Project Report

- Project Title: An Interpretable Machine Learning Approach for Breast Cancer Diagnosis

Team Details

Team Name: Model Mavericks

Team Members:

Kabeer, 24BCS10119 (Team Leader)

Sanskriti, 24BCS10247

Shrival Kumar, 24BCS10254

Tejas Kumat, 24BCS10299

Nishant Bhaleem, 24BCS10314

- Problem Statement

The early and accurate diagnosis of breast cancer is critical for improving patient survival rates. While machine learning models can predict whether a tumor is malignant or benign with high accuracy, their "black box" nature is a major barrier to clinical adoption. Doctors need to trust and understand why a model makes a specific recommendation. This project tackles this challenge by not only creating an accurate diagnostic model but also making its decision-making process transparent and explainable. The goal is to develop a trustworthy AI tool that can serve as a reliable aid to medical professionals.

- Data Source and Acquisition

  • Source: We will use the Breast Cancer Wisconsin (Diagnostic) Dataset, which is publicly available on Kaggle. This is a clean and well-structured version of the original dataset curated by the University of Wisconsin, ensuring its quality and relevance with more updated data having 5015 rows and 32 columns.

• Link: https://www.kaggle.com/datasets/shantanugarg274/breast-cancer-prediction-dataset?select=breast_cancer_updated.csv

- Initial Approach

Our initial approach was to begin with a baseline understanding of the dataset and apply simple machine learning models before moving on to more complex ones. Since the Breast Cancer Wisconsin dataset contains labeled data for tumor classification (benign vs. malignant), we treated this as a supervised binary classification problem.

The first steps of our approach included:

1. **Data Exploration**
   - Checked the dataset for missing values, duplicate rows, and inconsistent data.
   - Examined the distribution of the target variable to identify potential class imbalance between malignant and benign cases.
2. **Baseline Model Selection**

- We chose **Logistic Regression** as the first model, since it is simple, interpretable, and provides a good baseline for binary classification tasks.
- We did not initially apply any complex hyperparameter tuning or feature selection but instead trained the model with default parameters to get a benchmark performance.

3. **Evaluation Metrics**
    - From the beginning, we emphasized **recall** as a key metric, since misclassifying a malignant tumor as benign has more serious consequences than the reverse.
    - Along with recall, we tracked accuracy, precision, F1-score, and ROC-AUC to get a holistic view of model performance.
4. **Planned Next Steps**
    - If Logistic Regression performance was insufficient, the plan was to explore more powerful algorithms such as **Random Forest and XGBoost.**
    - After training predictive models, we also intended to integrate **explainability techniques** (e.g., feature importance, SHAP values) to make the model's decision-making transparent and trustworthy for clinical use.

Methodology and Solution Approach

Our methodology followed a structured machine learning workflow to achieve both **high predictive performance** and **interpretability**.

1. Problem Framing

The task was framed as a **binary classification problem**, predicting whether a tumor is **malignant (1)** or **benign (0)** based on diagnostic features. Recall was prioritized, as failing to detect malignant tumors has severe real-world consequences.

2. Data Preparation

- **Scaling**: StandardScaler was already applied to normalize features, which is essential for Logistic Regression.
- **Missing Values**: None were present.
- **Multicollinearity**: Variance Inflation Factor (VIF) analysis revealed that some features had extremely high VIF values (approaching infinity), indicating strong multicollinearity. However,

these features were **retained** because removing them led to a **significant drop in recall**, and multicollinearity in Logistic Regression primarily affects **coefficient stability** rather than predictive performance.

Model Training and Evaluation

1. Logistic Regression

- **Initial Training**: Trained with default parameters, yielding a recall of **0.94**.
- **Hyperparameter Tuning**: Adjusting regularization parameters slightly improved overall accuracy but **reduced recall to 0.93**, demonstrating the trade-off between optimizing for accuracy vs. recall in a medical dataset.
- **Threshold Tuning**: To prioritize recall, the decision threshold was lowered from 0.5 to **0.4**, which increased recall to **0.95** without significant loss in precision or overall accuracy.

