University of Miami

# Final Project Report:

# An Analysis of the Breast Cancer Coimbra Dataset

Michela Effendie

Michelle Manfrini

CSC 642: Statistical Learning with Applications

Dr. Vanessa Aguiar

May 7th, 2024

## Introduction

Breast cancer arises from the cells of the breast as a malignant tumor, also known as a collection of cancer cells. The abnormal cells thicken the breast, forming lumps and/or changing the skin on the affected area [5]. Primarily affecting women, breast cancer can cause complications that target almost every area of the body. It is one of the leading causes of death among women. The World Health Organization (WHO) [10] estimated that in 2018, there were 9.6 million cancer-related deaths, 2.09 million being breast cancer. This value will continue to rise unless prevented. Studies reveal that by avoiding or altering key risk factors, 30% to 50% of cancer deaths can be prevented [10]. Therefore, it is important to begin screening for breast cancer early in order to increase the chances of successful treatments.

Breast Cancer Index (BCI) is a genomic test that helps physicians predict a patient's risk of early-stage breast cancer [5]. To detect significant changes in BCI, it is imperative that robust predictive models are created as screening tools. These models are based on data gathered from patients during consultations and blood collections. Machine learning can assist medical professionals in identifying cases and improving treatments to reduce cancer cells [5].

In this project, we aim to use the Breast Cancer Coimbra data for analysis [1, 2]. The data includes multiple attributes related to breast cancer risk. For example, Monocyte Chemoattractant Protein-1 (MCP-1) reveals the levels of a cytokine involved in inflammation and Homeostatic Model Assessment (HOMA) assesses insulin resistance and beta-cell function. The dataset also includes the breast cancer diagnosis for each patient, allowing us to train a model that will use these variables to predict future diagnoses. Our goal is to develop a model that can accurately predict an individual's diagnosis based on the quantitative attributes.

Furthermore, it is important to verify if the final model is clinically relevant and applicable. By assessing the practical implications of the model's predictions, we will evaluate its potential utility in assisting healthcare professions identify early-stage breast cancer.

Research Questions

Using the Breast Cancer Coimbra dataset, we aim to classify the individuals into healthy controls and affected patients. It is a classification problem of which the binary dependent variable will be the basis. Our goal is to develop a model that can accurately predict an individual's diagnosis based on the quantitative attributes. We also strive to answer the following research questions:

1. Can the model accurately predict whether an individual has breast cancer based on the provided predictors?
2. Which attributes are significant in distinguishing between healthy and affected individuals?
3. How well does the model perform in terms of accuracy and reliability?
4. How clinically relevant and applicable is the model? Can it be used by healthcare professionals for early-stage breast cancer detection?

## State-of-the-art

The Breast Cancer Coimbra dataset has been studied extensively. A past paper by the University of Coimbra [8] proposed a model for detection based on breast cancer biomarkers. In Patricio et al. [8], predictive models were built with logistic regression, support vector machines,

and random forests. Each classification algorithm took in varying combinations of predictors that were deemed significant during a previous multivariate analysis in which a Receiver Operating Characteristic (ROC) analysis was conducted. The Gini coefficient estimated how important variables were as predictors for breast cancer. Starting with the more significant, these were: Glucose, Resistin, Age, BMI, HOMA, Leptin, Insulin, Adiponectin, MCP-1. Then, Monte Carlo Cross-Validation (MCCV) was performed to minimize bias.

The three models were trained and assessed based on the area under the curve from the ROC, the sensitivity, and specificity. The models were trained on about 70% of the data and 95% confidence intervals were computed. Since the cross-validation procedure did not fully eliminate bias, a power analysis was conducted as well. The study concluded that the best combination of sensitivity and specificity was acquired through support vector machines, specifically with the variables Glucose, Age, BMI, and Resistin. The model had a strong capacity to distinguish between classes with a sensitivity of 82-88% and specificity of 85-90%.

An additional study by Alfian et al. [3] also utilized support vector machines to build a predictive model. However, these investigators built an integrated model with extra-trees into web-based breast cancer prediction to improve the final performance. The extra-trees extracted important risk factors in addition to the SVM generating accurate predictions. The extra-trees identify the most significant features in order to enhance the predictions. Furthermore, the integration into web-based prediction added to the possibilities of real-life applications for medical professionals. Ultimately, they used 10-fold cross-validation to find an accuracy of 80.23%. Although lower than other studies, this paper found practical applications in their models. None of the existing papers analyzing this dataset offered any real-world applications, unlike the web-based application created in this study.

