SVM and Logistic Regression on Healthcare Datasets

Suyash Mhetre and AJ Walters

Wisconsin Breast Cancer Dataset:

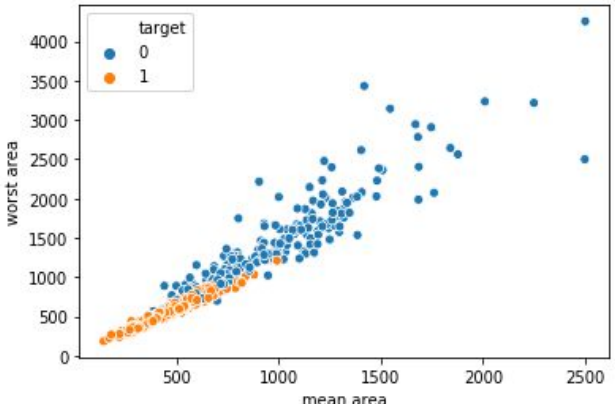
- Affects 1 in 8 women
- Survival rates of 27% to 99%, depending on stage of cancer
- Early diagnosis increases chance of survival
- Dataset compiled from images of nucleus of breast cancer cells
- 10 features:
  - Radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension
- 30 attributes total: mean, standard error, worst values for each feature
- Malignant: class 0, 37% of data
- Benign: class 1, 63% of data

F1 scores

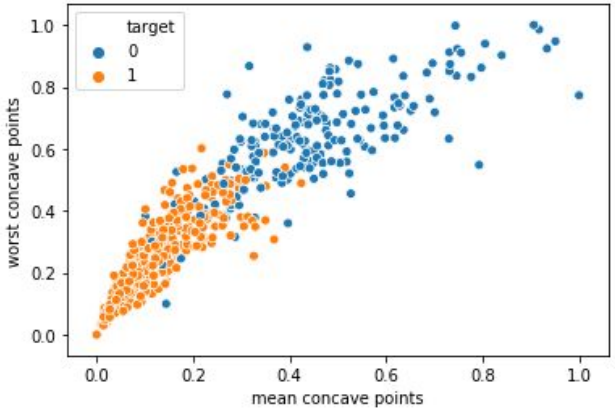
F1 Scores	SVM				Logistic Regression			
	1	2	3	4	1	2	3	4
No change	0.9179	0.9680	0.9310	0.8870	0.9767	0.9680	0.8926	0.8745
Normalization	0.9863	0.9191	0.9596	0.9327	0.9818	0.9114	0.9554	0.92511
Standardization	0.9908	0.9553	0.9550	0.9266	0.9863	0.9636	0.9455	0.9224

1: whole dataset. 2: mean area & worst area. 3: mean concave points & worst concave points. 4: mean concavity & mean concave points

