Breast Cancer Detection in the Philippines Using Machine Learning

APPROACH: A PILOT STUDY

Maria Maura S. Tinao

Ruth B. Rodirguez

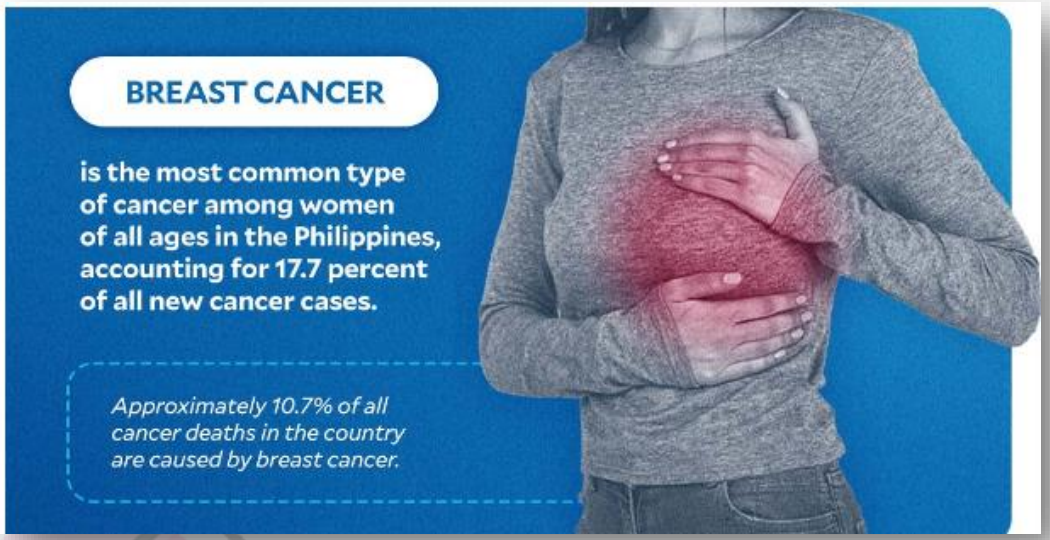
Eunelfa Regie F. Calibara

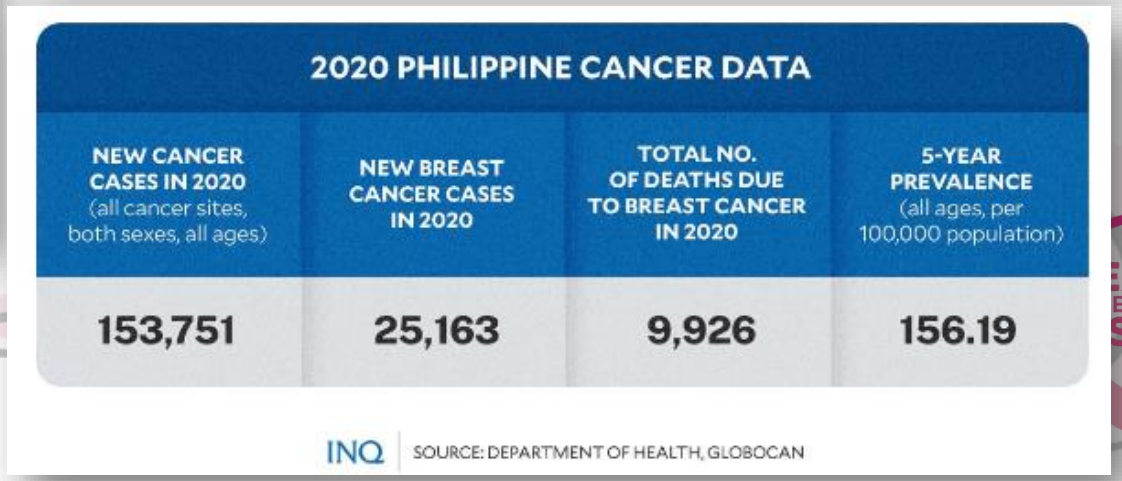
About Breast Cancer:

- The most common cancer and the main cause of death in the majority of cancer-affected women.
- The fifth (5th) most frequent reason for cancer death.
- Roughly, 2 million women are diagnosed with breast cancer each year and more than 500,000 women die from this disease worldwide.
- 45.4% of breast cancer patients are of Asian ethnicity
- According to World Health Organization:
- With 2.09 million cases, breast cancer had the highest incidence rate worldwide in 2018.
- A significant global health concern

In the Philippines, breast cancer:

- Ranks fifth in the highest incidence in East Asia and the Pacific region
- 17.7% of all cancer-related fatalities.





