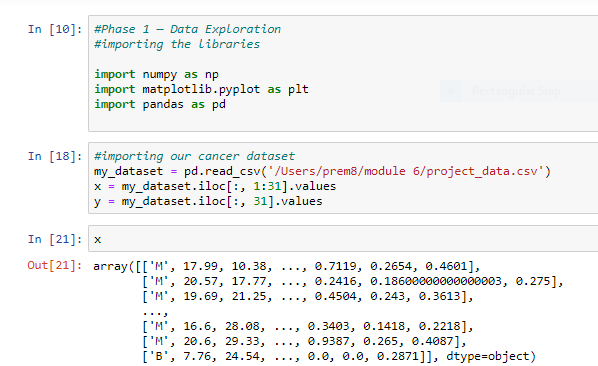
Detailed introduction of the project

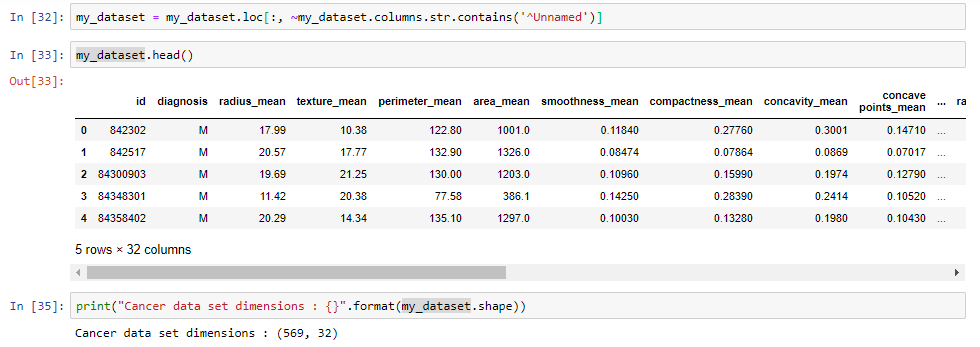
Breast cancer is among the 4 leading cancers in women worldwide (ie, lung, breast and bowel [including anus], stomach, and prostate cancers). The IARC statistics show that breast cancer accounts for 25% of all cancer cases diagnosed in women worldwide. Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today’s society.  
The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical features detection from complex BC datasets, machine learning (ML) is widely recognized as the methodology of choice in BC pattern classification and forecast modeling.

Data set description

We will be using *Jupiter Note* to work on this dataset. Lets us import important library and import our dataset to *Jupiter Note*.



We can examine the data set using the pandas’ **head()** method and also can find the dimensions of the data set using the panda dataset ‘shape’ attribute.



We can observe that the data set contain 569 rows and 32 columns. ‘Diagnosis’ is the column which we are going to predict, which says if the cancer is M = malignant or B = benign. 1 means the cancer is malignant and 0 means benign. We can identify that out of the 569 persons, 357 are labeled as B (benign) and 212 as M (malignant).

Propose a machine learning model

This is the most exciting phase in Applying Machine Learning to any Dataset. It is also known as Algorithm selection for predicting the best results.

Data Scientists use different kinds of Machine Learning algorithms to the large data sets.

In this case I would suggest supervised learning, before proceedings further let us have a brief about supervised learning.

**Supervised learning:** Supervised learning is a type of system in which both input and desired output data are provided. Input and output data are labeled for classification to provide a learning basis for future data processing. Supervised learning problems can be further grouped into **Regression** and **Classification** problems.

A **regression** problem is when the output variable is a real or continuous value, such as “salary” or “weight”.

A **classification** problem is when the output variable is a category like filtering emails “spam” or “not spam”

In our dataset we have the outcome variable or Dependent variable i.e. Y having only two set of values, either M (Malign) or B (Benign). So we will use Classification algorithm of supervised learning.

We have different types of classification algorithms in Machine Learning:-

1. Support Vector Machines
2. Random Forest Classification

**Support vector machines:**

**Random Forest Classification:**

Preliminary results