Breast Cancer (Wisconsin Dataset) Classification using Stacked Ensemble of Logistic Regression, SVM and Neural Networks

https://github.com/cmonappa1/wisconsinbreastcancer

1. Introduction

1.1 Overview

The problem starts with this important question, "Can you identify breast cancer before it occurs?" Early diagnosis significantly increases the chances of survival, but the key challenge is to be able to detect/classify tumors before it becomes malignant. A malignant tumor is called cancer and grows into surrounding tissues or spread to distant areas of the body. They function incorrectly when they are at different parts of the body and tend to not die naturally. For instance, a liver cancer cell would secrete protein dissolving enzymes when it is in the kidney, thus causing malfunction. Modern machine learning techniques dramatically improve the diagnosis; from experienced physicians' accuracy of 79% to the model's accuracy of 91% - 97%.

1.2 Motivation

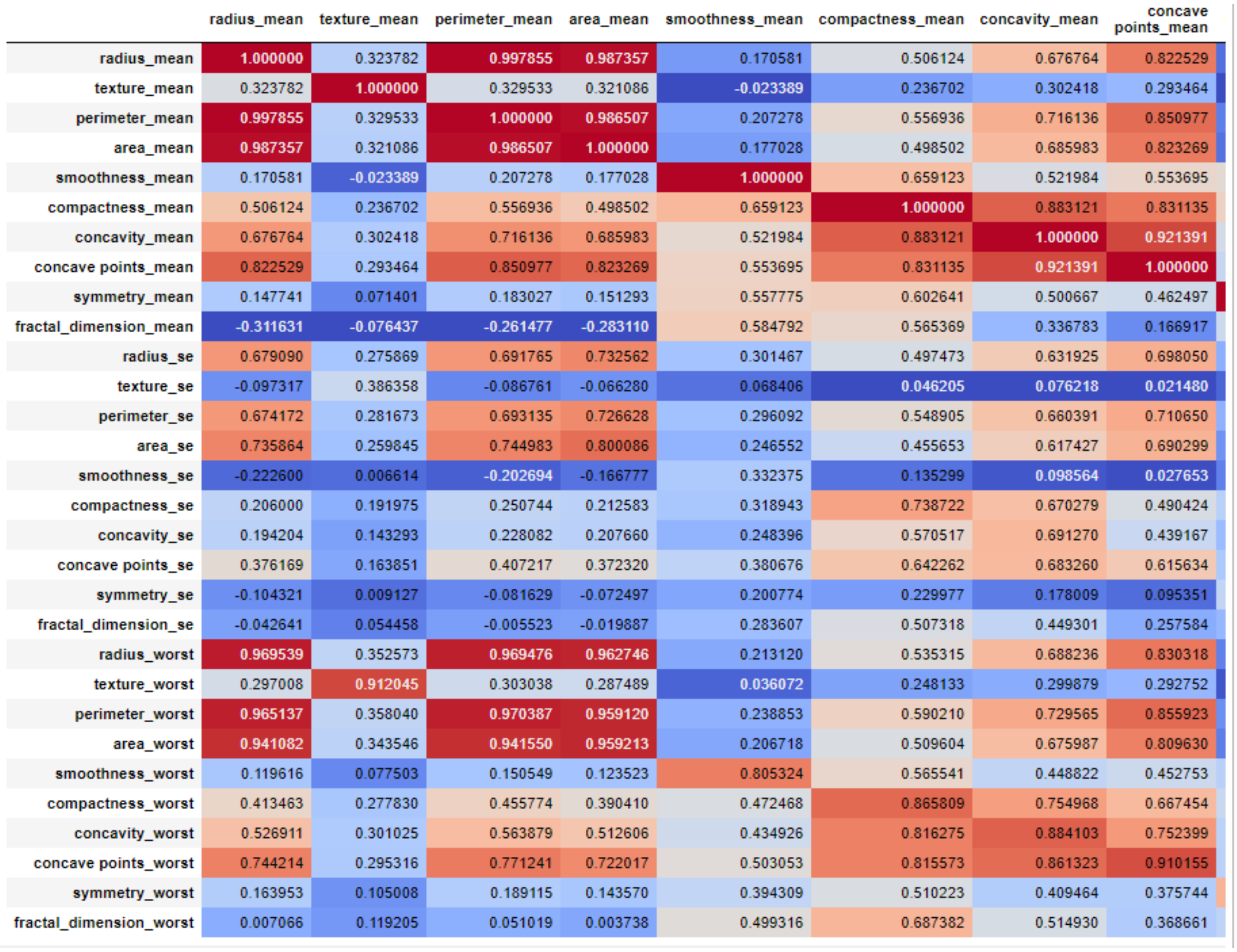
The problem has two main components, preprocessing the data matrix to bring out different features or enhance characteristics that are strong indicators of cancer conditions to implement an Ensemble super-learner to automatically classify the subjects into their respective classification segments. The preprocessing is done by feature engineering the salient features of the data matrix, with methods such as normalizing features, min max scaling amongst other techniques. We then will apply concepts taught in CS532 such as SVM, Logistic Regression and Neural Networks to partition, tune hyper-parameters and train the model to create a highly accurate model. The dataset is a numerical dataset from the Breast Cancer Wisconsin (Diagnostic), UCI Machine Learning Repository. Following is the description of the dataset.