EARLY DETECTION SAVES LIVES

Early detection is crucial because:

- prevents progression
- reduces the risk
- a favorable outlook depends significantly on early diagnosis
- improves survival rate

Machine Learning (ML) in Healthcare

Overview of Machine Learning in Medical Diagnosis

- A subset of artificial intelligence (AI) that involves the *development of algorithms* that enable computers to learn and *make predictions or decisions based on data.*
- The application of machine learning in the health sector aims to improve patient outcomes and provide medical insights that were previously unavailable.
- It offers a way to validate physicians' reasoning and decisions through predictive models

Significance of ML in breast cancer detection

- Particularly valuable for analyzing patient data to aid in the early diagnosis, prevention, and treatment of various illnesses, including breast cancer.
- In countries with low survival rates, like the Philippines, ML can accurately identify patterns and predict diseases in similar cases, addressing the delay in diagnosis and improving outcomes.
- Given its unique ability to identify critical features from complex breast cancer datasets, ML is widely regarded as the preferred approach for BC pattern classification and forecasting.
- In simple terms, ML has the potential to uncover hidden patterns in data, leading to improved cancer prediction and management.

Research Objectives

With the goal of predicting breast cancer risk early and explore potential correlations between characteristics, potentially improving health outcome, The following machine learning techniques used are:

