Machine Learning Engineer Nanodegree

Capstone Project

Charu Chhimpa
August 23rd, 2017

I. Definition

Project Overview

Advancement in the field of medicine have greatly improved our quality of life which can be clearly seen from the life expectancy rate. From 1816's stethoscope to today's handheld ultrasound machines, doctors have steadily adopted technology to advance healthcare. In past decades the healthcare community has taken major steps by adopting electronic healthcare measures. Machine Learning can contribute significantly in the field of healthcare, it has great usage in cases like early disease detection, finding signs of early breakouts of epidemics, using clustering to figure out regions of epidemics, or finding the best air quality zones in countries with high air pollution.

Heart disease is the leading cause of death in the world over the past 10 years. The World Health Organization reported that heart disease is the first leading cause of death in high and low income countries. There are several methods in the literature individually to diagnosis diabetes or heart disease. There is no automated diagnosis method to diagnose Heart disease for diabetic patient based on diabetes diagnosis attributes to our knowledge. Researches have been using several data mining techniques in the diagnosis of heart disease. Diabetes is a chronic disease that occurs when the pancreas does not produce enough insulin, or when the body cannot effectively use the insulin it produces. Most of these systems have successfully employed Machine learning methods such as Naïve Bayes and Support Vector Machines for the classification purpose. Support vector machines are a modern technique in the field of machine learning and have been successfully used in different fields of application.

Problem Statement

In this project we will train a model using Support Vector Machines to predict that a human being is suffering from a heart disease. We will consider several features to do this classification.

Task : To predict that a person is suffering from a heart disease or not.

Performance : Accuracy - No. of correct predictions.

Target Function : A function that give the weights to every feature of a patient and then finally tell that the patient is sufering from heart disease or not.

Target Function Representation: A Classification Model.

Metrics

Prediction Accuracy : The accuracy score will be used as an evaluation Metrics in this case. It would be calculated using the number of data points which are classified correctly.
As it is a binary classification problem, prediction accuracy turns out to be the best evaluation metrics.
If y_hat_i is the predicted value of the i-th sample and y_i is the corresponding true value, then the fraction of correct predictions over n_samples is defined as :

$$\texttt{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

II. Analysis

Data Exploration

I am using the UCI Heart Disease dataset for training the model. There are 14 features for every patient. In particular, there is data from 4 hospitals but I will use the Cleveland Database because it is properly processed. The dataset have 297 instances. The dataset link is :
http://archive.ics.uci.edu/ml/datasets/heart+Disease .
The different feature desciption :

1. age: continuous
2. sex: categorical, 2 values {0: female, 1: male}
3. cp (chest pain type): categorical, 4 values
   {1: typical angina, 2: atypical angina, 3: non-angina,4: asymptomatic angina}
4. restbp (resting blood pressure on admission to hospital): continuous (mmHg)
5. chol (serum cholesterol level): continuous (mg/dl)
6. fbs (fasting blood sugar): categorical, 2 values
   {0: <= 120 mg/dl, 1: > 120 mg/dl}
7. restecg (resting electrocardiography): categorical, 3 values
   {0: normal, 1: ST-T wave abnormality, 2: left ventricular hypertrophy}
8. thalach (maximum heart rate achieved): continuous
9. exang (exercise induced angina): categorical, 2 values {0: no, 1: yes}
10. oldpeak (ST depression induced by exercise relative to rest): continuous
11. slope (slope of peak exercise ST segment): categorical, 3 values
    {1: upsloping, 2: flat, 3: downsloping}
12. ca (number of major vessels colored by fluoroscopy): discrete (0,1,2,3)
13. thal: categorical, 3 values {3: normal, 6: fixed defect, 7: reversible defect}
14. num (diagnosis of heart disease): categorical, 5 values
    {0: less than 50% narrowing in any major vessel, 1-4: more than 50% narrowing in 1-4 vessels}

Below is the image descibing the dataset.

```
Total patients in dataset  297
Total patients which are having disease 137
Total patients which are not having disease  160
    age  sex  restbp   chol  fbs  thalach  exang  oldpeak   ca  hd  cp_1  \
0  63.0  1.0   145.0  233.0  1.0    150.0    0.0      2.3  0.0   0     1
1  67.0  1.0   160.0  286.0  0.0    108.0    1.0      1.5  3.0   1     0
2  67.0  1.0   120.0  229.0  0.0    129.0    1.0      2.6  2.0   1     0
3  37.0  1.0   130.0  250.0  0.0    187.0    0.0      3.5  0.0   0     0
4  41.0  0.0   130.0  204.0  0.0    172.0    0.0      1.4  0.0   0     0

   cp_2  cp_3  recg_1  recg_2  slope_1  slope_3  thal_6  thal_7
0     0     0       0       1        0        1       1       0
1     0     0       0       1        0        0       0       0
2     0     0       0       1        0        0       0       1
3     0     1       0       0        0        1       0       0
4     1     0       0       1        1        0       0       0
```

