

Identification of clinical, pathological and immunological variables associated with the presence of breast cancer using Machine Learning algorithms.

Leticia Maria Gil Ortiz, Raquel Leon, Gustavo M. Callico

Institute for Applied Microelectronics (IUMA)

University of Las Palmas de Gran Canaria

Las Palmas de Gran Canaria, Spain

leticia.gil101@alu.ulpgc.es

Abstract— This study evaluates the performance of machine learning algorithms for axillary metastasis prediction and molecular subtype classification, comparing them with traditional statistical approaches reported in the literature. Using a cross-validation strategies, different variables were analysed, including clinicopathological and immunohistochemical. The results show that models such as Efficient Logistic Regression and Medium Gaussian SVM offer comparable or superior performance to the reference models, especially in terms of sensitivity. These findings reinforce the potential of machine learning as a tool to support the diagnosis of breast cancer patients.

Keywords- breast cancer; machine learning; supervised learning; classification; regression; machine learning models; k-NN; SVM; Tree; feature selection;

I. INTRODUCTION

Breast cancer [1] is characterised by the uncontrolled and accelerated proliferation of glandular tissue cells. The breast consists of fatty tissue and lymphatic vessels, which are connected to lymph nodes. These nodes are responsible for defence mechanisms, capturing tumour cells, bacteria, and other harmful substances. Different types of breast cancer can be diagnosed depending on the area in which it develops, its size, and whether it has spread to other regions. Therefore, breast cancer may be classified as invasive, non-invasive, or phyllodes tumours with malignant potential.

Non-invasive breast cancer [2] has not spread into the surrounding breast tissue or beyond. Various techniques exist for tumour analysis [3]. Furthermore, molecular subtypes used in the assessment of recurrence include Luminal A, Luminal B, HER2-positive, and triple-negative breast cancer (TNBC) [4]. Among these, the Luminal A subtype is associated with the most favourable prognosis, while TNBC is linked to the poorest prognosis.

This Master's Thesis aims to apply various Machine Learning (ML) algorithms to a dataset comprising 181 patients diagnosed with different molecular subtypes of breast cancer. The objective is to classify these subtypes and to conduct a separate classification of patients based on the presence of axillary lymph node (ALN) metastases.

II. STATE OF ART

In study [5], the authors conducted a retrospective analysis using samples from 141 patients diagnosed with invasive breast carcinoma. These patients were divided into two groups: one with axillary lymph node metastasis (ALN+) and another without metastasis (ALN-). To accurately analyse and quantify various populations of immune cells, the researchers employed immunohistochemical techniques on Tissue Microarrays (TMAs) [6], followed by digital image analysis. These models were further validated using bootstrap resampling techniques, and their predictive performance was evaluated through Receiver Operating Characteristic (ROC) curves. Through these advanced statistical analyses, the authors identified specific immune cell patterns (such as CD68+ macrophages, S100+ dendritic cells, and CD123+ plasmacytoid dendritic cells) that were significantly associated with the presence of axillary metastases.

In study [4], the focus was on comparing the immune response in non-metastatic axillary lymph nodes (ALN-) between different molecular subtypes of breast cancer, specifically triple-negative breast cancer (TNBC) and Luminal A. This retrospective analysis included 144 patients with invasive carcinoma, of whom 88 were Luminal A and 56 were TNBC. Immunohistochemical techniques on TMAs [6] were again employed, and a digital analysis of eleven immune cell populations was conducted. Differences between the subtypes were assessed using various statistical tests: Student's t-test and Mann-Whitney U test for quantitative variables, and Chi-squared tests or Fisher's exact tests for categorical variables. To determine which immunological variables were independently associated with each subtype, univariate logistic regression models were applied, followed by multivariate analyses, further validated through bootstrap techniques and multiple imputation. The results revealed that interdigitating dendritic cells (S100+), cytotoxic T lymphocytes (CD8+), and natural killer (NK) cells (CD57+) in ALN- were significantly associated with the TNBC subtype, whereas mature dendritic cells (CD83+) were associated with the Luminal A subtype.