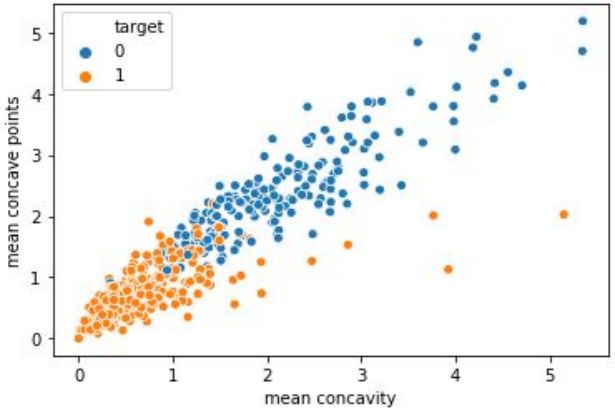
Feature selection: SelectKBest + chi2



Original data

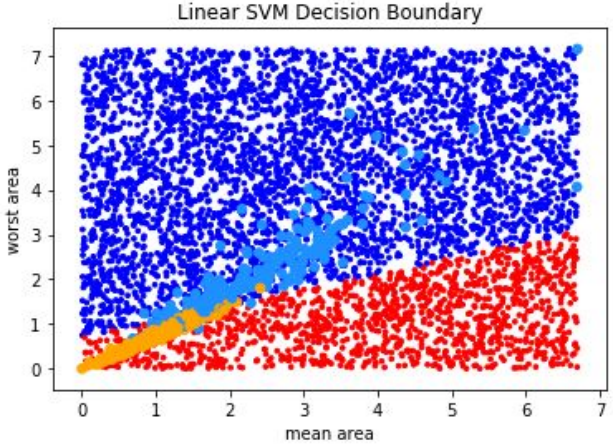


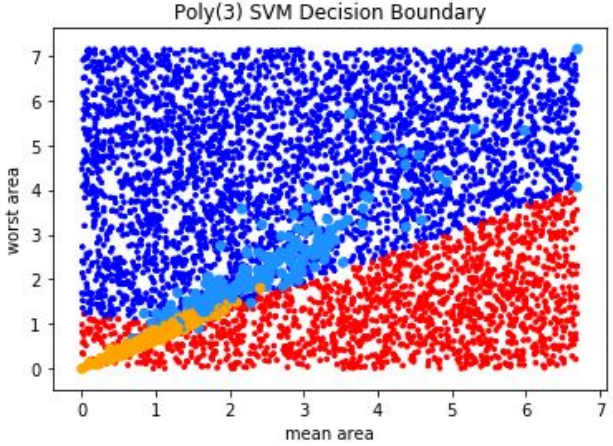
Normalized data

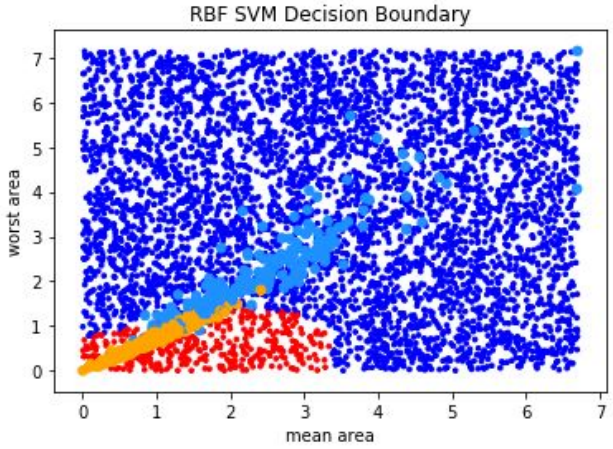


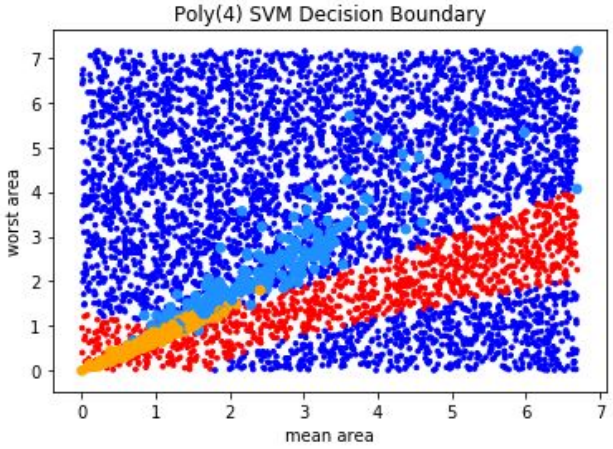
Standardized data

Visualizing Decision Boundaries (standardized data)