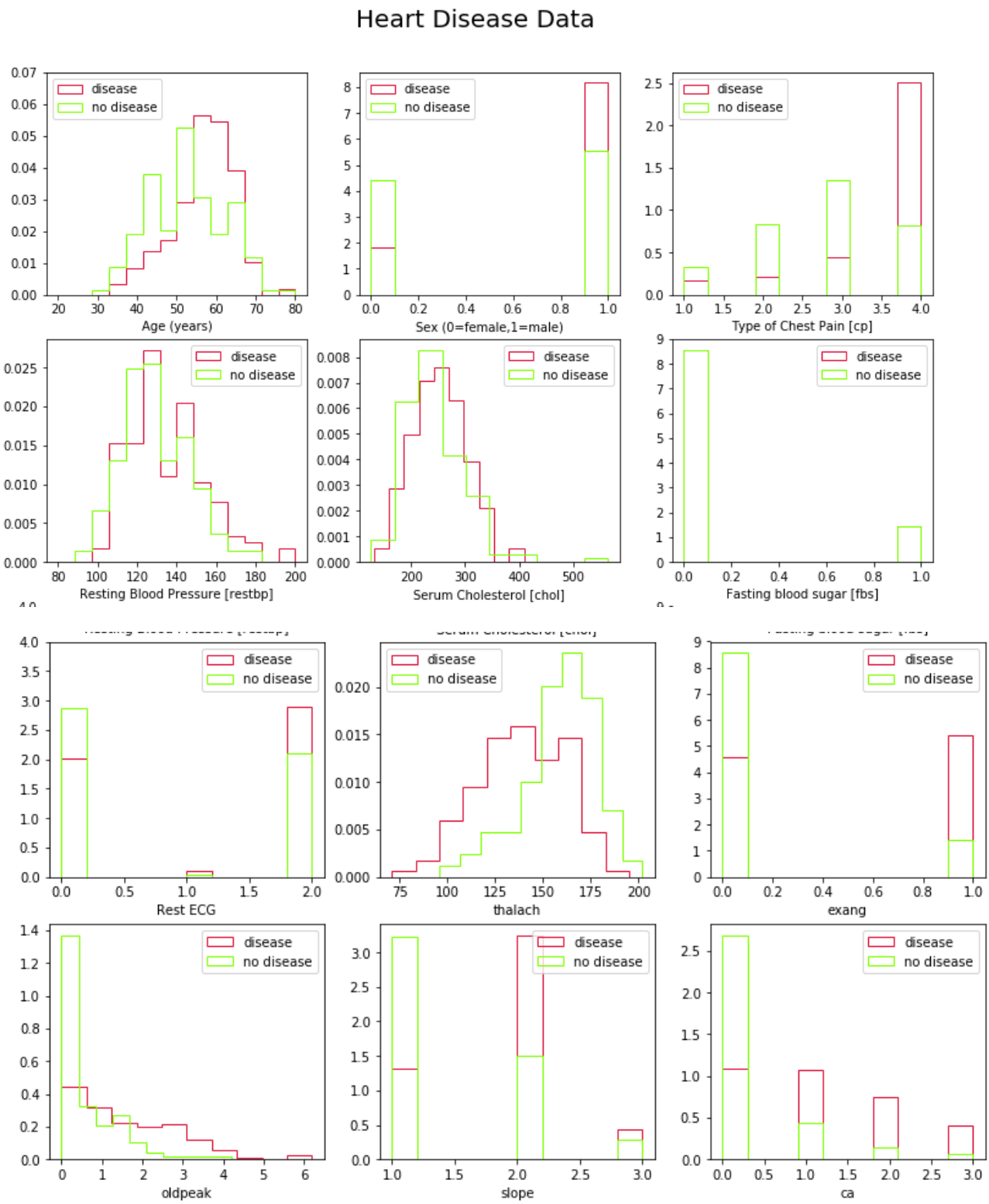
This second image describes various characteristics of the features :

