PREDICTING BREAST CANCER

Using the Wisconsin Breast Cancer Dataset

LESLIE MANRIQUE

PREDICTIVE ANALYTICS

professor Anasse Bari, Ph.D

December 20, 2016

Abstract

The increasing amounts of unstructured and structured medical data, such as those that exist in patient health records, doctor's notes, wearable devices, and Electronic Medical Records (EMR) pose the question to many medical professionals, "What can we do with all this data?" The answer lies in predictive analytics. The goal is to use available data to improve the lives of patients and to allow collaboration between different medical fields. Due to decreasing costs, and faster implementations, several biomedical industries exist that are utilizing big data to solve problems. One of the many use cases of predictive analytics in health care is projecting diagnosis of a certain disease. The importance in diagnosis using machine learning, is to expedite the treatments of patients and to reduce the cost. Also, if the diagnosis is fitted into a large dataset, we can potentially eliminate human error and have a more accurate diagnosis. In this paper, we will this discuss how predictive analytics is used to diagnose breast cancer by utilizing the Wisconsin Breast Cancer Dataset. We will focus on classification and evaluating several supervised classification models. I have utilized kaggler Buddhini Waidyawansa's IPython Notebook as a starting point to my analysis. I have built on top of their process and added my own analysis.

Dataset

The dataset is distributed by Kaggle[1] and is also available in the UCI Machine Learning Repository [2]. The features described are take from fine needle aspirations (FNA) of breast mass and the analysis of the cell nuclei. Fine Needle Aspiration is a minimally invasive biopsy that requires a needle inserted into abnormal tissue [3]. This procedure is done to either confirm or rule out cancer as a diagnosis. There are ten real valued features radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values),

perimeter, area, smoothness (local variation in radius lengths) , compactness (perimeter^2 / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, fractal dimension ("coastline approximation" - 1) [2]. 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 [1]. There are two class labels for the data, M for malignant and B for benign, as well as 569 records, with no missing values.

Figure 1

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	conc poin
100	862717	M	13.610	24.98	88.05	582.7	0.09488	0.08511	0.08625	0.04
101	862722	B	6.981	13.43	43.79	143.5	0.11700	0.07568	0.00000	0.00
102	862965	B	12.180	20.52	77.22	458.7	0.08013	0.04038	0.02383	0.01
103	862980	B	9.876	19.40	63.95	298.3	0.10050	0.09697	0.06154	0.03
104	862989	B	10.490	19.29	67.41	336.1	0.09989	0.08578	0.02995	0.01

Data Preparation

To better understand our data, we need to visualize how our classes (Malignant or Benign) are distributed in our data set. Figure 2 shows that of the 569 people represented in our data, about 350 people had a benign diagnosis. The rest had a malignant diagnosis. We then filter out our data and separate it into malignant and benign diagnosis', in order to reveal some insights. A careful analysis of figure 3 (we are only showing the means or the first 10 features) shows us that a greater distinction between diagnosis and nuclear features exists in cell radius, perimeter, area, compactness, concavity and concave points; meaning that these features show a preference for being malignant or benign. This is a good starting point to our feature selection, although we won't be analyzing features of just the means of the cell attributes. Our assumption is that the means can describe the rest of the features and perhaps if we do perform a feature selection, these features will be part of it. Note: Feature selection will be discussed below.

