Springboard Intensive Data Science Course

Breast Cancer Classification Project

Mentor: Pavitraa Parthasarathy
Student: Jing-Rui Li

Contents

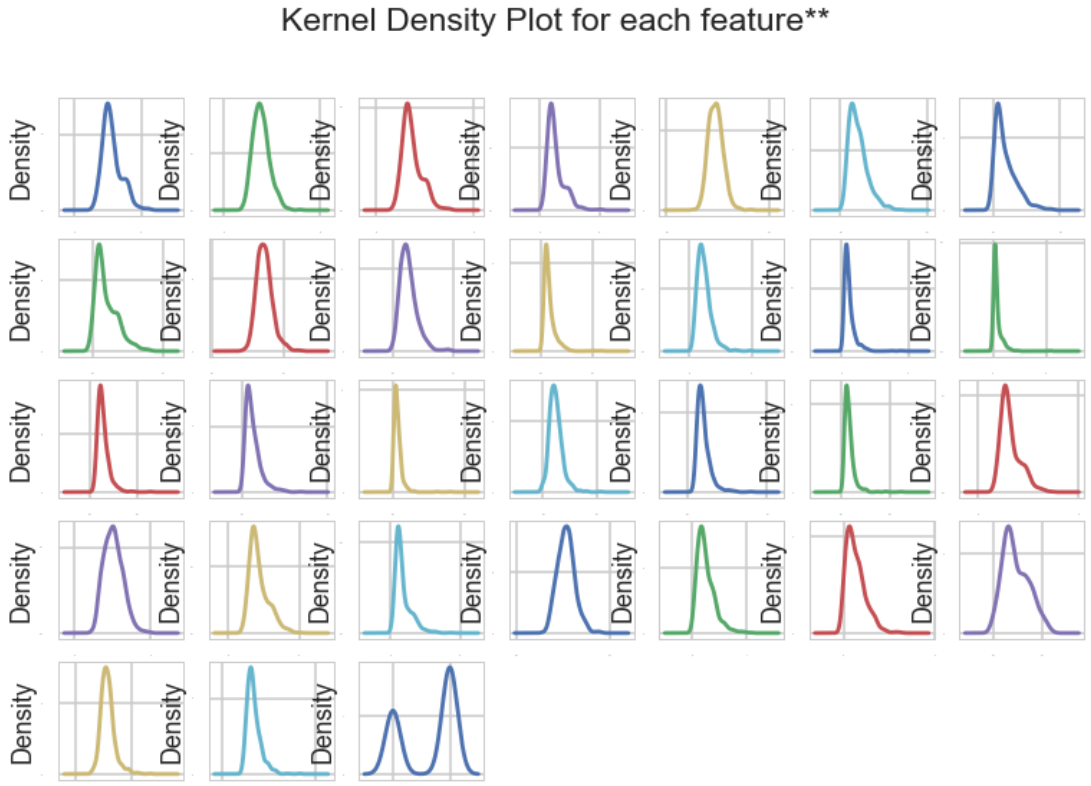
Background

Problem	Breast cancer is one of the most dangerous form of cancers in the world. If breast cancer is recognized and treated early, it is almost curable, with a 80 to 90% 5-year survival rate when treated with simple excision.
Goals	The Breast Cancer project aims to develop a system that takes breast features as inputs, and outputs the likelihood that the patient is malignant, in order to facilitate early diagnosis of breast cancer.
Solution	Compare the accuracy of different algorithms and find the trained classifier that perform the most accurate result. In the future, the trained model will be used to compute the likelihood that the lesion is malignant and return this percentage to the user.