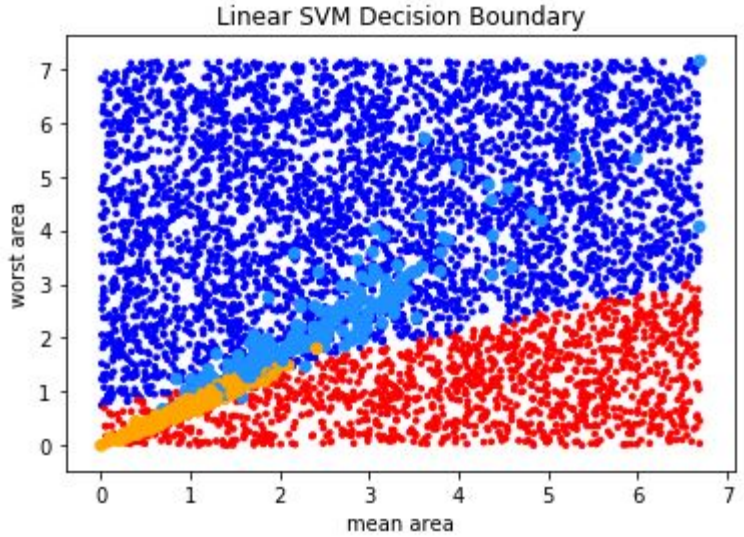


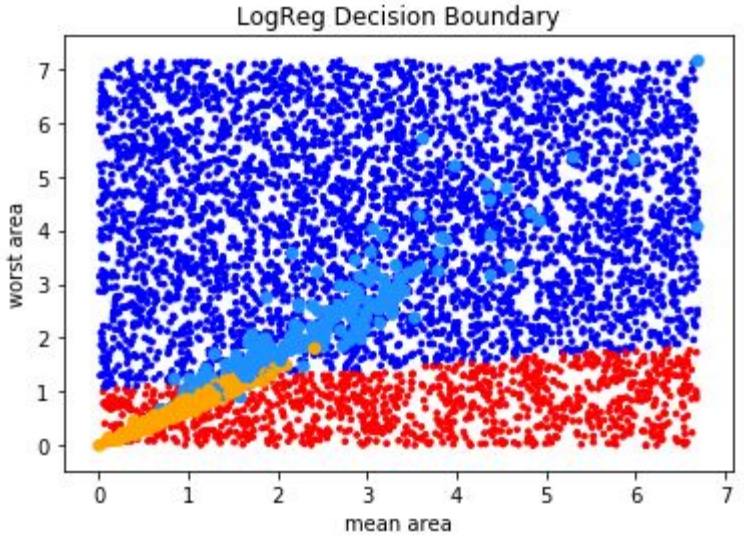




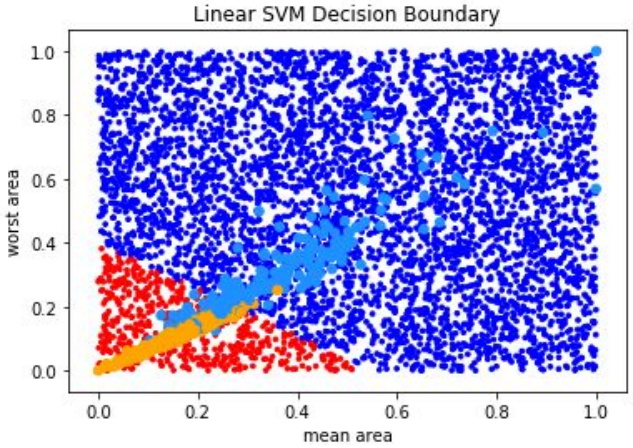


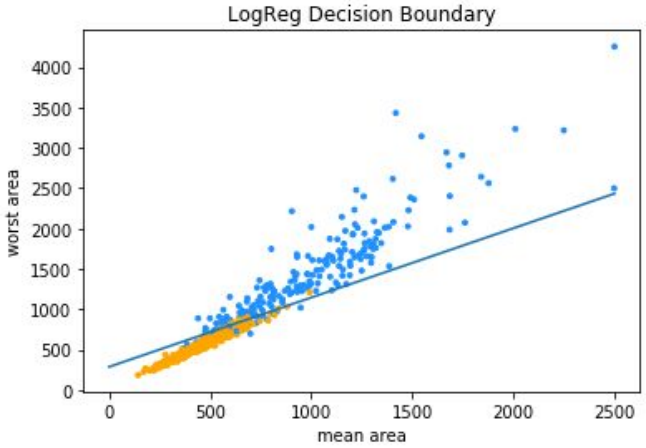
Visualizing Decision Boundaries (standardized data)