Reasoning: In medical diagnostics, **false negatives are more critical than false positives**, so threshold adjustment was an effective strategy to align the model with domain-specific objectives.

- ### Evaluation on Test Data

- **Accuracy:** 0.9292
- **Precision:** 0.9492
- **Recall:** 0.9387
- **F1-score:** 0.9439
- **ROC-AUC:** 0.9724

2. Random Forest Classifier

- ### Motivation

While Logistic Regression offered interpretability, Random Forest was explored to capture **non-linear feature interactions** and leverage ensemble learning for stronger performance.

- ### Model Training and Hyperparameter Tuning

- **Initial Training**: With default parameters, Random Forest achieved a recall of **0.98**, indicating very high sensitivity to malignant tumors.
- **Hyperparameter Tuning**:

  - Tuned parameters such as `n_estimators`, `max_depth`, `min_samples_split`, and `max_features`.
  - Used `class_weight="balanced"` to address mild class imbalance.
  - After tuning, recall **remained at 0.98**, but precision and overall accuracy improved, showing a better balance between false positives and false negatives.
- **Threshold Adjustment**: Reducing the classification threshold from **0.5 to 0.4** pushed recall further to **0.99**, ensuring almost all malignant cases were identified, though at the cost of slightly lower precision.

- **Evaluation on Test Data**

- **Accuracy:** 0.9651
- **Precision:** 0.9612
- **Recall:** 0.9841
- **F1-score:** 0.9725
- **ROC-AUC:** 0.9906

- **Feature Importance & Explainability**

- **Random Forest Feature Importance (Impurity-Based):**
  Key predictors were *compactness_mean (0.274)*, *symmetry_mean (0.146)*, *compactness_worst (0.070)*, and *perimeter-related features*.
- **Permutation Importance (Model-Agnostic):**
  Again, *compactness_mean* and *compactness_worst* emerged as the strongest
- predictors, confirming their robustness, though scores were smaller as this method directly measures the drop in accuracy when features are shuffled.
- **Why Both?**
  Impurity-based importance can be biased, while permutation provides a fairer view. Their agreement strengthens confidence in the identified features.

3. XGBoost

- **Motivation**

While Random Forest gave excellent performance, we explored **XGBoost (Extreme Gradient Boosting)** because:

- It is one of the most **powerful boosting algorithms**, known for **state-of-the-art accuracy** in structured/tabular data.
- Unlike Random Forest (bagging, parallel trees), XGBoost builds trees **sequentially** where each new tree **fixes the mistakes** of the previous ones.

- It also includes **regularization** (to prevent overfitting) and is highly optimized for speed and efficiency.

❖ Thus, it was a natural next step to test whether boosting can further improve recall and balanced performance.

- **Model Training and Hyperparameter Tuning**

→ **Initial Training:**

- With default parameters, XGBoost achieved high performance, with strong recall and ROC-AUC.

→ **Hyperparameter Tuning (via RandomizedSearchCV):**

- We tuned the following: `n_estimators` → number of boosting rounds (100–300).`learning_rate` → step size shrinkage (0.01–0.1).`max_depth` → maximum tree depth (3–7).`subsample` → fraction of samples used per tree (0.7–1.0).`colsample_bytree` → fraction of features used per tree (0.7–1.0).gamma → minimum loss reduction to split a node.
- Best parameters were selected based on **ROC-AUC** using 5-fold cross-validation.

→ **After Tuning:**

- Performance improved, particularly in terms of **balanced recall and precision**.
- Recall remained very high (critical for cancer detection), while precision improved compared to default settings.