III. METHODOLOGY

This section details the workflow to be followed to develop a classification model to make predictions from the breast cancer database. The workflow can be seen in **Error! No se encuentra el origen de la referencia.**, which includes the following phases that are developed in MATLAB:

- **Pre-processing:** data recoding (cleaning, renaming and transformation of variables).
- **Processing:** division of data into training and test sets.
- **Learning:** training and optimisation of hyperparameters using Machine Learning algorithms.
- **Evaluation:** assessment of model quality (accuracy, ROC, sensitivity and specificity) and obtaining the final model.

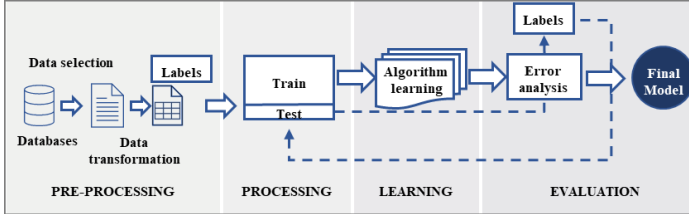


Figure 1. Design flow of the machine learning framework.

A. Machine Learning Algorithm

A ML algorithm is a computational procedure that builds predictive models from data. In the context of supervised learning, these algorithms learn a function that maps inputs to known outputs, with the objective of generalising to previously unseen data.

In this study, four supervised classification algorithms are employed:

1) Medium Tree

A decision tree of intermediate depth, which balances model complexity and generalisation capability [7]. Moreover, they allow capturing moderate non-linear relationships without overfitting, as can occur with trees that are too deep. Furthermore, they allow the extraction of explicit knowledge about the combinations of variables leading to classification, which is important in the clinical setting to interpret model decisions intuitively [8].

2) Efficient Logistic Regression

An optimised linear model that estimates class probabilities based on linear combinations of the input variables [9]. In addition, it incorporates efficient optimisation methods that allow the model to be trained more quickly and stably in the presence of many predictor variables. This model is useful in clinical contexts where the interpretability of the coefficients is essential to draw conclusions about the relevance of each variable [10][11].

3) Medium Gaussian SVM

A support vector machine with a Gaussian [12] kernel, suitable for non-linear classification problems [7]. This is because it has a good ability to project the data into a non-linear space without requiring an overly complex or over-fitted

configuration, which is suitable for clinical arrays with high dimensionality and small size [13].

4) Weighted k -Nearest Neighbours (k -NN)

A distance-based classifier that assigns class labels by weighting the closest neighbours [7]. In addition, it improves the sensitivity and stability of the model in situations where complex local structures exist or classes are partially overlapping, which are common conditions in real clinical data. In addition, the model scales well to small sets, does not require explicit training, and is especially suitable when the feature space has disparate densities between patient groups [14].

B. Feature Selection

In ML, feature selection algorithms are used to identify which variables (or features) are most relevant in predicting a response or outcome. These algorithms can improve the accuracy of models and help simplify them by selecting only the most useful features. There are several common methods used in the scientific literature [15], [16] such as:

ReliefF [17], which is based on the distance between similar observations, using nearest neighbours, and is most effective when all features are of the same type, either categorical or continuous.

The **Kruskal Wallis** algorithm [18] focuses on differences in medians and like ANOVA is suitable for categorical and continuous data. In addition, a score is assigned to each characteristic according to its importance.

C. Evaluation Metrics

To evaluate the performance of the model, several standard metrics were used [19]. Accuracy is the ratio of correct predictions to the total number of instances (1). Sensitivity, also known as recall, measures the model's ability to correctly identify actual positive cases (2), while specificity evaluates its ability to correctly detect actual negative cases (3). The F1-score represents the harmonic mean of precision and recall, providing a balanced measure between the two (4). Lastly, the AUC-ROC assesses the model's capacity to distinguish between classes by analysing the trade-off between true positive and false positive rates across various thresholds.

