BREAST CANCER PREDICTION

Development of a Classifier to Diagnose the State of a Sample.

Firoj Ahmmed Patwary

Full list of author information is available at the end of the article

Abstract

Goal of the Project: To develop a model that can accurately assess if the found tumor is malignant or benign.

Results of the Project: SVM classified tumors correctly 97.8% of the time, while Random Forest classified correctly only 95.3% of the time. The accuracy of SVM using PCA is 0.98.

My Key Learning: Biopsy, Principal Component Analysis (PCA), Confusion Matrix

Spent Time: Almost 15 Hours.

Project Evaluation Scale: 1: "Really Good Project".

Project Words: Total Words :900, Headers :9

1. Scientific Background of the Project

Nowadays breast cancer is one of the most frequent and dangerous cancer types. However, if early detected it can be successfully handled. Cell nuclei are important indicators for a tumor's classification as benign or malignant. Biopsies are used to deliver a sample of cells that are then assessed based on their size, shape and number.

2. Goal of the Project

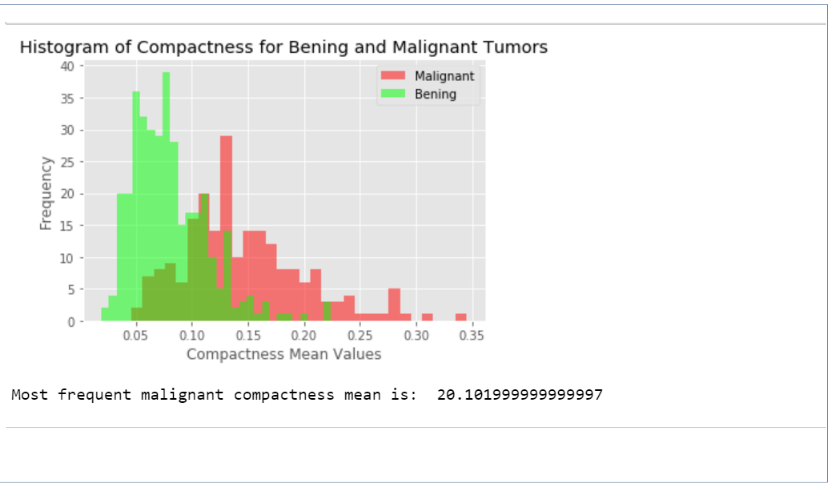
Our main objective of this project is to develop a model that can accurately assess if the found tumor is malignant or benign using the attributes of the given cell nuclei.

3. Description of the Data & Data Pre-processing Steps

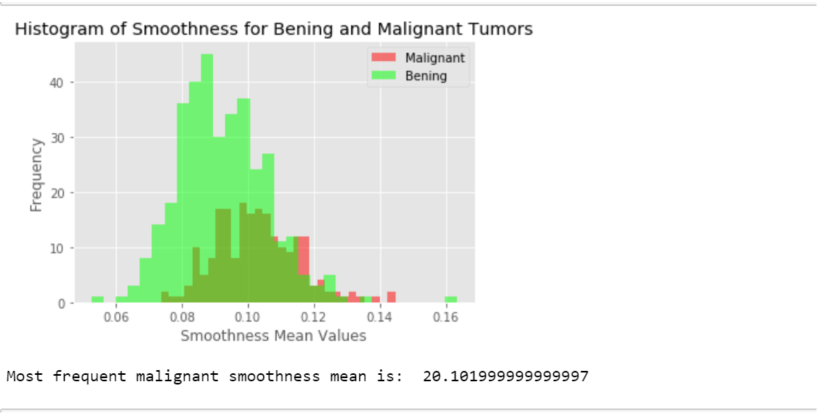
The data set (Breast Cancer Wisconsin (Diagnostic) Data Set (1995)) contains data of 569 instances/ patients with no missing values. The patients are being tested via biopsy to obtain data about their cell nuclei.

The nuclei had ten attributes that were measured using image processing and machine learning techniques (snakes"). Those ten attributes, which were to some extent indicators of either benign or malignant cancer, were radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. For every parameter it holds that a higher value of the attribute indicates a higher probability of a malignant tumor.

I found 30 correlation statistics of all 10 attributes (mean, standard error and worst), show that I got perfect correlation i.e. 1 between nucleus radius, parameter and area. Hence I could remove two of the three parameters. Also, I obtained a high correlation i.e. 0.7 between compactness, concavity and concave points and we decided to also remove two of those three parameters. To get idea about the attributes even indicate a higher or lower probability for malignant tumors, then I focused on histograms. These show that I can even remove three more attributes as they apparently bear no connection to malign tumors. Thus, the attributes I will use for the next step are area, compactness and texture, as they indicate malignant tumors the best and are not otherwise correlated to each other (here important to note, that we may use variations (mean, standard error or worst) of the selected parameters).



