Predictive Analysis of the Wisconsin Breast Cancer Dataset:

An Integrated Approach with PCA, KNN and Neural Networks

Universidad Politécnica Salesiana, Ecuador

Abstract

Breast cancer is one of the major public health problems worldwide, with numerous new cases reported every year affecting both women and men, although to a lesser extent in the latter. The high incidence and the clinical implications associated with late diagnosis highlight the urgent need to strengthen early detection methods. In this context, it is essential to rely on advanced techniques and tools, such as data analysis and machine learning models, which enable preventive diagnosis support through the early identification of patterns associated with malignancy. The Wisconsin Breast Cancer Diagnostic (WDBC) dataset has become a fundamental reference for the development and evaluation of computational techniques capable of supporting clinical diagnosis through automated analysis.

This work presents a comprehensive schematic process for the WDBC dataset in two phases: exploratory data analysis and predictive modeling. Phase 1 covers statistical analysis, correlation study, outlier detection using the IQR method, and comparative analysis between malignant and benign cases. Phase 2 focuses on the development and evaluation of predictive models, including k-nearest neighbours (KNN) and neural networks (MLP), using both the original features and reduced representations obtained via principal component analysis (PCA). The optimized KNN classifier without PCA reached an accuracy of 98.25 %, while the neural network classifier achieved an accuracy of 96.49 %. For the regression task on mean area, the best model obtained an R^2 coefficient of 0.954, demonstrating the effectiveness of the proposed approach for computeraided medical diagnosis. As future work, we plan to evaluate additional machine learning models, improve class-imbalance handling, incorporate interpretability methods, and extend the methodology to larger and more diverse clinical datasets.

Keywords: Breast Cancer Diagnosis, Machine Learning, Wisconsin Breast Cancer Dataset, Principal Component Analysis, K-Nearest Neighbors, Neural Networks, Computer-Aided Diagnosis

Introduction

This work follows the CRISP-ML(Q) (Cross-Industry Standard Process for Machine Learning with Quality Assurance) methodology [23], consistent with the framework adopted in previous research by our group [24]. This structured approach ensures quality assurance throughout the machine learning pipeline, from business understanding to model deployment and monitoring.

Breast cancer remains one of the leading causes of morbidity and mortality worldwide, and early diagnosis is essential to improving patient outcomes. Traditional diagnostic procedures such as visual assessment and fine-needle aspiration (FNA) cytology may exhibit interobserver variability and depend heavily on the expertise of the clinician. Consequently, there is a growing need for complementary computational tools that support a more standardized and reproducible interpretation of diagnostic information.

In recent years, computational analysis of clinical data has emerged as a valuable aid in medical decision-making. Machine learning models, in particular, are capable of identifying subtle morphological patterns associated with malignancy, contributing to more consistent diagnostic criteria and enabling the development of data-driven support systems.

The Wisconsin Breast Cancer Diagnostic (WDBC) dataset is one of the most widely used benchmarks for evaluating classification and predictive modeling techniques in biomedical research. Its set of detailed morphological features derived from digitized FNA images provides an ideal environment for investigating both statistical properties of tumor characteristics and the behavior of different learning algorithms. Despite extensive prior work, open questions remain regarding the relative importance of diagnostic features, the influence of preprocessing choices, and the impact of dimensionality reduction on model performance.

This study addresses these aspects through a comprehensive analysis that integrates exploratory statistics, class-based comparisons, overlap assessment, dimensionality reduction and supervised learning for both classification and regression tasks. The main contribution of this work is the formulation of a unified methodological pipeline that characterizes the structure and variability of the dataset while systematically evaluating multiple predictive approaches.

Main Contributions

- Development of an integrated analytical pipeline: A two-phase workflow is proposed, combining exploratory data analysis, preprocessing, dimensionality reduction, and predictive modeling specifically tailored to the WDBC dataset for breast cancer diagnosis.
- Implementation and comparative assessment of KNN and neural network (MLP) models: Both original-feature and PCA-reduced configurations are evaluated, allowing a systematic analysis of how dimensionality reduction impacts classification and regression performance.
- Hyperparameter optimization for improved model robustness: Techniques based on GridSearchCV, regularization strategies, and architecture tuning are applied to maximize predictive accuracy and minimize model variance.
- Open-source availability of the full implementation: All scripts, preprocessing workflows, and experiments are publicly accessible in the project repository: `https://github.com/natasha943/AnalisisCancerDeMama.git`

Structure of the Paper

The remainder of this document is organized as follows:

- Related Work: Reviews previous studies on computer-aided diagnosis and machine learning applications to clinical data, with emphasis on the WDBC dataset and dimensionality reduction techniques.