This dataset has also been analyzed in the past through a modified spherical k-means. K-means is a clustering technique utilized for data analysis where the information is split into clusters based on euclidean distance. In Rustam et al. [9], the experimenters improved the spherical k-means algorithm by adding a radial basis function kernel to improve efficiency and accuracy. Although the results from this study revealed that the spherical k-means was a successful classifier, we will test a different method. Instead of just using a clustering algorithm, we aim to classify the individuals into the two groups (healthy controls and affected patients). It is a classification problem of which the binary dependent variable will be the basis.

Ghani et al. [4] used recursive feature elimination for feature selection and combined many models such as decision trees, K-nearest neighbor, naive bayes, and artificial neural networks. The investigators implemented the hold-out validation method to find that the artificial neural network had the strongest performance with an accuracy of 80%. Katun et al. [6] implemented naive bayes, random forest, multilayer perceptron, and logistic regression. The hold-out validation method revealed that the multilayer perceptron model had the highest accuracy of 85%. Nanglia et al. [7] worked with an ensemble model and chi-square-based feature selection. Through K-nearest neighbors, decision trees, and support vector machines, the investigators created a stacking ensemble model. After applying a 20-fold cross-validation, the ensemble model achieved the greatest accuracy by a 78% difference.

## Materials and Methods

Dataset

There will be two datasets used in this project, both relating to breast cancer. The first dataset is the “Breast Cancer Coimbra” dataset derived from Kaggle which was authored by Vivek Agrawal. This particular dataset is a synthetic dataset from a deep learning model that was trained on the “Breast Cancer Coimbra” dataset, originally found in the UC Irvine Machine Learning Repository. The original dataset has 116 instances based on clinical observations from 64 patients with breast cancer and 52 healthy controls. The deep learning model Vivek used was able to create a synthetic dataset with 4,000 instances, creating the dataset found in Kaggle.

This synthetic dataset has feature distributions that closely resemble the original dataset, consisting of nine numerical variables and a binary categorical variable. The nine numerical variables are:

- Age (years): Represents the age of individuals in the dataset.
- BMI ($kg/m^2$): Body Mass Index, a measure of body fat based on weight and height.
- Glucose (mg/dL): Reflects blood glucose levels, a vital metabolic indicator.
- Insulin (μU/mL): Indicates insulin levels, a hormone associated with glucose regulation.
- HOMA: Homeostatic Model Assessment, a method assessing insulin resistance and beta-cell function.
- Leptin (ng/mL): Represents leptin levels, a hormone involved in appetite and energy balance regulation.
- Adiponectin (μg/mL): Reflects adiponectin levels, a protein associated with metabolic regulation.

- Resistin (ng/mL): Indicates resistin levels, a protein implicated in insulin resistance.
- MCP-1 (pg/dL): Reflects Monocyte Chemoattractant Protein-1 levels, a cytokine involved in inflammation.

The binary categorical variable is labeled with 1 for healthy controls and 2 for patients with breast cancer. The numerical variables in the dataset are originally gathered from routine blood analysis and they act as the groundwork for potential indicators of breast cancer. In addition, these particular datasets do not contain missing values and are in comma-separated values (csv) files.

For the purpose of this project, both datasets, the synthetic and original, will be used. The model will be trained on the synthetic dataset and validated on the original dataset. Through this method, it will be possible to determine the accuracy of the model's prediction on actual data from clinical observations.

## Methods

### *Exploratory Data Analysis*

Exploratory data analysis is conducted on both datasets to examine and analyze the data. The variables within the dataset were assigned to the numerical data type. Since this data type does not reflect the nature of every variable in the dataset, adjustments were made. The modification of data types are as follows:

- Age: integer
- BMI: numeric
- Glucose: integer

- Insulin: numeric
- HOMA: numeric
- Leptin: numeric
- Adiponectin: numeric
- Resistin: numeric
- MCP.1: integer
- Classification: factor

In order to understand the distribution of each variable in the datasets, a histogram for each numeric variable was plotted. In addition, the correlation matrix was calculated and visualized to examine the relationships between variables in both datasets. The correlation coefficients were analyzed to determine the degree of correlation between the variables.

The question regarding the homogeneity of the data arises. In order to answer this question, Principal Component Analysis (PCA) is conducted on the data. Since the aim of this project is to build a predictive model that can generalize to unseen data, it is important to examine and confirm that the synthetic and original data originate from similar distributions. Under circumstances that it does not, the model will fail to generalize well, reducing its performance. This method inspects the data for presence of homogeneity by comparing the distributions of principal components (PCs) between the synthetic data and the original data.

