**Title:** Breast Cancer Detection using Classification Algorithms

**Abstract:**

Disease Diagnosis is a major application of Big Data Analytics in health care. Machine Learning algorithms are widely used for disease diagnosis especially for breast cancer. My objective through this project is to apply three classification algorithms on a Breast Cancer Dataset to classify whether the tumor is benign or malign. The dataset has 699 observations with 10 variables. Algorithms will classify each data as benign or malign. I am considering this dataset as of now, I am also looking for a better data set simultaneously.

**Domain of Study:** Machine learning algorithms for Disease Diagnosis

**Algorithms:** I am planning to apply classification algorithms – Logistic Regression, Decision Trees, Random Forests and Support Vector Machine on a dataset and compare their performance.

Classification algorithms are supervised machine learning algorithms. They are used to teach machines how to group data based on particular criteria. In today’s world of big data, classification algorithms help to make sense of data and find patterns. For supervised learning, data contains predictor variables and outcome. For classification, dataset is divided into training and test samples. A predictor model is built using train data sample and tested for accuracy on test data sample.

Missing values in data have to be taken care of as they can lead to a biased model. Also we have to identify predictor, target variables. Also continuous, categorical variables have to be identified. For any classification algorithm, data has to be divided into train and test sets. It is usually done in 70:30 for train and test sets respectively.

One major application of classification algorithms is in mails/web pages/profiles classification (spam or not). Keywords, images, origin, header information etc., are used for these classifications. Other applications include disease classification, medical image classification, credit risk prediction (whether a person will repay the loan or not). In real world classification algorithms are used to build accurate models that can classify a new case into one of two groups (good or bad).

Confusion matrix which is cross-tabulation of actual and predicted status is used to calculate the accuracy of an algorithm. Also Sensitivity (true positive rate), Specificity (true negative rate) can be calculated and plotted as ROC curve to see the classifier ability.

**5. Data sources**

<http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>

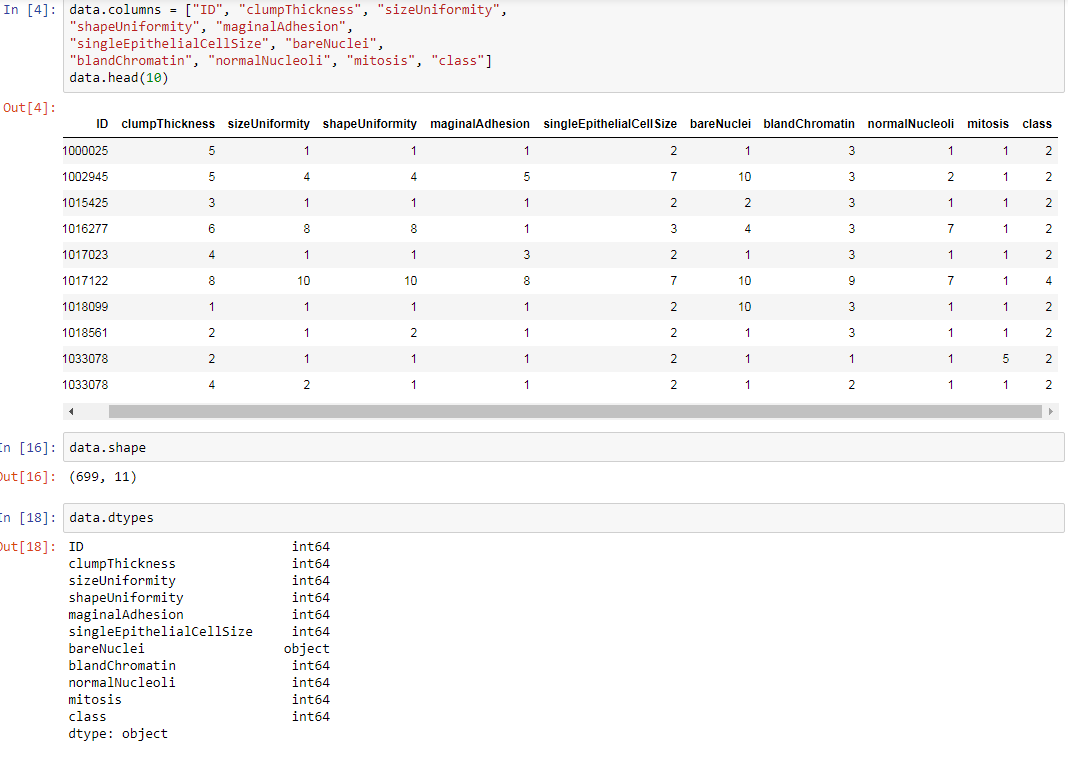
Wisconsin Breast Cancer dataset has 699 observations collected using fine-needle tissue from a mass under skin. It has 11 variables of which nine are predictor variables (cytological characteristics used to identify mass as benign or malign), ID and a class variable (has values 2 for benign, 4 for malignant). 458 of the samples are benign and 241 are malignant. There are 16 samples with missing data. Data file doesn't have column names and they are listed in a separate file.



