Assignment 3

Name: Mohamed-Obay Alshaer
Student Number: 300170489

Submission Date: March 21, 2025

Professor: Daniel Shpairo

SEG4300 - Applied Machine Learning

1 Wisconsin Breast Cancer Dataset Analysis

1.1 Dataset Selection

The Wisconsin Breast Cancer Dataset from Hugging Face (scikit-learn/breast-cancer-wisconsin) was selected for this binary classification task. The dataset contains 569 samples with 30 features derived from digitized images of fine needle aspirates of breast masses. Features represent characteristics of cell nuclei present in the images, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension (each with mean, standard error, and worst values). The target variable is 'diagnosis', with values 'M' (malignant) or 'B' (benign).

This dataset was chosen for its alignment with binary classification tasks, structured numerical format suitable for Scikit-learn models, manageable size (569 samples, 30 features), comprehensive documentation, and real-world medical relevance.

1.2 Model Selection and Training

Logistic Regression was selected as the classification model for this dataset based on:

- **Appropriateness:** Specifically designed for binary classification problems
- **Interpretability:** Provides feature coefficients, crucial in medical contexts
- **Efficiency:** Computationally efficient with less risk of overfitting
- **Performance:** Effective for linearly separable data
- **Baseline:** Establishes minimum performance expectations

1.3 Results Summary

The Logistic Regression model achieved impressive performance metrics:

- **Accuracy:** 96% correct classifications
- **Precision:** 95% of predicted malignant tumors were actually malignant
- **Recall:** 95% of all malignant tumors were correctly identified
- **F1 Score:** 95%, confirming balanced performance
- **AUC-ROC:** 0.99, indicating excellent discriminative ability

1.4 Feature Importance

Analysis of model coefficients revealed:

- Most influential positive predictors: concave points_worst, perimeter_worst, and concave points_mean
- Some features showed negative associations, potentially indicating benign characteristics
- Cell nucleus irregularity appears strongly associated with malignancy

1.5 Clinical Implications and Limitations

Clinical Insights:

- Concavity and perimeter measurements are key differentiators for tumor classification
- High recall minimizes missed malignant tumors, critical in cancer screening
- Model should support, not replace, clinical judgment

Limitations:

- Relatively small dataset (569 samples) may limit generalizability
- More complex models could potentially capture additional relationships
- Feature selection might improve model interpretability
- Cross-validation would provide more robust performance estimates

1.6 Conclusion

Logistic Regression proved to be an effective and interpretable model for breast cancer classification, delivering high performance while providing insights into feature importance. The model's strong recall is particularly valuable in medical contexts where missing malignant cases can have serious consequences. Future work could explore more complex models and feature selection techniques to potentially enhance performance and interpretability.