Moreover, the variance explained by each PC is calculated for additional analysis since it can reveal the magnitude of the total variability in the dataset captured by each PC. This information can be helpful to understand the underlying structure of the data because components that explain a large amount of variance may represent important patterns or

relationships. This analysis allows for a deeper understanding of the data to make better decisions in terms of feature selection and model building.

*Modeling*

Prior to fitting the data to the model, the datasets are split into train, validation, and test sets. The training and validation sets comprise 70 percent and 30 percent of the synthetic data respectively, while the test set contains 100 percent of the original data. The model will be trained to find patterns on the training set, evaluated on the validation set, and tested using the pattern it found on the test set.

In order to build the best performing model, the data is first fit into different classical machine learning models, namely:

- K-Nearest Neighbors Classifier (KNN): Classifies new data points based on the majority class of their 'k' nearest neighbors in the feature space. The choice of 'k' determines the number of neighbors considered for classification.
- Linear Discriminant Analysis (LDA): A linear classification algorithm that finds the linear combination of features that best separates classes. It calculates the optimal decision boundary by maximizing the between-class variance and minimizing the within-class variance.
- Quadratic Discriminant Analysis (QDA): Similar to LDA but allows for non-linear decision boundaries. Unlike LDA, QDA does not assume equal covariance matrices for different classes, meaning it calculates a separate covariance matrix for each class, leading to more flexible decision boundaries.

- Logistic Regression: Used for binary classification, it models the probability of the binary outcome using a logistic function, which maps the input features to the probability of belonging to one of the two classes. Despite its name, logistic regression is a linear model for classification, not regression.

Once the data are trained, validated, and tested on each model listed above, these base models' hyperparameters are tuned to increase its performance and predictive capabilities. During tuning, resampling is conducted using 5-fold cross-validation.

A parameter grid is created to tune the logistic regression model. The optimal model is selected based on the largest accuracy. Based on this optimal model, predictions are made on the validation data and probabilities are converted into class predictions. Then, this model is tested on the test set. The same method was used to tune the KNN model, but instead of tuning the alpha and lambda values, the value of k is tuned.

Since tuning parameters do not typically exist for LDA and QDA models, the hyperparameter tuning approach applied to the logistic regression and KNN models cannot be applied. LDA mainly relies on estimates of class means and covariance matrices; therefore, experimentation with different values of the shrinkage parameter in LDA can be conducted to potentially optimize the performance of the model. The optimization approach for the LDA model consists of training multiple LDA models with different values of the shrinkage parameter and evaluating the performances. On the other hand, the cross-validation approach is applied to the QDA model to evaluate the model's performance.

## Evaluation

In terms of evaluation, one of the methods applied is the hold-out approach. The two datasets, the synthetic and original datasets, are split into train, validation, and test sets. The original dataset is used as the test set throughout the project, and the synthetic dataset is split in a 70:30 ratio for the training and validation sets. The hold-out approach is implemented in the evaluation stage of the project since it is simple and easy to implement, but it has some disadvantages, which includes the high variability of the validation error, and since the training data is a subset of the observations used to fit the model, training is done on fewer observations which may affect the performance of statistical methods negatively.

Each base and optimized models' performance is measured through the following metrics:

- Confusion matrix
- Cross-validation
- Precision
- Recall
- Accuracy

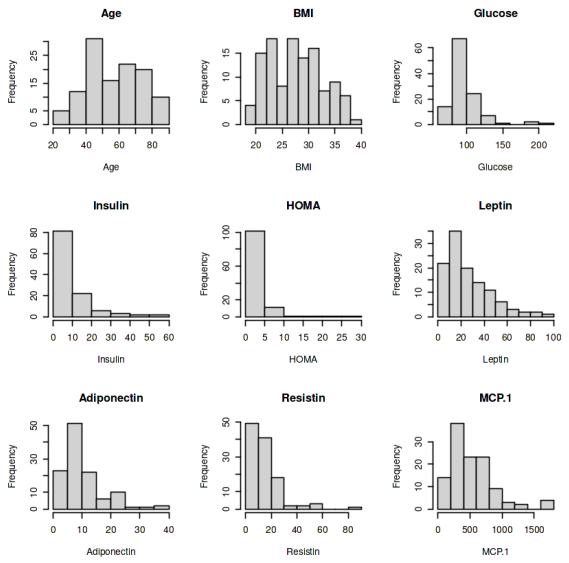
Accuracy measures the proportion of correctly classified samples out of the total set and creates a broad idea of how well a model performed. A high accuracy is ideal. Validation error is the proportion of incorrectly classified samples out of the total validation set. Also known as misclassification error, this metric measures the model's ability to generalize to new data that it was not trained on. We desire a minimal validation error in model training. Precision is the proportion of true positive predictions out of all instances labeled as positive. This is a significant metric in our work because the cost of false positives can be high. A high precision can reveal