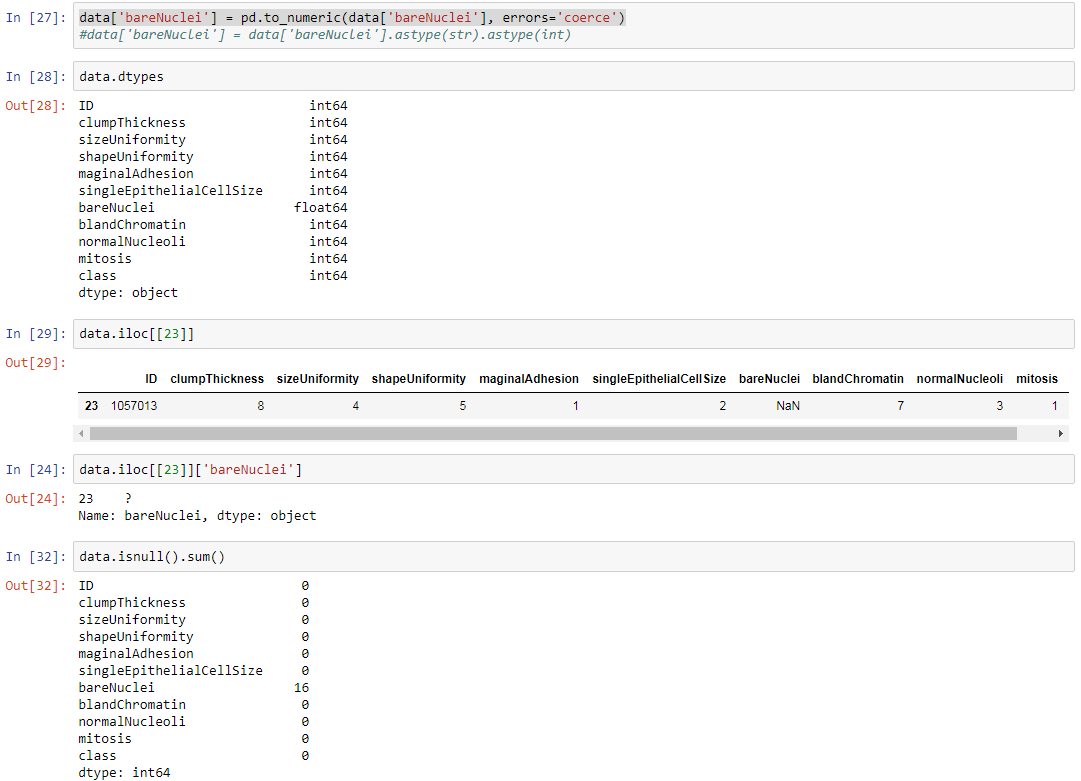
**Sample data** **(Source: from my local jupyter notebook)**

**Steps to transform the sample and make it ready for the analysis**

I have to add the column names to the data. I have to convert object data type of a column to integer data type. Also there are some missing values for which I have to do imputation using either mean/median or regression.



**Adding column headers to data (Source: My local jupyter notebook)**



**Converting object type column to integer and Listing null values (Source: My local jupyter notebook)**

**6. Graphics**

I’m planning to use Box Plots, Histograms to visualize continuous variables and frequency tables for categorical variables in dataset.

Also I am planning to use Confusion matrix, ROC curve for classification performance analysis and scatter plots to visualize results.

**7. What are your current challenges?**

Trying to learn/understand various classification algorithms. Also how they have to be implemented in python.

**8. At least 5 reference URLs with phrases or paragraphs you are planning to cite.**

**References:**

1. <https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/>

Steps of Data Exploration and Preparation and visualizations for data.

1. <https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/>

Understanding classification algorithms. Also I’m planning to cite introductory paragraphs of Logistic Regression, Decision Trees, Random Forests and Support Vector Machine that give a basic idea of algorithm.

1. <https://medium.freecodecamp.org/the-hitchhikers-guide-to-machine-learning-algorithms-in-python-bfad66adb378>

I will not cite any particular content from this site but I am planning to use this article as basis for my implementation of algorithms.

1. <https://machinelearningmastery.com/machine-learning-in-python-step-by-step/>

I am planning to use this article to plot visualizations for my dataset.

1. <https://www.dataquest.io/blog/machine-learning-python/>

<https://pythonprogramming.net/machine-learning-tutorial-python-introduction/>

I am referring these urls to understand algorithms and their implementation in python.

Project Blog URL: <https://medium.com/@sravanthiadusumilli/breast-cancer-detection-using-classification-algorithms-c169647e514d>