Predicting Breast Cancer Diagnosis

A Machine Learning Approach

By Rolddy SURPRIS and Naël Yssa Iben Ahmed ROBERT

Foreword

This project was completed as part of the final phase of our Data Science Bootcamp with Akademi. Throughout this intensive program, we developed essential skills in data analysis, statistics, machine learning, natural language processing, and neural networks. The capstone project represents the culmination of this training and demonstrates our ability to apply these concepts to a real-world problem.

We chose to focus our efforts on the early diagnosis of breast cancer, a subject of great importance both globally and locally. As data science students, we wanted our final work to go beyond theory and have potential social impact. Our goal was to explore how data-driven models can assist medical professionals in identifying breast cancer earlier and more accurately, particularly in environments where medical resources are limited.

This report reflects months of study, experimentation, and collaboration. It also represents our commitment to using technology for the greater good, contributing in a small but meaningful way to improving public health outcomes. We hope that this work will inspire further research and encourage the integration of data science solutions in healthcare systems that need them most.

Table of Contents

Introduction, background and objectives

Breast cancer remains one of the most common and deadliest diseases affecting women worldwide. Every year, millions of women are diagnosed, and many lose their lives because the disease is detected too late. In Haiti, the situation is particularly concerning due to limited healthcare infrastructure, a shortage of specialists, and restricted access to medical laboratories. Many patients live far from urban centers, and testing or receiving results can take weeks or even months, reducing the chances of timely treatment.

October is recognized as Breast Cancer Awareness Month, a period dedicated to education, prevention, and support for those affected. It is also a time to reflect on how advances in science and technology can improve early and accurate diagnosis. In recent years, data science and artificial intelligence have shown great potential in supporting medical decision-making. Machine learning models can identify complex patterns in medical data and assist healthcare professionals in detecting diseases earlier and with greater precision. These technologies can help reduce diagnostic delays and strengthen healthcare systems where resources are limited.

The purpose of this project is to apply data science techniques to breast cancer diagnosis and evaluate the effectiveness of machine learning algorithms in predicting tumor types. Specifically, it develops and compares several classification models using a publicly available dataset of tumor characteristics to determine which model provides the highest accuracy and reliability for early detection. The study also explores how such models could support medical systems in low-resource settings like Haiti.

By combining technology and human expertise, this project contributes to the ongoing fight against breast cancer. It demonstrates how data-driven approaches can assist in improving healthcare outcomes and give more women the opportunity to detect the disease early and begin treatment in time.

Dataset Description

The dataset used in this project is the Breast Cancer Wisconsin (Diagnostic) Dataset, a well-known and publicly available dataset from the University of Wisconsin–Madison. It was created by Dr. William H. Wolberg and his colleagues at the university's hospital, where they collected data from breast mass fine needle aspirate (FNA) samples. These samples were digitized and analyzed to measure the characteristics of cell nuclei present in breast tissue images. The dataset was later made available through the UCI Machine Learning Repository, one of the most reputable sources for academic datasets.

Each record in the dataset describes one breast tumor sample and includes a series of numeric features computed from a digitized image of a fine needle aspirate of a breast mass. These features capture important characteristics such as the texture, smoothness, compactness, symmetry, and fractal dimension of the cell nuclei. Based on these measurements, the tumors are labeled as either benign (non-cancerous) or malignant (cancerous).

This dataset has been widely used in machine learning research and education for tasks related to classification, feature selection, and model evaluation. It serves as a reliable benchmark for testing the performance of algorithms designed for medical diagnosis and prediction. Because it is clean, well-documented, and easy to interpret, it is particularly suitable for academic and applied data science projects.

We chose this dataset because it provides a clear and measurable way to explore how machine learning can contribute to early breast cancer diagnosis. Its structure allows for the application of various supervised learning techniques, including logistic regression, decision trees, support vector machines, and ensemble methods. Analyzing this dataset enables us to compare the performance of these models and evaluate their potential to support medical decision-making, especially in contexts where early detection can save lives.

Variables Overview

The Breast Cancer Wisconsin (Diagnostic) dataset contains 569 observations (rows) and 33 variables (columns).

