Breast Cancer Diagnosis Dataset Analysis

Final project Report
DSI YouJung Koo

1. Introduction

Purpose

Breast cancer is one of the most commonly diagnosed cancer types among women, so achieving a highly accurate breast cancer diagnosis classification model is necessary. Diagnosing breast cancer by measuring the size, shape, and texture of a cellular nucleus of the breast mass is very cost-effective and more accurate than using breast images of patients.

Wisconsin Breast Cancer dataset (WBCD)

The Wisconsin Breast Cancer dataset (WBCD) provides data of breast mass cellular nucleus computed from a digitized image of a fine needle aspirate of a breast mass. The extracted breast mass is stained to highlight the cellular nuclei.The team computed values for each of ten characteristics of each nuclei, measuring size, shape and texture. The mean, standard error and extreme values of these features are computed, resulting in a total of 30 nuclear features for each sample. [1]

Target Variable and Features

The Wisconsin Breast Cancer Diagnosis dataset is a classification dataset. The target variable is 'diagnosis'. 'B' is for benign (‘0’ after preprocessing), 'M' is for malignant(‘1’ after preprocessing). There are 569 data points and 30 feature columns. There are ten real-valued features computed for each cell nucleus: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension. The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. 27 features are iid, and 3 compactness features are dependent on the perimeter and area feature.

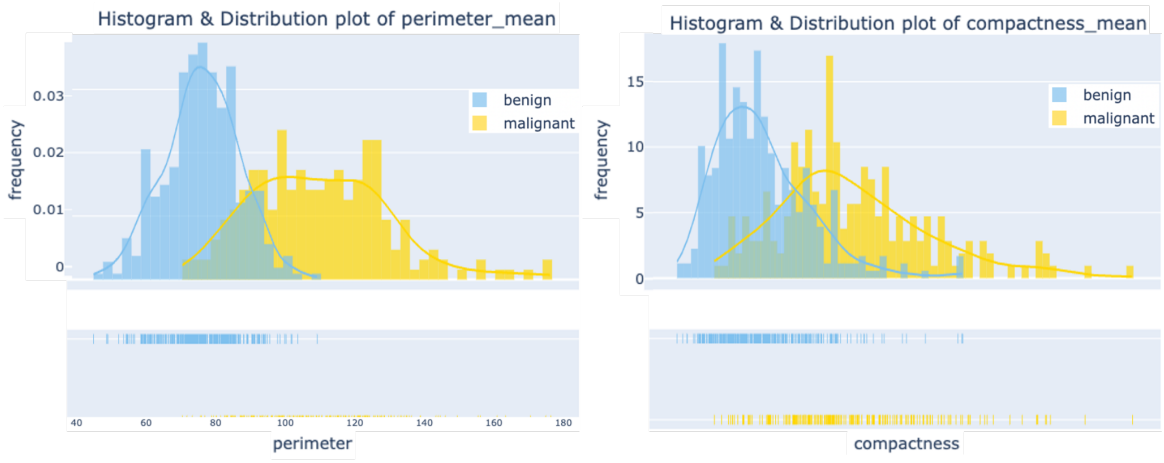
Current Research

Recent research used the Wisconsin breast cancer diagnosis dataset to improve machine learning models used in breast cancer diagnosis. [2],[3] These studies aim to guide physicians to adopt an effective model for a practical understanding and prognosis of breast cancer tumors. In one study, they found out that the SVM machine learning predictive model is efficient to deal with massive volumes of tumor data and shows a highest diagnosis accuracy of 99.3%. [2] Another study shows that KNN can show 99.42% accuracy, [3] showing higher accuracy than the 98% accuracy achieved using ANN. [4]