Dataset Introduction and Visualization

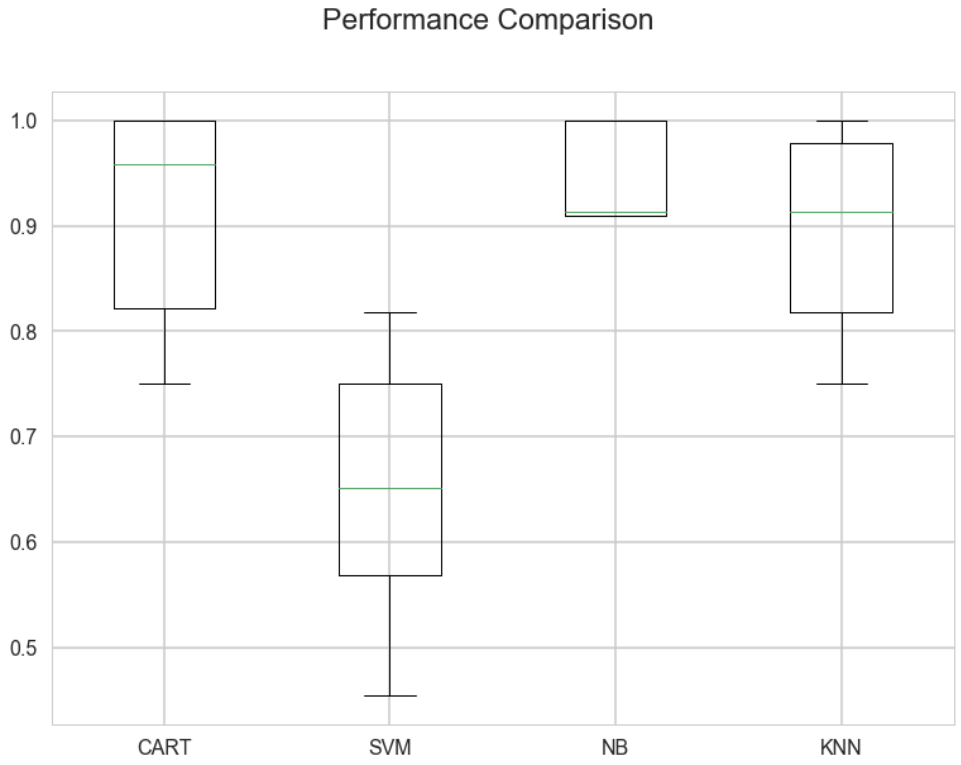
Data Set Characteristics

Number of Instances	Number of Attribute
569	30
212 - Malignant, 357 - Benign	Numeric, predictive attributes and the class