In all the above metrics, TP refers to true positives, TN to true negatives, FP to false positives, and FN to false negatives.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (3)$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

IV. MATERIALS

The breast cancer database used in this work, and to which ML techniques were applied, is composed of different biopsy samples from the Tumour Bank of the Anatomical Pathology Service of the Hospital de Tortosa Verge de la Cinta and the

Hospital Joan XXIII de Tarragona (Spain). This database collects protein parameters to obtain different indicators that provide information on patients diagnosed with invasive breast carcinoma. These parameters have been collected between 2000 and 2008.

Pathologists have collected in this database different variables (or attributes) such as the number of cases, age of the patients (Figure 2), as well as other medical parameters of the size or the existing proteins in the tumour. These data have been collected for different lines of research to understand the behaviour of a primary breast cancer tumour.

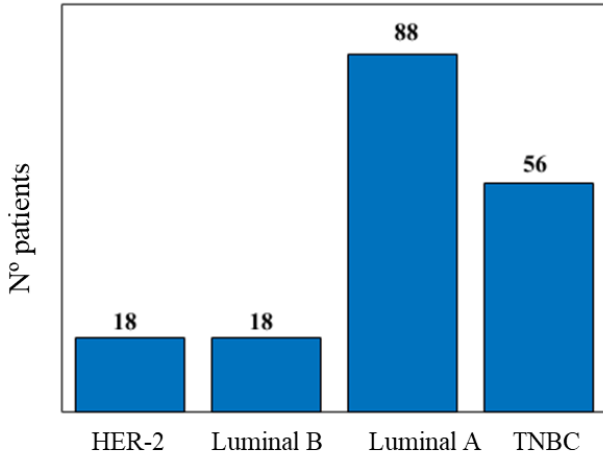


Figure 2. Number of patients by molecular subtype of breast cancer

Therefore, the **¡Error! No se encuentra el origen de la referencia.** shows a summary statistic of the selected variables, calculated on two different subsets of the total available records ($n = 181$). Two sample sizes have been considered as described above, i.e. one of 144 and one of 141 patient records. For each of the variables the mean and standard deviations are presented to identify possible changes in the distribution of the data when the sample size is modified. This comparison allows to evaluate the selected variables and to detect possible variations that could influence the results of the modelling using ML algorithms.

TABLE I. COMPARISON OF DESCRIPTIVE STATISTICS IN PATIENT SUBGROUPS.

Variable	Media		Standard deviation	
	n=144	n=141	n=144	n=141
age	60,20	60,41	11,83	11,35
tumour_diameter	21,40	22,36	15,70	15,84
lymph_number	13,875	14,32	6,38	6,64
lymph_number_metastasis	2,13	2,45	4,67	5,06
survival_120_months	102,46	103,21	30,17	28,34
period_recurrence_120_months	94,99	100,37	36,50	31,41

A. Subset 1: Metastasis identification

To replicate the methodology of the study [5], cases were excluded from the 181 available records due to insufficient information. Specifically, 41 records were eliminated in which the data corresponding to the variable of vascular invasion was missing. However, it was decided to include one additional case in which the vascular invasion value was missing but had HER2

amplification. This case was included because, according to the reference [5], the clinical relevance of this marker justifies its inclusion.

After applying these selection criteria, a final sample of 141 patient records was obtained and included in the statistical analysis. This is represented in the diagram in **¡Error! No se encuentra el origen de la referencia.**

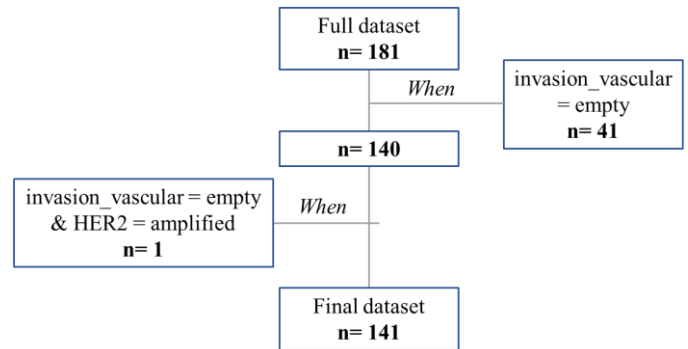


Figure 3. Flowchart for the selection of the DB sample in [5].

