data, algorithms, and human intuition propels us forward, seeking solutions that hold the promise of improved patient outcomes and a brighter horizon in the fight against breast cancer.

8.0 References

[1] Assignment files
[2] Sklearn Documentations