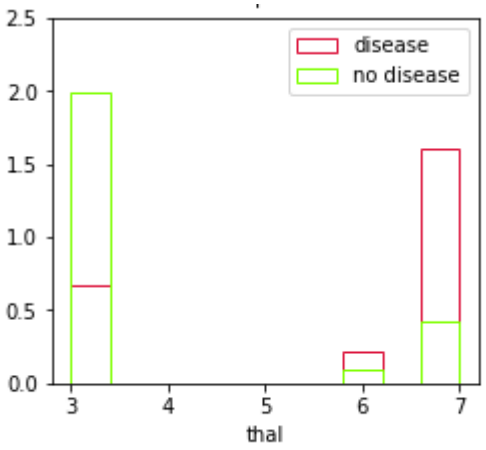
```
                age         sex      restbp        chol         fbs     thalach  \
count  297.000000  297.000000  297.000000  297.000000  297.000000  297.000000
mean    54.542088    0.676768  131.693603  247.350168    0.144781  149.599327
std      9.049736    0.468500   17.762806   51.997583    0.352474   22.941562
min     29.000000    0.000000   94.000000  126.000000    0.000000   71.000000
25%     48.000000    0.000000  120.000000  211.000000    0.000000  133.000000
50%     56.000000    1.000000  130.000000  243.000000    0.000000  153.000000
75%     61.000000    1.000000  140.000000  276.000000    0.000000  166.000000
max     77.000000    1.000000  200.000000  564.000000    1.000000  202.000000

             exang     oldpeak          ca          hd        cp_1        cp_2  \
count  297.000000  297.000000  297.000000  297.000000  297.000000  297.000000
mean     0.326599    1.055556    0.676768    0.461279    0.077441    0.164983
std      0.469761    1.166123    0.938965    0.499340    0.267741    0.371792
min      0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
25%      0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
50%      0.000000    0.800000    0.000000    0.000000    0.000000    0.000000
75%      1.000000    1.600000    1.000000    1.000000    0.000000    0.000000
max      1.000000    6.200000    3.000000    1.000000    1.000000    1.000000

             cp_3      recg_1      recg_2     slope_1     slope_3      thal_6  \
count  297.000000  297.000000  297.000000  297.000000  297.000000  297.000000
mean     0.279461    0.013468    0.491582    0.468013    0.070707    0.060606
std      0.449492    0.115462    0.500773    0.499818    0.256768    0.239009
min      0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
25%      0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
50%      0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
75%      1.000000    0.000000    1.000000    1.000000    0.000000    0.000000
max      1.000000    1.000000    1.000000    1.000000    1.000000    1.000000

           thal_7
count  297.000000
mean     0.387205
std      0.487933
min      0.000000
25%      0.000000
50%      0.000000
75%      1.000000
max      1.000000
```

Exploratory Visualization

The things that are explored in the previous section are :

Feature Distributions Compared for Disease and No Disease





Algorithms and Techniques

Algorithm : The Machine Learning Algorithm that we will useis Support Vecto Machines.

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection.

The advantages of support vector machines are: - Effective in high dimensional spaces.

- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

I have tried out several other algorithms too, but the accuracy that came out was lower than the model chosen (i.e. SVM). SVMs are well suited for this problem because we have so many features and SVMs performs quite well in high dimensional spaces. And even with SVM we can set a regularisation parameter using which we can avoid overfitting.

A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

The other algorithms that I have tried are:
- **Gaussian Naive Bayes** : This is our Benchmark model. Its explained above properly.

- **Logistic Regression** : The goal of logistic regression is to find the best fitting model to describe the relationship between the characteristic of interest (dependent variable = response or outcome variable) and a set of independen variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest:

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

where p is the probability of presence of the characteristic of interest.

- **Decision Tree Classifier** : The decision tree classifiers organized a series of test questions and conditions in a tree structure. In the decision tree, the root and internal nodes contain attribute test conditions to separate recordes that have different characteristics. All the terminal node is assigned a class lable Yes or No.

Benchmark

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

When dealing with continuous data, a typical assumption is that the continuous values associated with each class are distributed according to a Gaussian distribution. For example, suppose the training data contains a continuous attribute, x. We first segment the data by the class, and then compute the mean and variance of x in each class. Let μc be the mean of the values in x associated with class c, and let σ^2_c be the variance of the values in x associated with class c. Suppose we have collected some observation value v. Then, the probability distribution of v given a class x, p(x=v), can be computed by plugging v into the equation for a Normal distribution parameterized by μc and σ^2_c. That is,

$$p(x = v \mid c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

I'll be looking to maximize the accuracy of the predictions using Gaussian Naive Bayes model.