To replicate the methodology of [4], cases corresponding to HER2, Luminal B subtypes and those patients classified as null are excluded. After applying these criteria, the sample consists of 144 patients, belonging to the Luminal A and TNBC subtypes. This sample selection process is graphically illustrated in the diagram in **¡Error! No se encuentra el origen de la referencia.** The discarded records, according to cancer subtype, are detailed below:

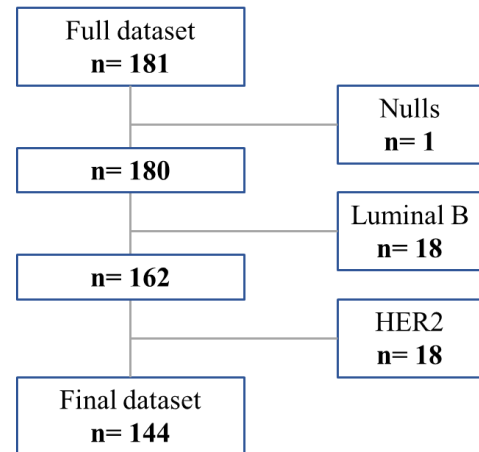


Figure 4. Flow chart for DB sample selection in [4].

V. EXPERIMENTS

The experiments conducted aimed to address different aspects of breast cancer diagnosis, focusing on the prediction of axillary metastasis and the classification of molecular subtypes. The following four experiments utilized various datasets and configurations to explore the relationships between different clinical variables and the presence of metastasis or molecular subtypes of breast cancer.

- Experiment 1: Focused on predicting the presence or absence of axillary metastasis using the full dataset of 181

samples. Ten variables were selected to compare performance with the study [5].

- Experiment 2: Predicted the presence or absence of axillary metastasis using a subset of 141 samples (Subset 1, Section A). To this end, 8 out of the 43 variables in the original dataset are excluded, and the results are compared with the models presented in study [5].
- Experiments 3.1 and 3.2: Also focused on predicting axillary metastasis, using the same subset of 141 samples (Subset 1, Section A). The variables selected are the same as those used in study [5].
- Experiment 4: Addressed the classification of molecular subtypes, specifically distinguishing between TNBC and Luminal A, using a dataset of 144 samples (Subset 2, Section B) [4].

In all cases, a 5-fold cross-validation was applied to the training set, which was obtained by randomly splitting the data into 80% for training and 20% for testing. This approach allowed for model optimisation and assessment of stability prior to applying the model to the independent test set.

For the experiments based on the 141 samples (Experiments 2, 3.1, and 3.2), the same train-test split was used across all runs, in order to ensure consistency in the comparison of results across different models and configurations.

A. Axillary Metastasis Prediction

In the following experiments, axillary metastasis prediction is performed using the ML algorithms previously described.

1) Experiment 1. Variable Selection (Top 10) with 181 registers

Two variable selection methods (ReliefF and Kruskal Wallis) were applied to identify the ten most relevant features using the complete dataset. During this process, variables containing information directly related to the target variable were removed to avoid bias in the prediction. The model was trained and evaluated using the cross-validation scheme described previously.

2) Experiment 2. With 35 variables and 141 registers

From the 43 available variables, those associated with clinical follow-up over time (censored at 120 months) and the variable “ganglios_numero_metastasis” were discarded, as they contained information directly related to the target variable. The aim was to classify patients with axillary metastases (ALN+) and without metastases (ALN-). Different classification models were trained and evaluated, the outputs of which were subsequently compared to analyse their relative performance.

3) Experiment 3.1. With 4 variables and 141 registers

Only clinicopathological and immunohistochemical variables were selected, following the methodological approach

applied in model 1 of the study [5], in order to replicate their experimental conditions. Patients were classified according to the presence (ALN+) or absence (ALN-) of axillary metastasis.

4) Experiment 3.2. With 4 variables and 141 registers

An optimised subset of four immunological and tumour-related variables was selected, based on the approach described for model 2 in study [5]. The aim was to assess the predictive capability of the model using a reduced number of highly informative features. The task involved classifying patients based on the presence (ALN+) or absence (ALN-) of axillary metastasis.