1.3 Methodology

To perform classification of the dataset, we will start by partitioning the dataset into a training and validation set in a 75% 25% split. A function will be written to ensure similar occurrences of malignant and benign classes in both the training and validation set. Data standardization and preprocessing is applied to both the training and validation set using Scikit-Learn. Model development algorithms are implemented to find an initial estimate of the model performance. For the experimentation, the training and validation dataset is divided into 10 subsets, where 8 were used to train the data, 1 to validate the data and the last revalidate the data with the best selected hyperparameter. The experimentation procedure is repeated 8 times for 8 different sets of hyperparameters. Then the average misclassification and R-squared error is calculated from the obtained solutions.

2. **Exploratory Data Analysis of the Wisconsin Breasts Cancer Dataset**

The dataset is a binary classed dataset with 569 data points and 30 features. The following is the correlation heatmap between all features.



	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean
radius_mean	1.000000	0.323782	0.997855	0.987357	0.170581	0.506124	0.676764	0.822529
texture_mean	0.323782	1.000000	0.329533	0.321086	-0.023389	0.236702	0.302418	0.293464
perimeter_mean	0.997855	0.329533	1.000000	0.986507	0.207278	0.556936	0.716136	0.850977
area_mean	0.987357	0.321086	0.986507	1.000000	0.177028	0.498502	0.685983	0.823269
smoothness_mean	0.170581	-0.023389	0.207278	0.177028	1.000000	0.659123	0.521984	0.553695
compactness_mean	0.506124	0.236702	0.556936	0.498502	0.659123	1.000000	0.883121	0.831135
concavity_mean	0.676764	0.302418	0.716136	0.685983	0.521984	0.883121	1.000000	0.921391
concave points_mean	0.822529	0.293464	0.850977	0.823269	0.553695	0.831135	0.921391	1.000000
symmetry_mean	0.147741	0.071401	0.183027	0.151293	0.557775	0.602641	0.500667	0.462497
fractal_dimension_mean	-0.311631	-0.076437	-0.261477	-0.283110	0.584792	0.565369	0.336783	0.166917
radius_se	0.679090	0.275869	0.691765	0.732562	0.301467	0.497473	0.631925	0.698050
texture_se	-0.097317	0.386358	-0.086761	-0.066280	0.068406	0.046205	0.076218	0.021480
perimeter_se	0.674172	0.281673	0.693135	0.726628	0.296092	0.548905	0.660391	0.710650
area_se	0.735864	0.259845	0.744983	0.800086	0.246552	0.455653	0.617427	0.690299
smoothness_se	-0.222600	0.006614	-0.202694	-0.166777	0.332375	0.135299	0.098564	0.027653
compactness_se	0.206000	0.191975	0.250744	0.212583	0.318943	0.738722	0.670279	0.490424
concavity_se	0.194204	0.143293	0.228082	0.207660	0.248396	0.570517	0.691270	0.439167
concave points_se	0.376169	0.163851	0.407217	0.372320	0.380676	0.642262	0.683260	0.615634
symmetry_se	-0.104321	0.009127	-0.081629	-0.072497	0.200774	0.229977	0.178009	0.095351
fractal_dimension_se	-0.042641	0.054458	-0.005523	-0.019887	0.283607	0.507318	0.449301	0.257584
radius_worst	0.969539	0.352573	0.969476	0.962746	0.213120	0.535315	0.688236	0.830318
texture_worst	0.297008	0.912045	0.303038	0.287489	0.036072	0.248133	0.299879	0.292752
perimeter_worst	0.965137	0.358040	0.970387	0.959120	0.238853	0.590210	0.729565	0.855923
area_worst	0.941082	0.343546	0.941550	0.959213	0.206718	0.509604	0.675987	0.809630
smoothness_worst	0.119616	0.077503	0.150549	0.123523	0.805324	0.565541	0.448822	0.452753
compactness_worst	0.413463	0.277830	0.455774	0.390410	0.472468	0.865809	0.754968	0.667454
concavity_worst	0.526911	0.301025	0.563879	0.512606	0.434926	0.816275	0.884103	0.752399
concave points_worst	0.744214	0.295316	0.771241	0.722017	0.503053	0.815573	0.861323	0.910155
symmetry_worst	0.163953	0.105008	0.189115	0.143570	0.394309	0.510223	0.409464	0.375744
fractal_dimension_worst	0.007066	0.119205	0.051019	0.003738	0.499316	0.687382	0.514930	0.368661