that the model is not falsely labeling instances as positive. Lastly, recall is the proportion of true positive predictions out of the true positive cases in the data. Also known as sensitivity or true positive rate, recall is significant when the cost of false negatives is high. In our work, a false negative could mean missing a breast cancer diagnosis and lead to losing a life. Therefore, it is very important to have a high recall score, which supports that the model is correctly identifying the majority of positive cases.

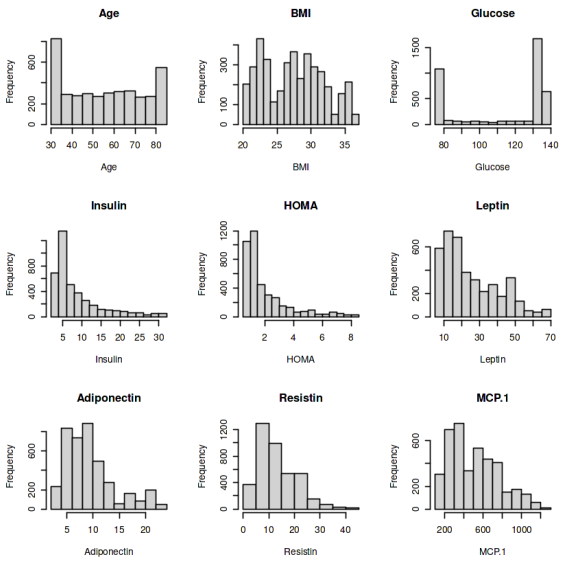
## Results

### *Exploratory Data Analysis*

The results of the exploratory data analysis revealed that the histograms of the numeric variables in both datasets exhibited distributions that deviate from the normal distribution. The insulin, HOMA, leptin, adiponectin, resistin, and MCP.1 histograms from both datasets display a right-skewed distribution. The BMI histogram for both datasets mimic the plateau distribution. The age and glucose histograms show different shapes for each dataset. The synthetic dataset's age and glucose histogram resemble the dog-food distribution while the original dataset resembles an edge-peak and right-skewed distributions respectively. Since none of the variables presents a normal distribution, the predictions made from the model trained and tested based on these datasets may be affected.

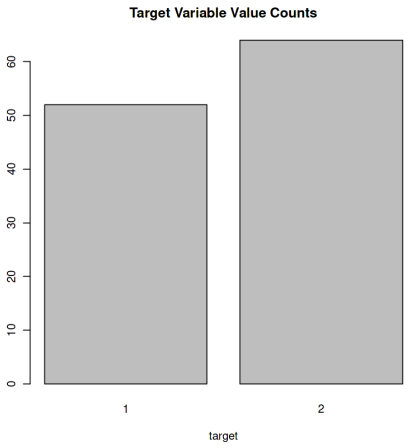



[Table 1] Original dataset histograms.

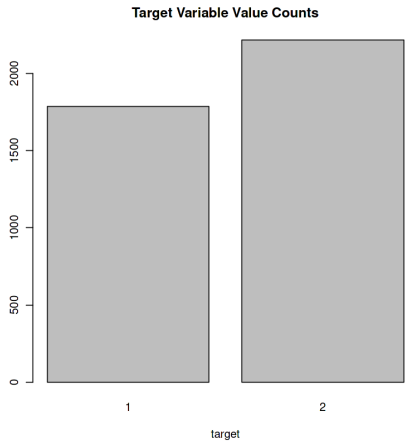



[Table 2] Synthetic dataset histograms.

The distribution of the binary categorical variable of both datasets seem to be fair. The number of healthy controls, noted by the number 1, and patients with breast cancer, 2, are calculated and displayed in a bar plot:




[Table 3] Original dataset binary variable distribution.

