# Early Prediction of Breast Cancer Malignancy Using Machine Learning

This repository contains code, data utilities, and a Streamlit app for a medical prediction project that classifies **breast cancer tumors** as *malignant* or *benign* using diagnostic features from the **Breast Cancer Wisconsin (Diagnostic)** dataset (available via scikit-learn).

## Project Structure

medical\_prediction\_project/

├── app/

│ └── streamlit\_app.py

├── data/

│ ├── raw/

│ │ └── breast\_cancer.csv

│ └── processed/

├── models/

│ └── (saved models here)

├── reports/

│ ├── metrics.json

│ └── report.md

├── src/

│ ├── data\_prep.py

│ ├── train.py

│ ├── evaluate.py

│ ├── inference.py

│ └── utils.py

├── requirements.txt

└── README.md

## Quickstart

1. **Create and activate a virtual environment (recommended):**

python -m venv .venv

# Windows:

.venv\Scripts\activate

# macOS/Linux:

source .venv/bin/activate

1. **Install dependencies:**

pip install -r requirements.txt

1. **Train models and evaluate:**

python src/train.py

This will:

* + Load the dataset from scikit-learn and export to data/raw/breast\_cancer.csv
  + Train Logistic Regression, Random Forest, and Gradient Boosting pipelines with cross-validation
  + Save the **best model** to models/best\_model.joblib
  + Write evaluation metrics to reports/metrics.json and save plots (ROC, PR, confusion matrix) to reports/

1. **Run the Streamlit app (optional deployment demo):**

streamlit run app/streamlit\_app.py

## Inference

After training, you can run predictions via:

python src/inference.py --sample

or pass a JSON file with feature values:

python src/inference.py --json sample.json

## Notes

* The dataset is bundled by exporting the sklearn.datasets.load\_breast\_cancer to data/raw/breast\_cancer.csv during training for transparency/reproducibility.
* The best model and preprocessing are stored in a single sklearn pipeline for safe inference (prevents data leakage).