B. Classification of molecular subtypes

1) Experiment 4. With 4 variables and 144 registers

The breast cancer subtype is predicted, distinguishing between TNBC and Luminal A, based on a set of four variables selected following the approach described in study [4].

VI. RESULTS

This section presents the results obtained for each of the implemented ML models, alongside those reported in studies [4] and [5].

A. Axillary Metastasis Prediction

Figure 5 present the ROC, Sensitivity (ALN+), and Specificity (ALN-) results for the different experiments conducted, as well as for the models proposed in [5]. This figure allows for a comparative analysis of classifier performance under various experimental configurations.

It is worth noting that, in Experiment 1, although Model-1 from [5] achieved the highest AUC-ROC value (0.94), the Medium Gaussian SVM (employing Kruskal-Wallis for feature selection) stood out by reaching a sensitivity of 88.2%, a result of particular clinical relevance for metastasis detection. In terms of specificity, the Medium Gaussian SVM demonstrated a level comparable to that of Model-1 from [5], highlighting the well-balanced performance of the proposed model in both sensitivity and specificity. In Experiment 2, although none of the proposed models surpassed Model-1 in AUC-ROC or specificity, the Medium Gaussian SVM once again exhibited strong performance, achieving a sensitivity of 82.4%, substantially higher than that of Model-2 from [5], while also maintaining competitive specificity. This trend continued in Experiment 3.1, where the Efficient Logistic Regression model outperformed Model-1 in AUC-ROC (0.963 vs. 0.94), while preserving similar levels of sensitivity and specificity. Finally, in Experiment 3.2, Efficient Logistic Regression also outperformed Model-2, achieving both a higher AUC-ROC (0.888) and an improved sensitivity (82.4%), along with comparable specificity, further reinforcing its robustness across different experimental configurations.

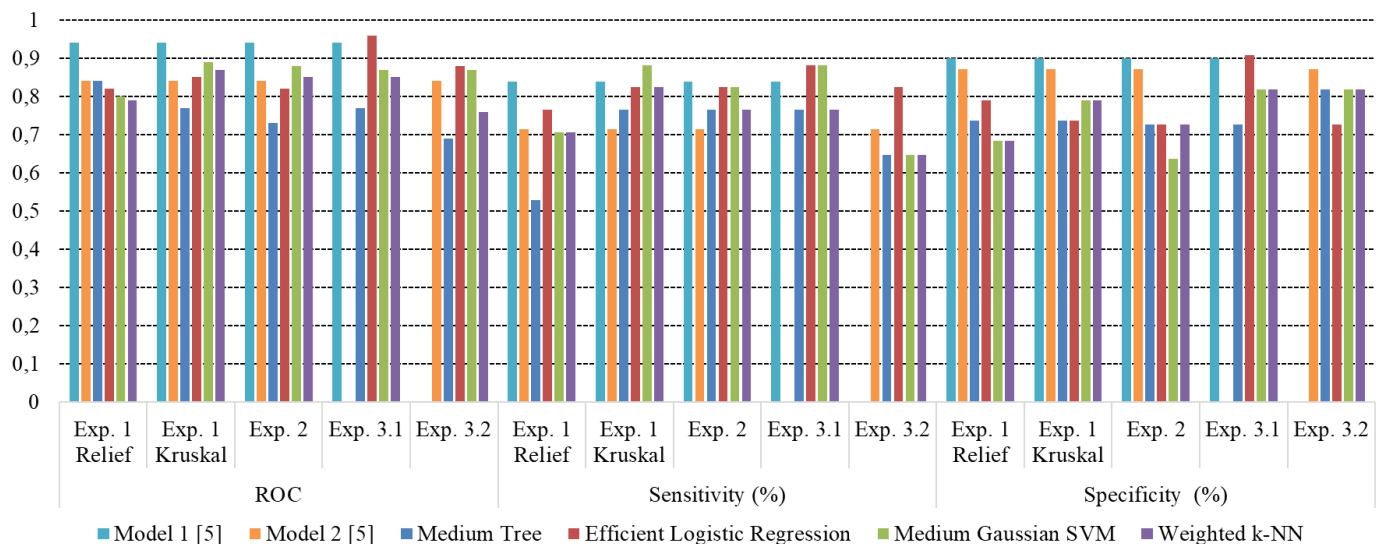


Figure 5. Comparison of experimental model performance (test) with published results [5] using ROC, sensitivity, and specificity Metrics.