K-Nearest Neighbor or KNN

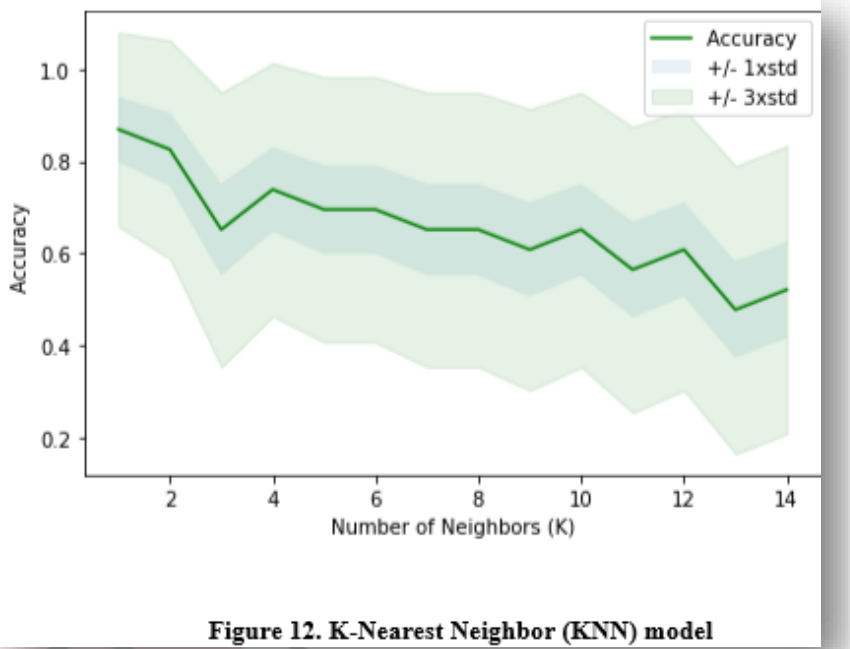


Figure 12. K-Nearest Neighbor (KNN) model

Decision Tree

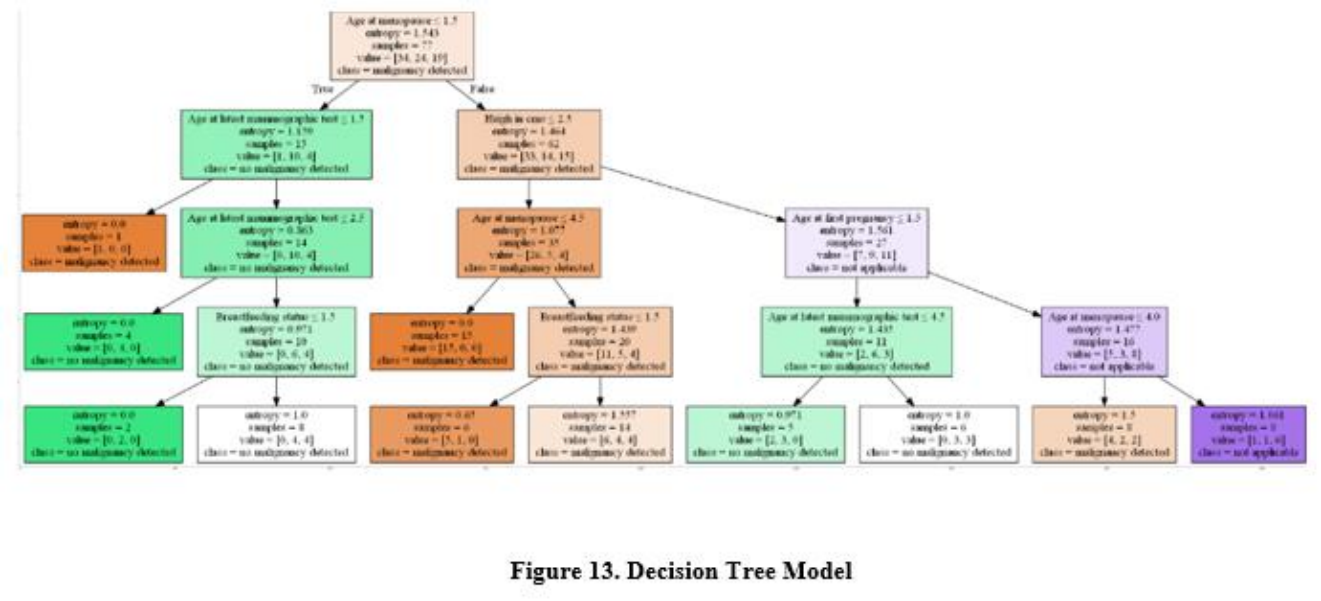


Figure 13. Decision Tree Model

Research Objectives

Support Vector Machine (SVM)

