Machine Learning Analysis Report

Breast Cancer Classification using Logistic Regression

Author: Gourav Karwasara
Date: February 11, 2026
Project: Udacity AI Programming with Python Nanodegree - Machine Learning Foundations

---

Overview

This project develops a binary classification model to predict breast cancer diagnoses (malignant vs. benign) from tumor cell measurements using the Wisconsin Breast Cancer Diagnostic dataset (Wolberg et al., 1993). The model employs Logistic Regression, a supervised learning algorithm well-suited for binary classification tasks with interpretable coefficients (Hastie et al., 2009). The dataset was obtained from the UCI Machine Learning Repository via Kaggle and contains 569 patient records with 30 computed features derived from digitized images of fine needle aspirate (FNA) samples. Through proper preprocessing and evaluation, the model achieves 96.49% accuracy and 99.60% ROC AUC on the test set, demonstrating strong predictive performance for this medical classification task.

Dataset Description

The **Wisconsin Breast Cancer (Diagnostic) Dataset** contains features computed from digitized images of fine needle aspirate (FNA) of breast masses, describing characteristics of cell nuclei present in the images (Wolberg et al., 1993). The dataset includes **569 patient records** (357 benign, 212 malignant) with **32 columns total**: 1 ID field, 1 diagnosis label, and 30 numerical features. Each feature represents measurements computed from the cell nucleus images, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

For each of the 10 base measurements, three variants are provided: mean, standard error (SE), and "worst" (mean of the three largest values), resulting in 30 total features. The target variable is `diagnosis`, encoded as M (Malignant) for cancerous tumors and B (Benign) for non-cancerous tumors. The dataset exhibits class imbalance with approximately 63% benign and 37% malignant cases. According to the UCI repository documentation, the dataset has no missing values and all features are real-valued numerical measurements (UCI Machine Learning Repository, 2024).

Dataset Source: Originally donated to the UCI Machine Learning Repository on October 31, 1995, and made available through Kaggle. The official DOI is 10.24432/C5DW2B.

Modeling Approach

Data Preparation

The data preparation process involved several critical steps to ensure data quality and appropriate format for machine learning. First, an exploratory inspection revealed an empty column (`Unnamed:`

32) containing only NaN values, which was removed as it provided no predictive information. The categorical target variable diagnosis was encoded from categorical labels (M/B) to binary numeric values (1/0), where 1 represents Malignant and 0 represents Benign—this binary encoding is required for Logistic Regression classifiers (Pedregosa et al., 2011).

Exploratory data analysis through histogram visualizations revealed that features like radius, perimeter, area, concavity, and concave points show strong separation between malignant and benign cases, while features like smoothness, symmetry, and fractal dimension exhibit more overlap. These patterns suggested the dataset contains discriminative information suitable for supervised learning.

The dataset was split into training (80%) and testing (20%) sets using stratified sampling to maintain the 63%-37% benign-malignant ratio in both subsets, preventing distribution shift between training and evaluation (Hastie et al., 2009). This resulted in 455 training samples and 114 test samples.

Preprocessing: Feature Scaling

Feature scaling was applied using **StandardScaler**, which transforms each feature to have mean=0 and standard deviation=1 (Pedregosa et al., 2011). This preprocessing step is essential for Logistic Regression because the optimization algorithm (gradient descent) converges faster when features are on similar scales, and regularization (if applied) treats all features equally regardless of their original measurement units (Hastie et al., 2009).

Critically, the scaler was fit **only on the training data** and then applied to both training and test sets—this prevents data leakage that would occur if test data influenced the scaling parameters (Hastie et al., 2009). A scikit-learn Pipeline was used to ensure this preprocessing sequence is applied correctly during both training and inference.

Model Selection: Logistic Regression

Logistic Regression was selected as the classification algorithm for several reasons:

1. **Appropriate for binary classification**: Logistic Regression models the probability of binary outcomes using the logistic (sigmoid) function, which maps linear combinations of features to probabilities between 0 and 1 (Hastie et al., 2009). This makes it particularly suitable for medical diagnostic tasks with two possible outcomes (Dreiseitl & Ohno-Machado, 2002).

2. **Interpretability**: The model produces coefficient weights for each feature, allowing medical practitioners to understand which tumor characteristics most strongly predict malignancy—a critical consideration in healthcare applications where model transparency supports clinical decision-making (Sidey-Gibbons & Sidey-Gibbons, 2019). Research has shown that Logistic Regression's interpretability makes it preferred over black-box models in clinical settings where physicians need to understand and validate predictions (Dreiseitl & Ohno-Machado, 2002).

3. **Well-calibrated probabilities**: Logistic Regression typically produces well-calibrated probability estimates that accurately reflect true outcome probabilities, which is essential for medical contexts where thresholds can be adjusted based on the cost of false negatives versus false positives (Niculescu-Mizil & Caruana, 2005).

