Breast cancer prediction using machine learning

SUML Project by Busra Gultekin,
Franciszek Waleruk

Table of Contents

Introduction

Project goal

The overarching goal of our project is to develop a sophisticated predictive model capable of accurately identifying the likelihood of breast cancer in individuals, with the primary objectives of enabling early detection for timely intervention, enhancing accuracy and precision in diagnosis by leveraging machine learning algorithms to discern subtle patterns in data, optimizing healthcare resources through targeted screening strategies based on personalized risk assessments, handling large datasets efficiently to extract meaningful insights, complementing traditional diagnostic methods to improve overall detection reliability, contributing to public health initiatives by facilitating early intervention programs and raising awareness, and fostering ongoing research and innovation in healthcare to continually advance models and techniques for more effective breast cancer prediction and diagnosis.

Selected technology

Pandas

Pandas is a powerful data manipulation and analysis library for Python. It provides data structures like DataFrames and Series, making it easy to clean, preprocess, and analyze datasets. In your project, Pandas would be crucial for loading and manipulating the breast cancer dataset, handling missing values, and preparing the data for machine learning models.

NumPy

NumPy is a fundamental library for numerical operations in Python. It supports large, multi-dimensional arrays and matrices, along with mathematical functions to operate on these arrays efficiently. In your project, NumPy would be used for numerical operations and handling the underlying data structures used by Pandas.

Matplotlib

Matplotlib is a comprehensive plotting library for creating static, animated, and interactive visualizations in Python. In the context of your project, Matplotlib would be essential for visualizing data distributions, relationships, and model performance metrics. It provides a wide range of customization options for creating informative plots.

Plotly

Plotly is an interactive plotting library that allows you to create interactive and dynamic visualizations. It can be used for building interactive charts and dashboards, enhancing the user interface of your project. In the context of breast cancer prediction, Plotly can be beneficial for creating visually engaging representations of data and model results, providing a more interactive experience for users or stakeholders.

Method

ML model parameters

In our project we used the Random Forest Classifier which is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

The parameters we used were as follows:

n_estimators:

N_estimators represents the number of decision trees to be created in the forest. Increasing the number of trees generally improves the performance of the Random Forest, but it comes at the cost of increased computational complexity. It helps in achieving a more robust and stable model by reducing overfitting and providing a better representation of the underlying patterns in the data.

criterion:

Criterion determines the function to measure the quality of a split in the decision tree. Two commonly used criteria are:

- gini: Gini impurity, a measure of how often a randomly chosen element would be incorrectly classified.
- entropy: Information gain, a measure of the reduction in uncertainty.

The criterion guides the decision tree's splitting process. Both options are effective, and the choice between them may depend on the specific characteristics of the dataset.

oob_score:

Out-of-Bag (OOB) score is a metric that provides an estimate of the model's performance on unseen data without the need for a separate validation set. During the training of each tree in the forest, some data points are not used (out-of-bag samples). The OOB score calculates the accuracy of the model on these out-of-bag samples. It serves as a built-in validation measure, helping to assess the performance of the Random Forest without the need for an additional validation set.

Description of functionality

The project is designed to utilize a comprehensive workflow for the accurate identification of breast cancer likelihood. The project begins with the collection and preparation of relevant datasets, involving data preprocessing to handle missing values, normalize features, and encode categorical variables. An essential step includes Exploratory Data Analysis (EDA) to gain insights into data characteristics and relationships between features.

Machine Learning model development follows, where suitable algorithms, such as the Random Forest Classifier, are chosen. The dataset is divided into training and testing sets for model evaluation, and the model is trained on historical data. Model evaluation involves assessing performance using metrics like accuracy, precision, recall, and F1-score, and hyperparameters are fine-tuned for optimization.

Upon model deployment, predictions on new data are made, providing a risk assessment or probability of breast cancer. Optionally, a user-friendly interface can be created for inputting data and displaying predictions. The project may also incorporate an Out-of-Bag (OOB) score, estimating the model's performance on unseen data during training.

Additional information

Trello team link:
https://trello.com/invite/b/XhET3kVH/ATTI257d2d3a7169d441827c367e959c32f21C34764C/project

Repository link:

https://github.com/krne1337/SUML-Project