III. Methodology

Data Preprocessing

The features which are having more than two values are converted to dummy variables and given a different name for each variable. After converting the variables there are total of 18 variables.

Implementation

I am using the UCI Heart Disease dataset for training the model. There are 14 features for every patient.

First we trained with the Benchmark Model, Gaussian Naive Bayes. The explanation for this model is given in the section above. The dataset was divided into training and testing datasets by using sklearn functions. The model was trained and finally accuracy was calculated on the testing data.

The accuracy it came out with is 78.33 %.

To select the final model we went through some other algorithms, that are Logistic Regression and Decision Tree Classifier. These algorithms after trainning and optimizing the models came out with an accuracy of 81 and 83% respectively on the test datasets.

After all this Support Vector Machines are trained using the sklearn SVC and the same process was repeated to break the dataset into Testing and Training.

For increasing the accurcay, Support Vector Machines are optimized on accuracy score, by using cross-validation. Here we use sklearn since it includes a cross_validation method. By using cross validation score we will find the set of features that yields the best accuracy score.

Try eliminating features with a non-significant coefficient, one by one, while keeping the model deviance as low as possible. We'll use this second method for the final results.

Finally the significant features are selected and the final accuracy that we came with after the training is 85%.

Refinement

Initially we trained on the Benchmark model, which came out with an accuracy of 78.33%.

Then after that we tested upon various other models like Logistic Regression and Decision Tree Classifiers. Sklearn was used to provide libraries for these models. The final accuracies these models came wih are 81 and 83% respectively.

Then the SVMs are trained using the SVC module of Sklearn and for increasing the accurcay, Support Vector Machines are optimized on accuracy score, by using cross-validation. Here we use sklearn since it includes a cross_validation method. By using cross validation score we will find the set of features that yields the best accuracy score.

Try eliminating features with a non-significant coefficient, one by one, while keeping the model deviance as low as possible. We used this second method for the final results.

Finally the SVM model is chosen because it came out with an accuracy of 85%, which was greater than every other model.

IV. Results

Model Evaluation and Validation

The final model came out with the accuracy of 85% . The final features are selected on the basis of cross validation score.
The model was tested on the test data set here is the image showing the result :

```
('the actual lables are : ', array([0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0,
       0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1,
       1, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0]))
('the predicted labels are : ', array([0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0,
       0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0,
       1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0]))
```

I also tested the model on 10 random patients from some other hospital dataset. Below shown are the features of 10 patients and they are arranged in the order of : ["age", "sex", "cp", "restbp", "chol", "fbs", "restecg", "thalach", "exang", "oldpeak", "slope", "ca", "thal"] :

1	63	1	1	145	233	1	2	150	0	2.3	3	0	6
2	67	1	4	160	286	0	2	108	1	1.5	2	3	3
3	67	1	4	120	229	0	2	129	1	2.6	2	2	7
4	37	1	3	130	250	0	0	187	0	3.5	3	0	3
5	41	0	2	130	204	0	2	172	0	1.4	1	0	3
6	56	1	2	120	236	0	0	178	0	0.8	1	0	3
7	62	0	4	140	268	0	2	160	0	3.6	3	2	3
8	57	0	4	120	354	0	0	163	1	0.6	1	0	3
9	63	1	4	130	254	0	2	147	0	1.4	2	1	7
10	53	1	4	140	203	1	2	155	1	3.1	3	0	7

The model predicted clas for all these patients was exactly correct. This shows the robustness of our model, that this model also works on the patients from some other hospital also.

The initial model of SVMs is tuned properly using cross validation score to get a an accuracy of 85%. We also tested the model on random patient features.