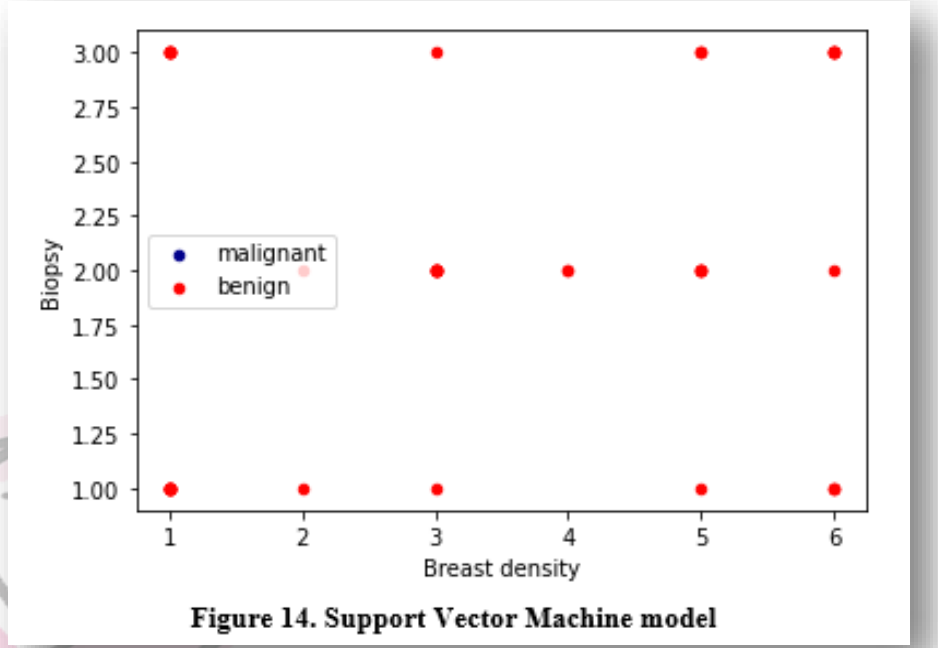
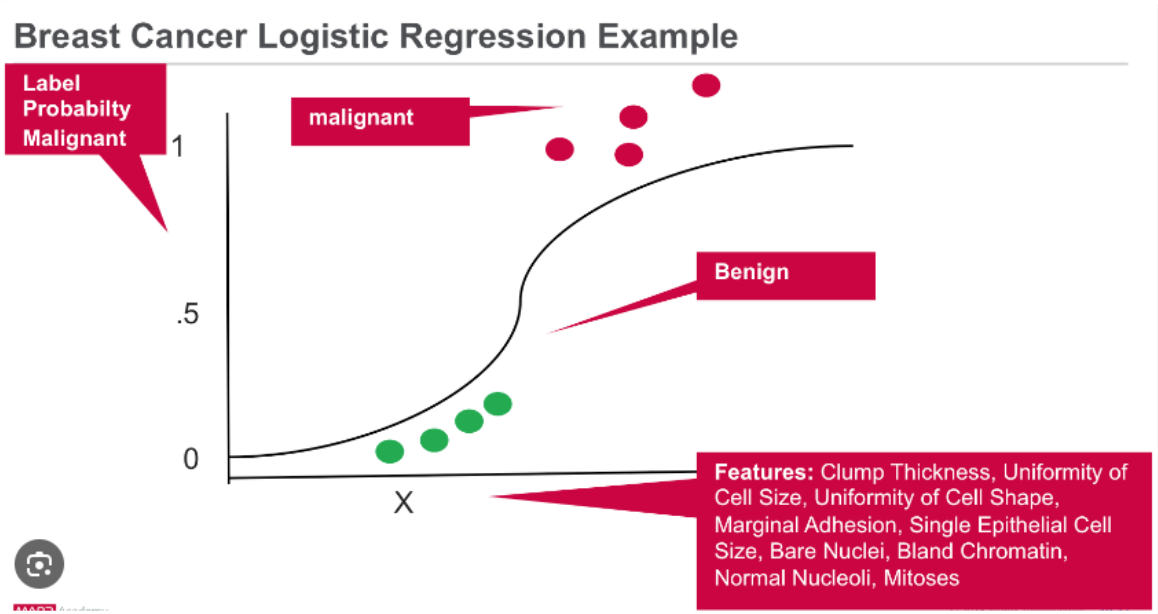


