Breast Cancer Diagnosis Classification

Yiling Peng (yp387)
Shuhan Ding (sd925)

I. Data Set Description

The dataset is from the Wisconsin Breast Cancer Diagnostic Dataset. Medical personnel acquired digital images of patients' breast lumps after fine needle aspiration (FNA) and extracted features from these digital images that describe the presentation of cell nuclei. Tumors can be classified as benign or malignant.

The dataset includes a total of 32 attributes, representing as following：

1) ID number;

2) Diagnosis (M = malignant, B = benign)

Ten real-valued features are computed for each cell nucleus:

a) radius; b) texture; c) perimeter; d) area; e) smoothness; f) compactness; g) concavity; h) concave points; i) symmetry; j) fractal dimension.

The mean-mean, standard deviation-se, and maximum values-worst (mean of the three largest values) of these ten features are calculated for each image separately.

Therefore, before data cleaning, we have an entire 569*32-dimensional data. Diagnosis result is treated as our perdition label and the rest of the features, which we analyzed and processed, are used to predict whether the tumor is benign or malignant. Our goal is to generate the final feature vectors by performing PCA, and then test and compare the models, expecting to achieve good prediction performance.

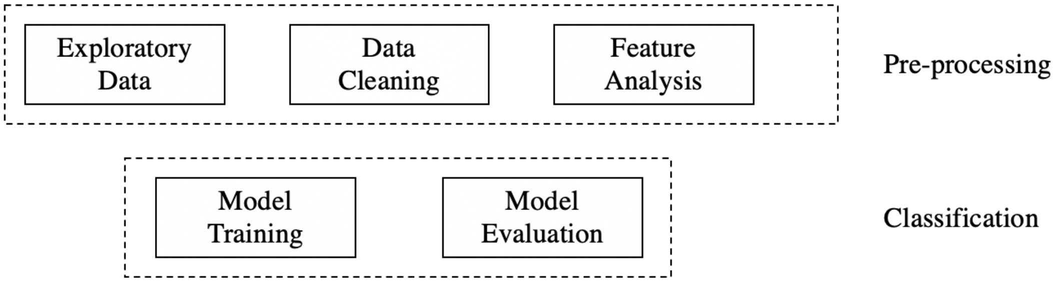


Figure 1. General structure of this project.

II. Data Analysis

A. Check for data

First, we filter out the attributes we need. Since "id" is a meaningless feature for our prediction, it is removed. Next, we check if there are still missing values. We found that in this dataset, no missing values exist. Therefore, the final size of data is 569*31.

B. Encode data label

We standardized non-numeric and numeric label values. In particular, in original dataset, malignant tumors were recorded as M and benign as B. For binary classification, we converted diagnostics parameters to values, 0 for benign and 1 for malignant. It is noted that in 569 patients, 357 were benign and 212 were malignant.

C. Explore significant features

As Figure 2. and Figure 3. shown, mean values of cell radius, perimeter, area, compactness, concavity, and concave points can be used in classification of the cancer. Larger values of these parameters tend to show a correlation with malignant tumors. On the other hand, mean values of texture, smoothness, symmetry, or fractural dimension does not show a particular preference of one diagnosis over the other. In any of the histograms there are no noticeable large outliers that warrants further cleanup.