- Proposed Method: Describes the three-phase machine learning workflow: Phase 1 (data preparation and exploratory analysis), Phase 2 (supervised modeling with KNN and neural networks), and Phase 3 (unsupervised clustering integration).

- Experimental Design: Presents the WDBC dataset characteristics, including sample distribution, feature description, and data preprocessing steps.

- Descriptive Statistics and Feature Analysis: Provides comprehensive statistical characterization, correlation analysis, and identification of most discriminative features between benign and malignant classes.

- Dimensionality Reduction with PCA: Details PCA implementation and explained variance analysis for both classification and clustering tasks.

- Classification and Regression Models: Covers development and evaluation of KNN and Neural Network models for both classification and regression tasks, with and without PCA, including performance comparison.

- Model Validation: Presents validation results using real test samples and synthetic cases to assess model reliability and generalization capability.

- Integration of Clustering: Details unsupervised clustering analysis using K-Means, including optimal cluster determination (Elbow Method) and visualization of natural data groupings.

- Conclusions: Summarizes key findings, discusses clinical implications, and outlines future research directions.

Related Work

In recent years, the use of machine learning techniques to support cancer diagnosis has received increasing attention. Several studies have explored the application of classification algorithms on the WDBC dataset and other clinical datasets related to breast cancer, using models such as SVM, decision trees, KNN, neural networks and ensemble methods [16, 18].

Comparative studies have shown that machine learning models can achieve competitive accuracy levels, provided that appropriate data preprocessing and careful feature selection are performed [16]. Other works have incorporated dimensionality reduction techniques, such as PCA and nonlinear methods, to improve efficiency and, in some cases, the robustness of the models [5, 6, 19].

Additionally, recent research has explored deep learning approaches applied directly to mammography or biopsy images, with promising results for detection and classification tasks [17, 21]. However, models based on tabular features remain relevant in scenarios where derived measurements are available, as in the case of WDBC.

Proposed Method

Phase 1 focused on preparing and exploring the dataset through validation, descriptive statistics, correlation analysis, outlier detection, and class comparison, resulting in a clean and relevant dataset for modeling. Phase 2 applied scaling, PCA, and predictive models such as KNN and neural networks, followed by optimization, performance evaluation, and real and synthetic predictions to validate the overall modeling process.Phase 3 integrated unsupervised clustering using PCA and K-Means, revealing a natural two-cluster structure aligned with benign and malignant classes, which reinforced the robustness and clinical relevance of the overall modeling approach.