1. id – Unique identifier for each patient/sample.
2. diagnosis – The target variable indicating the type of tumor: M for malignant and B for benign.
3. radius_mean – Mean of distances from the center to points on the tumor perimeter.
4. texture_mean – Standard deviation of gray-scale values; measures variation in texture.
5. perimeter_mean – Mean perimeter of the tumor.
6. area_mean – Mean area of the tumor.
7. smoothness_mean – Mean local variation in radius lengths; indicates how smooth the tumor boundary is.
8. compactness_mean – Mean of the ratio ((perimeter^2 / area - 1.0)); measures tumor compactness.
9. concavity_mean – Mean severity of concave portions of the contour.
10. concave points_mean – Mean number of concave portions of the contour.
11. symmetry_mean – Mean symmetry of the tumor.
12. fractal_dimension_mean – Mean “coastline approximation” of the tumor boundary; indicates shape complexity.
13. radius_se – Standard error of the radius.
14. texture_se – Standard error of the texture.
15. perimeter_se – Standard error of the perimeter.
16. area_se – Standard error of the area.
17. smoothness_se – Standard error of smoothness.
18. compactness_se – Standard error of compactness.
19. concavity_se – Standard error of concavity.
20. concave points_se – Standard error of concave points.
21. symmetry_se – Standard error of symmetry.
22. fractal_dimension_se – Standard error of fractal dimension.
23. radius_worst – Largest value of radius (mean of three largest values).
24. texture_worst – Largest value of texture.
25. perimeter_worst – Largest value of perimeter.
26. area_worst – Largest value of area.
27. smoothness_worst – Largest value of smoothness.
28. compactness_worst – Largest value of compactness.
29. concavity_worst – Largest value of concavity.
30. concave points_worst – Largest value of concave points.
31. symmetry_worst – Largest value of symmetry.
32. fractal_dimension_worst – Largest value of fractal dimension.
33. Unnamed: 32 – Empty column; contains no meaningful data.

Tools, Libraries, and Methodology

The analysis and modeling were conducted using Python within Jupyter Notebooks, organized in a GitHub repository. The repository includes:

- README – Provides an overview of the project, instructions, and documentation.
- (EDA) notebook – Used to explore the dataset, visualize distributions, identify correlations, and detect potential issues in the data.
- Data Cleaning notebook – Handles missing values, removes duplicates, and ensures that all variables are properly formatted for analysis.
- Modeling notebook – Implements predictive models, including logistic regression, multi-layer perceptron (MLP), and random forest, and evaluates them using relevant metrics.
- Optimization notebook – Focuses on tuning the parameters of the chosen model to improve performance, including threshold adjustments and hyperparameter optimization.

The project leverages a comprehensive set of Python libraries for data manipulation, visualization, modeling, and evaluation:

- Pandas – Data manipulation, cleaning, and analysis.
- NumPy – Numerical operations and array computations.
- Matplotlib – Basic data visualization and plotting graphs.
- Seaborn – Advanced statistical visualizations, heatmaps, and correlation plots.
- Scikit-learn (sklearn) – Machine learning algorithms and utilities:
  - Logistic Regression (sklearn.linear_model.LogisticRegression)
  - Multi-Layer Perceptron (sklearn.neural_network.MLPClassifier)
  - Random Forest (sklearn.ensemble.RandomForestClassifier)
  - Train-test split (sklearn.model_selection.train_test_split)
  - Cross-validation (sklearn.model_selection.cross_val_score)
  - Hyperparameter tuning (sklearn.model_selection.GridSearchCV, RandomizedSearchCV)
  - Standard scaling and normalization (sklearn.preprocessing.StandardScaler)
  - Label encoding (sklearn.preprocessing.LabelEncoder)
  - Evaluation metrics (sklearn.metrics.accuracy_score, precision_score, recall_score, f1_score, roc_auc_score, confusion_matrix, classification_report, roc_curve, auc)
- Statsmodels – Statistical modeling, regression analysis, and generating summary statistics.
- Python Standard Libraries

The project follows a structured workflow to analyze and model breast cancer data:

- Exploratory Data Analysis (EDA) – The dataset was initially explored to understand variable distributions, identify correlations, and detect anomalies. Visualizations were created to summarize relationships between features and the target variable.
- Data Cleaning – Missing values were handled, irrelevant or redundant columns were removed, and all variables were formatted correctly to prepare the data for modeling.
- Modeling – Three predictive approaches were implemented:
  - Logistic Regression – Baseline model for binary classification of tumors as malignant or benign. Evaluation metrics included precision, recall, F1-score, and ROC-AUC.
  - Multi-Layer Perceptron (MLP) – Used to capture potential non-linear relationships between features and the target variable.
  - Random Forest – Ensemble method combining multiple decision trees to assess performance improvement.
- Model Optimization – The best-performing model was further refined through parameter tuning and threshold adjustments to improve evaluation metrics such as recall.

This structured approach ensures a comprehensive analysis, from understanding the data to selecting and optimizing predictive models, leveraging Python, Jupyter Notebooks, and a suite of powerful libraries for reproducible and research-quality results.