Figure 2

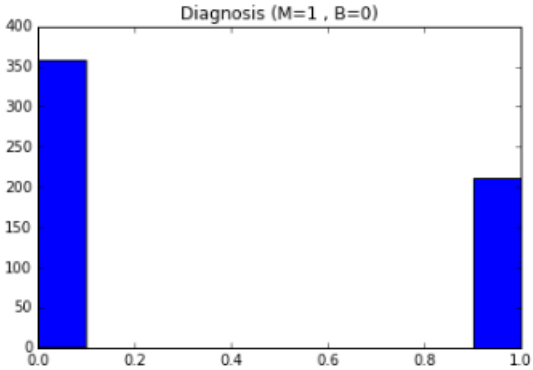
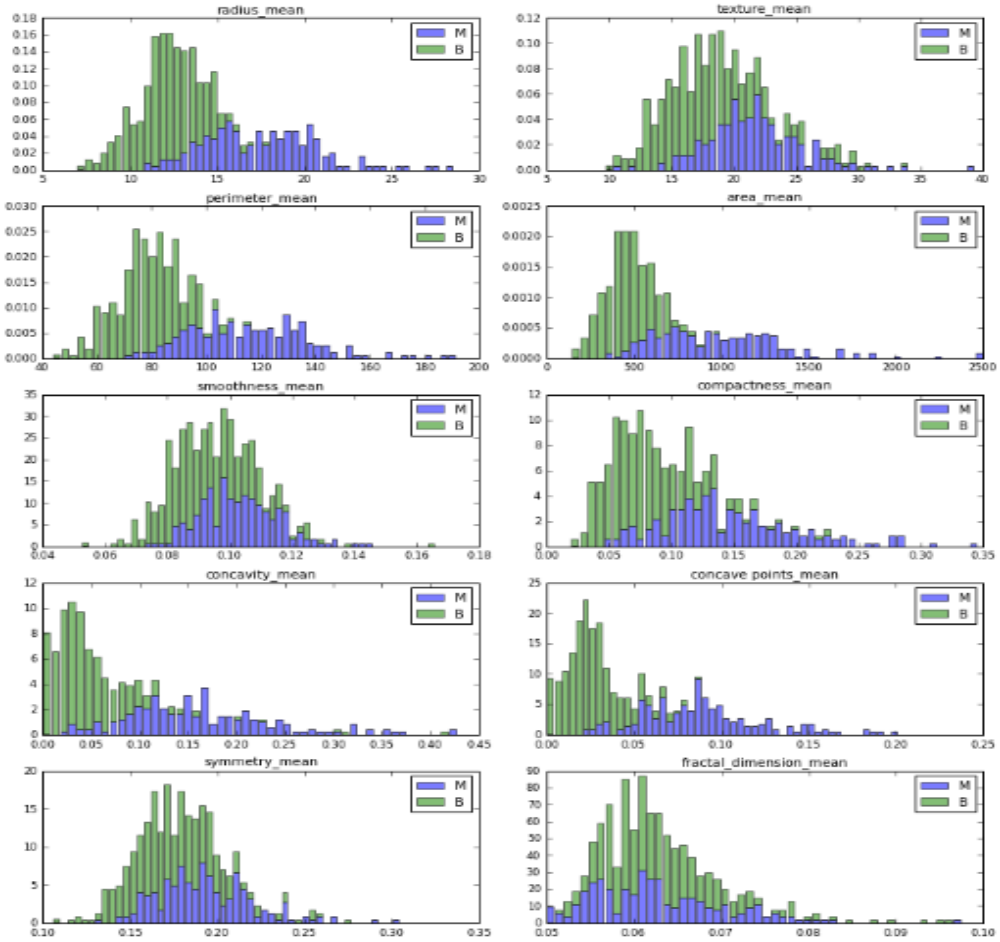


Figure 3

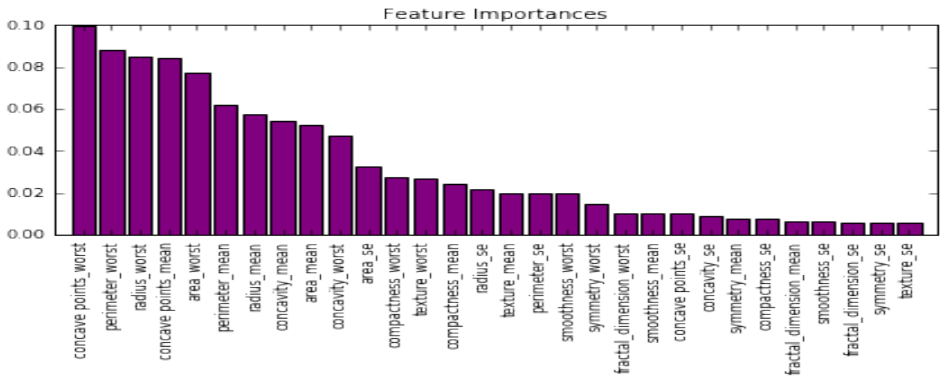


Data Reduction: Feature Selection

Feature selection is performed on data as a means of dimensionality reduction. However, reducing the data size and providing a smaller data set is not the only benefit of feature selection. Feature selection gives us a simpler model, with shorter computation time. Most importantly, feature selection allows us to lower our chances in over fitting our data. Models that are overfitted have poor predictive performance, because they are tailored to predicting the training model and perform badly on unseen data. To prevent overfitting, we will subset the data by figuring out what the most important features are. Our data will be training on the most important features.

