# CS 5830 - Project 7

Suyash Mhetre, AJ Walters

**INTRODUCTION**

Our analysis aimed to investigate logistic regression and support vector machines as useful methods of predicting disease. We used the Wisconsin Breast Cancer dataset and the Cleveland Heart Disease dataset, and applied both data science methods to determine which method produces a model with high predictive power measured by a high F1 score. This model could help physicians in diagnosing their patient's diseases and providing an early diagnosis.

**DATASET 1:** Wisconsin Breast Cancer Dataset. AJ Walters.

Breast cancer is a disease that affects 1 in 8 women, and has average survival rates ranging from 27% to 99%, depending on the severity of the cancer and how far it has spread. It is therefore imperative that breast cancer is diagnosed early. This dataset represents images of fine needle aspirates of a breast mass, with attributes computed for each cell nucleus. These 30 attributes include the mean, standard error, and worst values for features such as radius, area, and concavity. The dataset can be found here, and further information can be found here. The target variable is whether or not the cancer is malignant (0, 37% of the data) or benign (1, 63% of the data). There are 569 samples, provided through the sci-kit learn datasets package.

**ANALYSIS TECHNIQUE 1**

We first investigated the impact of normalization and standardization on feature selection and F1 scores. We used the sklearn.feature_selection SelectKBest and chi2 packages to identify which attributes would best correlate to the target variable. Standardization had to be shifted to be non-zero in order to use the chi2 feature selection tool. We then visualized the best combinations using seaborn. Next, we used the sklearn.model_selection train_test_split tool to train both SVM and Logistic Regression models on the original, normalized, and standardized data, using the whole dataset and various subsets. We used F1 scores from the sklearn metrics.f1_score package to identify the best models. To observe decision boundaries, we used matplotlib to predict the class on random data, then plot it overlayed with the original data. To characterize the ability of the model to generalize, we did 100x cross fold validation with a 30% hold out and plotted the F1 scores to determine if the average F1 score was acceptable. Linear, polynomial, and RBF kernels were attempted and visualized. A range of class weights were tested on the whole, standardized dataset to see if the F1 score could be improved. Finally, runtime performance of SVM and Logistic regression was compared.

**RESULTS 1**

Figure 1 presents the calculated F1 scores for SVM and Logistic regression under various conditions. It was identified that the highest F1 score was 0.9908 when the entire dataset was standardized and then fed into the SVM model. The highest F1 score for a subset of the data was 0.9680 for both SVM and Logistic Regression on the mean area & worst area. The three best pairs identified by SelectKBest are visualized in Figure 2. Figure 3 presents a visualization of the decision boundaries identified by SVM and Logistic Regression under various conditions. The linear SVM and scatter plot Logistic Regression for the mean area & worst area subset without normalization or standardization have the same decision boundary. See slides for more results.

Link to presentation

| F1 Scores | SVM | | | | Logistic Regression | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| No change | 0.9179 | 0.9680 | 0.9310 | 0.8870 | 0.9767 | 0.9680 | 0.8926 | 0.8745 |
| Normalization | 0.9863 | 0.9191 | 0.9596 | 0.9327 | 0.9818 | 0.9114 | 0.9554 | 0.92511 |
| Standardization | 0.9908 | 0.9553 | 0.9550 | 0.9266 | 0.9863 | 0.9636 | 0.9455 | 0.9224 |

*Figure 1. F1 scores for SVM and Logistic Regression. 1: whole dataset. 2: mean area & worst area. 3: mean concave points & worst concave points. 4: mean concavity & mean concave points*

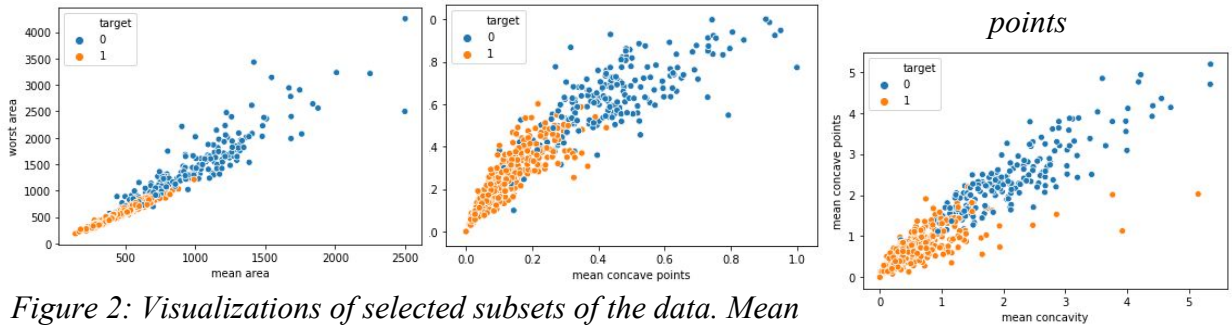


*Figure 2: Visualizations of selected subsets of the data. Mean area vs worst area using original data. Mean concave points vs worst concave points using normalized data. Mean concavity and mean concave points using standardized data.*

