Early Detection of Breast Cancer

CSE 572: Data Mining
Dr. Yanjie Fu

Submitted by:
Aditya Deshpande
1233720607

Background & Problem Statement

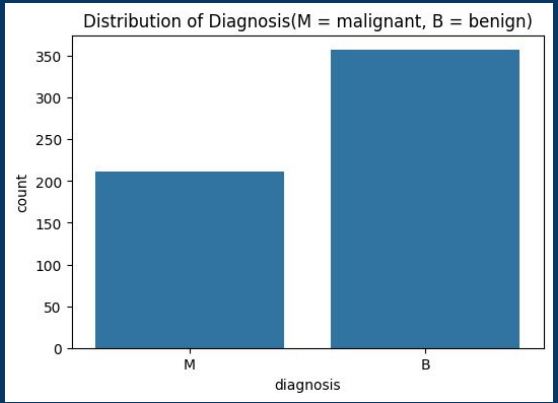
Breast cancer is a leading cause of cancer-related deaths, making early detection critical for survival rates.
However, accurate and timely diagnosis remains a challenge due to the complexity and variability of tumors.

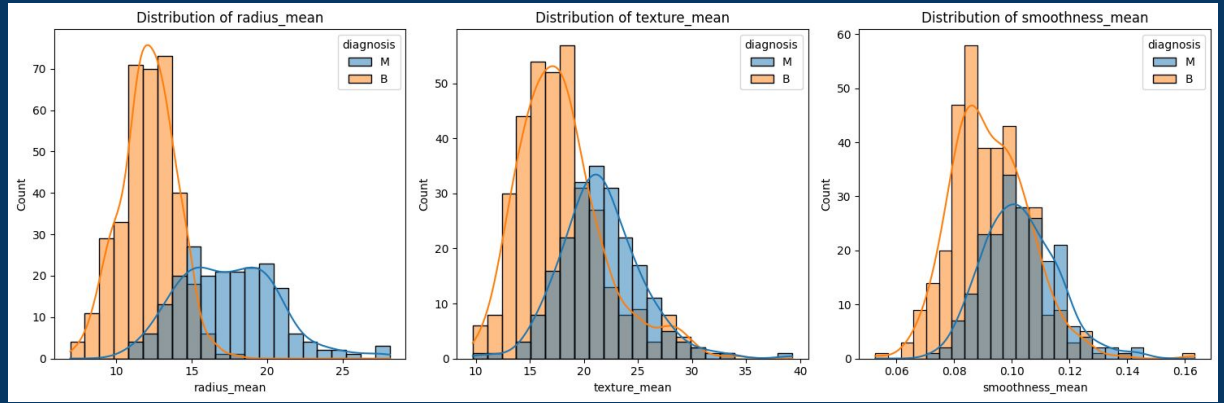
Our Project Goal:
To develop a highly accurate and interpretable machine learning model for early breast cancer diagnosis

Our Dataset: Wisconsin Diagnostic Breast Cancer (WDBC)

- **Origin:**
  - We are using the WDBC dataset from the UCI Machine Learning Repository.
- **Features:**
  - It contains 30 numerical features computed from digitized images of breast masses, such as radius, texture, and concavity.
- **Task:**
  - This data allows us to frame the problem as a binary classification task: predicting if a tumor is benign or malignant.

Initial Data Exploration: What We've Found So Far

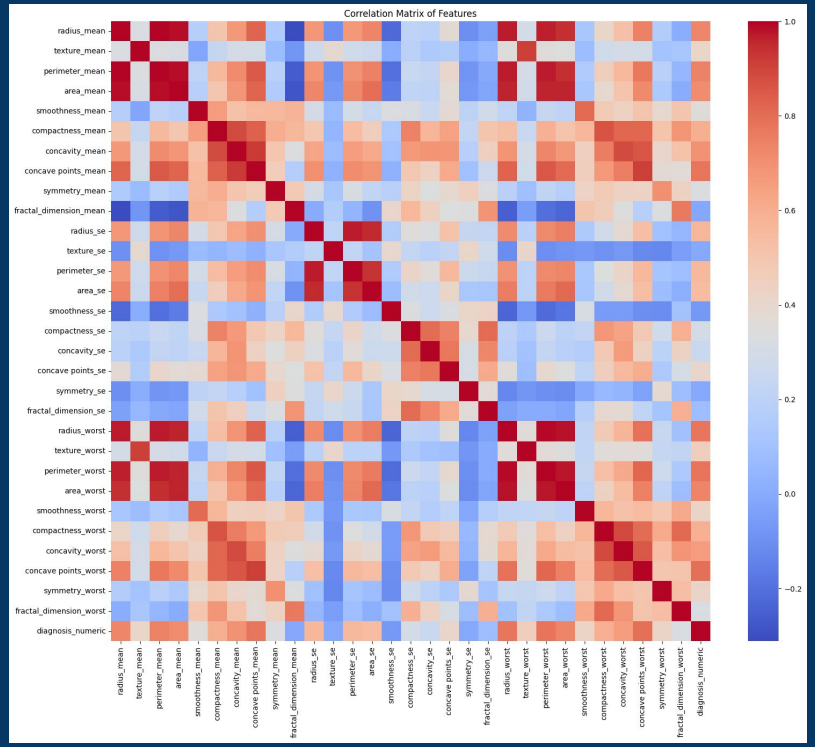




The class balance (e.g., "Our initial look shows the dataset is reasonably balanced, with X benign and Y malignant samples.").

Point out a key finding (e.g., "As shown, features like 'radius_mean' tend to have higher values for malignant tumors, confirming their predictive potential").

Initial Data Exploration: Identifying Relationships



Our analysis revealed strong multicollinearity between features like radius, perimeter, and area. This is an important data challenge we will need to manage in our modeling phase.

Planned Data Mining Pipeline

- **Data Preprocessing:**
  - We will handle feature scaling using standardization to prevent model bias. The data will be split into training and testing sets.
- **Baseline Modeling:**
  - We will start by implementing and evaluating classic models like Logistic Regression and SVM.
- **Advanced Modeling:**
  - We plan to explore more complex methods, such as Random Forest and potentially a stacked ensemble model, to improve performance.
- **Evaluation:**
  - Success will be measured not just by accuracy but critically by precision and recall to minimize false negatives.

Foreseen Challenges

Modeling Challenge

The primary challenge is minimizing false negatives. A misdiagnosis could delay critical treatment, so a high recall score is paramount.

Data Challenge

As seen in our EDA, high feature correlation needs to be addressed to ensure our model is robust and interpretable.

Interpretability Challenge

For a model to be trusted in a clinical setting, its predictions must be explainable. We plan to use SHAP analysis to make the model's decisions transparent.

Project Plan

Complete data preprocessing pipeline.

Implement and evaluate at least two baseline models.

Present initial performance results and analyze shortcomings.

Implement advanced models (Random Forest, Gradient Boosting).

Perform hyperparameter tuning to optimize results.

Compare all models and select the best candidate for the final report.

Finalize the best model and generate an interpretability report using SHAP.

Deliver a comprehensive write-up of the project.

Summary & Expected Outcomes

This project will deliver a complete data mining pipeline for breast cancer classification. We will compare multiple models to find the most accurate and reliable approach. Our final deliverable will be a high-performing, interpretable model that could serve as a valuable tool for clinicians.

THANK YOU !!