Figure 14. Support Vector Machine model

Logistic Regression



EARLY DETECTION SAVES LIVES

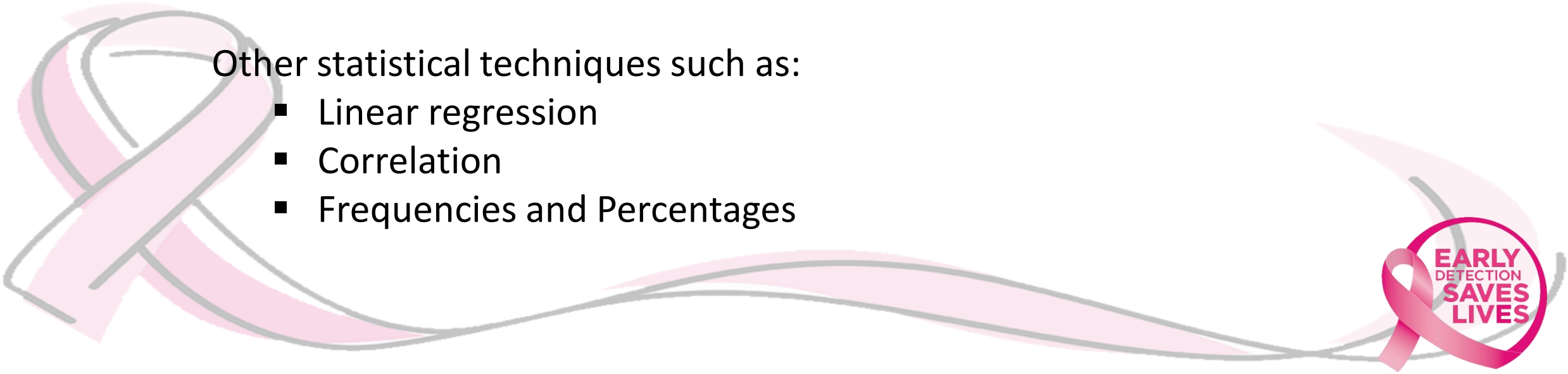
Research Method

Breast cancer features (Rabiel et al 2022) uses machine learning techniques like:

- K-Nearest Neighbor (KNN)
- Decision Tree
- Support Vector Machine (SVM)
- Logistic Regression

Other statistical techniques such as:

- Linear regression
- Correlation
- Frequencies and Percentages



Sample of the Study

- a challenge to find an active and up-to-date breast cancer registry in the Philippines
- Based from Rabiel, et al 2022 study, ten (10) important breast cancer features were chosen. These are: (1) age at the latest mammographic test; (2) Life event stress; (3) Marital status; (4) Smoker; (5) Biopsy; (6) Breast density; (7) Personal breast cancer history; (8) Personal history of other cancer; (9) Family breast cancer history; and (10) Family history of other cancer
- 112 randomly selected women from Olongapo City and Zambales
- use of anonymized data ensured that no written permission is needed, also for confidentiality and ethical considerations

Implementation tools

Utilized:

- Python programming language
- Jupyter Notebook application in Anaconda for the implementation of machine learning models.

  -a web-based, interactive computing environment that allows for the creation of human-readable documentation and data analysis.
- SPSS to compute linear regression
- Machine learning models include:

  K-Nearest Neighbors (KNN) - simple classification algorithm based on the similarity of data points;

  Decision Trees - hierarchical structures that make decisions based on feature values;

  Logistic Regression - models the probability of a data point belonging to a certain class

  Support Vector Machine (SVM) - a binary classification algorithm that finds a hyperplane to separate data points of different classes and;

  Confusion Matrix - table that presents a summary of the model's performance by comparing predicted class labels against actual class labels. The ultimate goal is to minimize false negatives and false positives to enhance the model's predictive capabilities

Results and Discussions

1.) Descriptive statistics

- The occurrence observed at the most recent mammographic examination observed are among those aged 45 to 54.

Table 1. Profile of the Respondents based on the ten (10) Distinct Characteristics of Breast Cancer (Rabiel, et al.2022)

Variables	Frequency	Percentage
1. Age at latest mammographic test		
18-24	5	4.5
25-34	16	14.3
35-44	5	4.5
45-54	33	29.5
55-64	10	8.9
TOTAL	**112**	**100.0**
2. Life event stress		
Family problem	40	35.7
Death of loved ones	14	12.5
Work related problems	20	17.9
Others	18	16.1
None	20	17.9
TOTAL	**112**	**100.0**
3. Marital status		
Single	41	36.6
Married	49	43.8
Widowed	14	12.5
Divorced	5	4.5
No answer	3	2.7
TOTAL	**112**	**100.0**
4. Personal breast cancer history		
Yes	13	11.6
No	99	88.4
TOTAL	**112**	**100.0**
5. Personal other cancer history		
Yes	34	30.4
No	78	69.6
TOTAL	**112**	**100.0**
6. Family history of breast cancer		
Yes	22	19.6
No	90	80.4
TOTAL	**112**	**100.0**
7. Family history of other cancer		
Yes	52	46.4
No	60	53.6
TOTAL	**112**	**100.0**
8. Smoker		
Yes	9	8.0
No	103	92.6
TOTAL	**112**	**100.0**
9. Breast density		
Fatty tissue	27	24.1
Heterogeneously dense	6	5.4
Extremely dense	11	9.8
Dense	22	19.6
No info	29	25.9
Glandular and fibrous tissue	17	15.2
TOTAL	**112**	**100.0**
10. Biopsy		
No malignancy detected	37	33.0
Malignancy detected	47	42.0
Not applicable	28	25.0
TOTAL	**112**	**100.0**