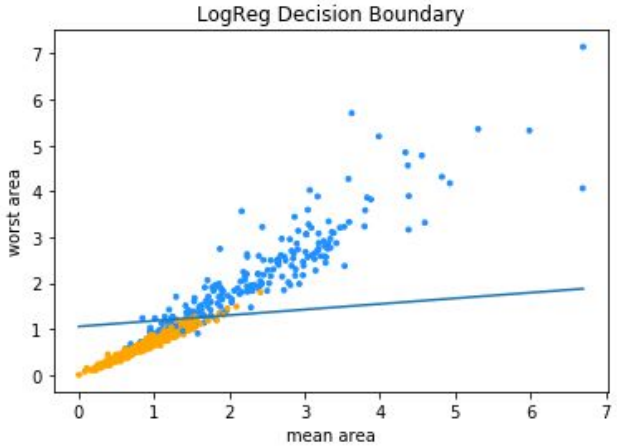




Visualizing Decision Boundaries





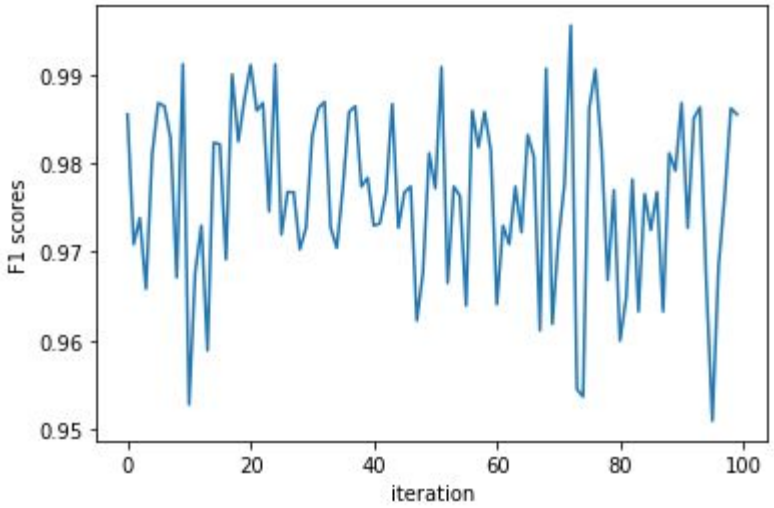


Normalized data

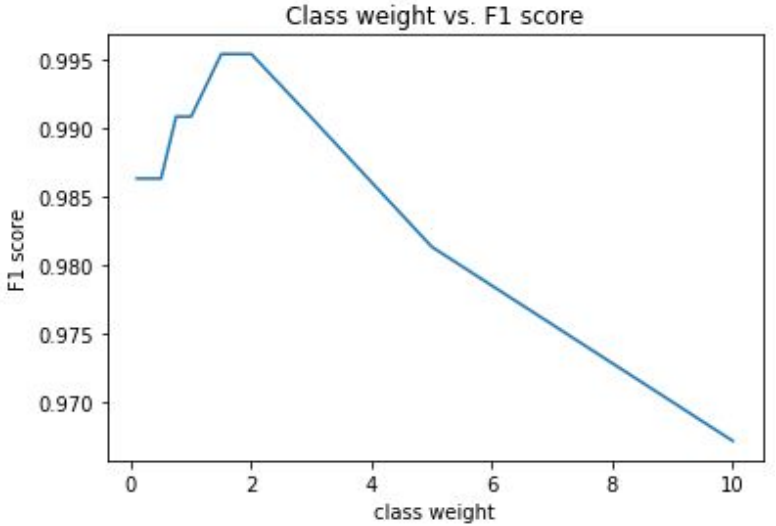
Original data

Standardized data

Generalization and Class Weights



100x, 30% hold out. Random seed.



30% hold out. Static seed.

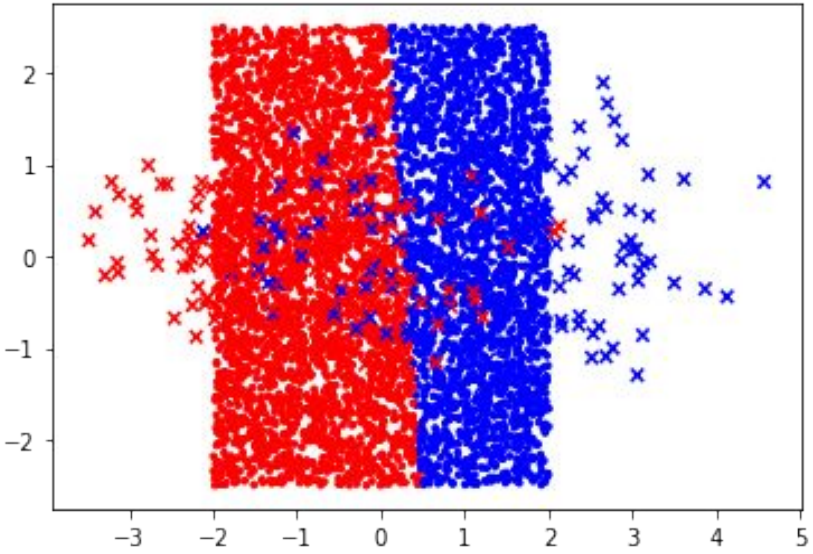
Cleveland Heart Disease Dataset

F1 Score

Logistic Regression: 86.23%
SVM: 83.67%

Decision Boundaries

Linear SVM



Logistic Regression