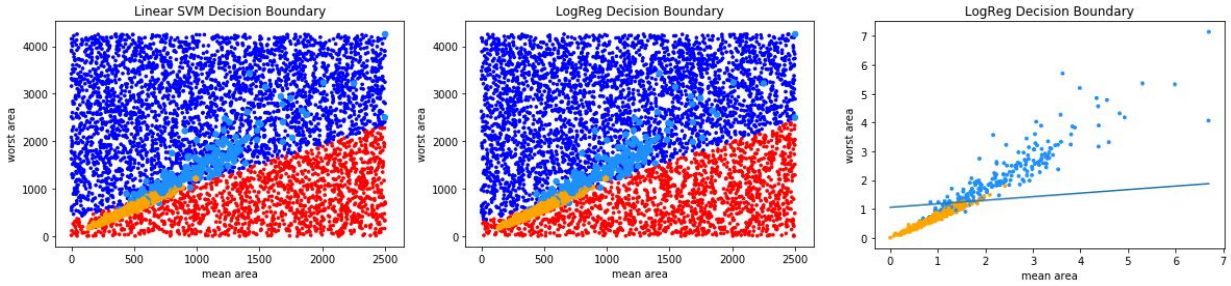


*Figure 3. Visualization of decision boundaries of linear SVM and Logistic Regression. Linear SVM and Logistic Regression identify the same boundary when the data is not normalized or standardized. The rightmost figure shows how the boundary changes when data is standardized.*

**DATASET 2:** Cleveland Heart Disease Dataset. Suyash Mhetre.

This dataset is from 1988 describing patients with and without heart disease. It has 14 attributes, including whether an individual has heart disease and at what level. We processed our data and based our classifier simply on the presence of heart disease, disregarding the different possible classifications of heart disease.

**ANALYSIS TECHNIQUE 2**

We first tried to analyze the significance of variables that can better analyze the target variable.For predicting whether the patient has heart disease we used Logistic Regression and SVM. We have standardized the dataset and used PCA for feature selection.

Link to presentation

# CS 5830 - Project 7

Suyash Mhetre, AJ Walters

**RESULTS**

The best F1-score for Logistic Regression is 86.23% and for SVM is 83.67%. *Figure1* shows the regression line for Logistic Regression and *Figure 2* shows the SVM decision boundary for age and thal.We found that there is a runtime difference between Logistic Regression and SVM.

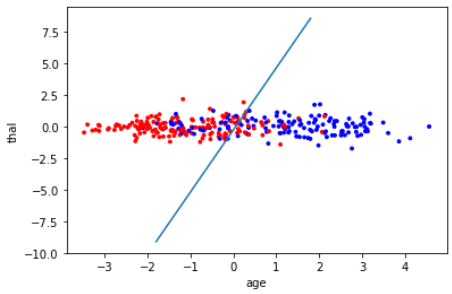


*Figure1*

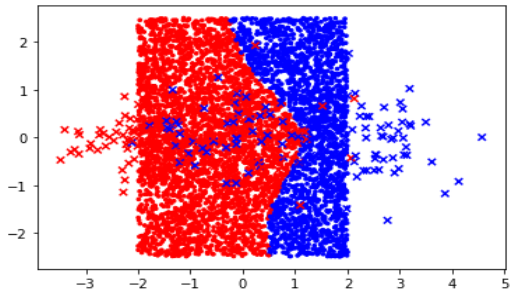


*Figure2*

**CONCLUSION**

A model that could identify malignant breast cancer samples was identified with an F1 score of 0.9908. This model was found to generalize well, with an F1 score over 0.95 across 100 iterations with a 30% hold out. Weighting the malignant class (0) 2:1 over the benign class (1) was shown to improve F1 scores above 0.995. On average Logistic Regression was found to have a slightly longer runtime: average LogReg/SVM runtime =1.014, n=10.

A model that could identify heart disease patients was identified with an F1 score of 0.8623 for Logistic Regression and 0.8367 for SVM.

Link to presentation