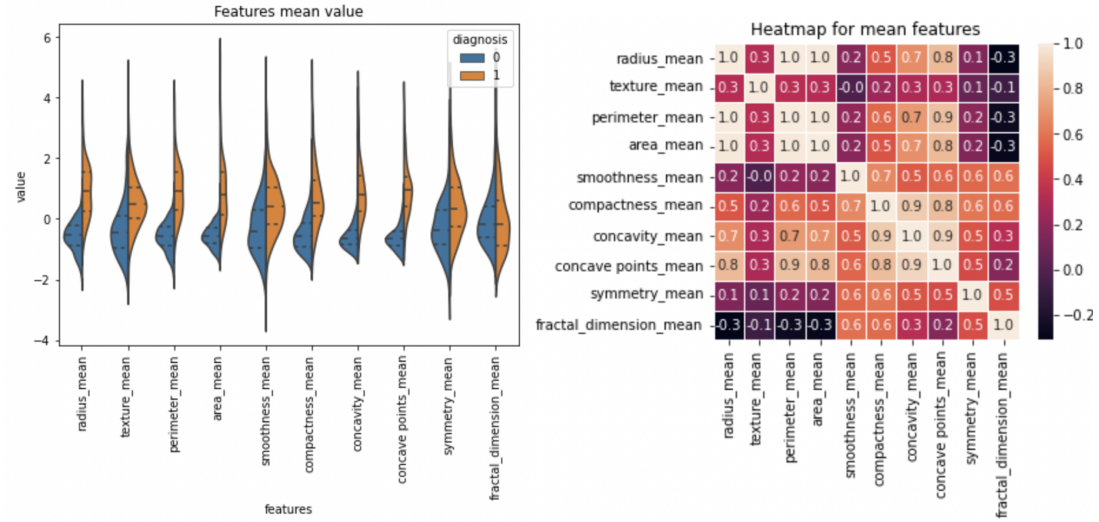
2. Explanatory Data Analysis

Feature Analysis

I used .describe to check for missing values and min max values of the dataset, and made a histogram plot, distribution plot, a violin plot, and heatmaps for all features. The perimeter_mean feature plot and compactness_mean feature plot gives an overall understanding about the size and shape features. The higher the length or size of the cell is, it’s more inclined to be cancerous (malignant). The y axis of the histogram is the target variabel diagnosis.

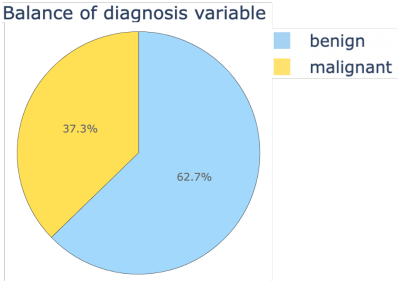


The violin plot of all the ten mean value features shows that radius, perimeter, area features show similar distributions. The y axis is the target variable, and 0 is for benign and 1 is for malignant. This will be a useful indicator in ranking the features according to importance. The heatmap of all ten mean features shows that the fractal dimension feature has the lowest correlation with other features.



Target Variable

The target variable in this dataset is 'diagnosis'. Each patient data is classified to two groups: 'B' or 'M', which stands for benign and malignant.



3. Methods

[Data preprocessing]

Dataset Split

For Logistic Regression, Decision Tree, KNN, SVM models, sklearn's train_test_split library is used to split the dataset into (Train+Validation) : Test = 8 : 2, which is the optimized ratio used in WBCD dataset research. [3] Then, apply 10 KFold Shuffle Split to the train and validation set, because the method empirically produces test error rate estimates that do not suffer from excessively high bias or very high variance. Doing cross validation using 10 KFold Split is effective

when dealing with a relatively small set of human data. For XGBoost models, the dataset is splitted into train, validation, and test set with a ratio of 6:2:2.

Preprocessor

All of the feature values in this dataset are continuous values. I used the Standard Scaler on all 30 features: the mean features and worst features to put the values in the same scale. does not have a bounding range, so the dataset distribution will not be affected by standardization.

Features

There is no change in the number of features, and there is no missing data. The 30 features and 569 data points are maintained.