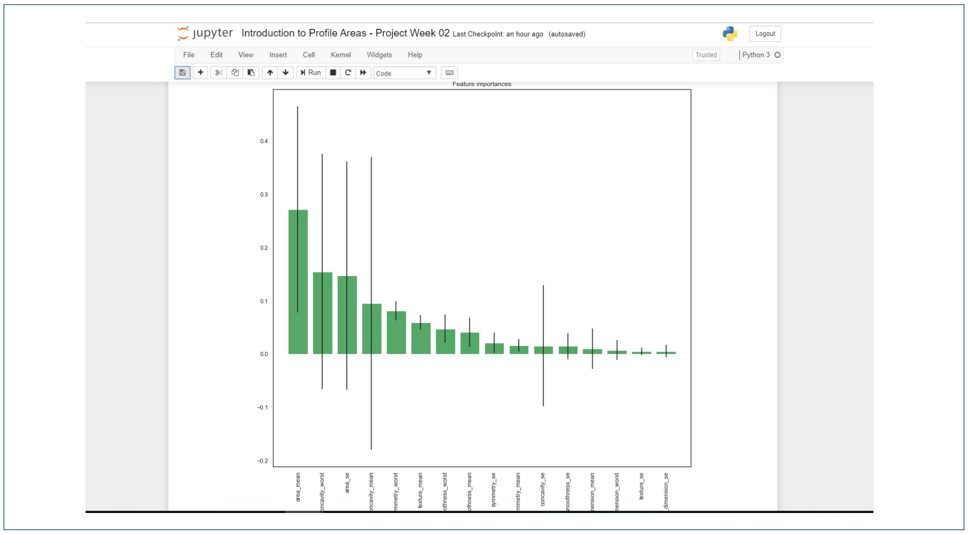
```
Most frequent malignant compactness mean is:  20.101999999999997
```



```
Most frequent malignant smoothness mean is:  20.101999999999997
```

4. Feature Selection & Selected Attributes Description

Now, feature selection used correlation heat map with swarmplot. In map heat figure, radius_mean, perimeter_mean and area_mean are correlated with each other so I will use only area_mean. Here, area_mean looks clear using swarmplot, the Benign and Malignant are separated very well in the plot but I cannot make exact separation among other correlated features. Also compactness_mean, concavity_mean and concave points_mean are correlated with each other. Therefore I only choose concavity_mean by using swarmplot and I did the same process for the all the features that are correlated. I drop other features that I did not select. So I am left with sixteen features that I used Random forest on to get an accuracy of 95.3%.



5. Description & Comparison of the Used Method(s)

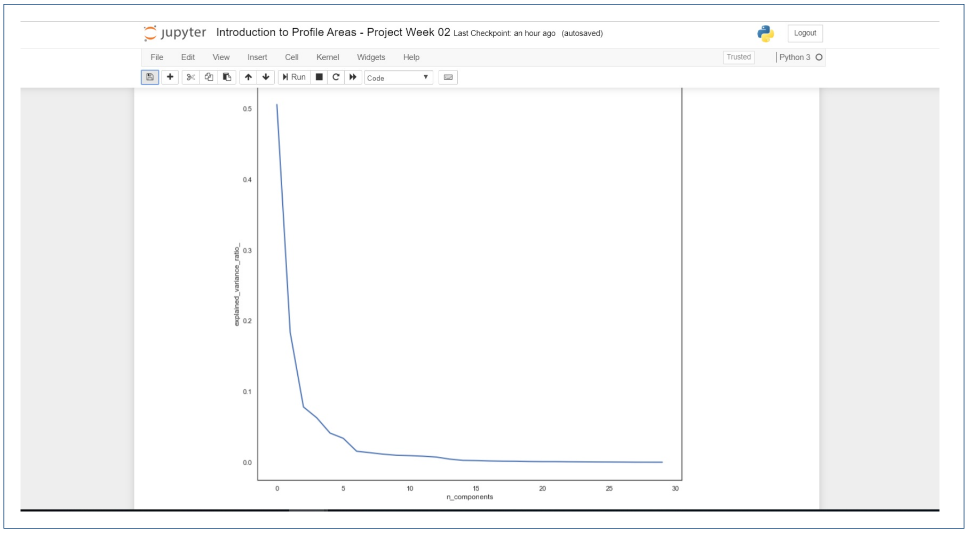
We know, Support Vector Machine (SVM) is a machine learning technique, that is frequently used for classification and regression analysis. SVM performs classification by finding the hyperplane that maximizes the margin between the two classes. Random forest consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest selects a prediction of the class, afterwards the class with the most ”votes” is the prediction of the model. Random forests are inherently multi class whereas Support Vector Machines need workarounds to treat multiple classes classification tasks. Usually this consists in building binary classifiers which distinguish
(i) between one of the labels and the rest (one-versus-all), or
(ii) between every pair of classes (oneversus- one). If you have a large data set (more than 10000 rows), prefer the random forest. I chose 16 different features for the Random forest classifier to get the accuracy of 95.3% and almost all the features(30 features) were used for SVM to get the accuracy of 97%.

6. Result

Here, I obtained better results by using Support Vector Machine than by using Random Forest classifier. Support Vector Machine classified tumors correctly 97.8% of the time, while Random Forest classified correctly only 95.3% of the time. The accuracy of Support Vector Machine using Principal Component Analysis(PCA) is 0.98 and, thus, better than using the raw features from before.

7. Discussion: why is this a typical project for a data-scientist or why not

First, we used correlation and statistics about the relation between multiple variables to assess whether some attributes are necessary for our classification. Here, we



found that we can actually work with less than the normal number of parameters. This process is typical for the work of a data scientist, as a closer look is taken into the data and the attributes. Then, Principal Component Analysis(PCA) is a helpful tool for data scientists to extract the best features of a data set. In this project, we used SVM and RF to classify a given data set into malignant or benign tumor's that indicate the type of breast cancer a patient has. As both algorithms are machine learning techniques, this is a pivotal part of Data Science.