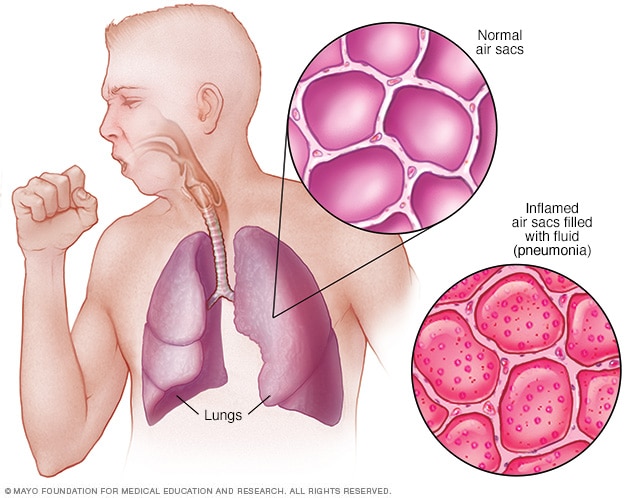
Pneumonia Detection Challenge

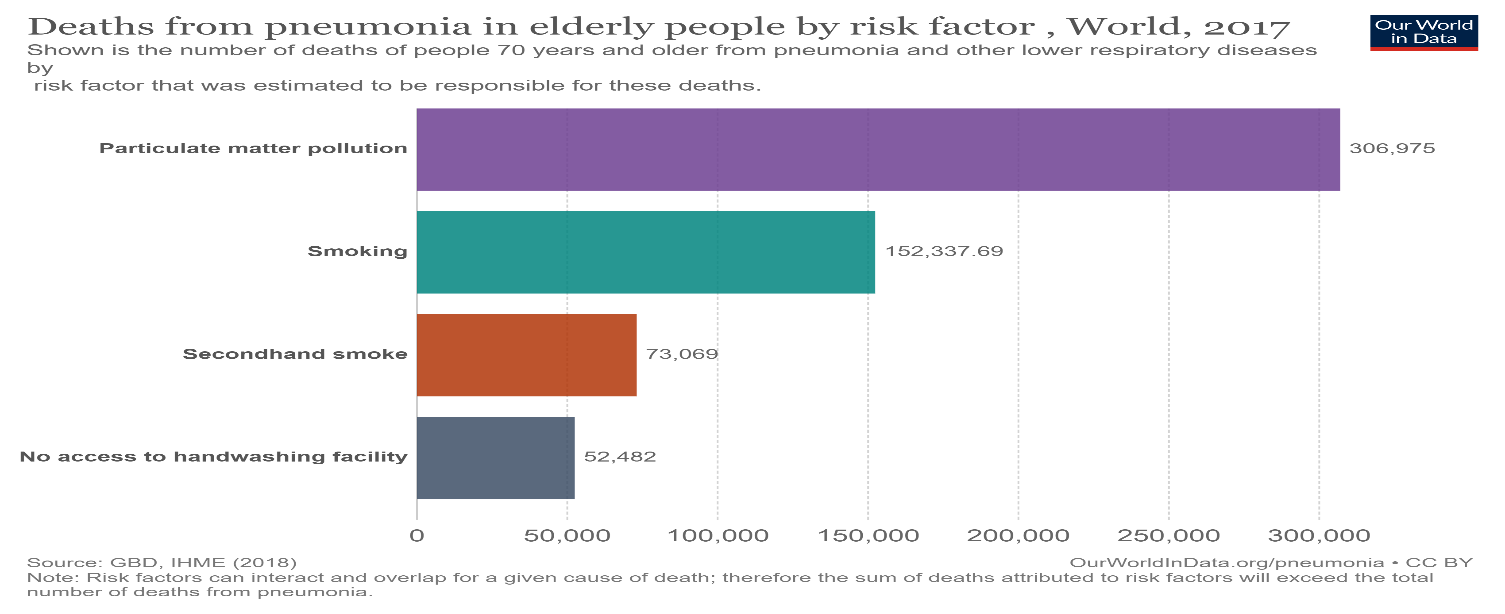
# **Project Definition:**

## **Project Overview:**

Pneumonia is an infection that inflames the air sacs in one or both lungs. The air sacs may fill with fluid or pus (purulent material), causing cough with phlegm or pus, fever, chills, and difficulty breathing. A variety of organisms, including bacteria, viruses and fungi, can cause pneumonia.



Pneumonia accounts for over 15% of all deaths of children under 5 years old internationally. In 2017, 920,000 children under the age of 5 died from the disease.



CXRs (**Chest Radiograph**) are the most commonly performed diagnostic imaging study. A number of factors such as positioning of the patient and depth of inspiration can alter the appearance of the CXR, complicating interpretation further. In addition, clinicians are faced with reading high volumes of images every shift.

Pneumonia detection requires review of a **chest radiograph** (CXR) by highly trained specialists and confirmation through clinical history, vital signs and laboratory exams. Pneumonia usually manifests as an area or areas of increased opacity on CXR.

