Project Documentation

Project Title

Breast Cancer Detection Using Random Forest: A Machine Learning Approach to Tumor Malignancy Classification

Abstract

This project focuses on the early and accurate diagnosis of breast cancer using machine learning, specifically a Random Forest Classifier. Leveraging a labeled dataset of tumor characteristics, the model was trained to distinguish between benign and malignant tumors. Data preprocessing, feature exploration, and visualization were conducted to understand key predictors. The model was evaluated using classification metrics such as precision, recall, and F1-score, showing high reliability and potential for aiding clinical decision-making. This tool can help in resource-constrained healthcare environments for early detection and screening.

Objectives

- Develop a model to classify breast tumors
- Explore feature relationships through EDA
- Evaluate classification performance
- Create an accessible diagnostic tool

Dataset Overview

- Source: UCI ML Repository
- Features: 30 numerical (e.g., radius, texture)
- Target: Diagnosis (Malignant = 1, Benign = 0)

Methodology

Project Documentation

1. Data Preprocessing: Removed irrelevant columns, checked for missing values, encoded labels.

2. EDA: Visualized class distribution, correlation matrix, and scatter plots.

3. Model Development: Used Random Forest, trained and tested with splits.

4. Evaluation: Accuracy, Precision, Recall, F1-score, and feature importance.

Results

- High accuracy and precision achieved

- Top features: radius_mean, texture_mean, concavity_mean

- Clear separation between classes in visual plots

Tools & Technologies

- Python, Jupyter Notebook

- Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn

Pain Points Addressed

- Delayed/inaccurate diagnosis

- Difficulty interpreting raw data

- Limited diagnostic access

- Need for reliable ML models

Conclusion

Random Forest was effective for breast cancer detection, showing strong predictive power and interpretability. It is reliable and accessible, making it suitable for supporting clinical decisions.

Project Documentation

Future Work

- Build a web app interface (e.g., Streamlit)

- Add more clinical variables

- Compare with other ML models or deep learning