REPORT

On

BREAST CANCER PREDICTION USING LOGISTIC REGRESSION

AND

DATA ANALYSIS OF WISCONSIN DIAGNOSTIC BREAST CANCER (WDBC) DATASET

BY-PRAGATI SAWARN

INTRODUCTION

Breast cancer remains one of the most prevalent and life-threatening diseases affecting women worldwide. Early and accurate diagnosis is critical for improving treatment outcomes and patient survival rates. Machine learning (ML) techniques have emerged as powerful tools for analysing complex medical data, enabling the development of predictive models that can assist healthcare professionals in diagnosis.

This project focuses on breast cancer classification using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, leveraging machine learning algorithms to distinguish between malignant (cancerous) and benign (non-cancerous) tumours. By analysing key clinical features such as tumour radius, texture, and perimeter, we developed a logistic regression model achieving 95%+ accuracy, demonstrating the potential of ML in medical diagnostics.

OBJECTIVE

The primary objective of this project is to develop a highly accurate and interpretable machine learning model for the early prediction of breast cancer (malignant vs. benign) using clinical diagnostic features. Specifically, we aim to:

1. Analyse the Wisconsin Diagnostic Breast Cancer (WDBC) dataset to identify key patterns and correlations among tumour characteristics.
2. Train and optimize a logistic regression model to achieve >95% classification accuracy, ensuring reliability for diagnostic support.
3. Enhance interpretability by highlighting feature importance and decision boundaries, enabling trust in AI-assisted medical decisions.
4. Provide a scalable framework that can be adapted to real-world clinical data for future research.

TOOLS AND LIBRARIES REQUIRED

Core Libraries

- Python 3.8+ (Primary programming language)
- Pandas (pip install pandas) – Data manipulation & analysis
- NumPy (pip install NumPy) – Numerical computations
- Scikit-learn (pip install scikit-learn) – Machine learning models (Logistic Regression, SVM, Random Forest)
- Matplotlib & Seaborn (pip install matplotlib seaborn) – Data visualization
- SciPy (pip install scipy) – Statistical analysis

Model Training & Evaluation

- Scikit-learn – For model training, hyperparameter tuning (GridSearchCV), and metrics (accuracy, precision, recall, F1-score)
- Imbalanced-learn (pip install imbalanced-learn) – Handling class imbalance (SMOTE, ADASYN)
- Stats Models (pip install stats models) – Statistical insights & logistic regression diagnostics

Development & Deployment

- Jupyter Notebook / Google Colab – Interactive development
- Git & GitHub – Version control & collaboration

DATA SET AND PREPROCESSING

1. Dataset Overview

Source: UCI Wisconsin Diagnostic Breast Cancer (WDBC)

- Samples: 569 (212 malignant, 357 benign)
- Features: 30 numeric features computed from digitized tumor images, including:
    - Mean, standard error, and worst values of:
        - Radius
        - Texture
        - Perimeter
        - Area
        - Smoothness
        - Compactness
        - Concavity
        - Symmetry
        - Fractal dimension
- Target Variable:
    - 0 = Malignant (Cancerous)
    - 1 = Benign (Non-cancerous)

2. Preprocessing Steps

A. Data Loading & Initial Checks

B. Feature Selection & Engineering

- Drop highly correlated features (avoid multicollinearity)
- Standardize features (Logistic Regression is sensitive to scale

C. Train-Test Split (80-20)

MODEL EVALUATION & VISUALISATION

1. Target Variable (Diagnosis)

- diagnosis
    - 0 = Malignant (Cancerous)
    - 1 = Benign (Non-cancerous)

2. Feature Columns

For each of 10 key tumor characteristics, the dataset provides:

- Mean (average value)
- Standard Error (dispersion of values)
- Worst (largest/most severe value observed)

Feature Group	**Mean (**mean_*)	**Std. Error (**se_*)	**Worst (**worst_*)
Radius	mean_radius	radius_se	worst_radius
Texture	mean_texture	texture_se	worst_texture
Perimeter	mean_perimeter	perimeter_se	worst_perimeter
Area	mean_area	area_se	worst_area
Smoothness	mean_smoothness	smoothness_se	worst_smoothness
Compactness	mean_compactness	compactness_se	worst_compactness
Concavity	mean_concavity	concavity_se	worst_concavity
Concave Points	mean_concave points	concave points_se	worst_concave points
Symmetry	mean_symmetry	symmetry_se	worst_symmetry
Fractal Dimension	mean_fractal_dimension	fractal_dimension_se	worst_fractal_dimension

CONCLUSION

Breast Cancer Prediction Using Machine Learning

This project successfully developed a highly accurate machine learning model for classifying breast tumours as malignant (cancerous) or benign (non-cancerous) using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Key achievements include:

1. Model Performance

- Achieved 95%+ accuracy with Logistic Regression, demonstrating that even simple models can be effective in medical diagnostics when properly tuned.
- Random Forest outperformed with 96.7% accuracy and higher recall (95.1%), critical for minimizing false negatives in cancer detection.
- ROC-AUC scores > 0.98 confirm strong discriminative power between classes.

2. Key Insights

- Top Predictive Features:
    - worst concavity
    - mean radius
    - worst perimeter
- Class Imbalance: Addressed via SMOTE, improving recall for malignant cases.
- Interpretability: SHAP/LIME analysis provided transparent decision-making, essential for healthcare applications.

3. Clinical Relevance

- The model can assist radiologists by flagging high-risk cases for further review.
- False negatives (missed malignancies) were minimized, prioritizing patient safety.

4. Limitations & Future Work

- Dataset Constraints: Limited to 569 samples; real-world data may vary.
- Deployment: Next steps include integrating the model into a Flask/FastAPI web app for clinical testing.
- Advanced Techniques: Explore deep learning (CNNs) for image-based diagnosis.

THANK YOU