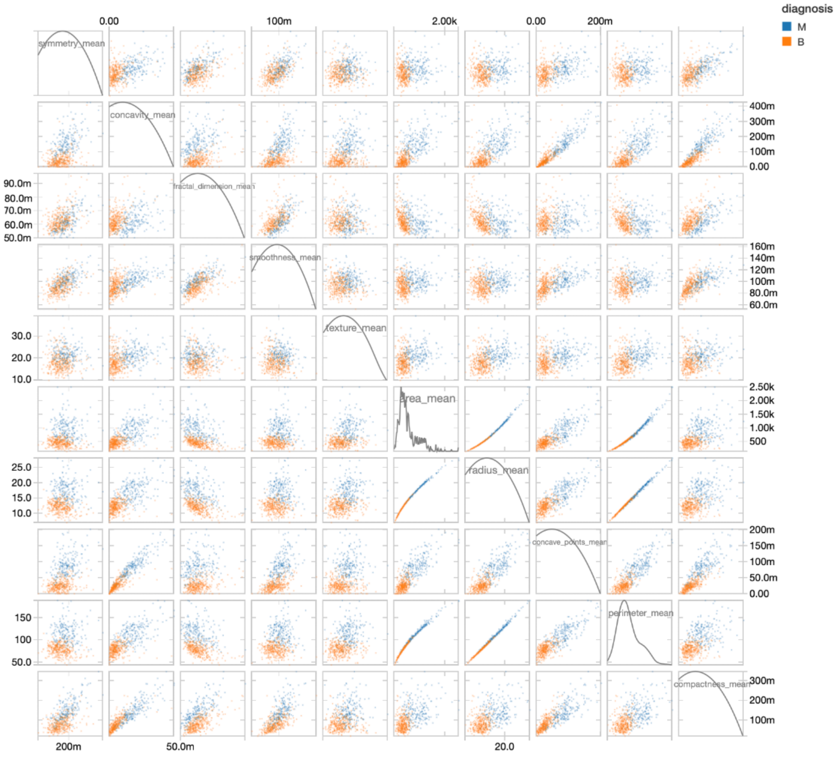
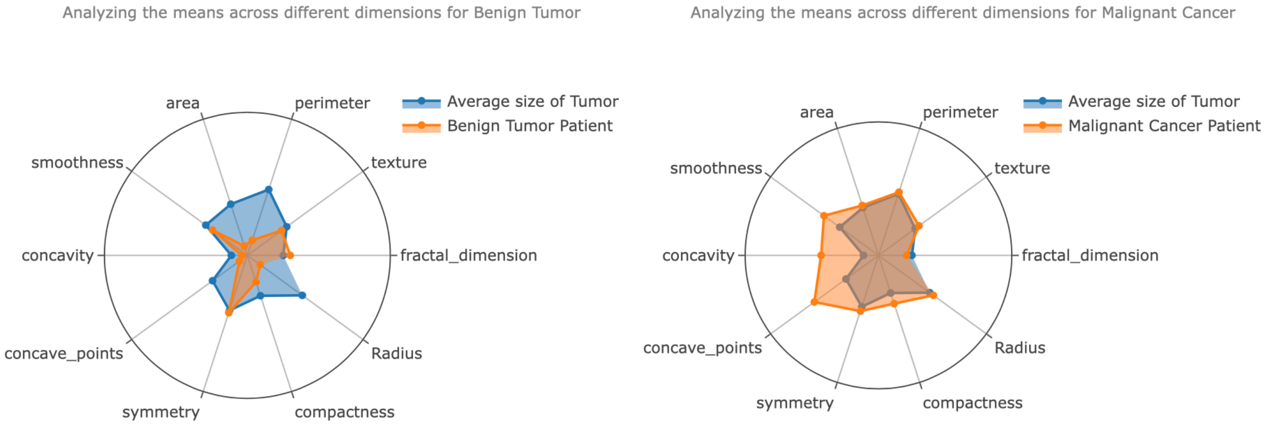


Figure 2. Plotpairs of mean data.


Analyzing the means across different dimensions for Benign Tumor

Analyzing the means across different dimensions for Malignant Cancer

Figure 3. Radar graph to determine optimum vs actual cell type for means.

In addition to this, we also analyze features using PCA. PCA did quite a decent job of visualizing our two target clusters. Although PCA is able to differentiate the classes very well, we found later that PCA technique also reduced some important features, which resulted in reduced accuracy. Therefore, we only used standard scaler on the dataset.