ML Pipeline

Five machine learning algorithms are used in this analysis: Logistic Regression, DecisionTree, KNN, SVM, and XGBoost classifier algorithm. For each algorithm, several parameters were tuned. (Table 1)

Logistic Regression	C : [100, 10, 1.0, 0.1, 0.01]
Decision Tree	max_depth : [3,4,5,6,8,9]
KNN	n_neighbors : [1, 5, 10, 20, 30, 100]
SVM	Gamma : [1e-3, 1e-1, 1e1, 1e3, 1e5], C : [1e-1, 1e0, 1e1]
XGBoost	reg_alpha : [0e0, 1e-2, 1e-1, 1e0, 1e1, 1e2], max_depth : [1,3,10,30,100]

<Table 1. Tuned parameter values for each machine learning algorithm>

The metric used to evaluate the models' performance is accuracy. Normally, cancer datasets are extremely imbalanced; benign data points take up about 90% of the dataset. The Wisconsin Breast Cancer dataset is a relatively balanced dataset with 357 class 'benign (0)' data points, and 212 data points for class 'malignant (1)'. Using accuracy instead of f1 score will give us a better understanding of the dataset.

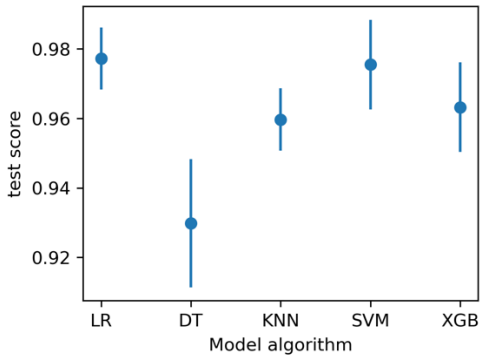
In this analysis, each model algorithms are assigned five different random state values when splitting. Also, except for the XGBoost model, 10 k fold method is applied to all models. At the end of the for loop, we save the best model and its validation and test score to a list. The random state results in test score difference between the models(difference range between 0.0263 to 0.0438). All the machine learning methods used are deterministic, thus, we get the same result every time we run the model.

Doing 10 k fold increased test accuracy for Logistic Regression, Decision Tree, KNN, SVM models. However, it decreased the test accuracy for XGBoost models. For XGBoost models, the train/validation/test datasets are divided using random seed and has an early stop round of 100.

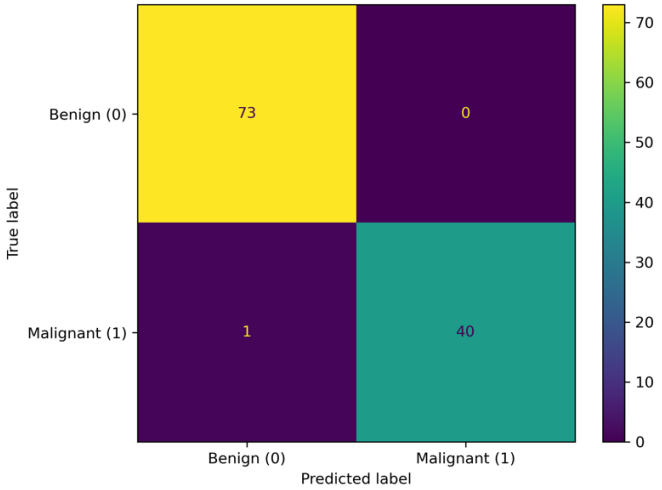
4. Results

	Mean test score	Standard deviation	(Mean test score – baseline) / standard deviation
Logistic Regression	0.9772	0.0089	36.6737
Decision Tree	0.9298	0.0184	15.2554
KNN	0.9596	0.0089	34.7126
SVM	0.9754	0.0129	25.3114
XGBoost	0.9632	0.0129	24.3588