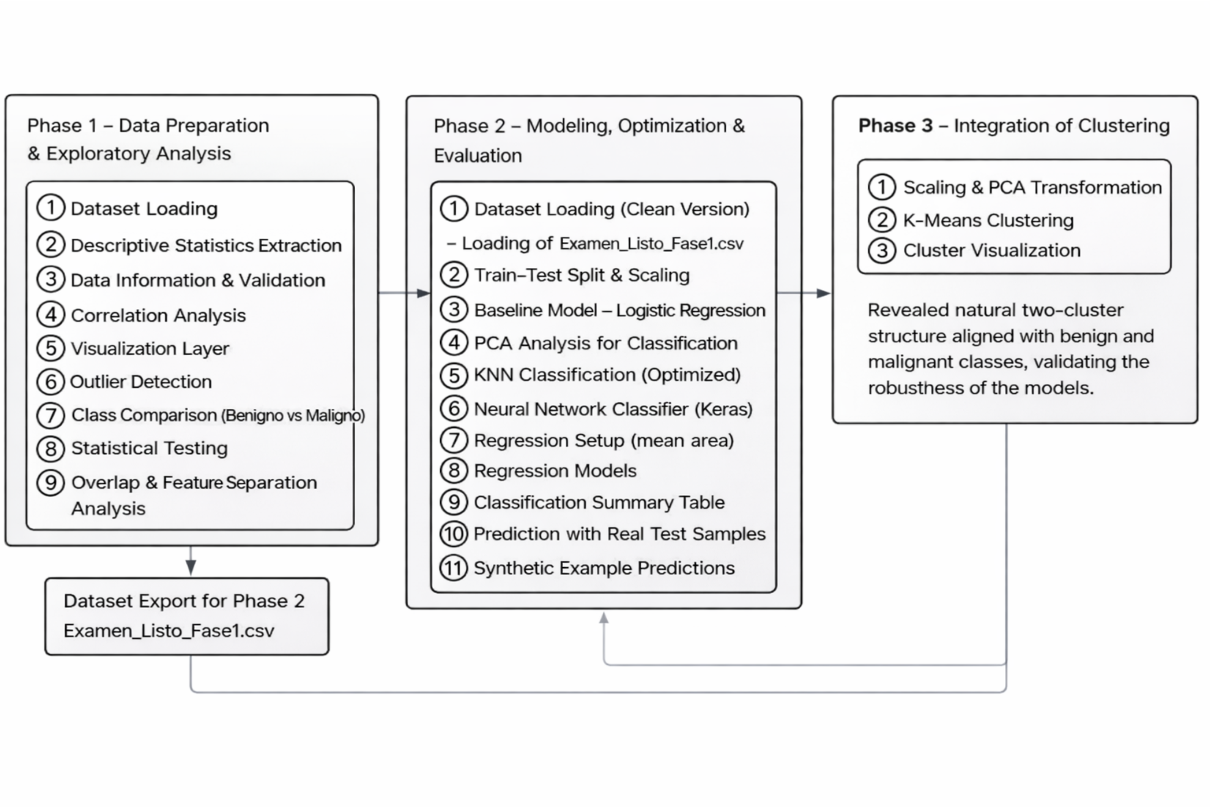

Phase 1 – Data Preparation & Exploratory Analysis
① Dataset Loading
② Descriptive Statistics Extraction
③ Data Information & Validation
④ Correlation Analysis
⑤ Visualization Layer
⑥ Outlier Detection
⑦ Class Comparison (Benigno vs Maligno)
⑧ Statistical Testing
⑨ Overlap & Feature Separation Analysis
Dataset Export for Phase 2
Examen_Listo_Fase1.csv
Phase 2 – Modeling, Optimization & Evaluation
① Dataset Loading (Clean Version)
– Loading of Examen_Listo_Fase1.csv
② Train–Test Split & Scaling
③ Baseline Model – Logistic Regression
④ PCA Analysis for Classification
⑤ KNN Classification (Optimized)
⑥ Neural Network Classifier (Keras)
⑦ Regression Setup (mean area)
⑧ Regression Models
⑨ Classification Summary Table
⑩ Prediction with Real Test Samples
⑪ Synthetic Example Predictions
Phase 3 – Integration of Clustering
① Scaling & PCA Transformation
② K-Means Clustering
③ Cluster Visualization
Revealed natural two-cluster structure aligned with benign and malignant classes, validating the robustness of the models.

Figura 1: Three-phase machine learning workflow including data preparation, supervised modeling, and clustering integration for breast cancer analysis.

Experimental Design

The Wisconsin Breast Cancer Diagnostic (WDBC) dataset is obtained through the datasets module of Scikit-learn and includes:

- **Samples:** 569 breast tumor cases.
- **Features:** 30 numerical variables derived from digitized images (measures of radius, texture, perimeter, area, smoothness, concavity, concave points, symmetry, fractal dimension, etc.).
- **Target variable:** Diagnosis (0 = Malignant, 1 = Benign).
- **Distribution:** 357 benign cases (62.7 %) and 212 malignant cases (37.3 %).

This dataset is widely used as a reference in the literature to evaluate classification models in the context of breast cancer diagnosis [1, 2, 16, 18].

Descriptive Statistics of Relevant Features

Feature	**Malignant (Class 0)**	**Benign (Class 1)**	**Difference**
Worst concave points	0.182	0.074	0.108
Worst perimeter	141.370	87.006	54.364
Mean concave points	0.088	0.026	0.062
Worst radius	21.135	13.380	7.755
Mean perimeter	115.365	78.075	37.290

Cuadro 1: Mean values by class for the most discriminative features.