Figure 1. Correlation Heatmap of 30 features

Similar, plotting a scatter plot on the features gives us an idea of the general trend of the dataset and the characteristics of the data. Figure 2 is the plot of the first 10 features for ease of display.

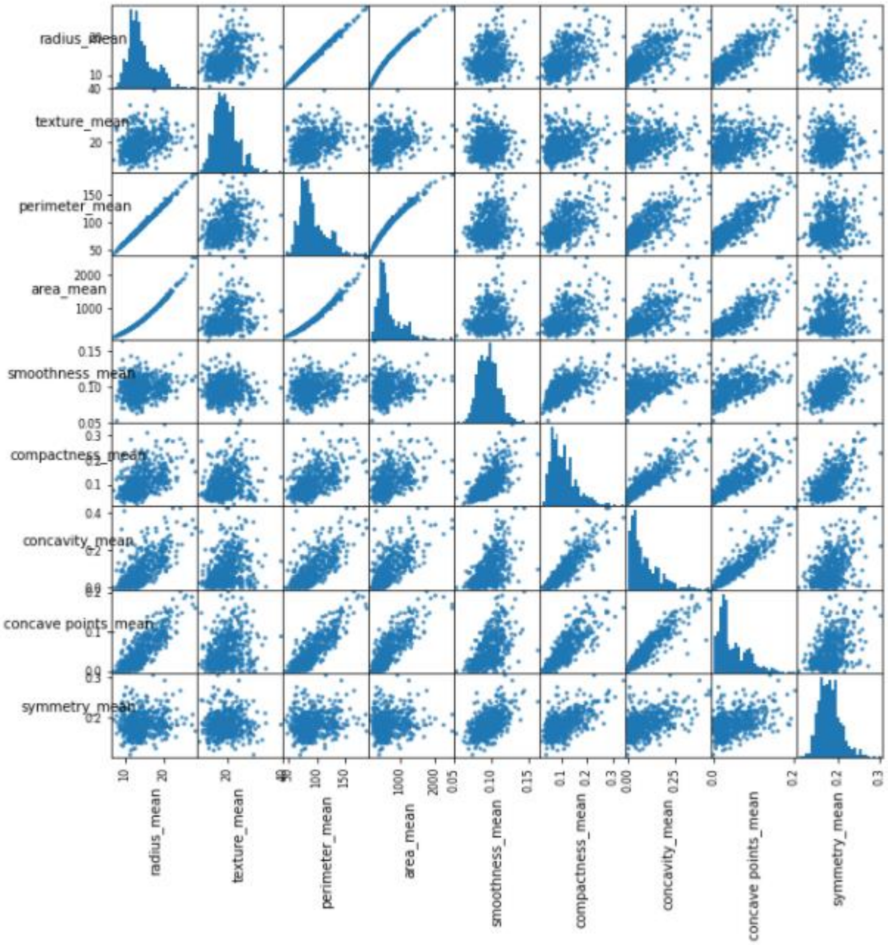


Figure 2. Scatter plots of the first 10 features

Deep diving in the effect of each feature on the class labels (Malignant and Benign), we see a trend. Figure 3 displays the overarching trend with each class label to each feature (first 10 as above)