Results and Discussions

- The most commonly reported stressor is family problems.
- The predominant status of occurrence are being married or single.
- The average score for personal breast cancer history is 1.88, indicating a moderate level of variability as indicated by the low standard deviation.
- A significant proportion of participants, specifically 69.6%, reported no prior occurrences of different forms of cancer.
- The average score for the family history of breast cancer is 1.80, exhibiting a considerable standard deviation of 19.6%. The vast majority of individuals, specifically 80.4%, do not possess a familial lineage associated with breast cancer.

Results and Discussions

- The average family history of other cancer score is 1.54, exhibiting a relatively small standard deviation of 0.501. A significant proportion, specifically 53.6%, of individuals surveyed do not possess a familial background of other cancer types.
- The average smoking score is 1.92, accompanied by a standard deviation of 0.273. The vast majority of individuals, specifically 92.0%, do not engage in smoking habits.
- The average breast density score is 3.63, accompanied by a standard deviation of 1.801. The aforementioned statistics indicate that a significant proportion of the populace has not encountered alternative forms of cancer.
- A biopsy is a medical process utilized to extract tissue samples for the purpose of examining them for any potential abnormalities, such as the presence of cancerous cells. The average biopsy score is 1.92, exhibiting a moderate degree of variability.

Results and Discussions

2.) Correlation

- Weak positive relationship between age, stress, personal history, and family history;
- A moderately negative correlation between personal breast cancer history and breast density and;
- A moderately negative correlation between personal other cancer history and breast density.

Table 2. Pearson Correlation of Distinct Characteristics with Breast Density

Distinct Characteristics correlated with breast density	Pearson Correlation	p-value
Age at the latest Mammographic test	0.153	0.108
Life Event Stress	-0.099	0.297
Personal Breast Cancer History	-0.250	0.008
Personal Other Cancer History	-0.181	0.057
Family History of Breast Cancer	-0.219	0.021
Family History of Other Cancer	0.290	0.002
Marital Status	0.057	0.553
Smoker	0.086	0.336
Biopsy	0.070	0.461

Results and Discussions

3.) Linear Regression - the relationship between predictor variables like age, stress, and breast cancer history with breast density.

- A linear regression analysis shows a weak positive linear relationship, with 8.1% of variance explained by these variables;
- The adjusted R-squared value is close to the original R-squared, suggesting that the inclusion of predictor variables has not substantially improved the model fit.
- The analysis results suggest that the overall model does not have a statistically significant effect in explaining the variance in breast density.
- The model's fit is not significantly improved, suggesting further refinement, additional predictors, and data quality improvements could enhance the model's performance.

MODEL SUMMARY

Model	R	R Square	Adjusted R square	Std. Error of the Estimate
1	.284a	.081	.009	1.793

a. Predictors: (Constant), Life event stress, Personal breast cancer history, Smoker, Marital status, Age at latest mammographic test, Family history other cancer, Family history of breast cancer, Personal other cancer history

ANOVAα

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	29.001	8	3.625	1.128	.351b
	Residual	330.990	103	3.213		
	TOTAL	359.991	111			

a. Dependent Variable: Breast Density
b. Predictors: (Constant), Life event stress, Personal breast cancer history, Smoker, Marital status, Age at latest mammographic test, Family history other cancer, Family history of breast cancer, Personal other cancer history

Table 3. Analysis of Variance Between Various Predictor Variables

EARLY DETECTION SAVES LIVES

Results and Discussions

4.) Machine Learning Models

Breast cancer is a significant healthcare challenge, and timely detection is crucial for improving patient prognoses.

This research evaluates the effectiveness of four algorithms - KNN, Decision Tree, SVM, and Logistic Regression - in breast cancer detection.

Results and Discussions

Model Evaluation Metrics - measure the proportion of correctly predicted positive instances.

- Jaccard Index - measures the proportion of correctly predicted positive instances relative to the total number of actual positive instances and false negatives.

  A high Jaccard Index indicates that the model's predictions align well with the true class labels.