The github repository can be accessed through this link: https://github.com/Rawldyh/Predicting-Breast-Cancer-Diagnosis.git

Exploratory Data Analysis (EDA) and Cleaning

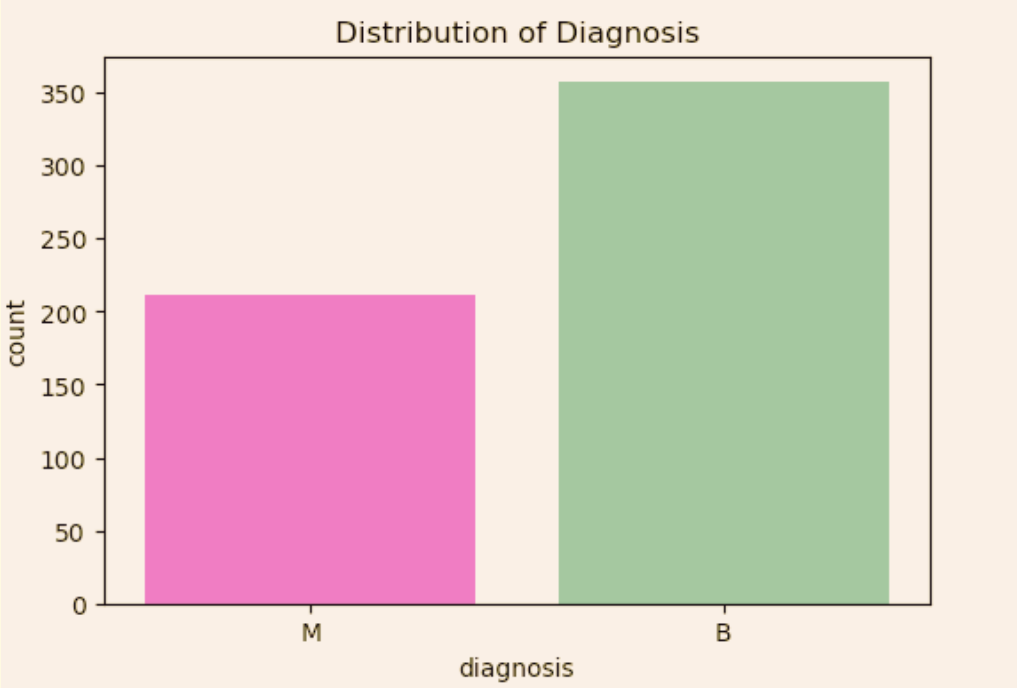
The initial step in the analysis involved exploring the Breast Cancer Wisconsin dataset to understand its structure, distributions, and relationships between features and the target variable.

1. Dataset Inspection

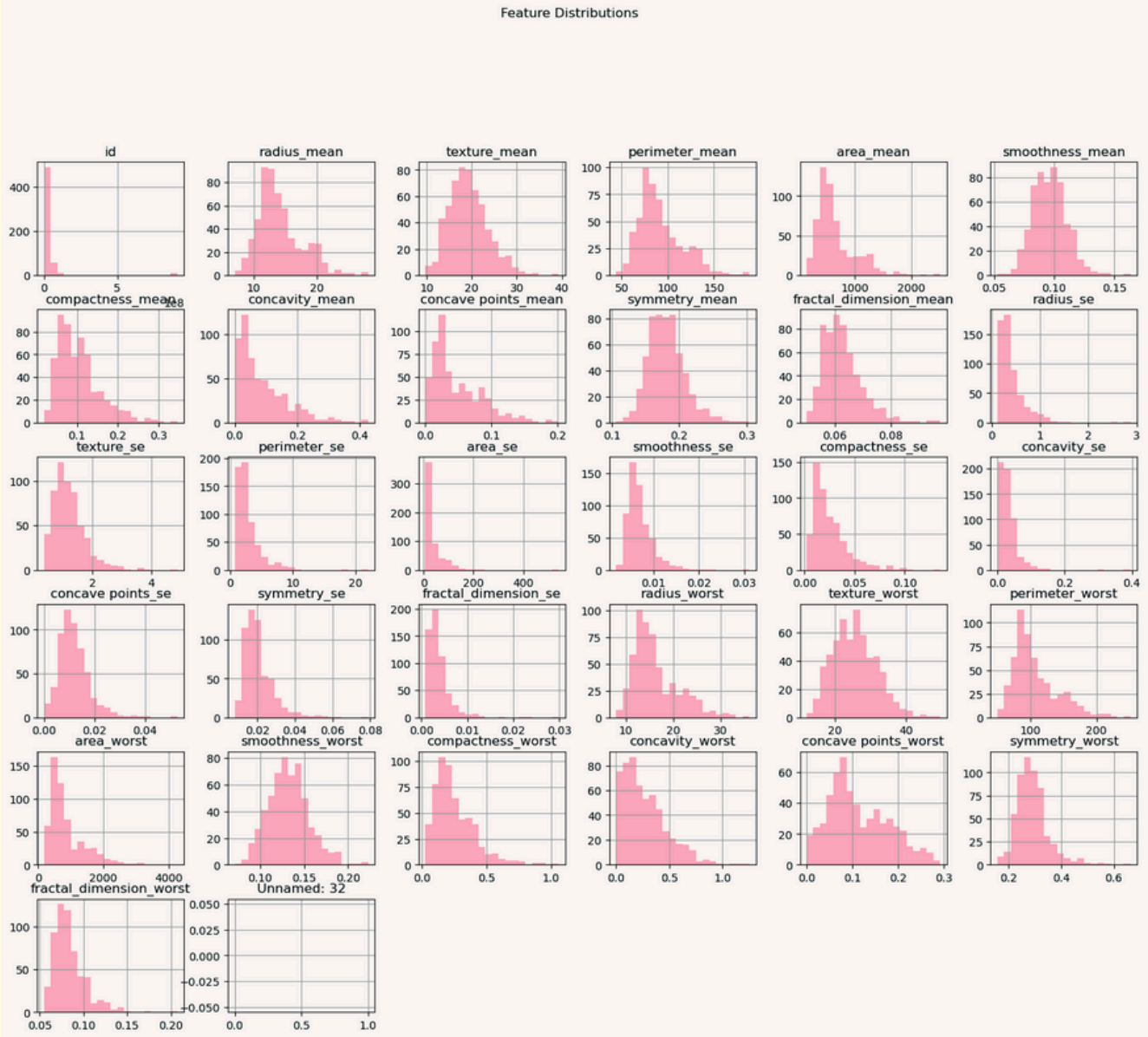
The dataset contains 569 observations and 33 variables. Inspecting the target variable, diagnosis, revealed the following distribution:

- Benign (B): 62.74%
- Malignant (M): 37.26%

This shows that the dataset is imbalanced, with more benign cases than malignant ones. While this imbalance can potentially influence predictive modeling, it also accurately reflects real-world medical scenarios, where most breast tumors are non-cancerous. Despite the dominance of benign tumors, the malignant subset provides critical information for identifying high-risk cases.



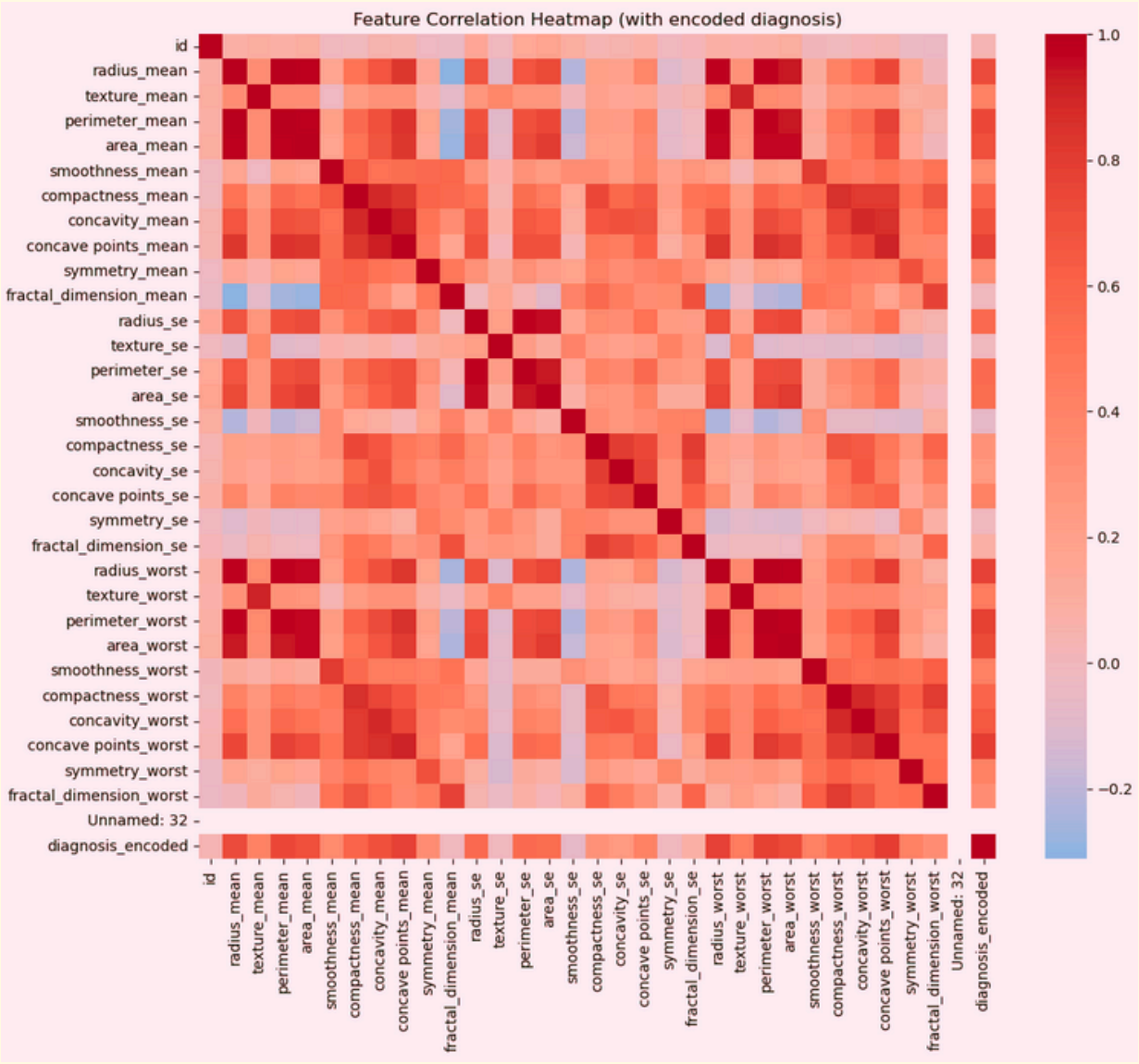
2. Feature Distribution



3. Feature Correlation Analysis

To identify the features most associated with malignancy, we computed the correlation of each numeric feature with the encoded diagnosis variable. The analysis highlights that tumor size (radius, perimeter, area) and shape irregularity (concavity, concave points) are the most critical factors associated with malignancy. In particular:

- Concave points_worst (0.79), perimeter_worst (0.78), and concave points_mean (0.78) indicate that tumors with more concave points and larger perimeters are highly indicative of malignancy.
- Radius_worst (0.78) and perimeter_mean (0.74) further emphasize the role of tumor size and boundary irregularity.
- Area-related features such as area_worst (0.73) and area_mean (0.71) suggest that malignant tumors occupy significantly larger regions compared to benign ones.
- Concavity features (concavity_mean 0.70, concavity_worst 0.66) show that indentation depth in tumor shapes is another key differentiator.



The exploratory analysis clearly indicates that size and irregularity features dominate the predictive signal for malignancy. These insights will guide feature selection and model interpretability in subsequent modeling steps.

4. Data Cleaning

Before modeling, the dataset was prepared to ensure compatibility with predictive algorithms. Cleaning steps were minimal but essential:

1. Dropping unnecessary columns – The id column and the empty Unnamed: 32 column were removed, as they do not contain predictive information.
2. Encoding the target variable – The diagnosis column was transformed into a numeric format, with malignant tumors encoded as 1 and benign tumors encoded as 0, to facilitate model training.

These steps ensured a clean, structured dataset ready for subsequent analysis and modeling.

Modeling and Evaluation

The objective of this step is to identify a classification model capable of reliably detecting malignant tumors while minimizing false negatives. In a medical context, false negatives are particularly critical because they could delay treatment and increase patient risk. Therefore, model evaluation prioritizes high recall for malignant cases alongside overall accuracy.

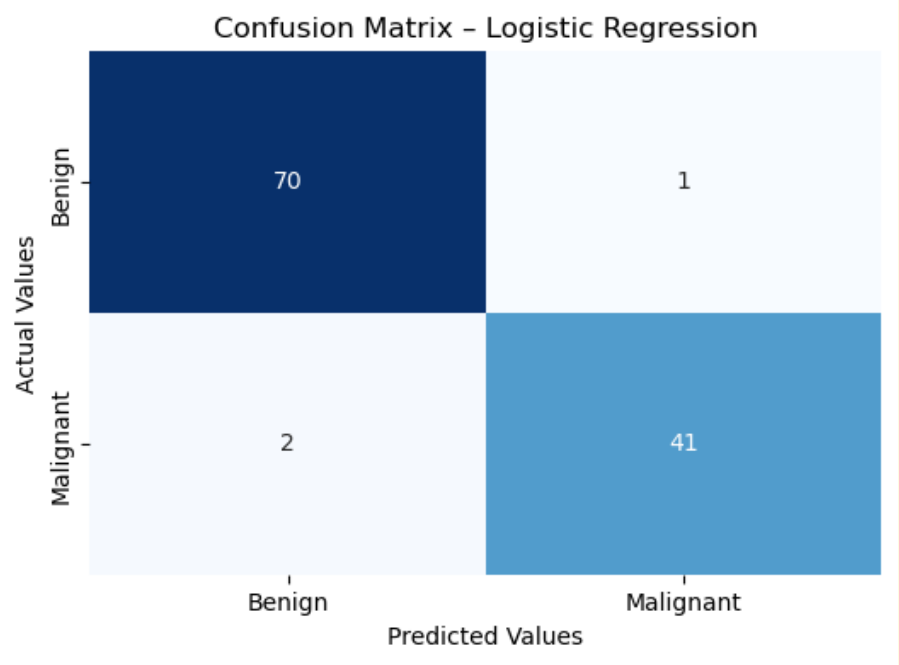
1. Logistic Regression

Logistic regression was used as the baseline model because of its simplicity and interpretability. The dataset was split into 80 percent training and 20 percent testing data. A pipeline combining data standardization with the logistic regression model was applied to train the classifier.

Performance metrics:

- Accuracy: 97.37 percent
- Confusion matrix: 70 true negatives, 41 true positives, 1 false positive, 2 false negatives
- Recall for malignant tumors: 0.95
- Precision for malignant tumors: 0.98

Logistic regression demonstrated strong performance, correctly classifying most tumors and minimizing critical errors. Only two malignant tumors were misclassified. This model provides a reliable baseline for further comparison.