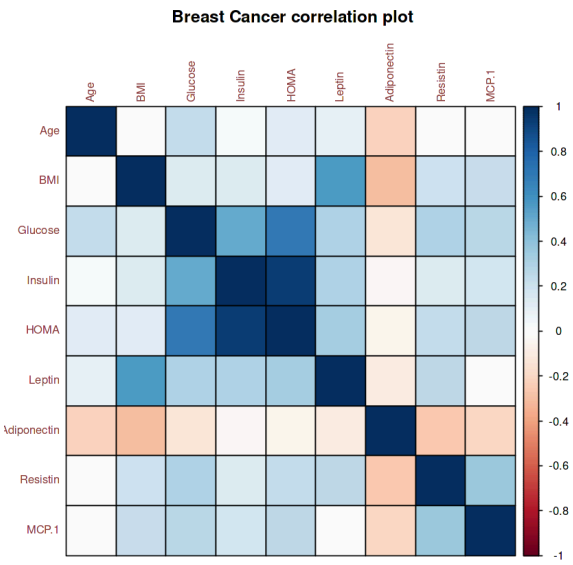



[Table 4] Synthetic dataset binary variable distribution.

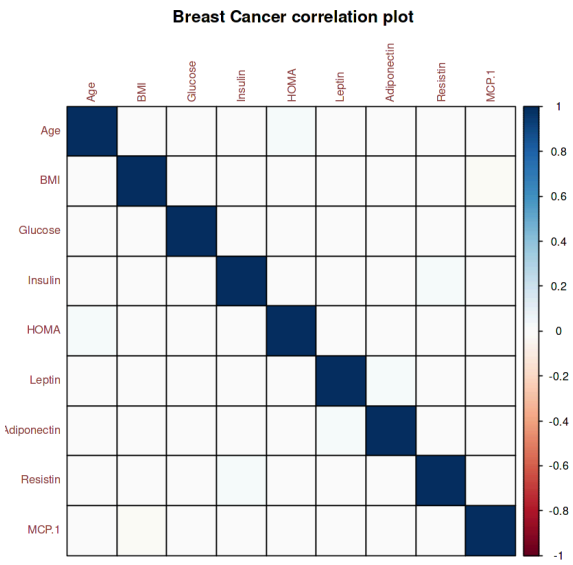
There are 1784 healthy controls and 2216 patients with breast cancer in the synthetic dataset, while there are 52 and 64 for each category in the original dataset. The number of observations in each category is not equal, but it does not appear to be problematic.

The summary statistics of the numeric variables of both datasets reveals the min, max, mean, median, first quartile, and third quartile of each variable. The mean and medians of the variables in both datasets appear to be similar in values. The difference lies in the min, max, first quartile, and third quartile values. The synthetic dataset's variables seem to have a smaller range of values, since the difference between the min and max values is closer than the original dataset's difference between these values. This analysis from the summary statistics may explain the performance of the models trained on the synthetic dataset when tested on the original dataset.

In order to understand the variables further, the correlation matrix is calculated and visualized:



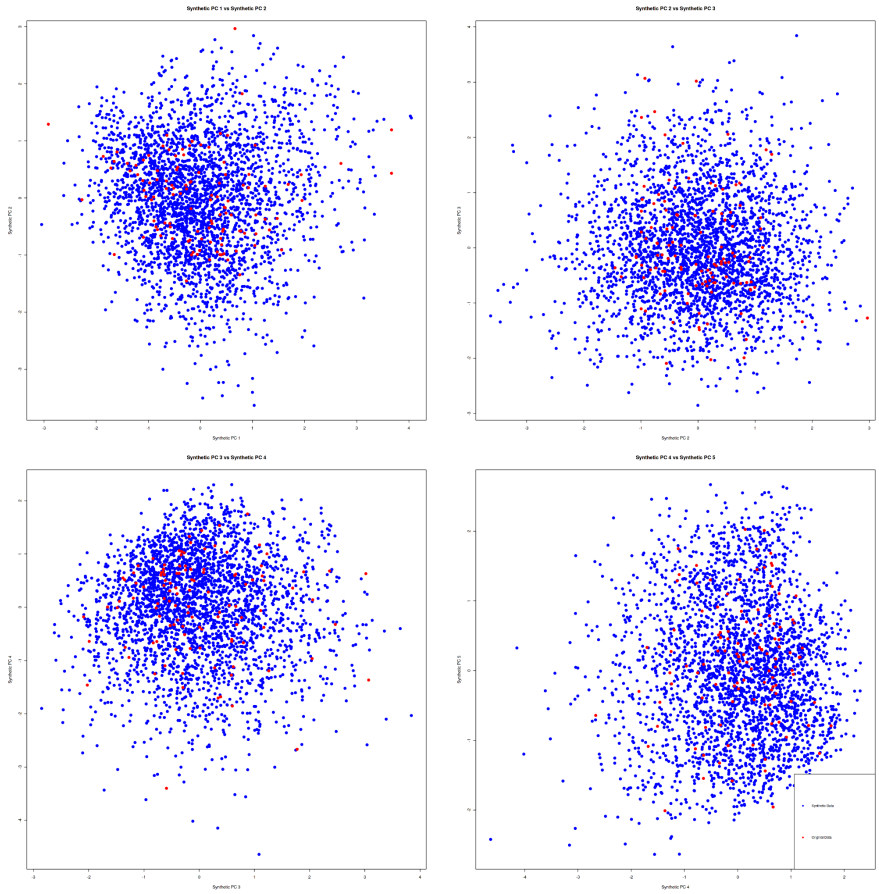

[Table 5] Correlation matrix for the original dataset.