- F1-Score - popular metric for assessing the performance of a classification model, especially when there is an imbalance between the classes.

  A high F1-Score indicates that the model is making accurate positive predictions while also capturing a large proportion of actual positive instances.

- *A low Log Loss indicates that the model is making accurate and confident predictions, an indication that the model is performing well and has a good understanding of the data.*

Results and Discussions

K-Nearest Neighbors (KNN)

- The Jaccard index, with a value of 0.3, indicates a considerable degree of similarity between the anticipated and actual examples.
- The F-1 score, which is 0.43, represents a satisfactory equilibrium between precision and recall. The accuracy rate of 0.7059 found in the Decision Tree algorithm implies that 70.59% of the breast cancer patients were correctly identified.

Decision Tree

- The Jaccard index, with a value of 0.19, indicates a comparatively lower degree of similarity when compared to KNN. The Jaccard index, with a value of 0.19, indicates a comparatively lower degree of similarity when compared to KNN.
- On the other hand, the F-1 score, which stands at 0.3, suggests a balance between precision and recall, similar to KNN.

	Algorithm	Jaccard	F1-Score	LogLoss
0	KNN	0.3	0.43	NA
1	Decision Tree	0.19	0.3	NA
2	SVM	0.26	0.38	NA
3	Logistic Regression	0.26	0.38	1.08

Summary of 4 Machine Learning Models Report

Results and Discussions

Support Vector Machine (SVM) algorithm

- The Jaccard index, with a value of 0.26, suggests a reasonable degree of similarity in the predictions made.
- On the other hand, the F-1 score, which is 0.38, indicates a well-balanced compromise between precision and recall.

Logistic Regression

- The Jaccard index, with a value of 0.26, indicates a moderate level of similarity between the predicted and real instances.
- The F1-score, which is 0.38, suggests a balanced trade-off between precision and recall.
- The Logloss value of 1.08 quantifies the cross-entropy loss between the predicted probability and the actual class labels. Lower values of Logloss indicate better calibration of the model.

	Algorithm	Jaccard	F1-Score	LogLoss
0	KNN	0.3	0.43	NA
1	Decision Tree	0.19	0.3	NA
2	SVM	0.26	0.38	NA
3	Logistic Regression	0.26	0.38	1.08

Summary of 4 Machine Learning Models Report

Conclusions:

- Breast cancer is a significant medical issue that affects healthcare systems worldwide. Early detection can lead to more effective treatment options and improved patient outcomes.
- Machine learning techniques are increasingly used in the healthcare industry to identify patterns and make predictions.
- This study showed how machine learning models can be used to detect breast cancer and how important it is to use performance review metrics. This research evaluates the performance of four machine learning algorithms: K-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), and Logistic Regression.
- The effectiveness of these algorithms is measured by metrics such as accuracy, precision, recall, Jaccard score, and F1 score.

Conclusions:

- The study found that the KNN model achieved the highest accuracy rate and F1 score, indicating its effectiveness in correctly classifying breast cancer cases while maintaining a balanced precision-recall trade-off.
- The Decision Tree, SVM, and Logistic Regression models showed lower performance metrics, with varying degrees of precision and recall.
- The research serves as an exploratory starting point for research endeavors in breast cancer detection using ML approaches, offering preliminary insights, generating hypotheses, identifying challenges, and laying the groundwork for more extensive investigations

Recommendations

- The study on breast cancer risk and detection requires a comprehensive understanding of factors such as genetic, hormonal, medical, socioeconomic, lifestyle, environmental, and psychosocial factors.
- The small sample size affects the use of machine learning models, such as K-Nearest Neighbors (KNN), Decision Tree, Logistic Regression, and Support Vector Machine (SVM).
- To improve model performance and dependability, larger datasets should be used. Combining multiple ML models, such as Random Forest or Gradient Boosting, or advanced techniques like deep learning, can improve handling of complex patterns and yield reliable predictions.

