Diagnosis of Breast Cancer Using Random Forests

Team 01

1.Introduction

Breast cancer diagnosis is a critical aspect of healthcare that involves the identification and assessment of breast abnormalities or potential cancerous cells within the breast tissue. It is a complex process that utilizes various medical techniques and technologies to determine the presence, stage, and characteristics of breast cancer. Early detection plays a crucial role in improving treatment outcomes and patient survival rates. In this context, we will explore the different methods and approaches used in the diagnosis of breast cancer, highlighting their significance in providing accurate and timely information for effective treatment planning.

2.Methodology

2.1.Data Collection

The experiment begins with the retrieval of the selected dataset from the UCI Machine Learning repository. The dataset contains 569 instances of tumors, 212 of which are malignant and 357 of which are benign growths.

2.2.Data Prepossessing

The dataset contains 32 columns, out of which the arbitrary ID column and unnamed column are dropped as they have no relevance to the experiment. The diagnosis column further becomes our target variable. Checking for missing data and it was found that there is no missing data. The Diagnosis feature is converted to binary values using one-hot encoding. Identify imbalance. The minority class is up-sampled to match the majority dataset count to resolve this. As there are not many outliers in the data, the 20% test set is sufficient to evaluate the final models objectively.

2.3.Feature Selection

2.3.1.Initial Feature Selection

The initial stage of feature selection is based on the Pearson correlation coefficient. For pairs of features we try to find correlation coefficient higher than 0.8. This dataset is used to train and evaluate the first set of models. It will be referred to as the initial dataset from now on. Since it's elusive to understand the digits in the heatmap, we showed the correlation table.