Percentage Difference Analysis

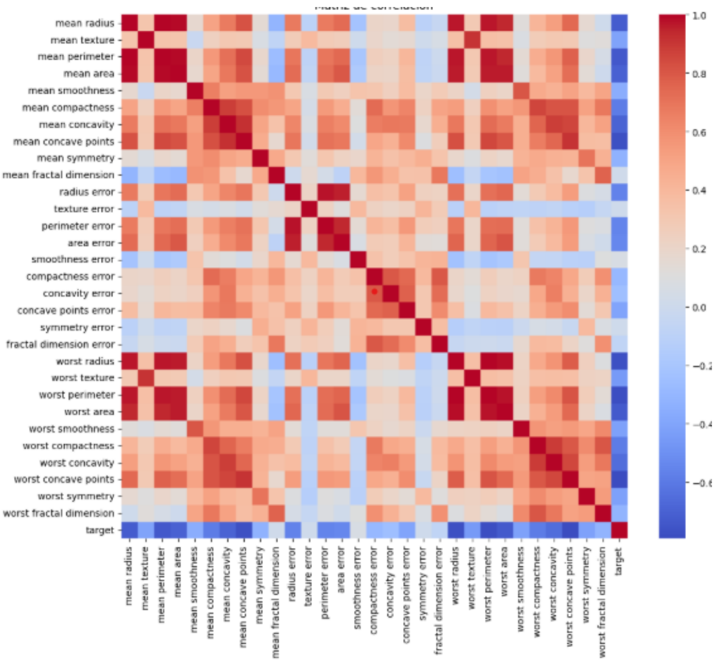


Figure 2: Correlation matrix of the selected features and the target variable. Strong positive correlations are observed among radius, perimeter and area measurements, while all features show a strong negative correlation with the target label (benign vs. malignant).

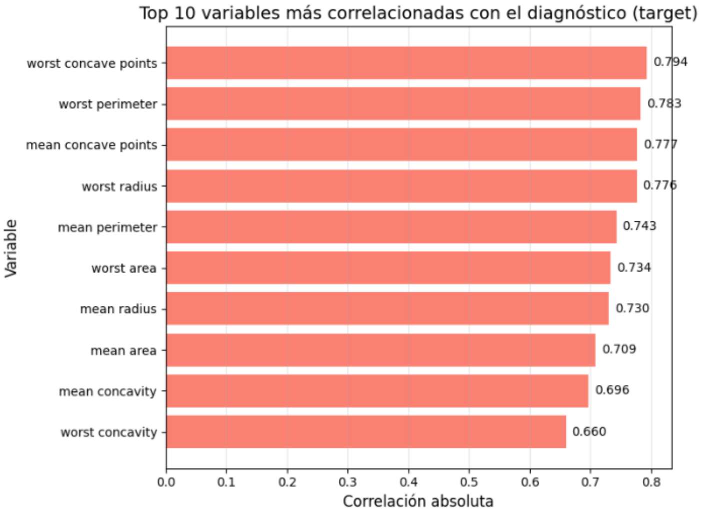


Figure 3: Top 10 features most strongly correlated with the diagnosis (target). Variables related to concavity, perimeter and radius exhibit the highest absolute correlations with the malignant/benign classification.

Modelling, Optimization and Evaluation

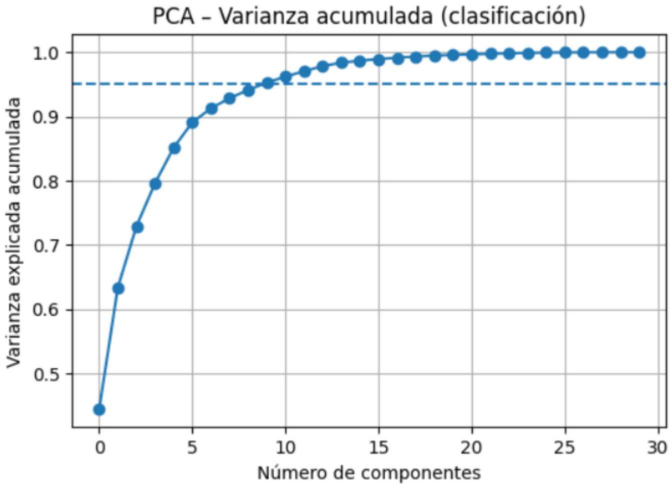


Figure 4: Accumulated explained variance of PCA components for the classification task. The first 10 components capture approximately 95 % of the total variance, justifying their selection for dimensionality reduction.

Classification and Regression Models

This section evaluates K-Nearest Neighbors (KNN) and Neural Networks for classification and regression, comparing performance with and without PCA.