4. **Proven effectiveness**: The algorithm has been widely validated on medical classification tasks and is recommended for structured tabular data like this dataset (Sidey-Gibbons & Sidey-Gibbons, 2019).

Model Configuration: The implementation used scikit-learn's `LogisticRegression` with default L2 regularization (C=1.0) and the LBFGS solver with maximum 5,000 iterations to ensure convergence (Pedregosa et al., 2011).

Model Assumptions

Logistic Regression makes several assumptions that were considered: - **Linear relationship**: Assumes a linear relationship between features and the log-odds of the outcome (Hastie et al., 2009) - **Feature independence**: Assumes features are not perfectly collinear, though some correlation exists in this dataset (e.g., radius, perimeter, and area are geometrically related) - **Binary outcome**: Appropriate for this two-class problem (malignant vs. benign)

Evaluation Metrics

Multiple evaluation metrics were selected to comprehensively assess model performance (Hastie et al., 2009):

1. **Accuracy**: Overall proportion of correct predictions—provides a general performance baseline.
2. **ROC AUC (Area Under the Receiver Operating Characteristic Curve)**: Measures the model's ability to discriminate between classes across all possible classification thresholds. ROC AUC is particularly appropriate for medical applications because it evaluates performance independent of any specific threshold, and values close to 1.0 indicate excellent discrimination (Hastie et al., 2009).
3. **Precision and Recall**: Precision measures the proportion of positive predictions that are correct (important for reducing false alarms), while recall measures the proportion of actual positives correctly identified (critical in medical screening to minimize missed diagnoses) (Hastie et al., 2009).
4. **Confusion Matrix**: Provides a detailed breakdown of prediction types (true positives, true negatives, false positives, false negatives), enabling analysis of specific error patterns relevant to clinical decision-making.

These metrics were chosen because they align with the binary classification task and address the medical context where both types of errors (false positives and false negatives) have different clinical consequences.

Results

The Logistic Regression model achieved strong performance on the held-out test set:

Primary Metrics

- **Accuracy**: 96.49% (110 correct out of 114 predictions)
- **ROC AUC**: 99.60%

Detailed Performance by Class

Benign (Class 0): - Precision: 95.95% - Recall: 98.61% - F1-score: 97.26% - Support: 72 samples

Malignant (Class 1): - Precision: 97.50% - Recall: 92.86% - F1-score: 95.12% - Support: 42 samples

Confusion Matrix Analysis

The confusion matrix revealed: - **True Negatives (Benign predicted as Benign)**: 71 - **False Positives (Benign predicted as Malignant)**: 1 - **False Negatives (Malignant predicted as Benign)**: 3 - **True Positives (Malignant predicted as Malignant)**: 39

Feature Importance

Analysis of the model's learned coefficients identified the most predictive features: 1. **smoothness_mean** (coefficient: 1.43) 2. **concavity_se** (coefficient: 1.23) 3. **texture_se** (coefficient: 1.06) 4. **concave points_se** (coefficient: 0.95) 5. **symmetry_worst** (coefficient: 0.91)

Both size-related features (radius, perimeter, area) and texture/shape features (concavity, smoothness) contribute significantly to the classification decision.

Visual Analysis

The ROC curve demonstrates excellent model performance with the curve hugging the top-left corner of the plot, corresponding to the 99.60% AUC score. The confusion matrix heatmap clearly shows the model's strong performance with the vast majority of predictions concentrated in the diagonal (correct predictions).

Results Interpretation and Clinical Context

The machine learning model developed for this project acts as a diagnostic support tool that can help identify whether a breast tumor is likely to be cancerous (malignant) or non-cancerous (benign) based on measurements taken from cell images.

How well does it work?

The model correctly classified 96.5% of cases in our test—meaning out of 114 patients it hadn't seen before, it made the right diagnosis for 110 of them. This is a very strong performance level for medical classification tasks.

What kinds of mistakes does it make?

When the model makes errors, they fall into two categories:

1. **False Alarms (1 case)**: The model predicted cancer when the tumor was actually benign. While this creates unnecessary worry and may lead to additional testing, it's the safer type of error in medical screening.
2. **Missed Diagnoses (3 cases)**: The model predicted benign when the tumor was actually malignant. This is the more concerning type of error because it could delay cancer treatment. Out of 42 actual cancer cases, the model missed 3 (about 7%).

What does this mean in practice?

The model shows excellent discrimination ability—it can reliably tell the difference between cancer and non-cancer cases based on cell measurements. However, it should be used as a **decision support tool** rather than a replacement for clinical judgment. The 3 missed cancer diagnoses remind us that no automated system is perfect, and medical professionals should use this as one input among many when making diagnostic decisions.

How does it make decisions?