Evaluation on Test Data

- **Accuracy:** 0.9691
- **Precision:** 0.9702
- **Recall:** 0.9810
- **F1-score:** 0.9755
- **ROC-AUC:** 0.9924

Feature Importance & Explainability

- **Built-in XGBoost feature importance** was used to rank top predictors.
- Important features overlapped with earlier models (e.g., compactness_mean, symmetry_mean, compactness_worst).
- This consistency across Logistic Regression, Random Forest, and XGBoost further confirms the **robustness of these features** for breast cancer classification.

Why XGBoost is Valuable?

- Random Forest → parallel bagging trees, reduces variance.
- XGBoost → sequential boosting trees, reduces bias.
- In our results, **XGBoost matched Random Forest in recall (~0.98–0.99)** but also gave slightly **better precision and F1-score**, offering a stronger trade-off between false positives and false negatives.

SHAP Analysis

1. **Overall Feature Importance (Bar Plot)**
   a. The SHAP bar plot shows that **`compactness_mean`** has the highest average contribution to the model's predictions, followed by **`symmetry_mean`, `compactness_worst`, and `perimeter_se`**.
   b. These features have the largest mean absolute SHAP values, meaning they consistently impact the model's decision towards malignant classification.
   c. In contrast, many features have relatively small contributions, suggesting that the model is primarily relying on a subset of key predictors.

2. **Feature Impact and Direction (Beeswarm Plot)**
   a. The beeswarm plot provides both **magnitude** and **direction** of feature influence.
   b. For example:
      i. **Low values of `compactness_mean`** (shown in red on the left side) push predictions strongly towards *malignant*.
      ii. **Low values of `symmetry_mean`** tend to push predictions towards malignant while higher values push towards benign.
      iii. Similarly, **`perimeter_se` and compactness_worst** at lower values increase the likelihood of malignancy.
   c. The color gradient (blue = low value, red = high value) makes it clear how different ranges of a feature affect prediction.

3. **Instance-Level Explanation (Waterfall Plot)**
   a. The waterfall plot illustrates the contribution of features for a **single prediction (an individual patient)**.
   b. Starting from the base value (average model output), features like **`compactness_worst` and `symmetry_mean`** increase the prediction probability toward malignant, while features like **`area_mean` and `smoothness_mean`** reduce it.

c. This breakdown shows *why* the model classified this specific tumor the way it did, which is critical for clinical interpretability.

Model Performance Analysis

The performance of all the three models was evaluated using recall, precision, F1-score, ROC-AUC, and Precision-Recall AUC, with particular emphasis on **recall**, as failing to detect malignant cases is clinically unacceptable compared to raising false alarms.

Logistic Regression:

- **Recall:** Initially **0.94**, showing strong sensitivity. After hyperparameter tuning, it dropped to **0.93**, suggesting better generalization but slightly reduced sensitivity. By lowering the threshold to **0.4**, recall improved to **0.95**, confirming the effectiveness of threshold adjustment in optimizing sensitivity.
- **Precision & F1-score:** Precision and f1-score remained same after threshold adjustment.
- **ROC-AUC:** The ROC-AUC was high (>0.97), showing excellent class separability.
- **Precision-Recall AUC:** Also, high (>0.95), further emphasizing that the model performed well in handling the class imbalance and focusing on malignant cases.

Random Forest:

- **Recall:** Random Forest achieved **0.98 recall** both before and after hyperparameter tuning, demonstrating robust performance and insensitivity to tuning. At a reduced threshold of **0.4**, recall further increased to **0.99**, nearly perfect detection of malignant tumors.
- **Precision & F1-score:** As with Logistic Regression, threshold reduction slightly decreased precision but increased recall, resulting in a balanced F1-score.
- **ROC-AUC:** The ROC-AUC was extremely high (>0.99), indicating that Random Forest separated malignant and benign cases almost perfectly.
- **Precision-Recall AUC:** Also exceeded 0.99, demonstrating superior performance in detecting malignant cases compared to Logistic Regression.

