Breast Cancer Prediction

Shayan Bhattacharjee

1. Introduction

Nowadays we can see the phrase 'breast cancer' popping out everywhere and it's no surprise with it being the 2nd highest type of cancer after lung cancer amongst the women. So the head of the well renowned private hospital has put a special emphasis on predicting the likelihood or probability of cancer which will help in providing personalized care and remedies to all the patients thereby accordingly focusing the attention towards serious cases and thereby better handling of the ailment. So, based on this business idea the approach is selected and it is planned to create a machine learning model to predict the probability. The target audience will obviously be the doctors and subsequently the patients and it will assist the already experienced doctors to better handle the cases.

2. Data Understanding

The first and foremost step will be to collect data in order to formulate a model. For this we will use the Breast Cancer patients' data of Wisconsin city. This was collected from the radiology and cancer detection centres of various hospitals spread across the city of Wisconsin. This will be a csv file containing different attributes which may or may not be used to train the model. These attributes are the independent variable based on which we need to model the probability of it growing into a cancer. It contains various attributes such as clump thickness, cell size, cell shape, marginal adhesion, bland chromatin, etc which will be used to predict the cancer and the relation between the attributes and its weightage will also be visualized, analyzed and studied. We will also use two models- KNN and SVM and compare the two models. We can find that most of the observations are good to train and test the machine learning model. There are some unbalanced labels which needs to be balanced in order to create an unbiased ML model.

For this project we used the following data:

https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data

3. Methodology

For conducting the project we divided the entire process into 3 parts:

3.1 Preprocessing of the dataset

After understanding the business idea and the data available we get an idea about the approach and the model to be followed. Now, for the data to be in sync with the model going to be deployed we accordingly modify our dataset. We visualize the weightage of each attribute and accordingly drop the ones that least effective to abstain from overfitting the model. We also look for the values in each columns and the ones with invalid values are dropped or if the attribute is of prime importance we replace those invalid values.