B. Clasification of molecular subtypes

Figure 6 the ROC, Sensitivity (TNBC), and Specificity (Luminal A) results for the various models evaluated in Experiment 4, as well as for the model proposed in [4]. This figure allows for a comparative analysis of classifier performance under different experimental configurations.

As shown in Figure 6, the model proposed in [4] achieved the highest AUC-ROC value (0.96), along with the highest specificity (91.0%) for the Luminal A class. On the other hand, the Medium Gaussian SVM and Medium Tree models achieved the highest sensitivity (90.9%) for detecting TNBC cases, matching the performance of other models such as Weighted k-NN, but with the Medium Gaussian SVM additionally standing out due to its higher specificity (83.3%). These results highlight the complementary value of the different models depending on the clinical priority (whether it is the identification of positive cases or the reduction of false positives) in the context of breast cancer subtype diagnosis.

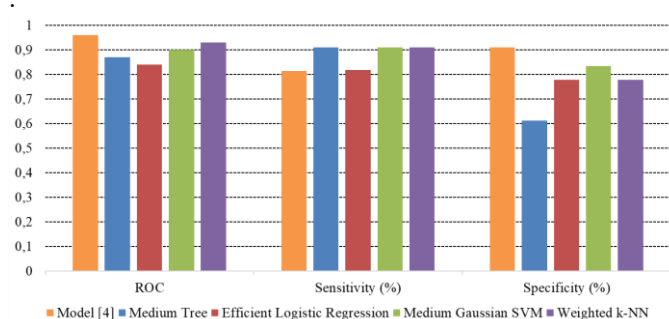


Figure 6. Comparison of the performance of experiment 4 (test) models with published results [4] using ROC metrics, sensitivity and specificity.

VII. CONCLUSIONS

Throughout the experiments, various ML models were evaluated in comparison with traditional statistical approaches described in studies [4] and [5]. One of the most relevant findings was the influence of variable selection on predictive performance. In particular, the Kruskal Wallis method showed better results than ReliefF, excelling in metrics such as sensitivity and AUC-ROC.

Models such as Efficient Logistic Regression and Medium Gaussian SVM obtained highly competitive results. In several cases, they outperformed Model 2 of [5], and even in experiment 3.1, Efficient Logistic Regression exceeded Model 1 in all metrics. A good balance between sensitivity and specificity was also observed in models such as Weighted k-NN, which is essential in the clinical context. In cancer subtype classification (Experiment 4), the proposed models showed high sensitivity in detecting TNBC, outperforming the reference model even with validation on an independent test set.

REFERENCES

- [1] C. para el C. y la P. de E. División de Prevención y Control del Cáncer, "¿Qué es el cáncer de mama? | CDC," División de Prevención y Control del Cáncer, Centros para el Control y la Prevención de Enfermedades, 26-Sep-2022. [En línea]. Disponible: https://www.cdc.gov/spanish/cancer/breast/basic_info/what-is-breast-cancer.htm. [Acceso: 28-Feb-2023]
- [2] American Cancer Society, "Carcinoma ductal in situ," Septiembre, 20, 2019. [En línea]. Disponible: <https://www.cancer.org/es/cancer/cancer-de-seno/acerca/tipos-de-cancer-de-seno/carcinoma-ductal-in-situ.html>. [Acceso: 28-Feb-2023]
- [3] National Cancer Institute, "Advances in Breast Cancer Research - NCI," National Cancer Institute, 28-Aug-2022. [En línea]. Disponible: <https://www.cancer.gov/types/breast/research>. [Acceso: 28-Feb-2023]