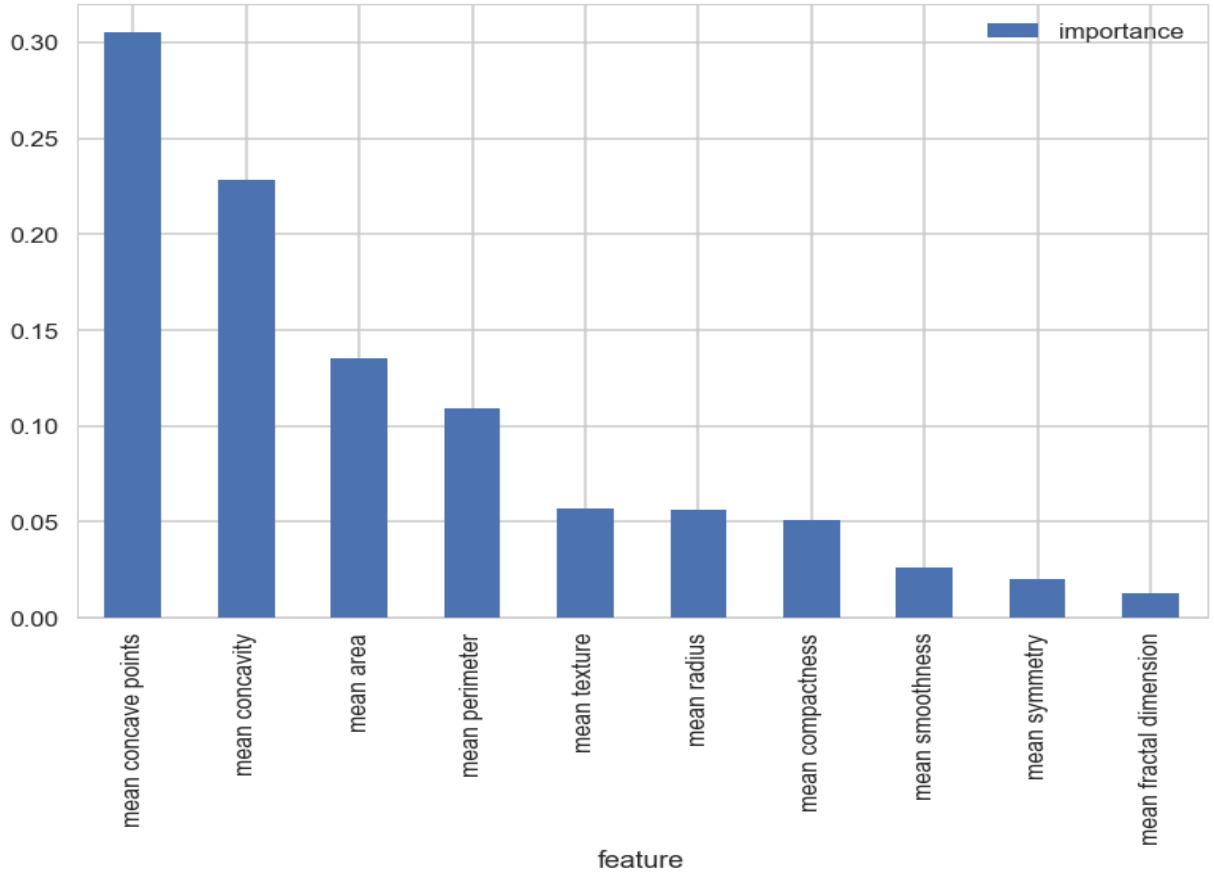


- ***The density plots above are for 30 features as follows : 'mean radius' ,'mean texture' ,'mean perimeter' ,'mean area','mean smoothness' ,'mean compactness', 'mean concavity','mean concave points' ,'mean symmetry' ,'mean fractal dimension','radius error' 'texture error' 'perimeter error' ,'area error','smoothness error' ,'compactness error' 'concavity error','concave points error', 'symmetry error', 'fractal dimension error','worst radius' 'worst texture', 'worst perimeter' 'worst area','worst smoothness' 'worst compactness' 'worst concavity','worst concave points' ,'worst symmetry' ,'worst fractal dimension','target(diagnosis)'.***

Algorithm Comparison

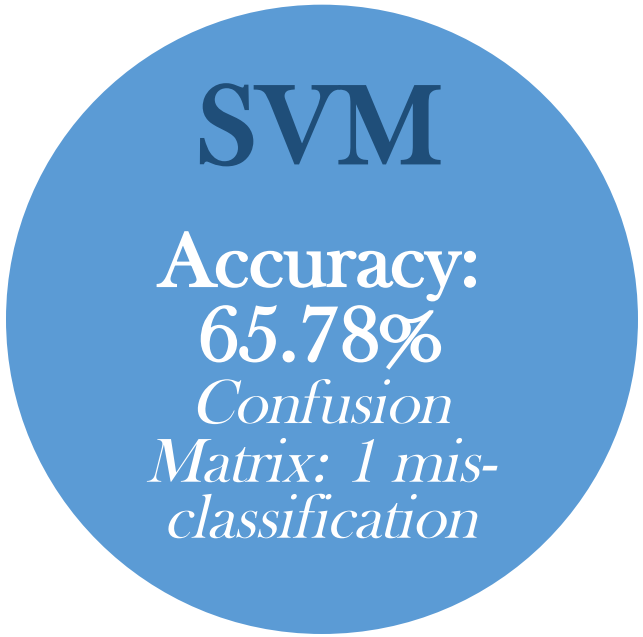


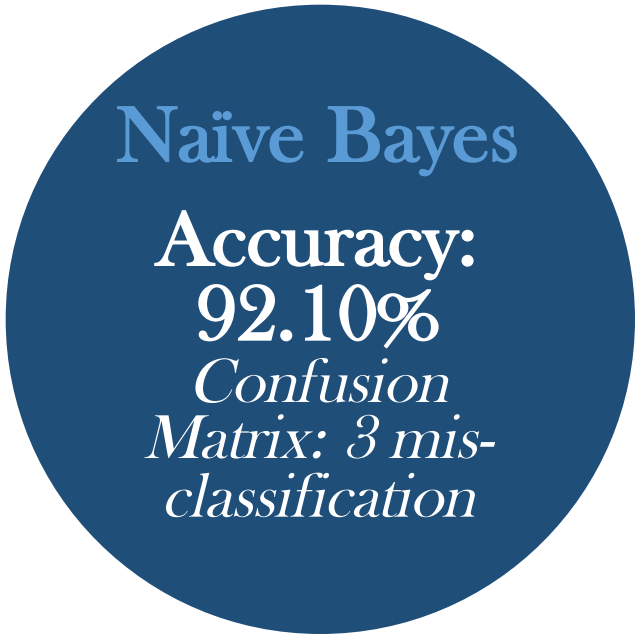
Ten-Fold Cross validation comparison between CART, SVM, NB and KNN algorithms



Importance ranking for the top features within the dataset using random forest algorithms

Results on Test Data





From the baseline processing, I selected SVM out of four algorithms as the most accurate model for forecasting the breast cancer data. Using the SVM tuning method, I further selected the parameters with RBF kernel and C=2.0. Finally, I use the selected parameters to estimate the accuracy of the test dataset and compare the result with Naïve Bayes algorithm.

Conclusion

- In this project, I compared the cross-validation score among four different algorithms. Through the comparison, I found that SVM did not perform well before standardize the data, however, it turned out to be the best estimation algorithm with cross validation score after standardization.
- I did tuning SVM and select RBF and C=2.0 as the best parameters and I used the best parameters and SVM algorithm on the test data and received an accuracy of 65.78%
- I did random forest analysis and found the importance ranking of the features