Classification: KNN without PCA achieved optimal performance (0.9825 accuracy, perfect 1.0000 recall for malignant cases) using $k = 3$ neighbors and distance weighting. Neural Networks (64/32 architecture, Adam optimizer, early stopping) reached 0.9649 accuracy. PCA reduced both models to 0.9561 accuracy, indicating loss of diagnostically critical features.

Regression: KNN Regressor achieved $R^2 = 0.9532$ without PCA and 0.9545 with PCA. Neural Network Regressor obtained $R^2 = 0.9546$ without PCA but dropped to 0.9202 with PCA, demonstrating sensitivity to dimensionality reduction.

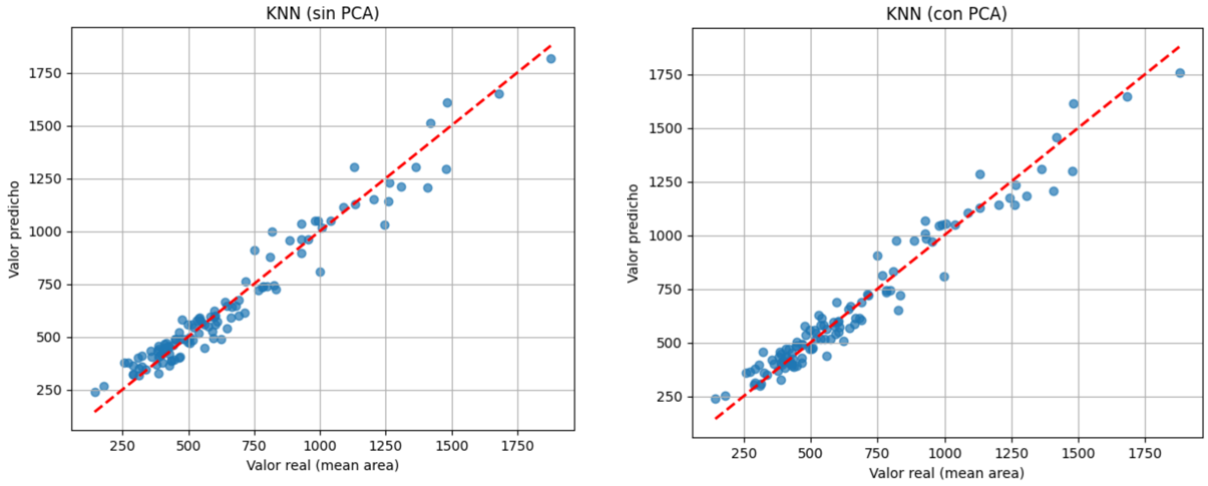
Table shows KNN without PCA as the best classifier (0.9825), while PCA consistently reduced classification accuracy (2.64 % average drop) but minimally affected regression. Distance-based methods outperformed neural networks, favoring KNN for clinical deployment.

Model	**PCA**	**Accuracy**
KNN Classifier	No	**0.9825**
KNN Classifier	Yes	0.9561
Neural Network Classifier	No	0.9649
Neural Network Classifier	Yes	0.9561
KNN Regressor	No	0.9532
KNN Regressor	Yes	0.9545
Neural Network Regressor	No	0.9546
Neural Network Regressor	Yes	0.9202

Cuadro 2: Accuracy comparison of all classification and regression models with and without PCA.

Regression – KNN Regressor and Neural Network (Keras)