XGBoost:

- **Recall: XGBoost** achieved **0.98 recall** both before and after hyperparameter tuning, demonstrating robust performance and insensitivity to tuning.
- **Precision & F1-score:** As with Logistic Regression, threshold reduction slightly decreased precision but increased recall, resulting in a balanced F1-score.
- **ROC-AUC:** The ROC-AUC was extremely high (>0.99), indicating that XGBoost separated malignant and benign cases almost perfectly, and also slightly better than Random Forest.

- **Precision-Recall AUC:** Also exceeded 0.99, demonstrating superior performance in detecting malignant cases compared to Logistic Regression and slightly better than Random Forest.

Feature Interpretability:

- Logistic Regression coefficients indicated feature contributions but were complicated by multicollinearity (infinite VIF values), which made individual coefficient interpretation less reliable.
- Random Forest provided robust feature importance rankings, with **compactness_mean, symmetry_mean, and perimeter measures** consistently emerging as the most influential predictors. SHAP analysis confirmed that low compactness_mean and high values of concave points_worst strongly pushed predictions toward malignancy.
- Unlike Logistic Regression, XGBoost naturally handled multicollinearity better, and compared to Random Forest, it provided slightly sharper separation of malignant vs. benign cases, with SHAP values capturing both linear and non-linear feature interactions effectively.

	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.929212	0.938776	0.949206	0.943962	0.972411
Random Forest	0.965105	0.961240	0.984127	0.972549	0.990695
XGBoost	0.969093	0.970173	0.980952	0.975533	0.992361

- Assessment

  The primary objective of this project was to build a predictive system capable of accurately distinguishing between **malignant and benign tumors**, while also ensuring that the model's decisions are **transparent and explainable** for clinical adoption.

  - **Random Forest** achieved outstanding predictive performance with a recall of **0.99** and an AUC close to **1.0**, meaning it was able to detect nearly all malignant cases. This fulfills the clinical requirement of minimizing false negatives, where missing a cancer diagnosis could have serious consequences.

- **Logistic Regression**, while slightly less powerful (recall ~0.96), provided greater interpretability through coefficients and SHAP analysis, making it easier to communicate the reasoning behind predictions to clinicians.
- **XGBoost** slightly outperformed the Random Forest, achieving an **Accuracy of 0.9691, Recall of 0.9810, F1-score of 0.9755, and ROC-AUC of 0.9924**. This indicates that XGBoost not only matches but even surpasses Random Forest in terms of predictive performance, while being more computationally efficient and better at handling complex feature interactions.
- The use of **SHAP explainability** enhanced trust in both models by highlighting feature contributions (e.g., compactness_mean, symmetry_mean, perimeter-related features). This bridges the gap between high-performing machine learning models and the need for medical transparency.

In conclusion, the models **successfully addressed the problem**:

1. They achieved **high sensitivity (recall)**, critical for early cancer detection.
2. They offered **interpretability**, enabling doctors to understand the "why" behind predictions.
3. They demonstrated robustness through validation and hyperparameter tuning.

- Overall, the combination of **Random Forest and XGBoost** successfully addresses the problem. Random Forest offers interpretability and robustness, while XGBoost provides state-of-the-art accuracy and efficiency. Logistic Regression, while slightly behind in performance, remains a reliable interpretable baseline model. Together, these models demonstrate that machine learning can be a powerful tool in early breast cancer diagnosis, with XGBoost emerging as the most promising candidate for deployment in real-world scenarios.

Recommendations and Next Steps

- **Feature Refinement:** Apply advanced feature selection or dimensionality reduction to address multicollinearity and redundancy.
- **Class Imbalance Handling:** Compare SMOTE/ADASYN and cost-sensitive learning to further stabilize recall.
- **Model Generalization:** Validate on external datasets and diverse demographics to ensure robustness.
- **Explainability:** Expand SHAP analysis into clinician-friendly dashboards for better adoption.