[Table 6] Correlation matrix for the synthetic dataset.

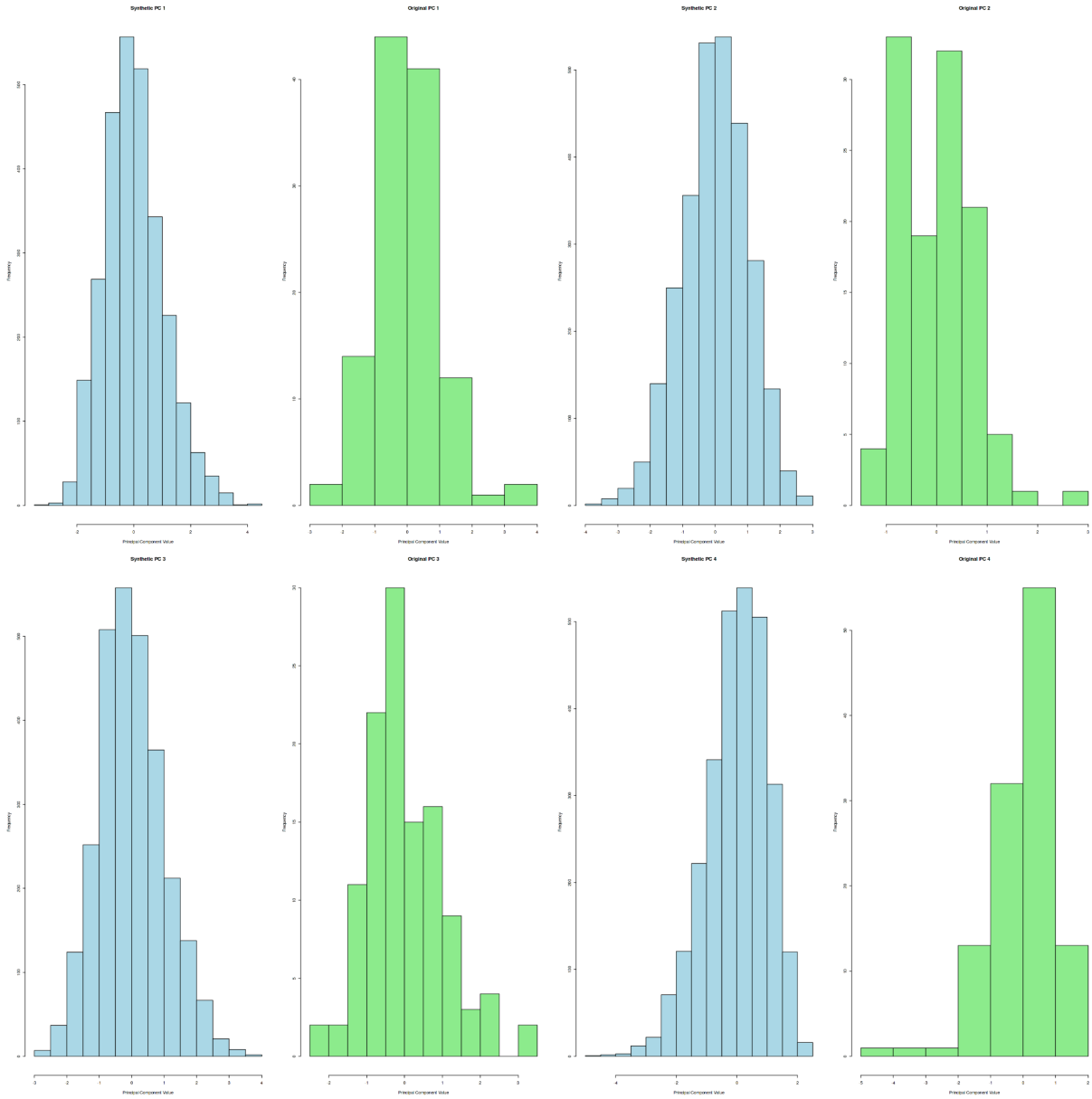
The variables in the synthetic dataset seem to have low correlation since its correlation coefficients are close to zero. The opposite is true for the original dataset. Its variables exhibit medium to strong correlations. This result is surprising, considering the synthetic dataset is generated from a deep learning model trained on the original dataset.