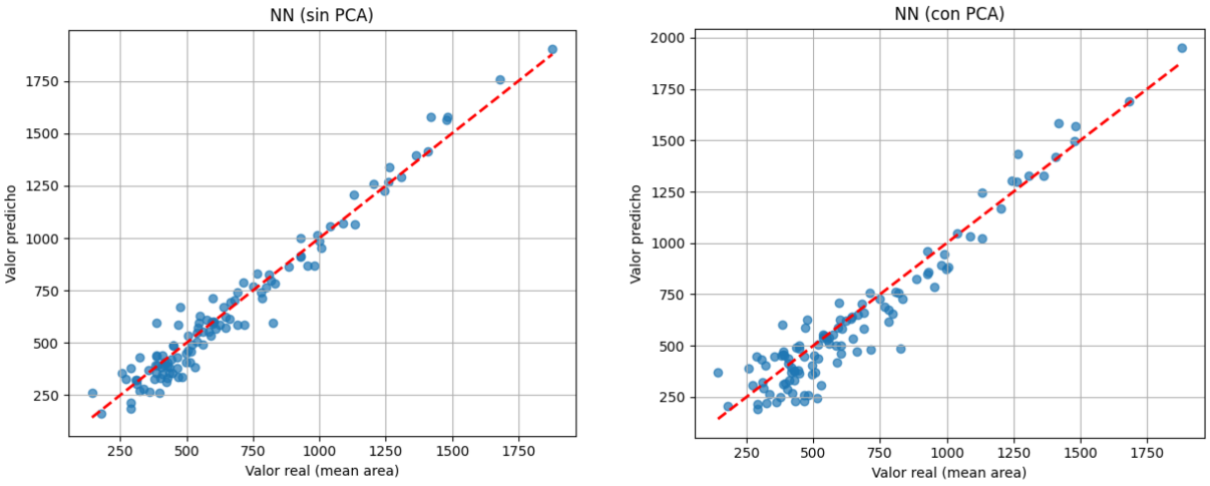
This section evaluates the ability of two regression models to predict the value of the mean area feature from the remaining variables in the dataset. Two approaches were implemented: a model based on k-nearest neighbours (KNN Regressor) and a neural network with an MLP architecture developed in Keras. Both models were trained using previously scaled data and their performance was compared in configurations with and without PCA, in order to analyse the impact of dimensionality reduction on prediction accuracy.



(a) Predicted vs. true values for the KNN regressor without PCA. The model closely follows the ideal regression line.

(b) Predicted vs. true values for the KNN regressor with PCA. PCA introduces a slight increase in dispersion.

Figure 5: Comparison of KNN regression performance with and without PCA, highlighting how dimensionality reduction affects prediction accuracy.



(a) Neural network regression without PCA. Predictions show strong alignment with the true values.

(b) Neural network regression with PCA. Performance remains solid, though PCA introduces slight precision loss.

Figure 6: Comparison of neural network regression performance with and without PCA, highlighting the impact of dimensionality reduction on predictive accuracy.

Prediction with Real and Synthetic Examples

Model reliability was validated using real test samples and synthetic cases. Two real samples (one benign, one malignant) were correctly classified by both KNN and Neural Network with perfect confidence (probabilities 1.0000 and 0.0000 respectively). Synthetic examples with extreme morphological profiles were also correctly predicted: the benign case received probability 0.9984, while the malignant case obtained 0.0000. These results confirm robust discriminative capability and reliable generalization across clinical and edge-case scenarios.

Integration of Clustering

Cluster Visualization

Once the optimal number of clusters (K=2) was determined, K-Means clustering was applied to the reduced PCA dataset using the first 10 principal components. The resulting cluster assignments provide an unsupervised perspective on the natural grouping structure within the breast cancer data.

Figure 7 presents a two-dimensional visualization of the clustering results, projecting the data onto the first two principal components.

