An Analysis of the Breast Cancer Coimbra Dataset

CSC 642 - Final Project

Michelle Manfrini
Michela Effendie

Introduction

- Breast cancer arises from the cells of the breast as a malignant tumor, also known as a collection of cancer cells.
- 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.
- 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**.
- Machine learning can assist medical professionals in identifying cases and improving treatments to reduce cancer cells

Our Goal

- We used the Breast Cancer Coimbra data for analysis, which includes multiple attributes related to breast cancer risk such as MCP-1 and HOMA
- The dataset allowed us to train a model using these variables to predict future breast cancer diagnoses based on the quantitative attributes.
- **Our goal** → develop a model that could accurately predict an individual's diagnosis based on the quantitative attributes.

Research Questions

We were faced with a classification problem of the main target was to classify the individuals into healthy controls and affected patients.

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

Patricio et al ...

built predictive models with logistic regression, support vector machines, and random forests. 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.

Ghani et al ...

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%.

Dataset

Two datasets:

1. The original “Breast Cancer Coimbra” dataset found in the UC Irvine Machine Learning Repository, with 116 instances based on clinical observations from 64 patients with breast cancer and 52 healthy controls.
2. The derived dataset from Kaggle which is a synthetic dataset from a deep learning model that was trained on the original. The synthetic dataset has 4,000 instances.

Dataset

Ten numerical variables and a binary categorical variable:

1. Age (years)
2. BMI (kg/m^2)
3. Glucose (mg/dL)
4. Insulin (μU/mL)
5. HOMA: Homeostatic Model Assessment
6. Leptin (ng/mL)
7. Adiponectin (μg/mL)
8. Resistin (ng/mL)
9. MCP-1 (pg/dL)

The binary categorical variable is labeled with 1 for healthy controls and 2 for patients with breast cancer. 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, were used. The model was trained on the synthetic dataset and validated on the original dataset. Through this method, it was possible to determine the accuracy of the model's prediction on actual data from clinical observations.

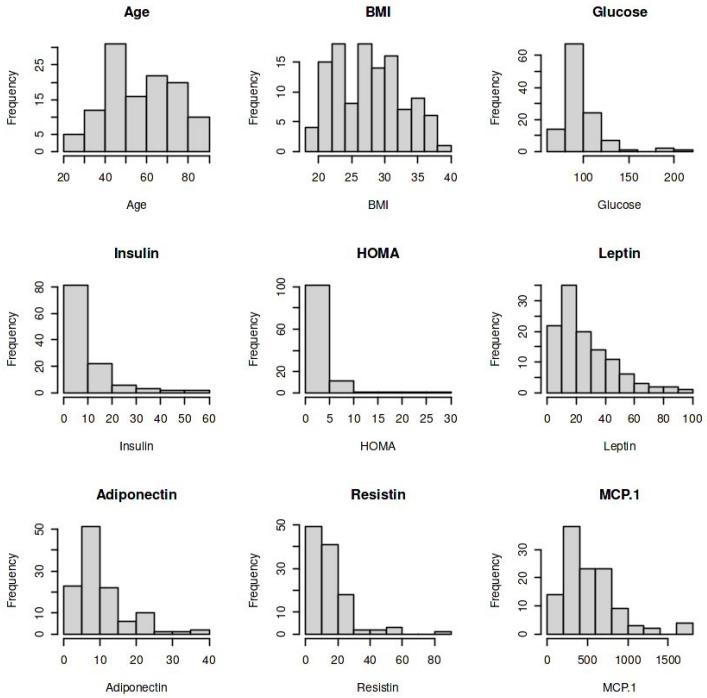
Methods - Exploratory Data Analysis

- Data type modification
- Histogram plotting
- Correlation matrices
- Principal Component Analysis (PCA)

