Predicting Breast Cancer Diagnosis Based on Cellular Features

Introduction

About 1 in 8 women (around 12%) will develop breast cancer in her lifetime. Furthermore, it is predicted that approximately 276,000 new cases of breast cancer will be diagnosed and 42,000 deaths due to breast cancer will occur in year 2020 in the US alone. However, early diagnosis and accurate detection of breast cancer plays and important role in improving the survival rate of breast cancer and have decreased deaths by 1.3% every year from 2013 to 2017. [1]

Breast Cancer is usually diagnosed by surgical biopsy, FNA (Fine Needle Aspiration) for visual interpretation and screening techniques like Mammography, X-rays and MRI’s. Surgical Biopsy is 100% sensitive but is expensive, time consuming and needs surgical intervention. Other visual techniques are error prone and might lead to misdiagnosis.

Objective

Benign and Malignant Cells are very distinct in terms of cell features and morphology. UCI Machine Learning Repository has a dataset describing cell features as analyzed by computes vision diagnostic system of tumor samples from patients. The goal of my project is to develop a classification model that can classify benign tumors from malignant ones based on these cellular features.

Features

There are in all 9 features in the dataset as listed below:

1. Clump Thickness

2. Uniformity of Cell Size

3. Uniformity of Cell Shape

4. Marginal Adhesion

5. Single Epithelial Cell Size

6. Bare Nuclei

7. Bland Chromatin

8. Normal Nucleoli

9. Mitoses

Target Variable – Malignant or Benign.

Findings

I performed classification using Logistic Regression and Random Forest and found that Random Forest had the highest F1 score (95.55%).

Modelling Methodology

EDA:

From the pairplot the features seemed to be highly separable by Class which might be expected from highly distinct morphology of cancer and non-cancerous cells. From the correlation table only ‘Uniformity of Cell Shape’ and ‘Uniformity of Cell Size’ show high correlation when compared to others.

Firstly, I performed Logistic Regression and Random Forest with optimized parameters for all features. As for breast cancer diagnosis I wanted to get a sweet spot between precision and recall so I chose to evaluate F1 score for my model selection. With all the features I got very similar F1 score of 0.93 for both the models. I also analyzed the feature importance in my model.

Looking more closely at the high correlation score between ‘Uniformity of Cell Shape’ and ‘Uniformity of Cell Size’ I decided to omit one feature to avoid multicollinearity and see how that would affect the model. I also chose to remove the ‘Mitosis’ feature as it was least important amongst all the features.

The logistic regression scores after feature reduction were the same that says that we got same scores for a simplier model. However, for Random Forest the F1 score increased to 0.95 and so was selected as the final model.

Notebook Navigation

All the analysis is summarized into one notebook from start to end. The initial part is Data Cleaning followed by EDA which is followed by Modelling.

References:

1. <https://www.breastcancer.org/symptoms/understand_bc/statistics>