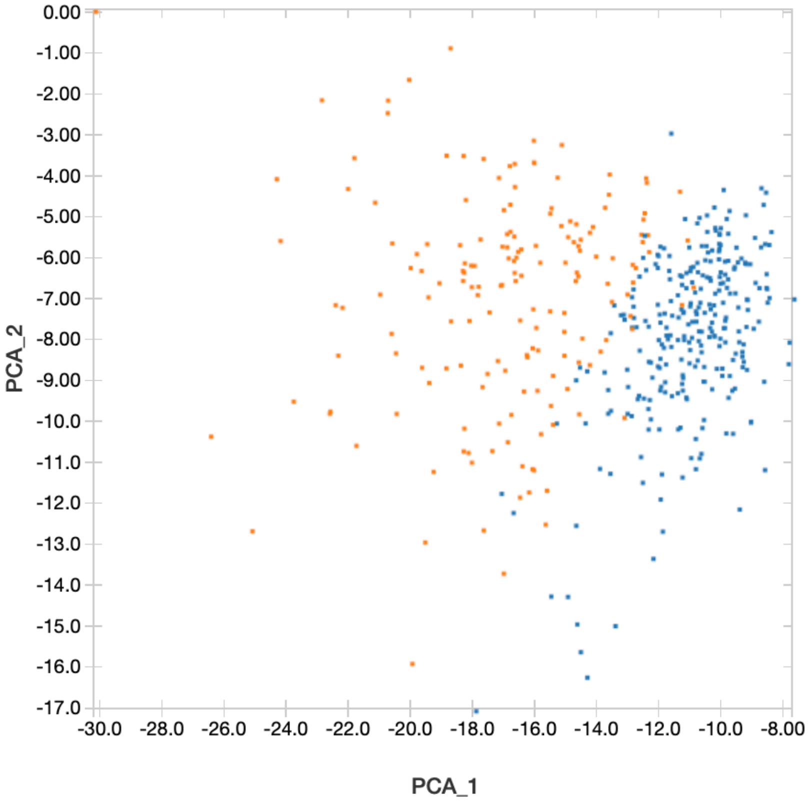


Figure 4. PCA of two clusters (yellow for Malignant and blue for Benign).

III. Classification

A. Splitting the data set

Since this dataset is not ordered, we make a simple 8:2 split of the dataset to create a training dataset and a test dataset. And we also 8:2 split the training dataset into a training and validation dataframe.

B. Classification model

We trained and tested several classifiers that are widely used in binary classification problems, including Random Forest, Support Vector Machine, Logistic Regression, Light GBM and XG BOOST. And we summarized the results for these 5 models, in terms of accuracy on the training, validation and test sets shown in Table 1. It is found that the performance of all models was acceptable due to the organized data features in the previous part. SVM preformed best, even reaching the 99% accuracy in the test set.

Table 1. The accuracy of training, validation, and test sets with different classifiers.

Accuracy	XG BOOST	Random Forest	Support Vector Machine	Logistic Regression	Light GBM
Training	0.97	0.99	0.97	0.97	1.0
Validation	0.95	0.94	0.95	0.93	0.95
Test	0.90	0.96	0.99	0.97	0.97

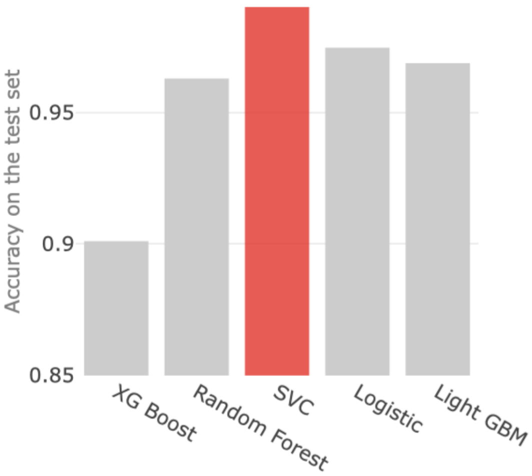


Figure 5. Ranking the top 5 models in terms of test set accuracy.

IV. Conclusion

Overall, we have implemented pre-processing of the data and feature analysis. Through PCA analysis, we fouded that top 5 predictors are 'concave_points_mean', 'area_mean', 'radius_mean', 'perimeter_mean', 'concavity_mean'. Besides, we have trained and tested the Random Forest, Support Vector Machine, Logistic Regression, Light GBM and XG BOOST Classifiers, achieving 99% accuracy. The prediction is quite accurate, but for medical usage, the interpretability needs to be improved.

V. Future Work

Next, we will adjust the significant parameters of the models. Currently we are using the model with direct function calls and some default parameters.

And we now compare only one evaluation parameter, accuracy. We also intend to use specificity, sensitivity, confusion matrix, etc. to evaluate the model results. We believe that the results will be more objective and reliable.

Besides, we are also considering whether to use the existing features to propose a new feature to be involved in the subsequent classification. Since our dataset involves a small dimensionality, adding features may improve the underfitting.

Refernce:

[1] Bhattacharjee, Nabanita and R. Parekh. "Skin texture analysis for medical diagnosis." IEEE Multimedia 19.2 (2012): 28-37.

[2] Permuter, Haim J. Francos and I. Jermyn. "A study of Gaussian mixture models of color and texture features for image classification and segmentation." Pattern Recognition 39.4 (2006): 695-706.

[3] Ibrahim, Norhayati, et al. "Automated detection of clustered microcalcifications on mammograms: CAD system application to MIAS database." Physics in Medicine & Biology 42.12 (1997): 2577.

[4] Carter, Jane V., et al. "ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves." Surgery 159.6 (2016): 1638-1645.