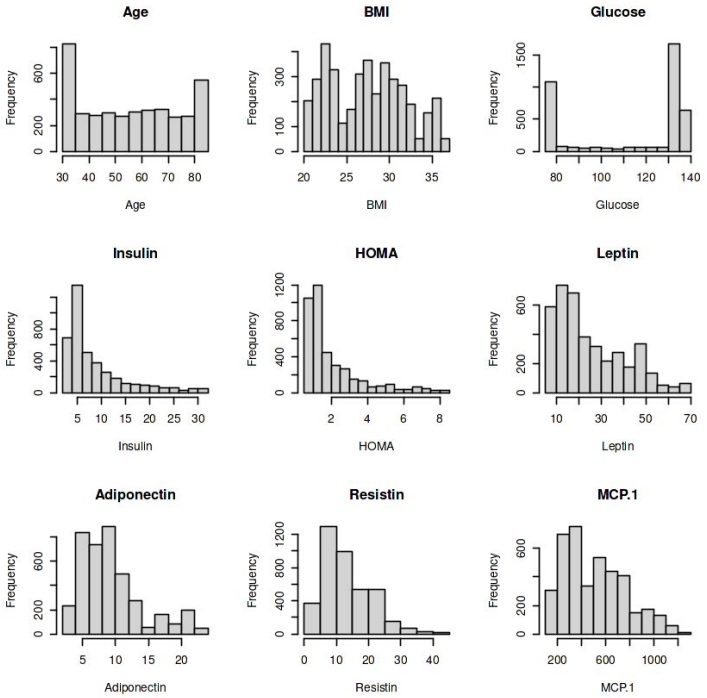
Methods

1. Train/Val Split: training and validation sets comprise 70% and 30% of the synthetic data respectively, while the test set contains 100 percent of the original data.
2. Data is fit into different classical machine learning models:
    - K-Nearest Neighbors Classifier (KNN)
    - Linear Discriminant Analysis (LDA)
    - Quadratic Discriminant Analysis (QDA)
    - Logistic Regression
3. Hyperparameters tuning and resampling using 5-fold cross-validation
4. Evaluation: hold-out approach, confusion matrix, cross-validation, precision, recall, accuracy