Figure 4 is a bar graph of our 30 features, in descending order. This was obtained by using an extremely randomized trees (extra-trees) classifier. Extra trees are similar to a random forest model (discussed later) but instead picks decision boundaries at random, rather than the best boundary. This is good for our purposes, because extra trees have a good performance. We observe that the bottom third portion of the graph includes many features of type standard error (se). The rest is a mix of worst and mean, with worst being more prevalent in the first third portion of the graph. Concave points worst is the most discriminative feature in our dataset. We will use the first 10 features in our sorted important features list, as our features for developing our classification models.

Figure 4



Modelling

To create our data models, we use the standard procedure of splitting our data set into a ratio of 70:30. 70% of our data will be dedicated to our training set and 30% of our data is dedicated to our test set. We employ 2 measures of evaluating the performance of our dataset. One is a simple accuracy percentage based on the comparison of the training data with the predictive values. The other is K-fold cross validation. For our model we use 10 folds. K-fold cross validation splits our data set into 10 bins. The model will be running k times where for each run 1/10 of the bins will be used as a test data set and the rest as training sets. The classifier will be applied to the training set and then evaluated on test set. Each time the model is run, the average is taken of the 10 different testing set performances. The reason we use K-fold cross validation is to make sure our data is not being over fitted, we will compare this with the accuracy.

Logistic Regression Model

The fact that we have two class labels (malignant and benign) means we can use logistic regression. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more metric (interval or ratio scale) independent variables [4].

The results for our analysis, accuracy and cross validation score are as follows:

```
Accuracy : 95.226%
Cross-Validation Score : 92.500%
Cross-Validation Score : 93.750%
Cross-Validation Score : 92.500%
Cross-Validation Score : 93.125%
Cross-Validation Score : 93.500%
Cross-Validation Score : 94.167%
Cross-Validation Score : 94.286%
Cross-Validation Score : 94.375%
Cross-Validation Score : 94.430%
Cross-Validation Score : 94.987%
```

We can see that our cross validation score is approximately 95% for our 10-fold cross validation and our accuracy is also about 95%.

K-Nearest Neighbors (KNN)

Using the K-Nearest Neighbors algorithm, we hoped to improve our results from the logistic regression model. We iterated through a list of possible K nearest neighbor values and found that our optimal K is 5, by comparing accuracies of our training data. This means that for every value predicted, it is assigned to a single nearest neighbor.

```
number of neighbors:  5
Accuracy : 94.975%
Cross-Validation Score : 95.000%
Cross-Validation Score : 96.250%
Cross-Validation Score : 94.167%
Cross-Validation Score : 93.125%
Cross-Validation Score : 92.500%
Cross-Validation Score : 93.750%
Cross-Validation Score : 93.214%
Cross-Validation Score : 93.438%
Cross-Validation Score : 93.597%
Cross-Validation Score : 94.237%
```

Our accuracy is 94.975% and our cross-validation score is not as high, but very similar.