Justification

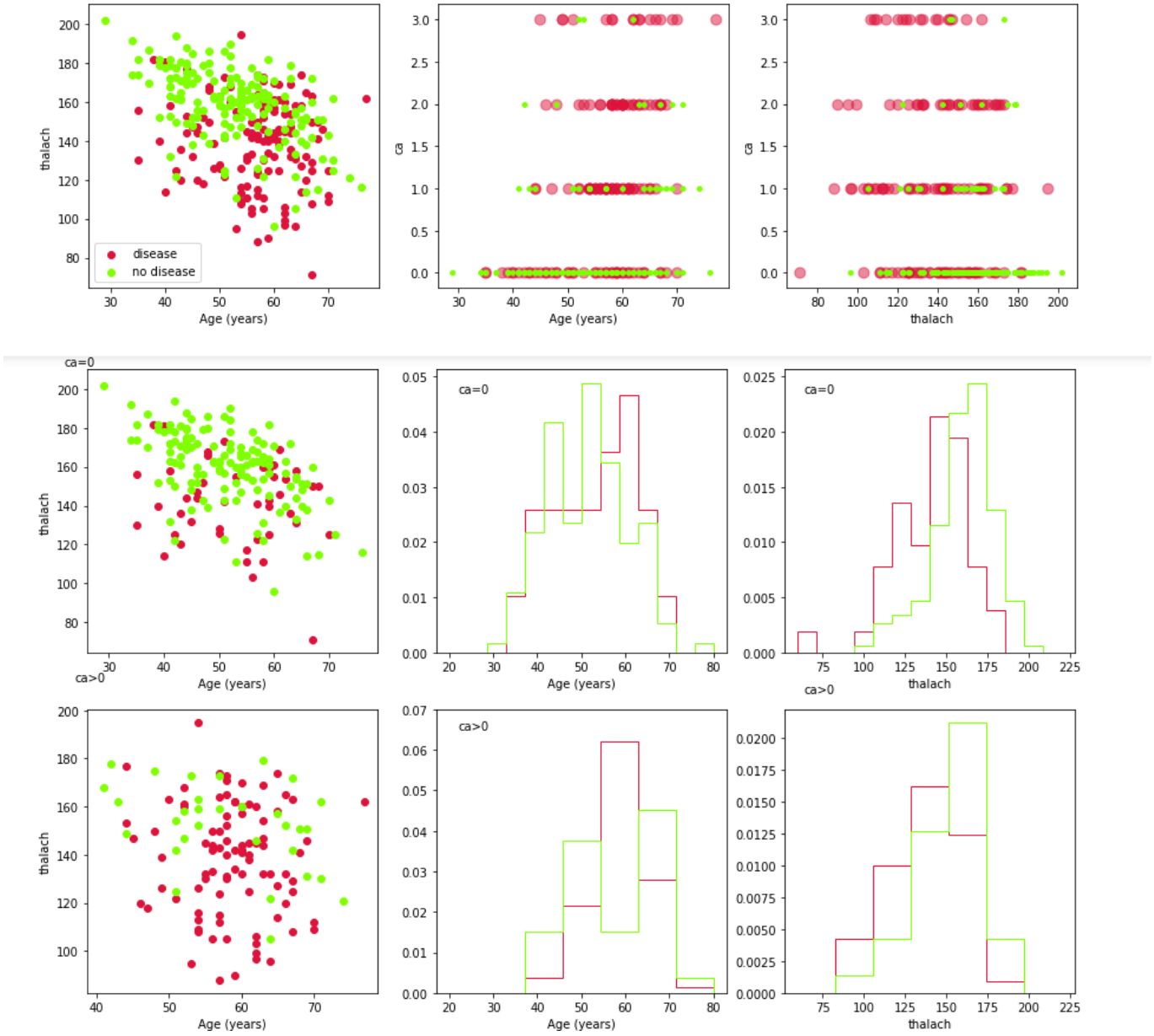
The final model performs really well in comparison of the Benchmark model. Our benchmark was to acheive atleast 78% accuracy and our model acheives nearly 85% accuracy with a standard deviation of 0.008% . This clearly beats the benchmark model and is significant enough to solve the problem.

V. Conclusion

Free Form Visualization

I visualized a different feature in this dataset that there are correlations between the features age, thalach and ca. Here are some graphs of the correlations.

Correlations



Reflection

The major challenge I faced during this process was to select an appropriate model. So, I tested upon various models and Iopted for the best one i.e. Support Vector Machines.

The project was started with acheiving an accuracy atleast equals to Benchmark model, Gaussian Naive Bayes i.e. 78.33%. But by using Support Vector Machines and by optimizing the model using cross validation score we were able to arrive at an accuracy of 85%. Final features were selected on the basis of cross validation score, keeping the model deviance as low as possible.

Improvement

We can improve the model by further improving the parameters. We can even use grid search to get more optmized results.

References

- http://archive.ics.uci.edu/ml/datasets/heart+Disease
- http://psrcentre.org/images/extraimages/84.%201211402.pdf
- http://scikit-learn.org/stable/modules/svm.html