Results - EDA

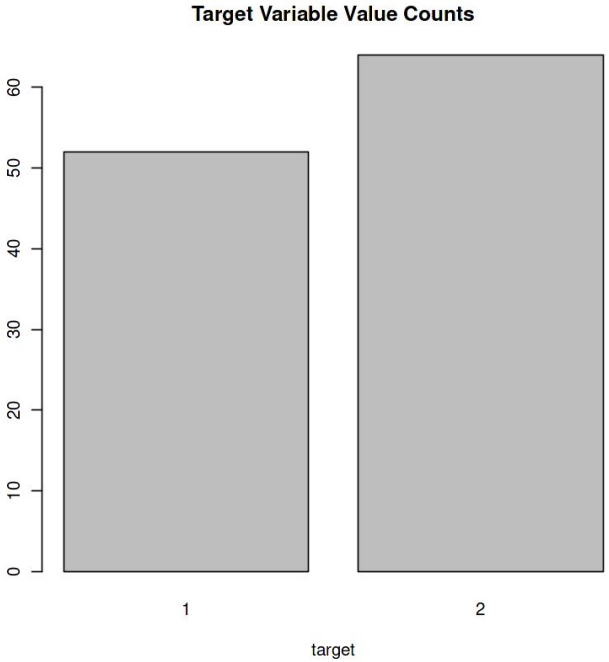


Original dataset

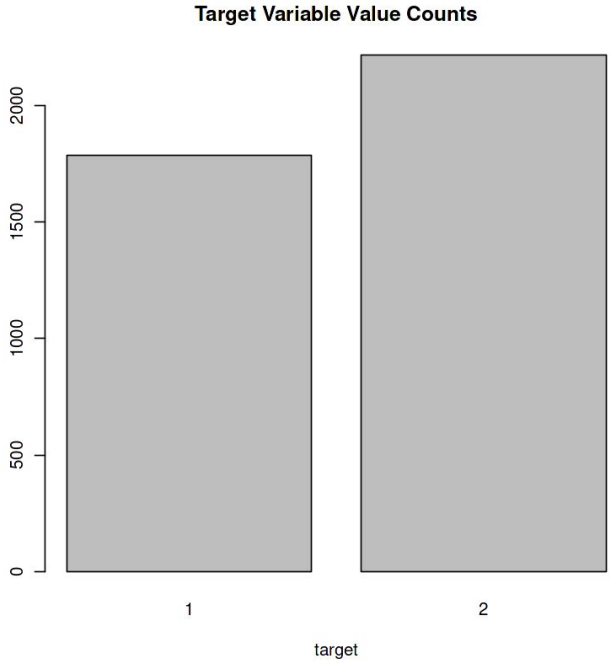


Synthetic dataset

Results - EDA



Original dataset binary variable distribution

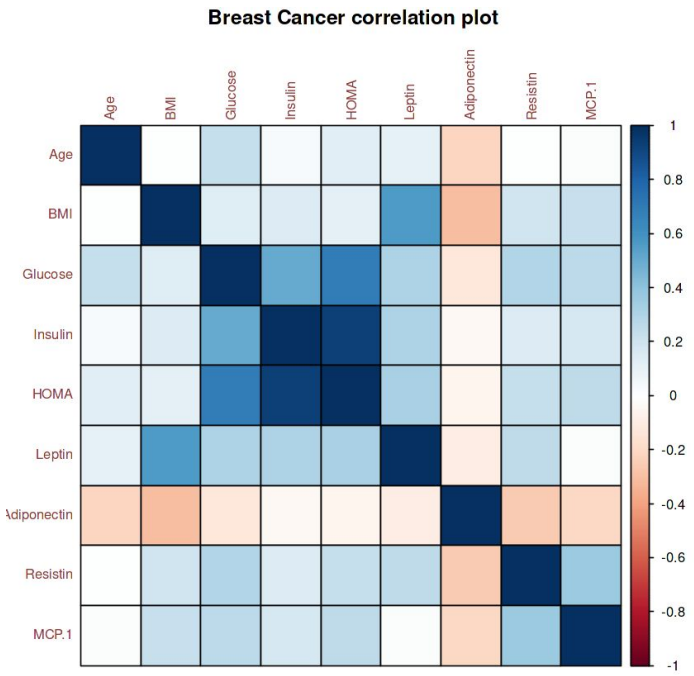


Synthetic dataset binary variable distribution

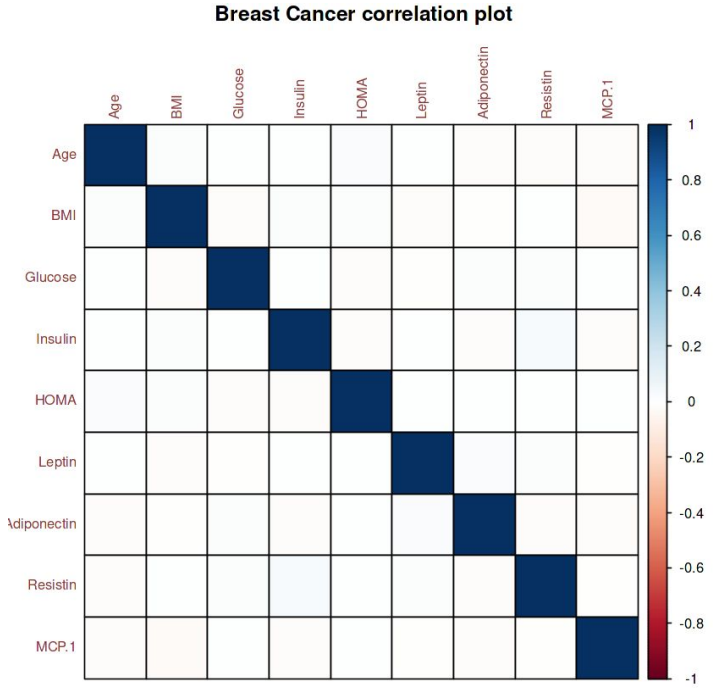
Results - EDA

- Synthetic dataset: 1784 healthy controls and 2216 patients with breast cancer
- Original dataset: 52 healthy controls and 64 patients with breast cancer
- imbalance does not appear to be problematic
- mean and medians of the variables in both datasets appear to be similar
- min, max, first quartile, and third quartile values differ
- synthetic dataset variables seem to have a smaller range of values