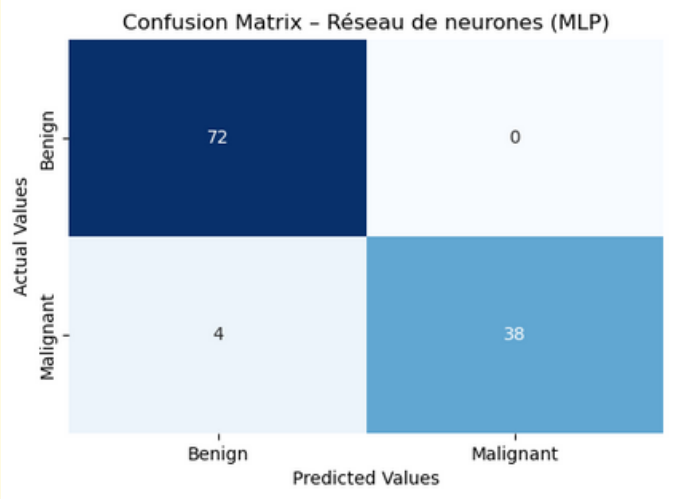


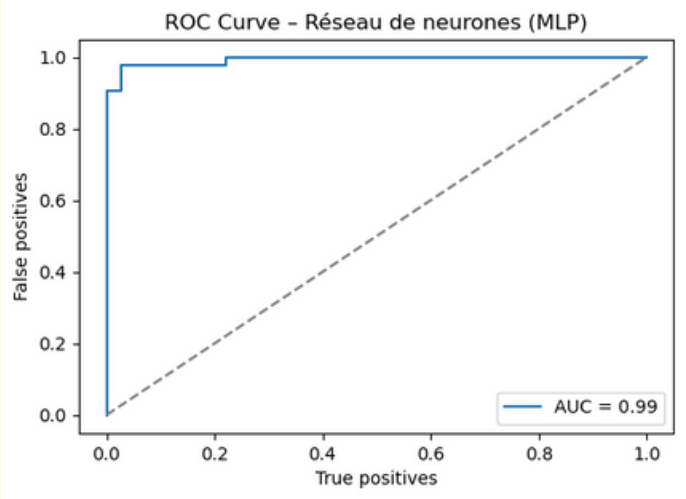
2. Multi-Layer Perceptron

A multilayer perceptron (MLP) was implemented to capture potential non-linear relationships between features and the target variable. The network consisted of two hidden layers with 64 and 32 neurons, ReLU activation functions, and the Adam optimization algorithm. Data were standardized before training.
Performance metrics:

- Accuracy: 96.49 percent
- Confusion matrix: 72 true negatives, 38 true positives, 0 false positives, 4 false negatives
- Recall for malignant tumors: 0.90
- Precision for malignant tumors: 1.00
- ROC-AUC: 0.97

The MLP correctly identified all benign tumors and achieved perfect precision for malignant cases. However, it missed more malignant tumors than logistic regression. This result indicates that a more complex model does not necessarily improve performance for critical medical outcomes.



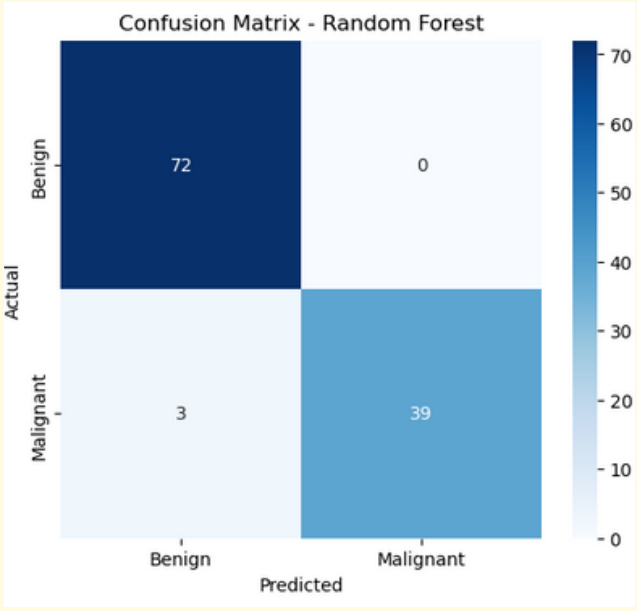


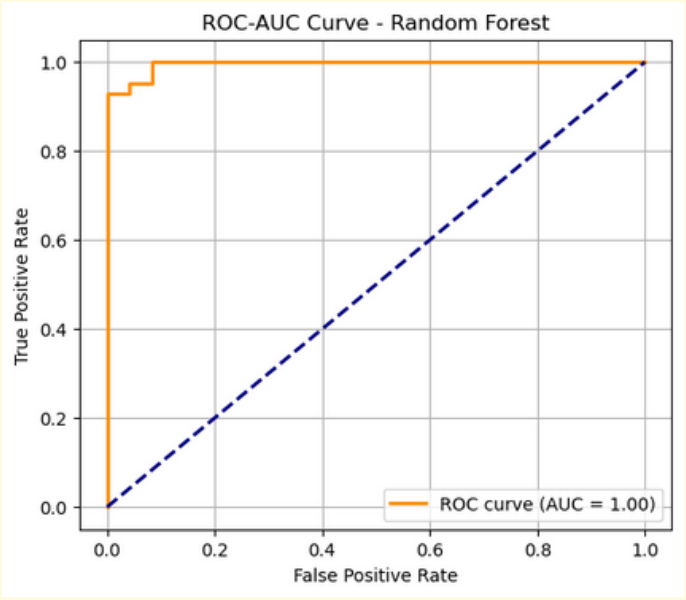
3. Random Forest

A Random Forest classifier with 100 trees and a maximum depth of five was evaluated to examine the benefits of ensemble learning.
Performance metrics:

- Accuracy: 97.37 percent
- Confusion matrix: 72 true negatives, 39 true positives, 0 false positives, 3 false negatives
- Recall for malignant tumors: 0.93
- Precision for malignant tumors: 1.00
- ROC-AUC: 0.97

Random Forest achieved high overall accuracy and correctly identified all benign tumors. It missed three malignant tumors, which is slightly worse than logistic regression in terms of recall.





Model Comparison

Model	Accuracy	Precision (Malignant)	Recall (Malignant)	False Negatives
Logistic Regression	97.37%	0.98	0.95	2
MLP	96.49%	1	0.9	4
Random Forest	97.37%	1	0.93	3

Logistic regression was selected as the final model because it achieved the highest recall for malignant tumors and the lowest number of false negatives. Its simplicity and interpretability make it suitable for clinical applications where minimizing missed cancer diagnoses is essential. Further parameter tuning will be conducted to optimize performance without reducing interpretability.

Model Optimization and Results

After identifying logistic regression as the optimal baseline model for clinical purposes, the next step focused on hyperparameter tuning to maximize recall for malignant tumors while maintaining interpretability. The primary clinical goal is to minimize false negatives, ensuring malignant cases are rarely missed.

Hyperparameter Tuning

A grid search was conducted to evaluate combinations of regularization strength (C = 0.1, 1, 10), solver type (liblinear, saga), L2 penalty, and class weighting (balanced or none). Recall was used as the scoring metric. The best configuration was:

- C: 10
- Penalty: L2
- Solver: saga
- Class weight: balanced

This setup achieved a cross-validated recall of 0.965, indicating that the model is highly sensitive to malignant cases.

Evaluation on Test Data

The optimized model was evaluated on the held-out test set. Using the standard threshold of 0.5, the results were:

- Accuracy: 97.4%
- Recall (malignant): 0.95
- Precision (malignant): 0.98

The confusion matrix revealed two malignant cases were misclassified as benign, and one benign case was misclassified as malignant. Lowering the decision threshold to 0.4 did not change these results, indicating that the model's probability estimates are well-calibrated.

Threshold analysis was further extended from 0.5 down to 0.1. Recall remained at 0.95 across all thresholds, while precision decreased only at the extreme cutoff of 0.1 due to additional false positives. This demonstrates that threshold adjustment alone cannot recover the two misclassified malignant cases, as their predicted probabilities were firmly in the benign range.

Threshold	Accuracy	Precision (Malignant)	Recall (Malignant)	Notes
0.5	0.974	0.976	0.952	Default threshold
0.45	0.974	0.976	0.952	Slightly lower cutoff
0.4	0.974	0.976	0.952	Custom threshold tested
0.35	0.974	0.976	0.952	Stable performance
0.3	0.974	0.976	0.952	Minimal impact on classification
0.25	0.974	0.976	0.952	Still no improvement in recall
0.1	0.921	0.846	0.952	Extreme threshold; recall maintained,

Logistic Regression with Engineered Features

Feature engineering and interaction terms were introduced to capture non-linear relationships and subtle patterns. These included:

- Ratio features: area-to-perimeter, concavity-to-smoothness, radius worst-to-mean
- Logarithmic transformations for skewed variables
- Polynomial interactions (degree 2, interaction-only)

The resulting model achieved the same performance as the baseline optimized logistic regression:

- Accuracy: 97.4%
- Recall (malignant): 0.95
- Precision (malignant): 0.98

This indicates that the persistent false negatives are not due to the absence of feature interactions but reflect inherent limitations in the current feature set.

Analysis of Misclassified Cases

The two malignant tumors that were consistently misclassified were smaller than the average malignant tumor and had mixed irregularity signals. One case exhibited extreme irregularity, while the other showed only mild irregularity. Engineered features did not alter their classification. This confirms that some malignant tumors naturally resemble benign cases in measured features, demonstrating the biological unpredictability of cancer.

Clinical Implications

The logistic regression model is stable, interpretable, and accurate, reliably detecting the majority of malignant tumors. The small number of false negatives highlights a fundamental reality: no algorithm can fully account for biological variability. Machine learning should therefore serve as a decision support tool, augmenting clinical judgment rather than replacing it. Clinicians remain essential for integrating model outputs with patient-specific information and broader medical context.