<Table 2 : Mean values of each Machine Learning algorithm>

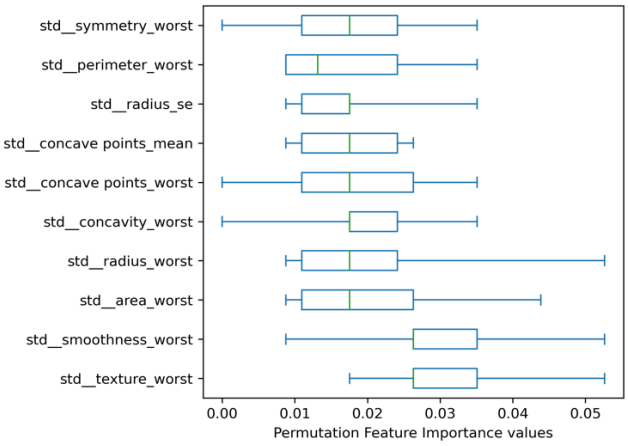


<Table 3 : Model performance & measure uncertainties due to splitting>

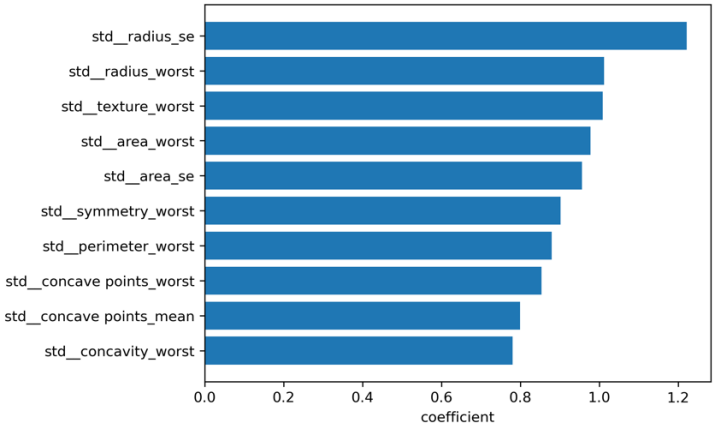


<Table 4 : Test set confusion matrix using the best performing Logistic Regression model>

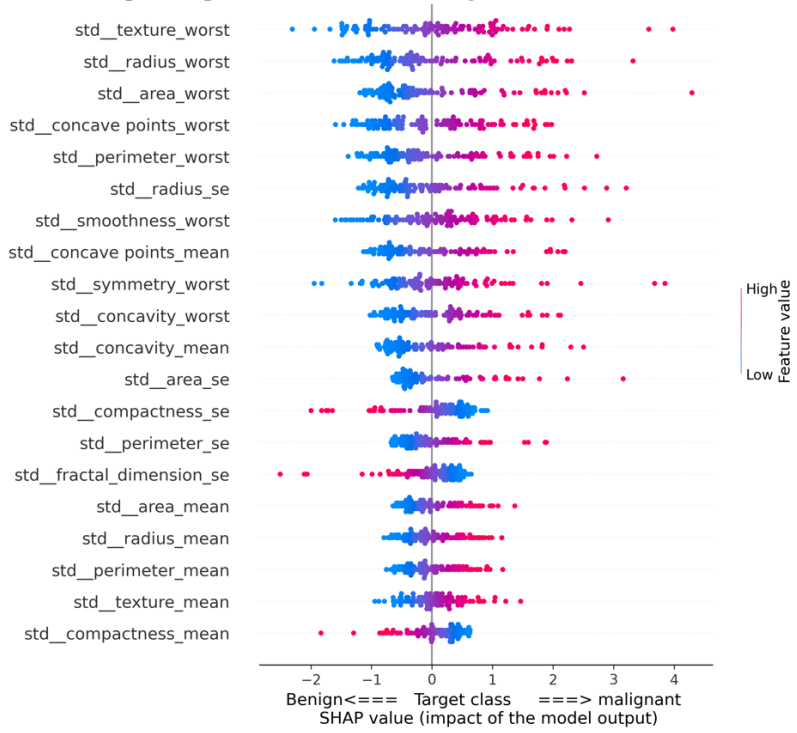
The baseline accuracy of the entire dataset is 0.6274, calculated by dividing the number of benign data points by the number of all data points. All models showed a high accuracy score, and out of the models, the Decision Tree model showed the lowest mean score, and the Logistic Regression model showed the best prediction overall. Logistic Regression model showed the best prediction score. The best Logistic Regression model showed an accuracy of 0.9912. It successfully predicted all test data points correctly except for one data point.