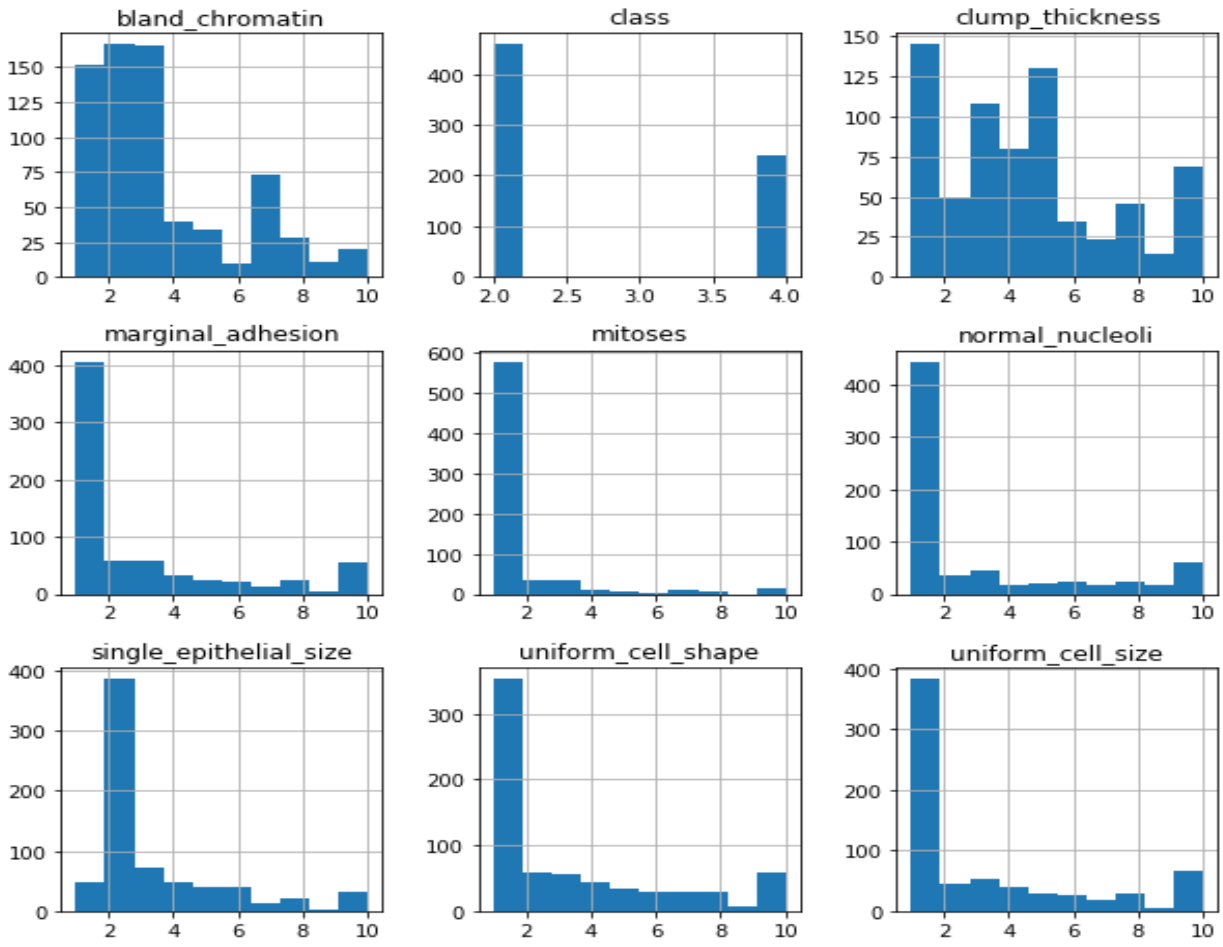

Fig: Attributes Distribution

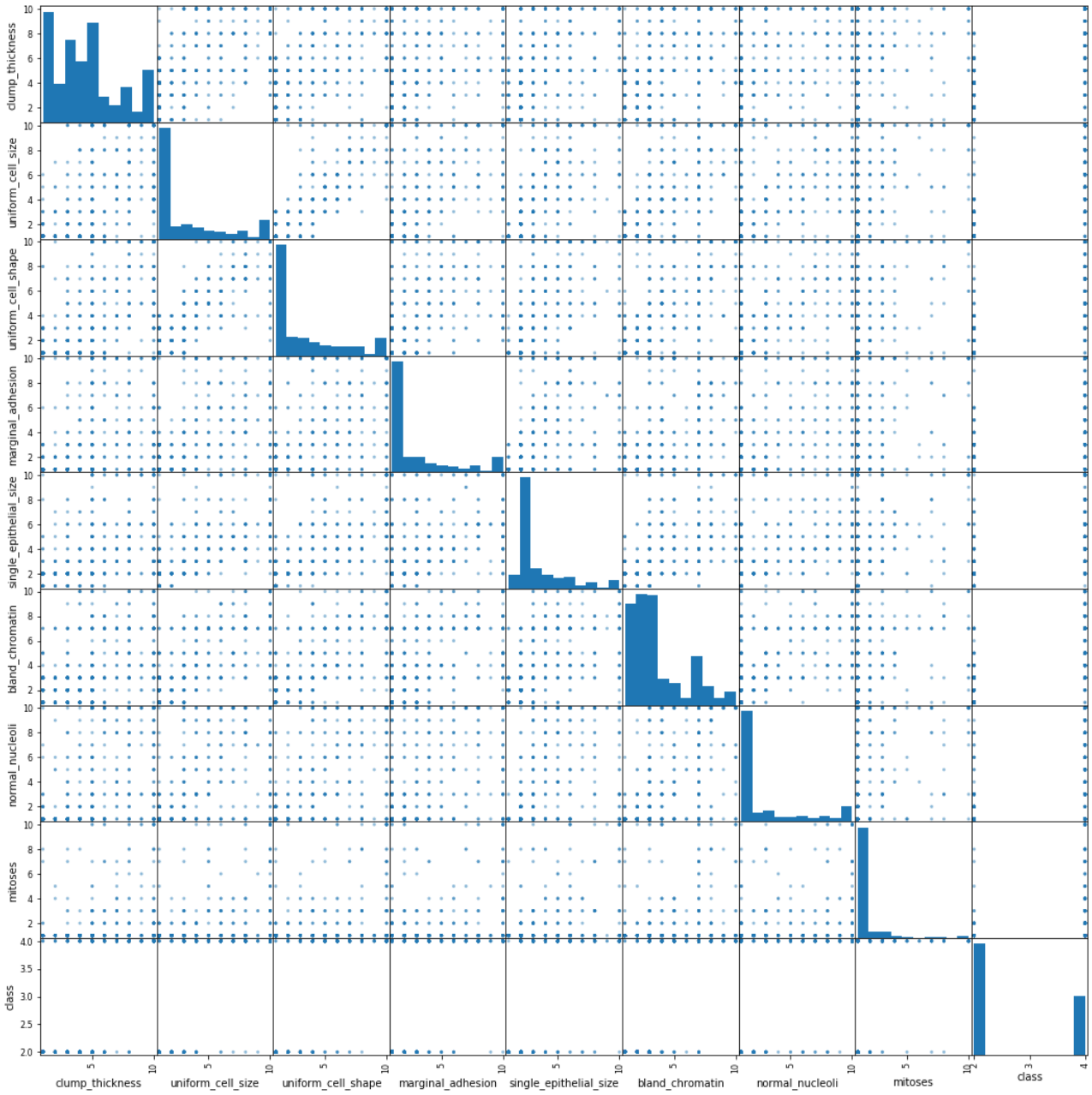

Fig: Scatter Plot Matrix

3.1 Modelling

The next step is to go ahead with the model(s) that we agreed upon during the introduction period. In this project for the prediction of breast cancer we planned to go ahead with 2 models-

i) K Nearest Neighbours (KNN)

ii) Support Vector Machine (SVM)

We split the dataset into training and testing dataset and used the 'Class' column to make the model. After careful visualization we came with k=5 as the optimum value and prepare the KNN and SVM model with the training dataset. Upon getting a slightly different result the next process was to predict the accuracy with the testing dataset and go ahead with the model with better accuracy and precision.

3.3 Calculating the accuracy and final prediction

We then tested both the KNN and SVM models with the testing dataset to calculate the accuracy. For the sake of better understanding we calculated out the precision, recall f1-score and the support of each models. Based on the values obtained we could clearly see the SVM with a better precision and recall score with the rest parameters remaining same. We thus could finally chose a better and close model and go ahead with the final prediction which will act a precedent for further predictions and patient analysis.

4. Results

The result section again deals with basically the 2 types of results we came up with this project.