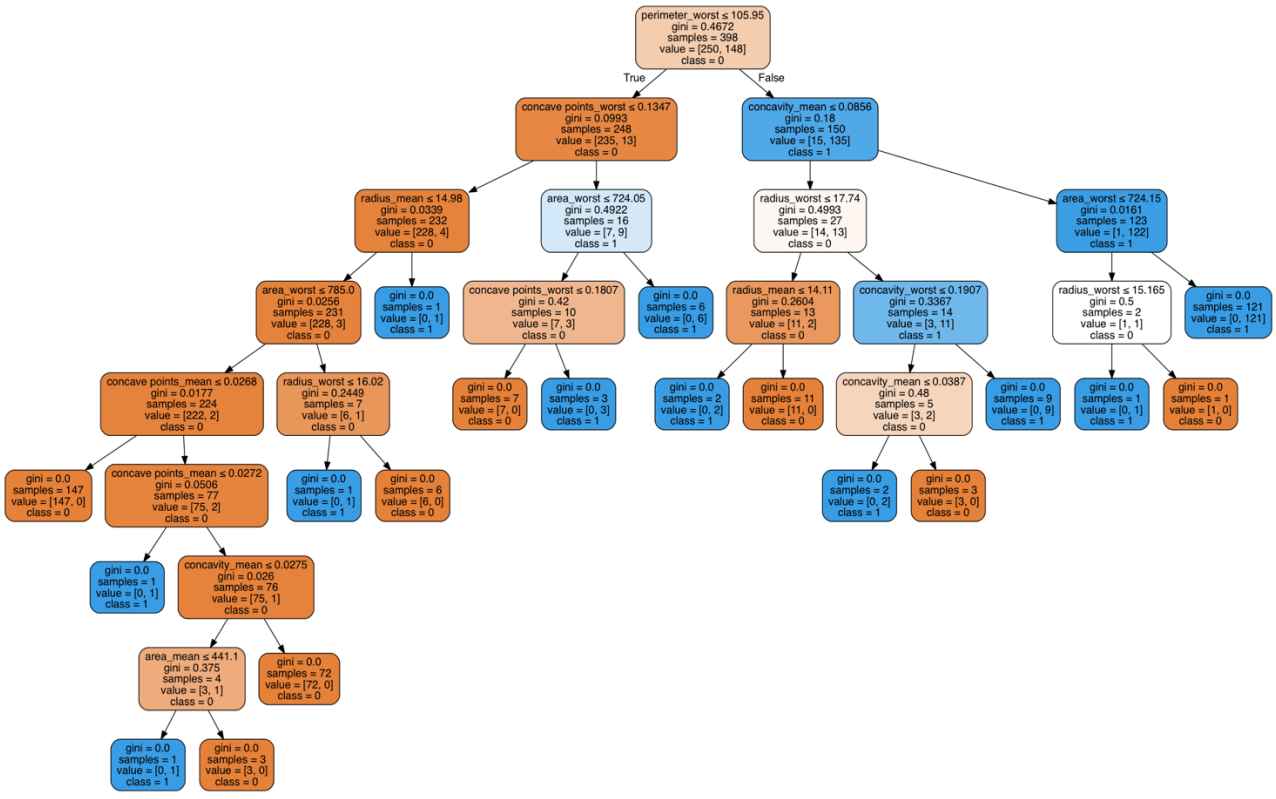
Decision Trees

A decision tree is a flow chart like graph whose nodes represent data attributes. Each leaf node in a decision tree represents the class label. Decision tree is constructed recursively by measuring the predictive power of each of it's data attributes. Therefore, we can follow the nodes of a decision tree, till we hit a leaf in order to predict a label for our data. The decision tree we constructed is shown in figure 5. Class = 0 represents a benign diagnosis while class = 1, represents a malignant diagnosis. When we trained our algorithm we received the following results:

```
Accuracy : 100.000%
Cross-Validation Score : 92.500%
Cross-Validation Score : 95.000%
Cross-Validation Score : 93.333%
Cross-Validation Score : 93.750%
Cross-Validation Score : 92.500%
Cross-Validation Score : 92.917%
Cross-Validation Score : 93.214%
Cross-Validation Score : 93.438%
Cross-Validation Score : 93.312%
Cross-Validation Score : 93.468%
```

Our accuracy is 100% but our cross validation score across the 10 folds is much less at 93%. A high accuracy compared to a lower cross validation score could signify overfitting in our training data, as cross validation attempts reduce our chances in selecting an overfitted model. The discrepancy is therefore large enough to conclude an overfitted model.

Figure 5



Random Forest

Random Forest correct the fault of decision trees as they overfit datasets. From our results, we can see that this is true because of our decreased accuracy. We also have a similar cross validation score.

```
Accuracy : 96.482%
Cross-Validation Score : 95.000%
Cross-Validation Score : 92.500%
Cross-Validation Score : 93.333%
Cross-Validation Score : 93.125%
Cross-Validation Score : 91.500%
Cross-Validation Score : 92.500%
Cross-Validation Score : 92.857%
Cross-Validation Score : 93.125%
Cross-Validation Score : 92.749%
Cross-Validation Score : 93.474%
```

Support Vector Machine

We used a linear kernel in our SVM classifier, this means that in order to classify our data points, the SVM will "draw" a line that separates our data into the two classes. From our results below, we see that it's accuracy was the same as our random forest. However, our algorithm did better in our cross validation scoring. The cross validation score is also not too different from the accuracy. So far, SVM is the best classifier. It even performs better than logistic regression in both accuracy and cross validation score.

```
Accuracy : 96.482%
Cross-Validation Score : 92.500%
Cross-Validation Score : 93.750%
Cross-Validation Score : 94.167%
Cross-Validation Score : 95.000%
Cross-Validation Score : 94.000%
Cross-Validation Score : 94.583%
Cross-Validation Score : 95.000%
Cross-Validation Score : 95.000%
Cross-Validation Score : 94.701%
Cross-Validation Score : 95.231%
```

