PROJECT REFLECTION REPORT: BREAST CANCER CLASSIFICATION

INTRODUCTION

This project aimed to classify breast cancer tumors as malignant or benign using the Breast Cancer Wisconsin (Diagnostic) Dataset. The primary goal was to develop, fine-tune, and compare various machine learning models to identify the most effective classifier for this critical diagnostic task. Given breast cancer's prevalence, accurate and reliable classification is paramount.

METHODOLOGY AND KEY STEPS

The project followed a standard machine learning pipeline:

1. Data Importation and Initial Inspection: The dataset was loaded, and an initial review identified and removed irrelevant 'id' and 'Unnamed: 32' columns. The 'diagnosis' column was mapped from categorical ('M', 'B') to numerical (1, 0) for model compatibility.

2. Data Cleaning and Preparation: Descriptive statistics provided early insights. Exploratory Data Analysis (EDA) included a pie chart highlighting initial class imbalance, histograms showing feature distributions, and box plots revealing numerous outliers. Pair-plots and a correlation heatmap demonstrated significant multicollinearity among features. Outlier Removal, Outliers were addressed using a quantile-based method, reducing the dataset size.

3. Data Preprocessing: This phase was crucial for model readiness:
    - Feature Selection: To mitigate multicollinearity, highly correlated features (correlation >0.92) were dropped, resulting in a refined set of 22 features.

- Handling Class Imbalance: The SMOTE technique was applied to balance the 'diagnosis' classes, increasing the dataset to 674 samples (337 per class), preventing biased model training.
- Data Splitting and Scaling: The data was split into 80% training and 20% testing sets. `StandardScaler` was then applied to standardize features, ensuring equitable contribution during model learning.

4. Model Building and Evaluation: Four classification models were developed and evaluated:

• Logistic Regression: Achieved an accuracy of 0.9852

• Random Forest Classifier: Recorded an accuracy of 0.9556

• XGBoost Classifier: Performed strongly with an accuracy of 0.9704.

• CatBoost Classifier: Also achieved an accuracy of 0.9704

Confusion matrices provided detailed performance insights. Random Forest feature importances highlighted `concave points_worst`, `area_worst`, and `concave points_mean` as the most influential predictors.

PROJECT OUTCOMES AND CONCLUSION

This project successfully navigated an end-to-end machine learning pipeline, emphasizing the importance of thorough data preparation, imbalance handling, and feature engineering. While Logistic Regression showed the highest accuracy, the XGBoost model was chosen for predictions due to its ability to capture complex patterns, which is critical in sensitive diagnostic applications where misclassifications carry significant weight. This underscores that model selection should consider not just metrics, but also the practical implications and robustness

required for the specific domain.

LIMITATIONS

Dataset Specificity: The model's performance is currently validated only on the Wisconsin Diagnostic Breast Cancer dataset. Its generalizability to other datasets or real-world clinical data may vary.

Outlier Handling: While beneficial for some models, the quantile-based outlier removal might have inadvertently discarded rare but important data points.

Hyperparameter Optimization: Although multiple models were tested, a more extensive and systematic hyperparameter tuning process (e.g., using GridSearchCV or RandomizedSearchCV) was not thoroughly performed for each model.

Interpretability: While XGBoost offers high predictive power, its black-box nature can make understanding individual predictions challenging, which is a significant consideration in medical diagnoses.

RECOMMENDATIONS

External Validation: To ensure robustness, the trained models should be validated against independent, external datasets of breast cancer diagnoses.

Advanced Feature Engineering: Explore creating more sophisticated features or applying dimensionality reduction techniques beyond correlation-based selection.

Explainable AI (XAI): Incorporate XAI techniques (e.g., SHAP, LIME) to enhance the interpretability of the XGBoost model, providing clinicians with insights into why a particular prediction was made.

Cost-Sensitive Learning: Given the critical nature of medical diagnoses, consider implementing cost-sensitive learning to differentially penalize false negatives (missing a malignant case) more heavily than false positives.

Investigate other Ensemble Methods: Explore advanced ensemble techniques or neural networks, if justified by data complexity and available computational resources.