- For the sake of proper prediction we calculated the precision, recall and F1- score of both the KNN and SVM model with the SVM getting a score of 1 against the 0.98 and 0.93 against 0.96 of KNN algorithm for precision and recall respectively. But with the macro average of 0.97 of SVM compared to 0.96 for KNN in recall we went ahead with SVM as a more accurate algorithm for this model.

- Using the SVM algorithm we finally predicted the probability of getting a breast cancer depending on various attributes for the particular unknown patient and was found to be 0.9571 or 95.71% or very highly likely.

5. Discussion

Breast Cancer is quite a burning topic and coming up with a way to model an accurate prediction will definitely add as an extra bonus for the experienced doctors to better analyze and appropriate for the serious patients. In order to prevent overfitting we also had to visualize various attributes and select the one that had more weightage. Based on the idea we came up with 2 approaches- KNN and SVM and we could model using any one of them but to be on a better side we first tried to analyze both the models and finally go ahead with the best one. On comparing we found SVM to be better one and thus went ahead with this algorithm for our machine learning model.

6. Conclusion

We were therefore successfully able to create a machine learning model using the SVM algorithm to accurately predict the probability of a patient to get breast cancer based on attributes by training our model with attributes like- clump thickness, uniform cell shape, marginal adhesion, bland chromatin, single epithelial size, mitosis etc.