Naïve Bayes

Naïve Bayes is a probabilistic classifier that applies Bayes theorems to make independent assumptions about features. It basically uses probabilities to determine the class labels of training data. Here, our naïve Bayes classifier, performed well, but did not achieve as high marks as our linear SVM.

```
Accuracy : 93.970%
Cross-Validation Score : 92.500%
Cross-Validation Score : 91.250%
Cross-Validation Score : 91.667%
Cross-Validation Score : 91.875%
Cross-Validation Score : 91.500%
Cross-Validation Score : 92.917%
Cross-Validation Score : 92.857%
Cross-Validation Score : 93.125%
Cross-Validation Score : 93.319%
Cross-Validation Score : 93.731%
```

Evaluations

Table 1 demonstrates the accuracy and 10-fold cross validation scores for each of our models. From our accuracy plot (figure 6) we see that Decision Trees had the highest accuracy. However, such high accuracy can be a result of overfitting, so we need to make sure that we are using the best scoring measure. We evaluate our model with 10-fold cross validation (figure 7) and realize that decision tree is no longer the best model. In fact, SVM and logistic regression rank much higher. Because SVM also performed highly in accuracy, and did slightly better than SVM in cross validation score, we conclude that this is our best model.

On our test data, the accuracy and cross validation scores of SVM decrease but still hit the 90% mark.

```
Accuracy : 94.737%
Cross-Validation Score : 94.444%
Cross-Validation Score : 94.281%
Cross-Validation Score : 92.266%
Cross-Validation Score : 89.788%
Cross-Validation Score : 90.654%
Cross-Validation Score : 90.251%
```

```
Cross-Validation Score : 90.803%
Cross-Validation Score : 91.217%
Cross-Validation Score : 90.232%
Cross-Validation Score : 90.621%
```

Table 1

	Accuracy	**10-Fold Cross Validation Score**
Logistic Regression	95.226%	94.987%
K-Nearest Neighbors (K=1)	94.975%	94.237%
Decision Trees	100.000%	93.468%
Random Forests	96.482%	93.474%
Linear SVM	96.482%	95.231%
Naïve Bayes	93.970%	93.731%