Due to the analysis derived from the correlation matrix, further investigation needs to be done. For simplicity, the distribution of PCs are plotted, since visual inspection is more practical to notice patterns or clusters:




[Table 7] PCA scatterplot.

Homogeneous data would display points from both datasets distributed similarly without forming distinct clusters in a scatterplot. Based on this definition, the scatterplots generated in this method seem to suggest that the datasets are homogeneous, since there are no clusters and points from both datasets are spread out. In order to confirm this analysis, histograms are generated to provide a different lens:




[Table 8] PCA histograms.

| PC | Variance_Explained | Proportion_Variance_Explained |
|---|---|---|
| <int> | <dbl> | <dbl> |

| | | |
|---|---|---|
| 1 | 1.0554052 | 0.1172672 |
| 2 | 1.0518554 | 0.1168728 |
| 3 | 1.0336469 | 0.1148497 |
| 4 | 1.0165072 | 0.1129452 |
| 5 | 1.0005579 | 0.1111731 |
| 6 | 0.9842718 | 0.1093635 |
| 7 | 0.9646444 | 0.1071827 |
| 8 | 0.9609623 | 0.1067736 |
| 9 | 0.9321489 | 0.1035721 |

[Table 9] PCA variance explained.

The shapes of both the synthetic and original dataset are very similar. They exhibit similar distribution shapes with centers that are close or aligned. In addition, the spread of the histograms mimic each other, further indicating homogeneity. Based on the results, the variance explained and proportion of variance explained do not have noticeably large values.

In order to obtain the best model possible for each approach, we conducted hyperparameter tuning. The details were discussed in the previous "Methods" section. To assess the performance of each model, we focused on the accuracy, validation error, precision, and recall score. The results are displayed in the tables below:

*Summary of Model Metrics on Validation and Test Set Before Hyperparameter Tuning*

| | **Test Error** | **Validation Error** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
|---|---|---|---|---|---|---|
| **Logistic Regression** | 0.552 | 0.565 | 0.435 | 0.891 | 0.523 | 0.659 |
| **KNN** | 0.457 | 0.454 | 0.546 | 0.953 | 0.550 | 0.701 |
| **LDA** | 0.509 | 0.437 | 0.563 | 0.891 | 0.523 | 0.659 |
| **QDA** | 0.526 | 0.459 | 0.541 | 0.750 | 0.516 | 0.611 |

*Cross-Validation Summary of Model Metrics on Test Set After Hyperparameter Tuning*

| | **Accuracy** | **Test Error** | **Precision** | **Recall** | **F1 Score** |
|---|---|---|---|---|---|
| **Logistic Regression** | 0.448 | 0.552 | 0.984 | 0.548 | 0.704 |
| **KNN** | 0.517 | 0.482 | 0.640 | 0.554 | 0.594 |
| **LDA** | 0.491 | 0.508 | 0.891 | 0.523 | 0.659 |
| **QDA** | 0.474 | 0.526 | 0.750 | 0.516 | 0.611 |

*Confusion matrix of models after hyperparameter tuning: Validation Set*

| Logistic Regression | KNN | LDA | QDA |
|---|---|---|---|
| val.y<br>glm_pred 1 2<br>0 0 2<br>1 522 676 | val.y<br>knn_pred 1 2<br>1 214 280<br>2 308 398 | val.y<br>lda_pred 1 2<br>1 0 2<br>2 522 676 | val.y<br>qda_pred 1 2<br>1 100 129<br>2 422 549 |

*Confusion matrix of models after hyperparameter tuning: Test Set*

| **Logistic Regression** | **KNN** | **LDA** | **QDA** |
|---|---|---|---|
| test.y<br>glm.pred_test 1 2<br>0 0 7<br>1 52 57 | test.y<br>knn.pred_test 1 2<br>1 19 23<br>2 33 41 | test.y<br>lda.class 1 2<br>1 0 7<br>2 52 57 | test.y<br>qda.class 1 2<br>1 7 16<br>2 45 48 |

The results above reflect the comprehensive evaluation of various models before and after hyperparameter tuning for optimization of the models' performance. The summary of model metrics on the validation set before hyperparameter tuning brings attention to the LDA model. It has the highest accuracy score and its other metrics seem to be on par with the other models, but the confusion matrix seems to not support this idea. The LDA model has a high tendency to predict one class over the other which means that it won't perform well with other datasets.