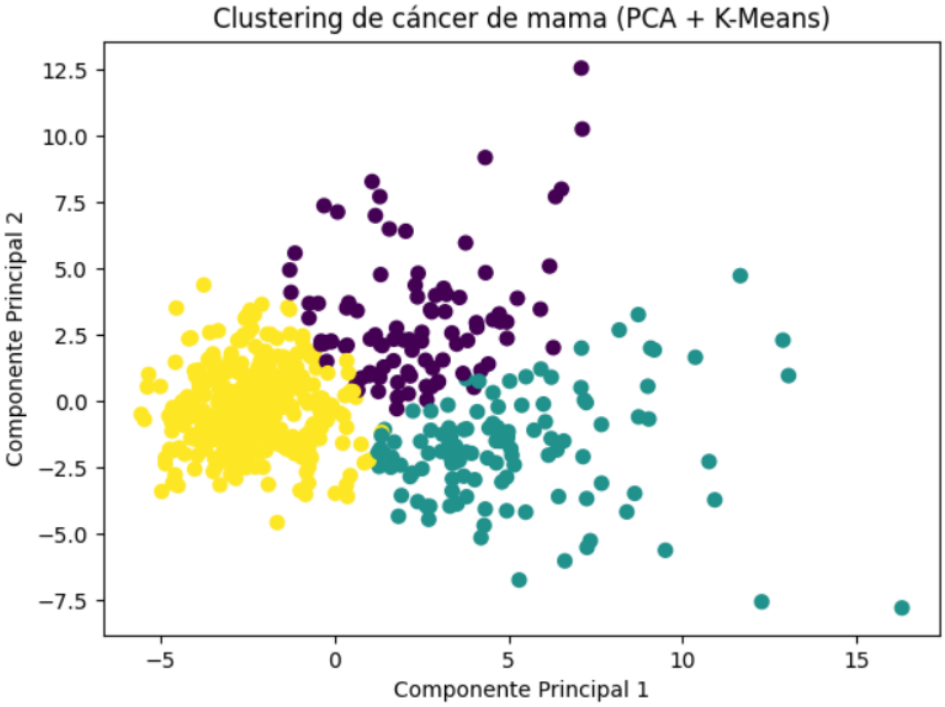


Figure 7: Two-dimensional visualization of K-Means clustering results. The scatter plot shows clear spatial separation between clusters (yellow and teal) on the first two principal components, demonstrating effective identification of natural groupings corresponding to benign and malignant tumor classes.

Scaling and Dimensionality Reduction (PCA)

To prepare the data for clustering analysis, feature scaling was first applied using StandardScaler to normalize all variables to a common scale with zero mean and unit variance. This preprocessing step is crucial to prevent features with larger magnitudes from dominating the analysis.

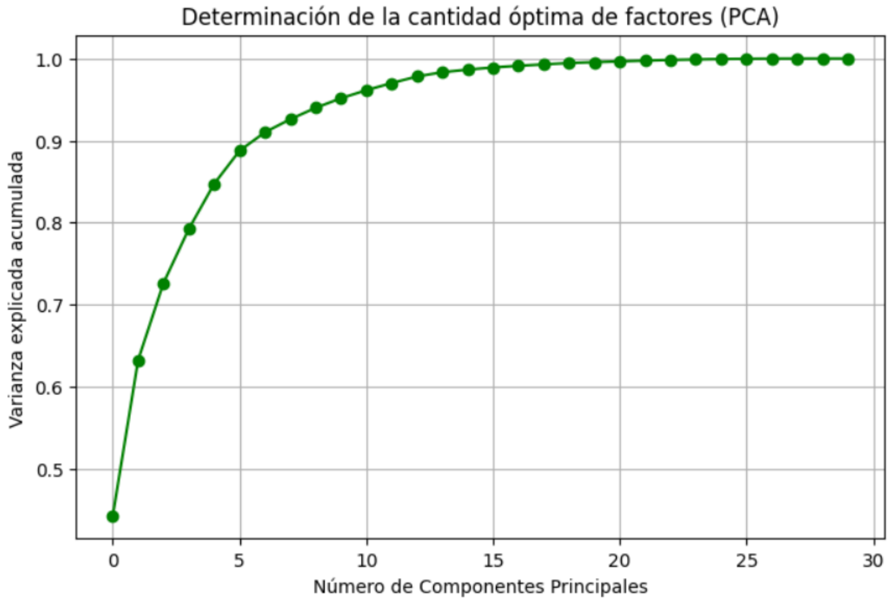


Figure 8: Cumulative explained variance of PCA components for clustering analysis. The curve shows that 10 principal components are sufficient to capture 95 % of the total variance, enabling effective dimensionality reduction while preserving the essential structure of the data.

Determining Optimal Number of Clusters (Elbow Method) After dimensionality reduction, the next critical step is to determine the optimal number of clusters (K) for the K-Means algorithm. The Elbow Method is a widely used heuristic that evaluates the within-cluster sum of squared distances (inertia) for different values of K.

The method works by fitting K-Means models with varying numbers of clusters (from 1 to 10) and computing the inertia for each configuration. Inertia measures the compactness of clusters: lower values indicate tighter, more cohesive clusters.