Recommendations

- A nationwide breast cancer detection study offers enhanced representation, allowing for the identification of regional variations and the development of tailored models.
- Collaborating with medical institutions, researchers, and experts from multiple regions can lead to more comprehensive data collection, thorough analysis, and meaningful insights.
- Nationwide studies can also influence healthcare policies and clinical guidelines, and contribute to a broader understanding of breast cancer detection methods globally

References:

[1] A. Fanizzi, T.M.A. Basile, L. Losurdo, R. Belloti, U. Bottigli, R. Dentamaro, V. Dedonna, a. Fausto, R. Massafra, M. Moschetta, O. Popescu, P. Tamborra, S. Tangaro & D. La Forgia, "A Machine Learning Approach on Multiscale Texture Analysis for Breast Microcalcification Diagnosis" BMC Bioinformatics, 21 (1), March 2020. https://doi.org/10.1186/s12859-020-3358-4

[2] A. Tapas, "Only 1% of Filipino Women Screened for Breast, Cervical Cancer Study, Manila Times, July 2023.

[3] Coursera, "What is Machine Learning in Health Care?: Applications and Opportunities", June 2023. https://www.coursera.org/articles/machine-learning-inhealth-care

[4] D. Sandeep & G.N. Beena Bethel, "A Survey on Accurate Breast Cancer Detection and Classification Using Machine Learning Approach. E3S Web of Conference, 309, 01116. https://doi.org/10.1051/e3sconf/202130901116

[5] D.Z. Pazzibugan, "Only 1% of PH Women Screened for Breast, Cervical Cancer", Philippine Daily Inquirer, July 2023.

[6] Foresee Medical, "Benefits of Machine Learning in Healthcare, September 2023. https://www.foreseemed.com/blog/machine-learning-inhealthcare

[7] H. El Masari, N. Gherabi, S. Mhammedi, Z. Sabouri, H. Ghandi & F.Qanouni, "Effectiveness of Applying Machine Learning Techniques and Ontologies in Breast Cancer Detection", Procedia Computer Science, pp. 2392-2400, 2023.

[8] J.E. Mendoza, "Breast Cancer Cases in PH 'Staggering' – Health Expert. Philippine Daily Inquirer, February 2023.

[9] K. Mali, "Data Science Blogathon", August 2023. https://datahack.analyticsvidhya.com/blogathon/

[10] Kasuso – Philippine Foundation for Breast Cancer Care, Inc., "About Breast Cancer, (n.d.). https://www.kasuso.org/about-breast-cancer

[11] M.A. Elsadig, A. Altigan, & H.T. Elshoush, "Breast Cancer Detection Using Machine Learning Approaches: A Comparative Study, International Journal of Electronics and Communications, 13, pp.736- 745.

[12] N. Arya, "Nearest Neighbors for Classification", KDnuggets, April 2022. https://www.kdnuggets.com/2022/04/nearest-neighborsclassification.html

[13] N.S. Chauhan, "Decision Tree Algorithm Explained", KDnuggets, February 2022. https://www.kdnuggets.com/2020/01/decision-treealgorithm-explained.html

[14] S. Jessica, "How Does Logistic Regression Work?", KDnuggets, July 2022. https://www.kdnuggets.com/2022/07/logistic-regressionwork.html

References:

15] S.M.A. Reza Rabiel, S. Sohrabel, M. Esmaeli, & A. Atashi, “Prediction of Breast Cancer Using Machine Learning Approaches, Journal of Biomedical Physics & Engineering, 12 (3), 2022.

[16] S. Nag & J. Nag, “A Comparative Analysis of Machine Learning Approaches for Prediction of Breast Cancer. Journal of Emerging Investigators, 3, pp.1-9, 2021.

[17] W. Yue, Z. Wang, H. Chen, A. Payne & X. Liu, “Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis”, May 2018. https://doi.org/10.3390/designs2020013

[18] Y.X. Lim, Z.L. Lim, P.J. Ho & J. Li, “Breast Cancer in Asia: Incidence, Mortality, Early Detection, Mammography Programs, and Risk-Based Screening Initiatives”, August 2022. https://pubmed.ncbi.nlm.nih.gov/36077752/

[19] Z. Salod & Y. Singh, “Comparison of the Performance of Machine Learning Algorithms in Breast Cancer Screening and Detection: A Protocol. Journal of Public Health Research, 8(4), pp. 112-118,