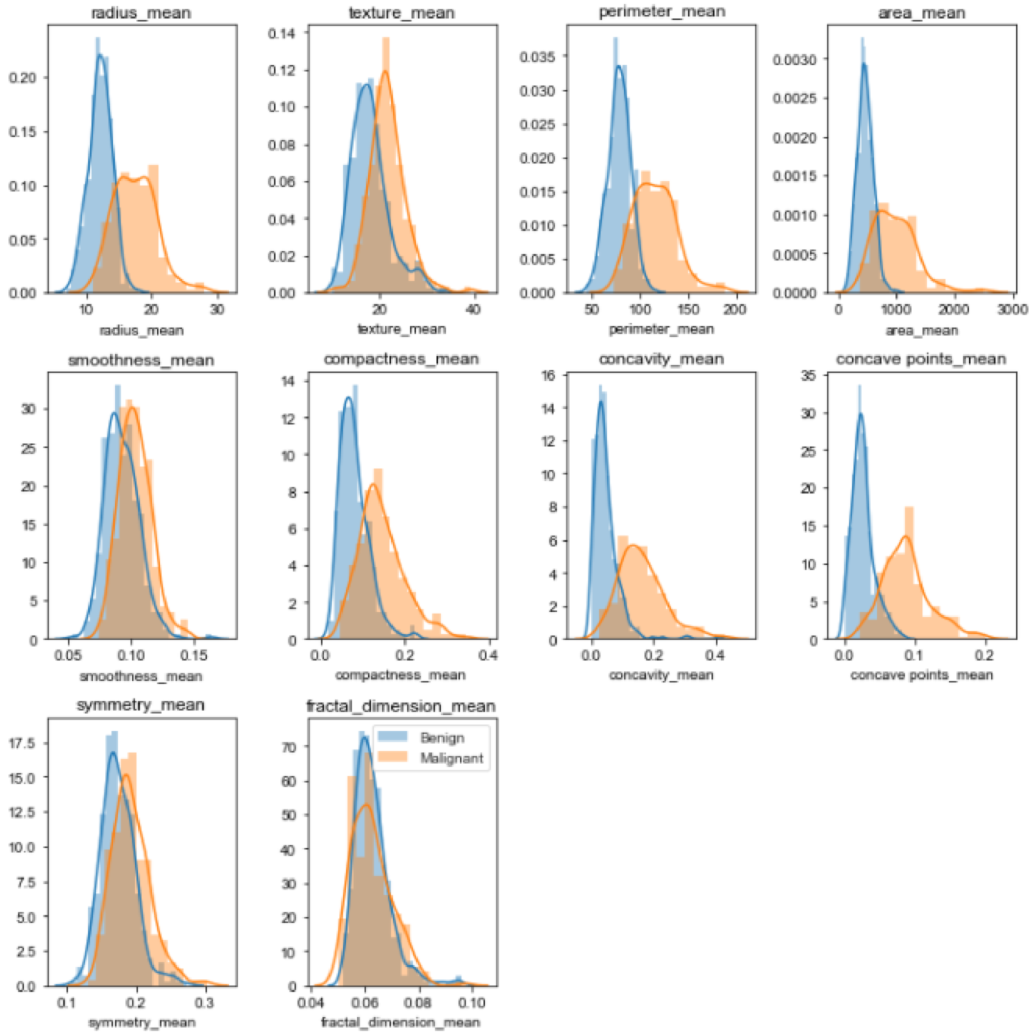


Figure 3. Plots of the first 10 features compared with each class label (B/M)

The ratios of the class labels are as following, Benign: 357, Malignant: 212. Each of them have an effect on each feature which can be differentiated. For instance,(Figure 4) malign cells have greater radius than benign cells and (Figure 5) malign cells have greater concave points_mean than benign cells.