The model learned that certain tumor characteristics are strongly associated with cancer, particularly: - Larger cell nuclei (size measurements like radius, perimeter, area) - Irregular surface texture - More concave portions in the cell outline - Higher variation in cell shape

When a tumor shows high values in these measurements, the model assigns a higher probability of malignancy. This aligns with medical knowledge that cancerous cells often have larger, more irregular nuclei than healthy cells.

Limitations and Potential Bias

Model Limitations

1. **Class Imbalance**: The dataset contains more benign cases (63%) than malignant cases (37%), which may cause the model to be slightly biased toward predicting benign outcomes. While stratified sampling mitigates this during training, the imbalance could affect decision threshold optimization (Hastie et al., 2009).
2. **Default Decision Threshold**: The model uses a standard 0.5 probability threshold for classification. In medical contexts, this threshold should be adjusted based on the relative costs of false negatives (missed cancers) versus false positives (unnecessary follow-up procedures). The 3 false negatives observed suggest that lowering the threshold to increase sensitivity (recall for malignant class) may be appropriate for clinical deployment (Dreiseitl & Ohno-Machado, 2002).
3. **No Hyperparameter Tuning**: The model uses default Logistic Regression hyperparameters without systematic optimization through cross-validation or grid search. While performance is strong, tuning regularization strength and other parameters could potentially improve generalization.
4. **Limited Feature Engineering**: The model uses the raw features provided in the dataset without domain-specific feature engineering or polynomial interactions that might capture nonlinear relationships between measurements.
5. **Single Model Architecture**: Only Logistic Regression was evaluated. Comparison with ensemble methods (Random Forest, Gradient Boosting) or support vector machines might reveal whether more complex models offer performance gains.

Potential Sources of Bias

1. **Geographic and Demographic Bias**: The Wisconsin Breast Cancer dataset was collected from a single medical institution in Wisconsin, USA, in the early 1990s. The population demographics, imaging equipment, and medical practices from that specific time and location

may not represent current global populations. Model performance could differ when applied to:

- Different demographic groups (age, ethnicity, genetic backgrounds)
- Different geographic regions with varying cancer incidence patterns
- More recent imaging technology with different measurement characteristics
- Different healthcare settings with varying diagnostic protocols

2. **Sample Size for Malignant Cases**: With only 212 malignant cases in the full dataset (and 42 in the test set), the model's performance on rare cancer subtypes or edge cases may not be fully characterized. Medical datasets often struggle with limited positive cases due to the relative rarity of cancer (Sidey-Gibbons & Sidey-Gibbons, 2019).

3. **Measurement Bias**: The features are computed from FNA (fine needle aspirate) images, which depend on:

   - Image quality and standardization
   - Operator technique during image acquisition
   - Consistency in feature computation algorithms

   Variations in any of these factors could affect model reliability in real-world deployment.

4. **Outcome Label Quality**: The Malignant/Benign labels depend on the accuracy of the original diagnoses, which themselves may have been subject to diagnostic errors or ambiguous cases.

Mitigation Strategies

To address these limitations and biases, the following approaches are recommended:

1. **Threshold Optimization**: Conduct a cost-benefit analysis to determine the optimal classification threshold that minimizes the overall cost to patients, likely emphasizing higher sensitivity (recall) to reduce false negatives.

2. **External Validation**: Test the model on independent datasets from different institutions, time periods, and geographic regions to assess generalization (Sidey-Gibbons & Sidey-Gibbons, 2019).

3. **Diverse Training Data**: Collect and incorporate data from more diverse patient populations to improve representativeness.

4. **Ensemble Approach**: Combine multiple model architectures and use techniques like cross-validation to improve robustness.

5. **Clinical Integration**: Deploy the model as a decision support tool that augments (rather than replaces) expert clinical judgment, with clear communication about its limitations.

6. **Continuous Monitoring**: If deployed in clinical settings, continuously monitor performance on new cases and retrain periodically to account for population drift or changes in imaging technology.

7. **Explainability**: Provide feature importance visualizations and case-specific explanations to help clinicians understand individual predictions and identify potential errors.

References

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics, 35*(5-6), 352–359. https://doi.org/10.1016/S1532-0464(03)00034-0

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer. Freely available at: https://hastie.su.domains/ElemStatLearn/

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning (ICML '05)*, 625–632. https://doi.org/10.1145/1102351.1102430

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. Available at: https://jmlr.org/papers/v12/pedregosa11a.html

Scikit-learn Developers. (2024). *Logistic Regression*. scikit-learn Documentation. https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

Sidey-Gibbons, J. A. M., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: A practical introduction. *BMC Medical Research Methodology, 19*(1), Article 64. https://doi.org/10.1186/s12874-019-0681-4

UCI Machine Learning Repository. (2024). *Breast Cancer Wisconsin (Diagnostic) Dataset*. University of California, Irvine, School of Information and Computer Sciences. https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). *Breast Cancer Wisconsin (Diagnostic)* [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B

End of Report