Conclusion and Perspectives

In this study, we developed a logistic regression model to support the detection of malignant breast tumors. Our focus was not only on accuracy but also on clinical reliability, ensuring that the model would minimize the risk of missing high-risk cases. The results showed that the model is robust, interpretable, and capable of flagging the vast majority of malignant tumors correctly.

When we explored model tuning, threshold adjustments, and feature engineering, we observed that the few remaining misclassified tumors were not due to flaws in the model or limitations of machine learning. Instead, they reflect the inherent unpredictability of cancer and the medical field itself. Biological processes are complex and dynamic, and some tumors do not follow expected patterns. This insight reminds us that models can support, but never replace, the nuanced judgment of a skilled clinician.

This work demonstrates the immense potential of machine learning in medicine. Tools like the one we built can serve as reliable assistants, highlighting high-risk cases, standardizing evaluations, and providing interpretable insights that help clinicians make better-informed decisions. When integrated thoughtfully into healthcare workflows, such models can accelerate detection, improve patient safety, and contribute to more effective treatment strategies.

Looking forward, we are eager to continue exploring this intersection of data science and medicine. We hope to work with broader and richer datasets, integrate multiple data modalities, and build more advanced models that can capture subtle patterns in disease progression. Our goal is to participate actively in this ongoing data revolution, contributing to innovations that help healthcare systems work faster, more efficiently, and with greater precision. Through these efforts, we aim to use the power of data to create meaningful, tangible improvements in patient care, while always keeping human expertise at the center of decision-making.

Appendix

A. Dataset Information

Dataset Name: Breast Cancer Wisconsin (Diagnostic) Dataset
Source: UCI Machine Learning Repository

B. Models Used

Logistic Regression

Logistic Regression is a linear classification model that predicts the probability of a sample belonging to a certain class (benign or malignant). It calculates a weighted sum of the features and applies the logistic (sigmoid) function to produce a probability between 0 and 1.

Multi-Layer Perceptron (MLP)

MLP is a type of artificial neural network that can model complex, non-linear relationships. It consists of layers of interconnected nodes (neurons), where each node applies a weighted sum and activation function to its inputs. MLPs can learn patterns in data that linear models may not capture.

Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions. Each tree is trained on a random subset of the data and features. The final prediction is obtained by averaging (regression) or majority voting (classification) across all trees, which generally improves accuracy and reduces overfitting.

Logistic Regression with Feature Engineering

An extension of logistic regression that includes additional features such as ratios and interaction terms to capture more complex relationships between tumor characteristics.

C. Evaluation Metrics

In this project, we used the following metrics to assess model performance:

Metric	Description	Documentation Link
Accuracy	Proportion of correctly classified samples among all samples.	accuracy_score
Precision	Proportion of predicted positives that are true positives.	precision_score
Recall (Sensitivity)	Proportion of actual positives correctly identified by the model.	recall_score
F1-score	Harmonic mean of precision and recall. Balances both false positives and false negatives.	f1_score

Metric	Description	How to Interpret	Documentation Link
ROC-AUC	Area under the ROC curve;	High ROC-AUC means the model	roc_auc_score
Precision-Recall Curve	Precision and recall values	Shows how precision and	precision_recall_curve
Confusion Matrix	2×2 table showing true positives,	Reveals detailed classification	confusion_matrix

Disclaimer

This project and the analyses presented herein are conducted strictly for educational and research purposes. The dataset used in this study, the Breast Cancer Wisconsin (Diagnostic) dataset, is a publicly available resource intended for academic exploration of data science and machine learning techniques.

The models developed and results obtained in this work are not intended for clinical or diagnostic use. They do not replace medical expertise, professional judgment, or laboratory testing. Any interpretation or application of these findings in real-world medical contexts must be done under the supervision of qualified healthcare professionals.

All interpretations, analyses, and conclusions are the responsibility of the authors and do not reflect the official position of any medical institution or data provider.

Acknowledgments

We would like to express our deepest gratitude to Akademi for offering this intensive and enriching bootcamp experience. These past months have been a journey of growth, discovery, and dedication, during which we gained not only technical knowledge but also the confidence to apply data science to real-world challenges.

A special thank you goes to Castelline Tilus, CEO of Akademi, whose vision and leadership made this opportunity possible. Your commitment to empowering learners and fostering excellence has been a true inspiration.

We also extend our heartfelt appreciation to our instructors, Jérôme Wedter and Geovany Batista Laguerre, for their guidance, patience, and encouragement throughout the program. Your passion for teaching and your constant support played a key role in our learning journey.

Finally, thank you to all those who contributed, directly or indirectly, to this project, This work stands as a reflection of the collective effort, mentorship, and inspiration that Akademi embodies.

Contact

Rolddy SURPRIS
Naël Yssa Iben Ahmed ROBERT
rolddysurpris@gmail.com
rnaelyssa@gmail.com

https://github.com/Rawldyh/Predicting-Breast-Cancer-Diagnosis.git