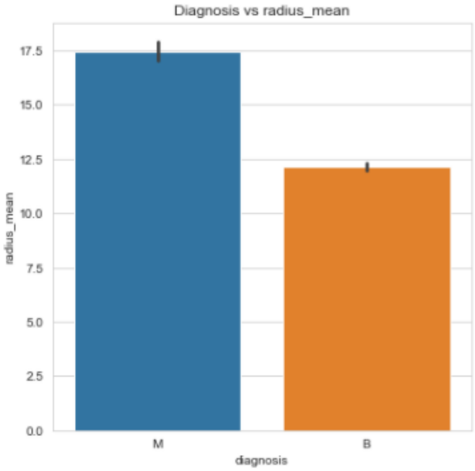


Figure 4. Class Vs Radius Mean

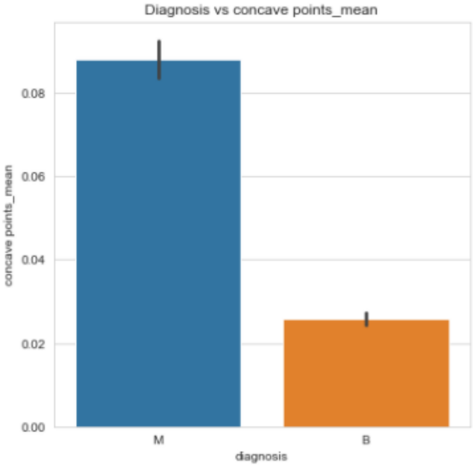


Figure 5. Class Vs Concave points_mean

Running a sample variable important test (Sci-kit Learn package), we can identify the most important features (Figure 6). As a result, we can understand the effect low-rank approximation along with application of LASS regularization, which we will experiment in later stages.

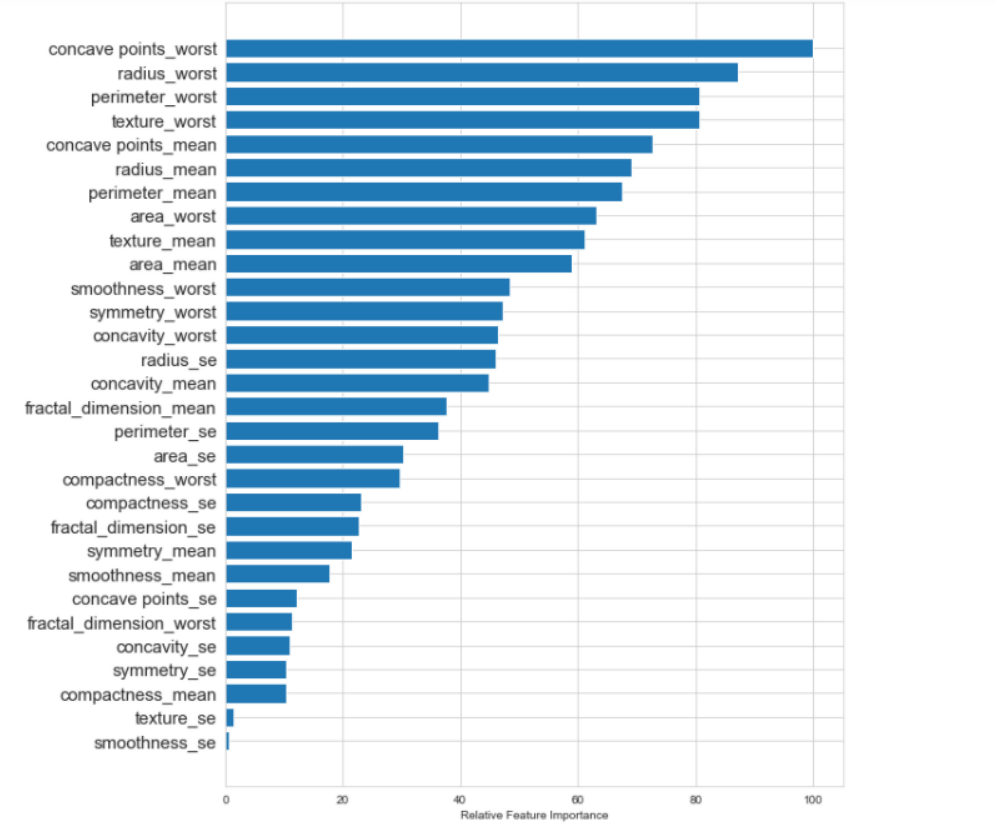


Figure 6. Variable importance

2. Preliminary Experimentation Results

Technique	Misclassification Rate (%)	R-Squared
LASSO Regression	59.61	43.55
Ridge Regression	54.19	22.60
Low Rank Approximation	53.77	21.07

The Above experimentation proves that least squared losses do not work well with classification problems.

3. Next Steps

As per this timeline, the project has moved in the rate as expected. For the next update, the below techniques will be implemented moving forward on the dataset.

1. Implementation of Hinge Loss

2. KNN

3. SVM

4. Neural Networks

4. References

[SSBD (https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/)] S. Shalev-Shwartz and

S. Ben-David, Understanding Machine Learning, CUP, 2014

[Data (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))]

Breast Cancer Dataset