<Table 5 : Top 10 Permutation Importances using Logistic Regression test set>



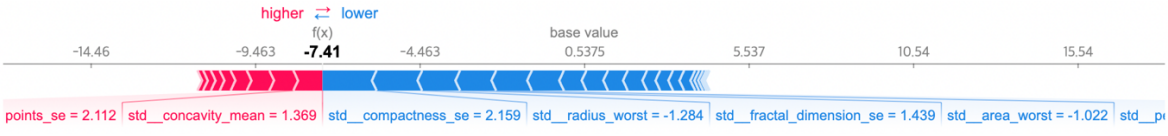
<Table 6 : Top 10 Important features according to coefficient value>



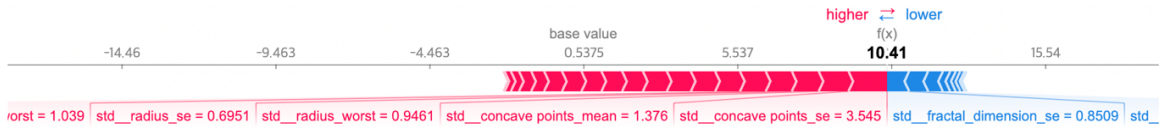
<Table 7 : Impact of each feature in Logistic regression prediction>

Permutation importance, coefficient values, and Shap value are used to determine which feature contributes the most to the prediction. 'std_texture_worst' is the most impactful feature according to table 5 and table 7. 'std_radius_se' is the most impactful feature according to table 6. In all the three tables (5,6,7), worst features (the features with names that end with _worst) appear frequently. Also, table 7 shows that worst features positively contribute to the prediction, giving a positive impact to the prediction (making the prediction to be malignant). With this result, we can assume that the worst features have a distinct correlation with cancer diagnosis. On the other hand, mean features show lower feature importance.

The third most important feature according to table 7 is the area_worst feature. This suggests that by solely looking at the size of the breast cell, we can get an initial idea of whether the patient has breast cancer or not.



<Table 8 : Shap local value for index 0 >



<Table 9 : Shap local value for index 50 >

The true prediction for index 0 is class 0, and for index 50 it is class 1. In table 8, it is remarkable that 'std_ compactness_se' is contributing more to the prediction than 'std_concavity_mean' even though 'std_concavity_mean' is more important in a global scale. In table 9, the same thing happens with the 'std_concave points_se' feature and 'std_fractal_dimension_se' feature.

5. Outlook

The best model is the Logistic Regression algorithm model, and it uses the default C value (Inverse of regularization strength). The only parameter tuned is the C value, so the default Logistic Regression model is the best model. Tuning an additional parameter, such as penalty, can improve the model. Also, considering that the Logistic Regression model shows the best result, it could have produced good results when using a different kind of linear classification model.

Also, the reason why the 10k fold method was not used in XGBoost model is because it showed low accuracy. Using train_test_split library showed a better result. Reducing the number of folds or changing the parameter could've increased performance.

Worst features (mean of the three largest values) show higher feature importance than mean and se features, and they positively impact the prediction (higher chance of the patient being diagnosed as having cancer). The provided dataset consists of 569 data points, which is relatively small, so retrieving more patient data would provide a better understanding of the data and contribute to achieving higher accuracy score.

6. Reference

[1] *Machine Learning for Cancer Diagnosis and Prognosis*. (n.d.). Retrieved October 20, 2022, from https://pages.cs.wisc.edu/%7Eolvi/uwmp/cancer.html

[2]Rasool, A.; Bunterngchit, C.; Tiejian, L.; Islam, M.R.; Qu, Q.; Jiang, Q. Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis. Int. J. Environ. Res. Public Health 2022, 19, 3211. https://doi.org/10.3390/ijerph19063211
[3] Mushtaq, Z., Yaqub, A., Sani, S., & Khalid, A. (2019, October 22). Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets. *Journal of the Chinese Institute of Engineers*, *43*(1), 80–92. https://doi.org/10.1080/02533839.2019.1676658
[4] Abdel-Ilah, L., and H. Šahinbegovi. 2017. "Using Machine Learning Tool in Classification of Breast Cancer." *CMBEBIH 2017, IFMBE Proceedings* 62 (1): 3–8. doi:10.1007/978-981-10-4166-2.

7. Github repository

https://github.com/youjungkoo/breastcancer