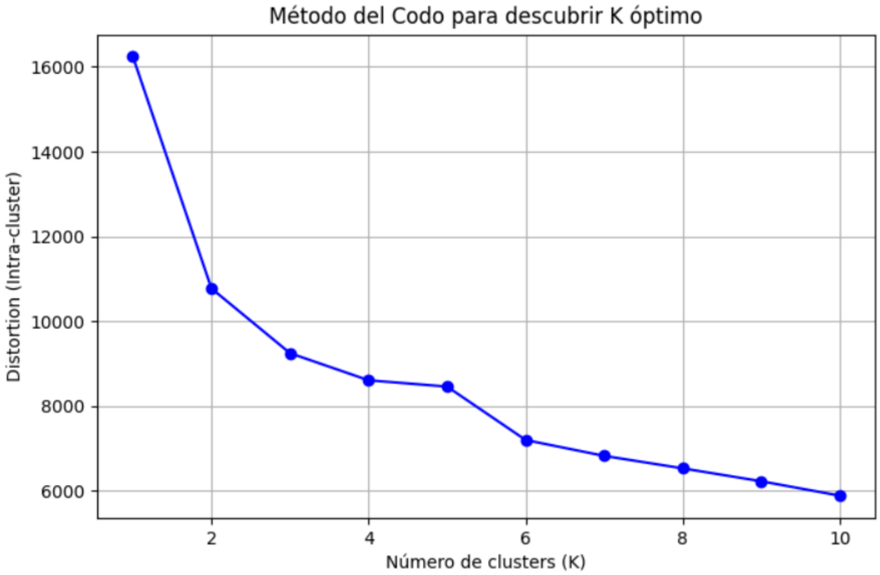


Figure 9: Elbow method for determining the optimal number of clusters. The plot shows the within-cluster distortion (inertia) as a function of K. The elbow point at K=2 indicates the optimal cluster configuration, suggesting a natural binary partition in the data that corresponds to the benign and malignant classes.

Conclusions

This study demonstrates that machine learning models provide robust diagnostic support for breast cancer classification, with KNN achieving 0.9825 accuracy and perfect recall for malignant cases. Size and shape features (radius, area, concave points) showed discriminative power exceeding 200 % between classes, while unsupervised K-Means clustering independently validated the binary structure, reinforcing supervised predictions through convergent evidence.

The clinical impact encompasses three critical contributions: automated pattern recognition enables rapid preliminary screening, reducing diagnostic time; standardized morphological assessment minimizes interobserver variability in FNA cytology interpretation; computational prioritization optimizes specialist workload by directing expert attention to high-risk cases.

Models without dimensionality reduction outperformed PCA variants, indicating that feature preservation is essential for optimal performance. Neural networks demonstrated superior generalization capability and lowest regression errors, validating their potential for clinical deployment.

Future work should prioritize multicenter validation to ensure robustness across institutional settings and patient demographics, explore ensemble methods and deep learning architectures to reduce false negatives, and extend unsupervised validation through hierarchical and density-based clustering for improved clinical interpretability.

Referencias

[1] Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. *IS&T/SPIE International Symposium on Electronic Imaging.*

[2] Wolberg, W. H., & Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences*, 87(23), 9193–9196.

[3] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[4] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

[5] Jolliffe, I. (2002). *Principal Component Analysis.* Springer.

[6] Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, 26(3), 303–304.

[7] Chollet, F. (2015). Keras. GitHub repository: `https://github.com/keras-team/keras`.

[8] Abadi, M., et al. (2016). TensorFlow: Large-scale machine learning on heterogeneous systems. `https://www.tensorflow.org/`

[9] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press.

[10] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323, 533–536.

[11] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3), 175–185.

[12] Leys, C., et al. (2013). Detecting outliers: Do not use standard deviation around the mean, use the MAD instead. *Journal of Experimental Social Psychology.*

[13] Iglewicz, B., & Hoaglin, D. (1993). *How to Detect Outliers.* ASQC.

[14] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS).*

[15] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the predictions of any classifier. *KDD.*

[16] Dhahri, H., et al. (2019). Diagnosing breast cancer using machine learning: A comparative analysis. *Procedia Computer Science*, 164, 131–140.

[17] Chougrad, H., Zouaki, H., & Alheyane, O. (2018). Deep convolutional neural networks for breast cancer screening. *Computer Methods and Programs in Biomedicine.*

[18] Kourou, K., et al. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, 8–17.

[19] Jamieson, A. R., et al. (2010). Exploring nonlinear feature space dimension reduction and data representation in breast CADx. *Medical Physics*, 37(1), 131–145.

[20] Topol, E. (2019). *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again.* Basic Books.

[21] Esteva, A., et al. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25, 24–29.

[22] Checking your browser - reCAPTCHA. (n.d.-b). https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

[23] Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K.-R. (2021). Towards CRISP-ML(Q): A machine learning process model with quality assurance methodology. *Machine Learning and Knowledge Extraction*, *3*(2), 392–413. https://doi.org/10.3390/make3020020

[24] Inga Ortega, E. M., García Herranz, N., Robles-Bykbaev, V. E., & Gallego Diaz, E. (Eds.). (2025). *Systems, Smart Technologies, and Innovation for Society: Proceedings of CITIS 2024*. Springer. https://doi.org/10.1007/978-3-031-87065-1_41