- [4] C. López et al., "Differences in the Immune Response of the Nonmetastatic Axillary Lymph Nodes between Triple-Negative and Luminal A Breast Cancer Surrogate Subtypes," *American Journal of Pathology*, vol. 191, no. 3, pp. 545–554, 2021 [En línea]. Disponible: <https://doi.org/10.1016/j.ajpath.2020.11.008>
- [5] C. López et al., "The Immune Response in Nonmetastatic Axillary Lymph Nodes Is Associated with the Presence of Axillary Metastasis and Breast Cancer Patient Outcome," *American Journal of Pathology*, vol. 190, no. 3, pp. 660–673, 2020 [En línea]. Disponible: <https://doi.org/10.1016/j.ajpath.2019.11.002>
- [6] V. W. Facchetti F, Cigognetti M, Fisogni S, Rossi G, Lonardi S, "Neoplasia hematodérmica CD4+/CD56+. Diagnóstico histopatológico, fisiopatología y avances recientes de un tumor originado en células dendríticas plasmocitoides," 2016. [En línea]. Disponible: <https://www.elsevier.es/es-revista-vigilia-sueno-270-pdf-S169988550770051X>. [Acceso: 06-Apr-2023]
- [7] Mathworks, "Statistics and Machine Learning Toolbox TM User 's Guide R 2024 a," pp. 1–12004, 2024.
- [8] M. Zorman, H. P. Eich, P. Kokol, and C. Ohmann, "Comparison of three databases with a decision tree approach in the medical field of acute appendicitis," *Studies in Health Technology and Informatics*, vol. 84, no. February, pp. 1414–1418, 2001.
- [9] E. C. Zabor, C. A. Reddy, R. D. Tendulkar, and S. Patil, "Logistic Regression in Clinical Studies," *International Journal of Radiation Oncology*Biophysics*, vol. 112, no. 2, pp. 271–277, 2022 [En línea]. Disponible: <https://www.sciencedirect.com/science/article/pii/S0360301621026468>
- [10] E. Y. Boateng and D. A. Abaye, "A Review of the Logistic Regression Model with Emphasis on Medical Research," *Journal of Data Analysis and Information Processing*, vol. 07, no. 04, pp. 190–207, 2019.
- [11] L. E. Barker, "Logit Models From Economics and Other Fields," *Technometrics*, vol. 47, no. 1, pp. 97–98, 2005 [En línea]. Disponible: https://www.cambridge.org/ua/files/2013/6690/1022/1208_default.pdf
- [12] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, "A Practical Guide to Support Vector Classification," *BJU international*, vol. 101, no. 1, pp. 1396–1400, 2008 [En línea]. Disponible: <http://www.csie.ntu.edu.tw/%7B~%7Dcjlin/papers/guide/guide.pdf>
- [13] S. L. Lin, "Application of machine learning to a medium gaussian support vector machine in the diagnosis of motor bearing faults," *Electronics (Switzerland)*, vol. 10, no. 18, 2021.
- [14] M. Ashraf, K. Le, and X. Huang, "Iterative weighted k-NN for constructing missing feature values in Wisconsin breast cancer dataset," *Proceedings - 3rd International Conference on Data Mining and Intelligent Information Technology Applications, ICMIA 2011*, pp. 23–27, 2011.
- [15] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018 [En línea]. Disponible: <https://www.sciencedirect.com/science/article/pii/S0925231218302911>
- [16] A. Ghosal, A. Nandy, A. K. Das, S. Goswami, and M. Panday, "A Short Review on Different Clustering Techniques and Their Applications BT - Emerging Technology in Modelling and Graphics," 2020, pp. 69–83.
- [17] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *Journal of Biomedical Informatics*, vol. 85, pp. 189–203, 2018 [En línea]. Disponible: <https://www.sciencedirect.com/science/article/pii/S1532046418301400>
- [18] M. A. Jamil and S. Khanam, "Influence of One-Way ANOVA and Kruskal–Wallis Based Feature Ranking on the Performance of ML Classifiers for Bearing Fault Diagnosis," vol. 12, no. 3. Springer Nature Singapore, 2024 [En línea]. Disponible: <https://doi.org/10.1007/s42417-023-01036-x>
- [19] "Validation Techniques for Machine Learning Models: Examples." [En línea]. Disponible: <https://vitalflux.com/machine-learning-validation-techniques-interview-questions/>. [Acceso: 18-Jul-2025]