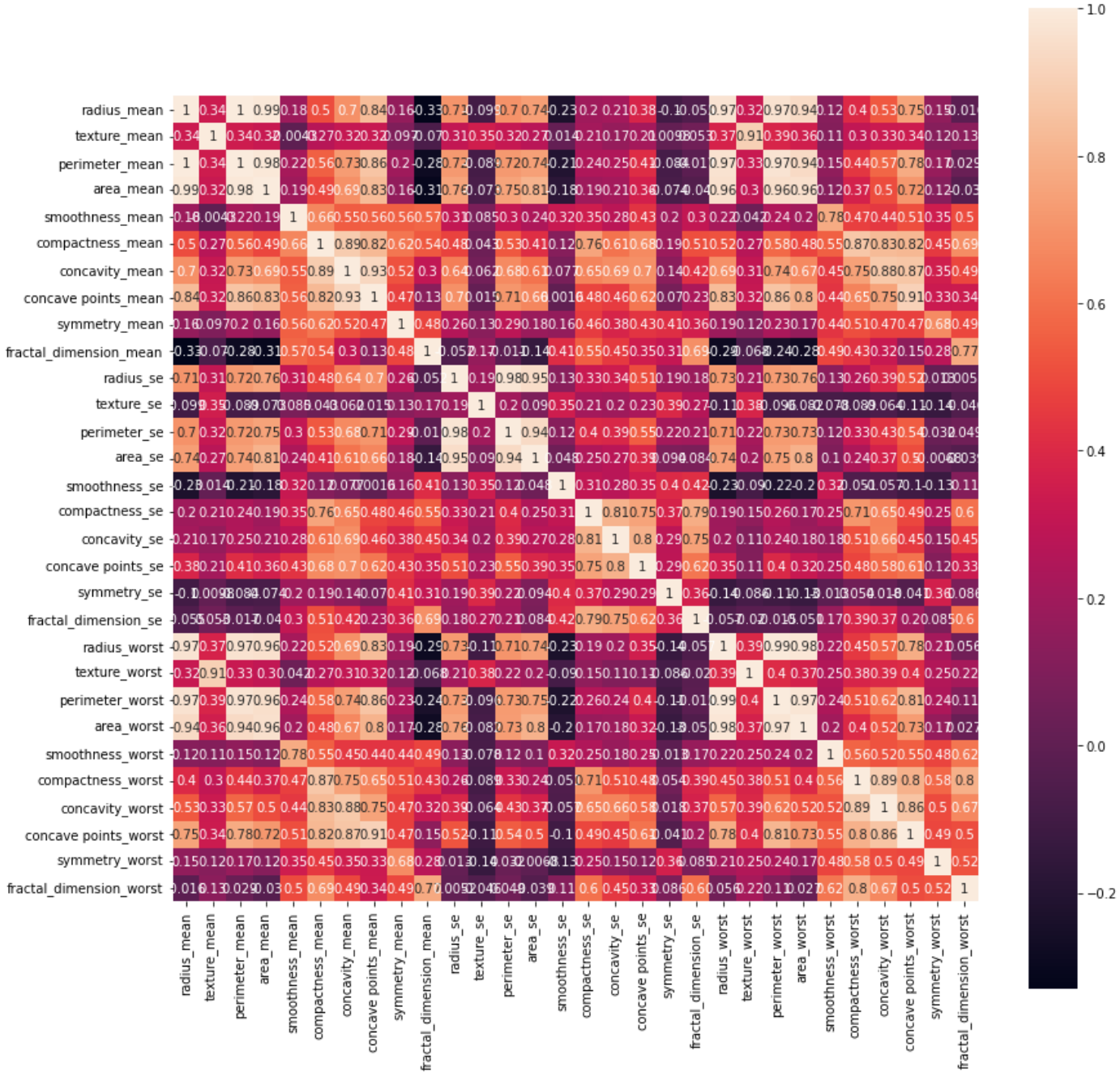

Fig. 1. Feature correlation map

Initial dataset contains 16 features:
features = {texture_mean, area_mean, smoothness_mean, concavity_mean, symmetry_mean, fractal_dimension_mean, texture_se, area_se, smoothness_se, concavity_se, symmetry_se, fractal_dimension_se, smoothness_worst, compactness_worst, symmetry_worst, fractal_dimension_worst}

2.3.2.Further Feature Selection

1.Recursive Feature Elimination

uses an external estimator (Extra Trees Classifier).The features are ranked by importance using the external estimator, and the least important features are removed (lower rankings have better performance). This procedure is then repeated recursively till eight features remain.

Top features = {texture_mean, area_mean,concavity_mean,area_se, smoothness_worst, compactness_worst, symmetry_worst, fractal_dimension_worst}

2.Logistic regression

In the second method, feature importance are determined using a Logistic Regression model [19]. The eight features with the highest importance are then selected.

Selected features logistic = {texture_mean, area_mean,concavity_mean,area_se, smoothness_worst, compactness_worst, symmetry_worst, fractal_dimension_worst}

3.Univariate Selection

By using ANOVA F-value, we select the best 8 features. The set of features contained in the minimal dataset is detailed in set F1(selected features univariate) as follows:
F1 = {texture_mean, area_mean, smoothness_mean, concavity_mean, area_se, smoothness_worst, compactness_worst, symmetry_worst}

Note that the further selected features here are only 7, while in the paper they are 8. This is mainly because data science algorithms include random initialization and random splitting. Differences in available system resources (CPU, GPU) or changes in the software environment (updates to libraries, changes in dependencies) can impact the results as well.

However, all 7 features selected here are selected in the paper as features of the minimal dataset. So the produced results are very similar to those in the paper so far.

2.4.Model Development

Five models are selected for a preliminary comparison- Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), Multilayer Perceptron (MLP), and K-Nearest Neighbors (KNN). They are trained on the initial dataset with default parameters, and their results are compared. The models are 5-fold cross-validated using recall value as the scoring function.

The Random Forest model performs the best with a cross-validation recall score of 0.965. Thus, the Random Forest model is selected as our primary model.

The four remaining models are trained using their default parameters on the initial and the minimal dataset. Random Forest's hyperparameters are precisely tuned to maximize the primary model's performance.

Hyperparameter	Value	Description
n_estimators	100	Number of trees
criterion	entropy	Function to measure split quality
max_depth	10	Max depth of a tree
max_features	0.2	Number of features to consider while splitting
min_samples_split	2	Number of samples needed to split an internal node

2.5.Model Evaluation

The five models are trained and evaluated on both the initial and minimal datasets. As accuracy is not a sufficient metric in the medical field, the models are compared based on a multitude of measures. All five models are trained and evaluated using these metrics on both the initial and minimal datasets. Their performances are collated and analyzed in the Results section.

2.6.Results and Discussion

we conclude that Random Forest performed the best while KNN performed the worst.