Figure 6

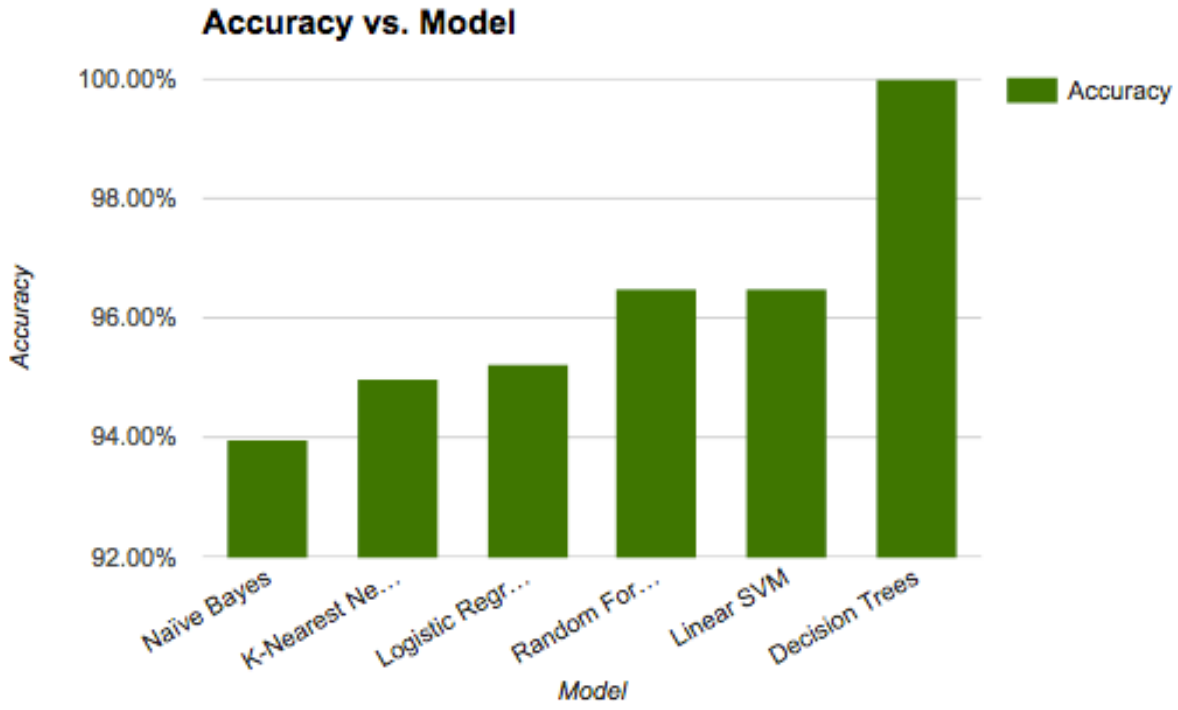
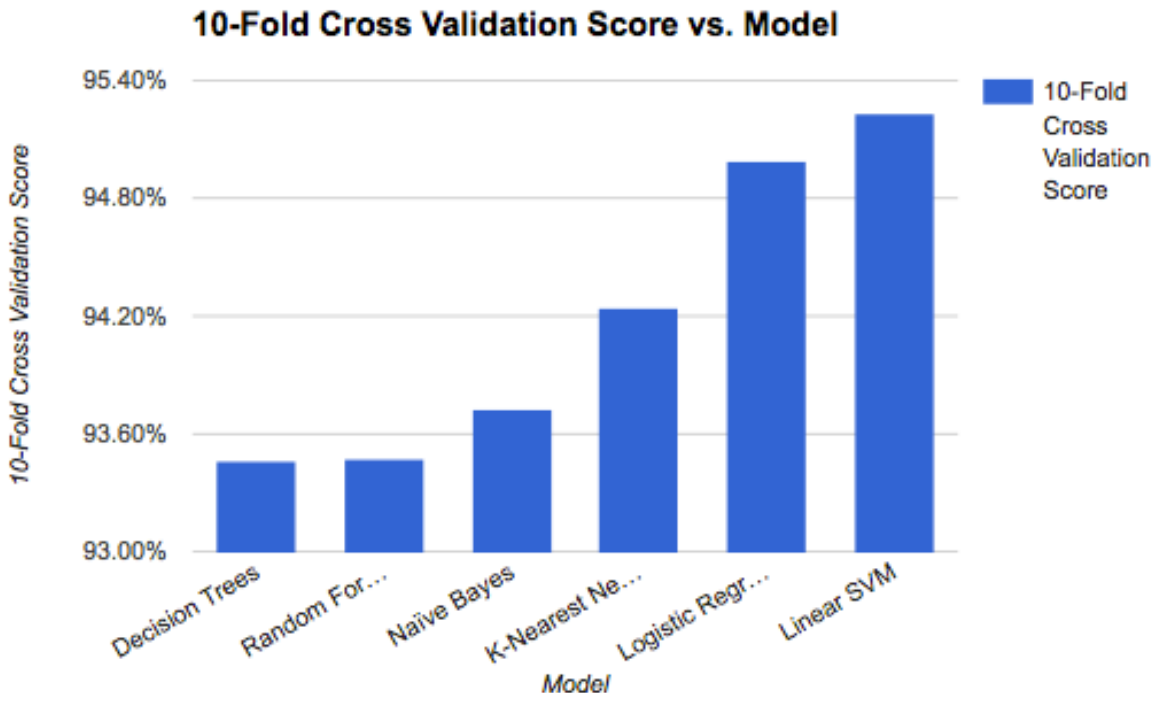


Figure 7



Conclusion

Though our algorithm did not reach above 90%, there are still other classifiers we can look into and there is still some fine tuning to be done on current measures. Perhaps the best improvement we can make is to obtain more data. Our dataset was pretty small and only included 569 rows. If we work with more data, we can have a larger training set to work with. This can in turn reduce the possibility of overfitting and leave us with more accurate results when actually deploying our algorithm to physicians or cancer researchers.

Citations

[1] UCI Machine Learning. "Breast Cancer Wisconsin (Diagnostic) Data Set | Kaggle."Breast Cancer Wisconsin (Diagnostic) Data Set | Kaggle. Kaggle, n.d. Web. 18 Dec. 2016.

[2] Lichman, M. (2013). UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)]. Irvine, CA: University of California, School of Information and Computer Science.

[3] "Fine Needle Aspiration Procedure: What to Expect." WebMD. WebMD, n.d. Web. 18 Dec. 2016.

[4] @statssolutions. "What Is Logistic Regression? - Statistics Solutions." Statistics Solutions. N.p., n.d. Web. 19 Dec. 2016.