Results - EDA

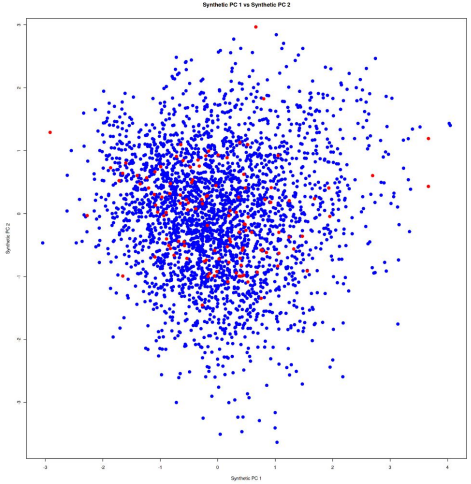


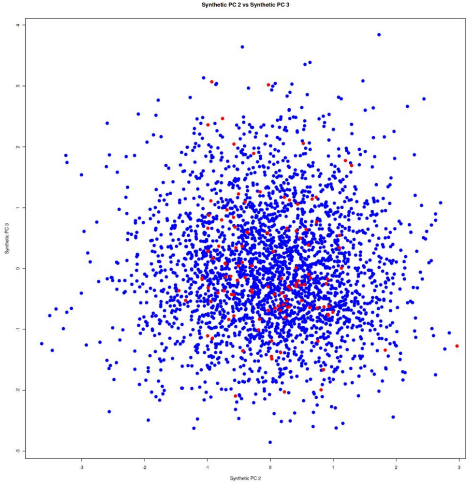
Original dataset

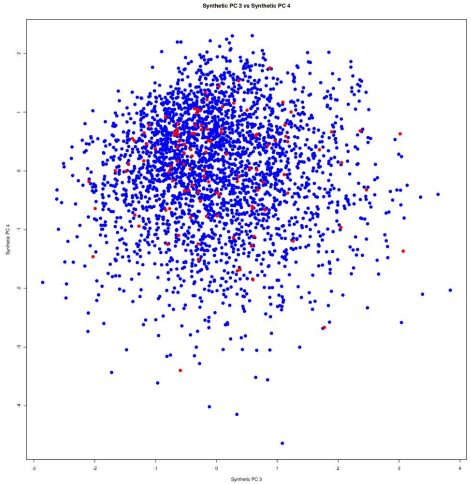


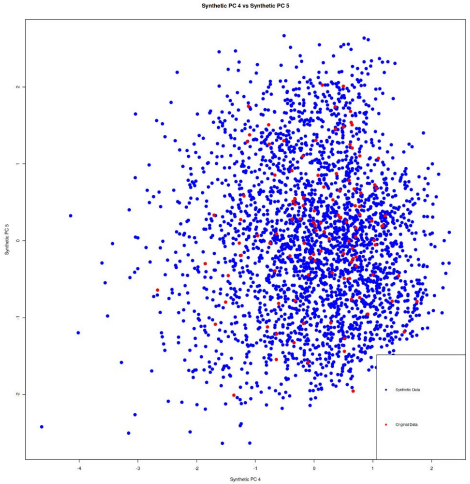
Synthetic dataset

PCA

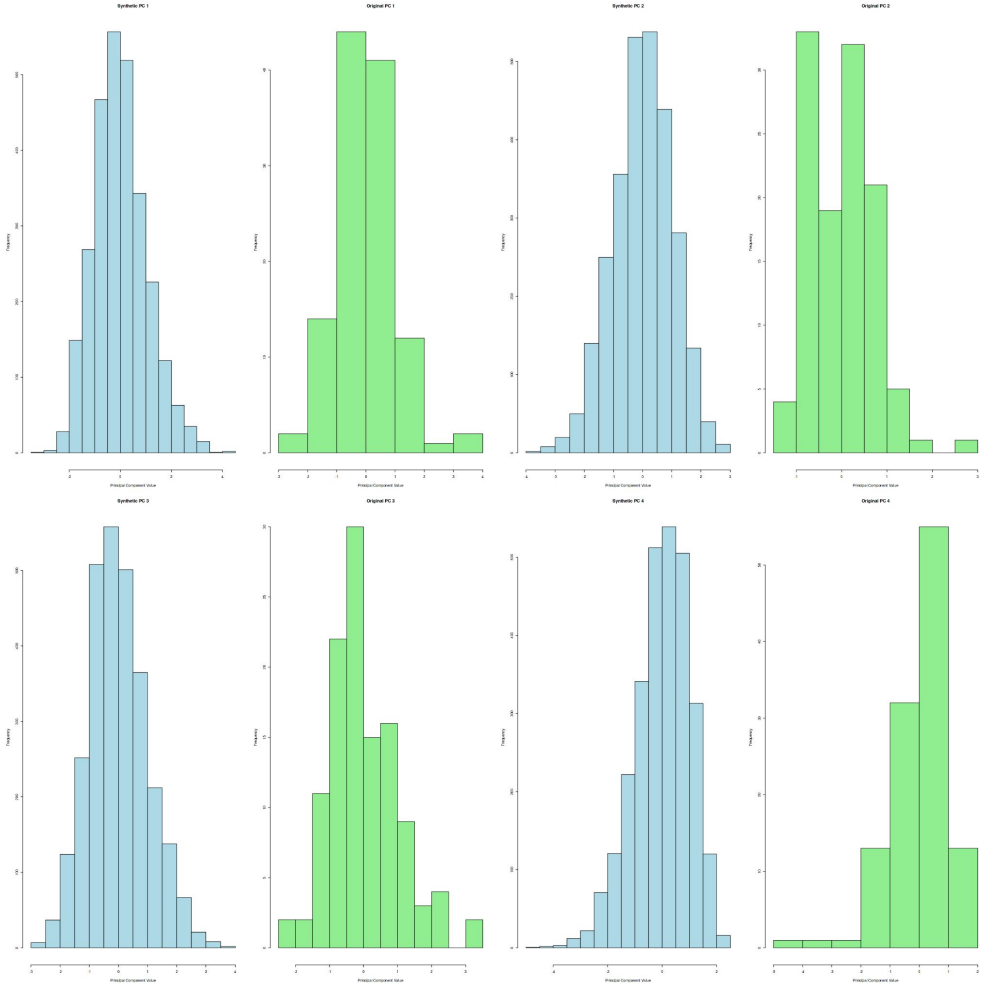

Synthetic PC 1 vs Synthetic PC 2
Synthetic PC 2
Synthetic PC 1


Synthetic PC 2 vs Synthetic PC 3
Synthetic PC 3
Synthetic PC 2


Synthetic PC 3 vs Synthetic PC 4
Synthetic PC 4
Synthetic PC 3


Synthetic PC 4 vs Synthetic PC 5
Synthetic PC 5
Synthetic PC 4
Synthetic Data
Original Data

PCA


Synthetic PC 1
Original PC 1
Synthetic PC 2
Original PC 2
Synthetic PC 3
Original PC 3
Synthetic PC 4
Original PC 4
Frequency
Principal Component Value

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: Test Set

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

Results

Pre-Tuning Model Metrics:

- LDA model shows highest accuracy score.
- Metrics seem comparable to other models, but confusion matrix reveals bias.
- High tendency of LDA to predict one class, affecting generalizability.

Post-Tuning Model Analysis:

- LDA maintains strong performance after tuning.
- KNN outperforms LDA in accuracy, recall, and test error.
- Logistic regression, despite lower accuracy, achieves highest precision and F1 score.

Confusion Matrix Insights:

- LDA and logistic regression show bias towards predicting positive cases.
- KNN and QDA exhibit more balanced predictions across classes.

Conclusions

- No single model significantly outperforms others.
- Trade-offs present in model selection.
- KNN emerges as top performed considering all evaluation metrics, including confusion matrix.

Future Work

- Future studies should assess additional prediction models, feature selection methods, and more expansive clinical datasets.
- 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.
- More research must be conducted to verify the relationship between these quantitative attributes and a true diagnosis of breast cancer.

Bibliography

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.