2.6.1.Initial Dataset

Model	Accuracy (%)	Precision	Recall	F1 Score	ROC-AUC
Random Forest	97.89	0.97	0.99	0.97	0.98
Support Vector Machine	96.78	0.98	0.95	0.96	0.96
Decision Tree	97.05	0.96	0.98	0.97	0.97
Multilayer Perception	96.49	0.96	0.96	0.96	0.96
K-Nearest Neighbors	94.95	0.94	0.96	0.95	0.94

2.6.2.Minimal Dataset

Model	Accuracy (%)	Precision	Recall	F1 Score	ROC-AUC
Random Forest	98.17	0.96	0.99	0.98	0.98
Support Vector Machine	96.07	0.98	0.93	0.95	0.96
Decision Tree	97.06	0.95	0.98	0.97	0.97
Multilayer Perception	94.81	0.95	0.93	0.95	0.95
K-Nearest Neighbors	94.53	0.93	0.95	0.94	0.94

The Random Forest model continues to achieve the best metrics, with an accuracy of 98.45% and a perfect recall score.

Fig. 2 and Fig. 3 depict the ROC curves for the RF model on the initial and minimal datasets.

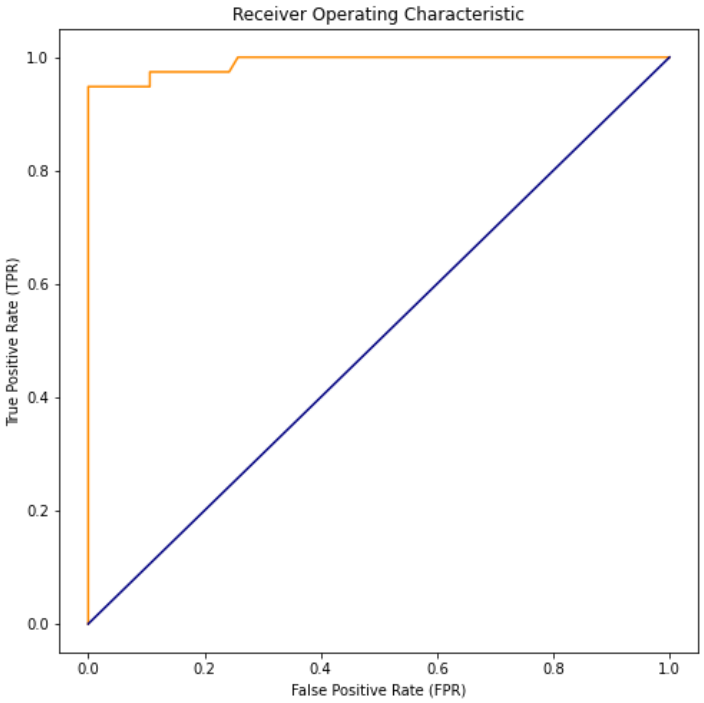

Fig. 2.ROC curve for initial dataset

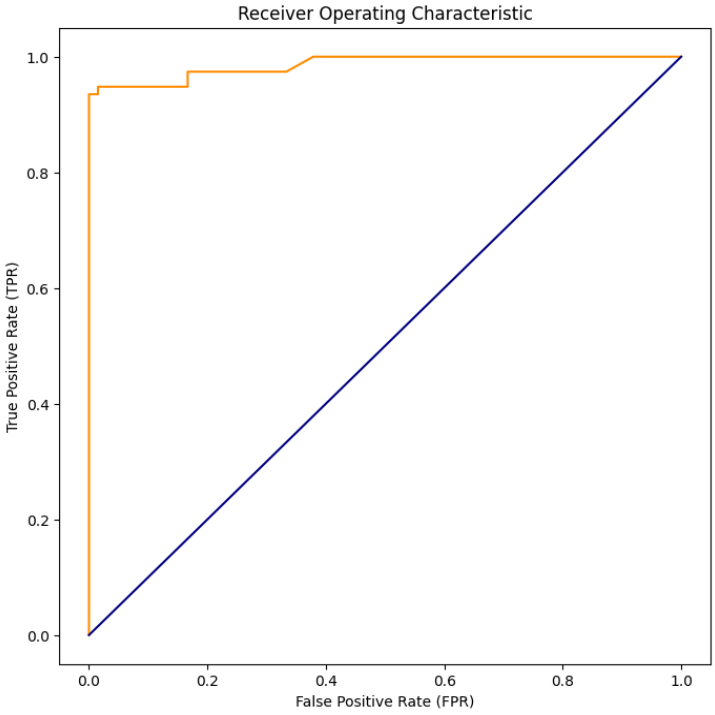

Fig. 3.Plot the ROC curve for minimal dataset