After cross-validation and hyperparameter tuning, the results continue to reveal that the logistic regression model has the strongest all around performance. KNN model outperforms the LDA model in terms of accuracy, recall, and test error. The logistic regression model seems to

have a lower accuracy rate than all the other models, but achieved the highest precision and F1 score. Although the LDA seems to be the best performing model in terms of all metrics, its confusion matrix seems to disagree. The model has a high tendency to predict that patients have breast cancer, and this situation is also evident in the logistic regression model.

The KNN and QDA models did not outperform the LDA model, but they seem to be predicting more evenly, with a lower tendency to predict a certain class. Based on these observations, there does not seem to be a model that performs significantly better than the other models. There are trade-offs when selecting the "best" model, but the KNN seems to be the highest performing model based on all the evaluation metrics including the confusion matrix; it does not have a high tendency to predict one class over the other.

In an attempt to improve our results, we trained, validated, and tested an additional logistic regression model on only the original Breast Cancer Coimbra dataset. In doing so, we acquired an accuracy of 80%. Therefore, future studies should build models utilizing only the original dataset.

## Conclusion and Future Work

Our project proposed four breast cancer prediction models based on logistic regression, K-Nearest Neighbors Classifier (KNN), Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA). Results reveal that none of the models significantly outperform the other models. Although KNN seems to be the highest performing model based on all the evaluation metrics including the confusion matrix, it does not have a high tendency to predict one class over the other. Further research should assess varying prediction models to determine which is the best at predicting breast cancer. The models should also be trained, validated, and tested using the single original Breast Cancer Coimbra dataset. We did not have success with the synthetic data derived from Kaggle.

Breast cancer is one of the leading causes of death among women, making it important to begin screening for breast cancer early in order to increase the chances of successful treatments. A robust model can assist medical professionals in identifying cases and reducing breast cancer risk. Future studies should assess additional prediction models, feature selection methods, and more expansive clinical datasets. In addition, more research must be conducted to verify the relationship between these quantitative attributes and a true diagnosis of breast cancer.

**Citations**


1. "Breast Cancer Coimbra." UCI Machine Learning Repository, archive.ics.uci.edu/dataset/451/breast+cancer+coimbra. Accessed 25 Jan. 2024.
2. Agrawal, Vivek. "Breast Cancer Coimbra." Kaggle, 7 Jan. 2024, www.kaggle.com/datasets/atom1991/breast-cancer-coimbra.
3. Alfian, G.; Syafrudin, M.; Fahrurrozi, I.; Fitriyani, N.L.; Atmaji, F.T.D.; Widodo, T.; Bahiyah, N.; Benes, F.; Rhee, J. Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method. *Computers* **2022**, *11*, 136. https://doi.org/10.3390/computers11090136
4. Ghani, M.U.; Alam, T.M.; Jaskani, F.H. Comparison of Classification Models for Early Prediction of Breast Cancer. In Proceedings of the 2019 International Conference on Innovative Computing (ICIC), Lahore, Pakistan, 9–10 November 2019; pp. 1–6.
5. Jin Yue, Na Zhao, and Liu Liu. "Prediction and Monitoring Method for Breast Cancer: A Case Study for Data from the University Hospital Centre of Coimbra." Cancer Management and Research, vol. 12, 2020, pp. 1887-1893, DOI: 10.2147/CMAR.S242027.
6. Khatun, T.; Utsho, M.M.R.; Islam, M.A.; Zohura, M.F.; Hossen, M.S.; Rimi, R.A.; Anni, S.J. Performance Analysis of Breast Cancer: A Machine Learning Approach. In Proceedings of the 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2–4 September 2021; pp. 1426–1434.

7. Nanglia, S.; Ahmad, M.; Khan, F.A.; Jhanjhi, N. An enhanced Predictive heterogeneous ensemble model for breast cancer prediction. *Biomed. Signal Process. Control* **2021**, *72*, 103279.
8. Patrício, M., Pereira, J., Crisóstomo, J. *et al.* Using Resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer* 18, 29 (2018). https://doi.org/10.1186/s12885-017-3877-1
9. Rustam, Zuherman, and Ajeng Leudityara Fijri. "IOPscience." Journal of Physics: Conference Series, vol. 1490, no. 1, IOP Publishing, 2020, p. 012028, iopscience.iop.org/article/10.1088/1742-6596/1490/1/012028.
10. World Health Organization. "Cancer." World Health Organization, www.who.int/health-topics/cancer#tab=tab_1. Accessed 24 Jan. 2024.