However, the diagnosis of pneumonia on CXR is complicated because of a number of other conditions in the lungs such as fluid overload (pulmonary enema), bleeding, volume loss (atelectasis or collapse), lung cancer, or post radiation or surgical changes. Outside of the lungs, fluid in the pleural space (pleural effusion) also appears as increased opacity on CXR. When available, comparison of CXRs of the patient taken at different time points and correlation with clinical symptoms and history are helpful in making the diagnosis.

## **Problem Statement:**

The dataset is made up of chest radiographs in the form of DICOM images and are separated into testing and training sets. The dataset also contains two CSV files: one contains the identifications of patients (of the DICOM files), the coordinates of the bounding boxes for those patients with pneumonia and a target column indicating if a patient has pneumonia or not. The second CSV file contains detailed information that categorizes the DICOM images into three categories: No Lung Opacity/Normal, Normal, and Lung Opacity. Only “Lung Opacity” category indicates that a patient has pneumonia. A patient’s chest radiograph can contain 0 or more cases of lung opacity. Each case is represented by a row in the CSV files.

### **Existing System:**

Properly diagnosing pneumonia can be a tall order because it requires the review of chest radiographs (CXR) by highly trained specialists. The specialist confirms a case of pneumonia by also examining the patient’s clinical history, vital signs and laboratory examination results. Pneumonia usually reveals itself in the lungs as an area(s) of increased opacity on CXR, however, diagnosis on CXR can be complicated by several other lung conditions.

CXRs are the most commonly performed diagnostic imaging study. Several factors such as positioning of the patient and depth of inspiration can alter the appearance of the CXR, complicating interpretation further. In addition, clinicians are faced with reading high volumes of images every shift.

### **Proposed System:**

The project will attempt to develop a model that will detect visual signals of pneumonia in medical images. It will automatically locate lung opacities on CXRs. To detect Pneumonia we need to find the Inflammation of the lungs. An algorithm needs to be built to detect a visual signal for pneumonia in medical images. Specifically, the algorithm needs to automatically locate lung opacities on chest radiographs.

## **Project Flow:**

The Project flow involves following steps.

1. Exploratory Data Analysis- This step involves exploring and understanding the data
2. Data Preparation- Preparing the Data to feed it to the model.
3. Algorithms and Techniques - Trying out different models on the train data to get the model which performs well
4. Performance Tuning- Tuning the model with different Parameters.
5. Finalising the Model and Parameters- Build the final model with the best parameters and running the model with the Test Data.

# **Exploratory Data Analysis:**

Exploratory Data Analysis (EDA), also known as Data Exploration, is a step in the Data Analysis Process, where a number of techniques are used to better understand the dataset being used.

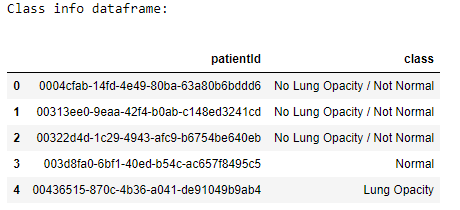
## **2.1 Understanding the Files:**

There are two set of files provided for Train and Test Dataset respectively. Images provided for both Train and Test data is in the DCM format. The DCM file extension is used for DICOM which stands for Digital Imaging and Communications in Medicine.

For Train data there are two CSV files which has been provided:

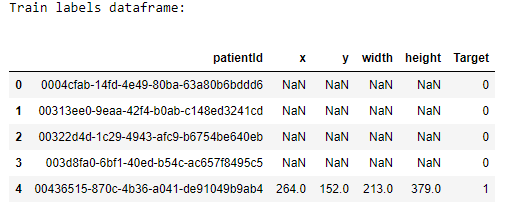
1. ***stage\_2\_detailed\_class\_info.csv*** - provides the class information for each Patient Id.

Class info data frame has 30227 rows and 2 columns

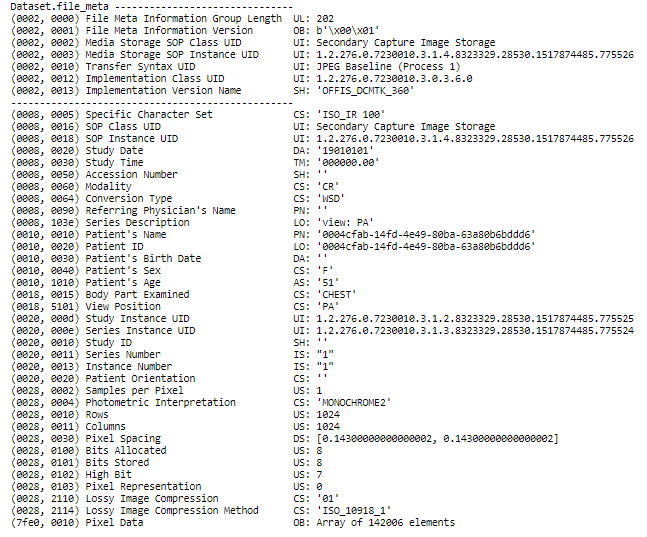


1. ***stage\_2\_train\_labels.csv*** – provides the Bounding Boxes details where the masses are found.

Train Labels data frame has 30227 rows and 6 columns. Bounding boxes are defined as follows: x, y, width and height



Each dcm file holds the metadat of patient these needs to be read and checked for useful information.

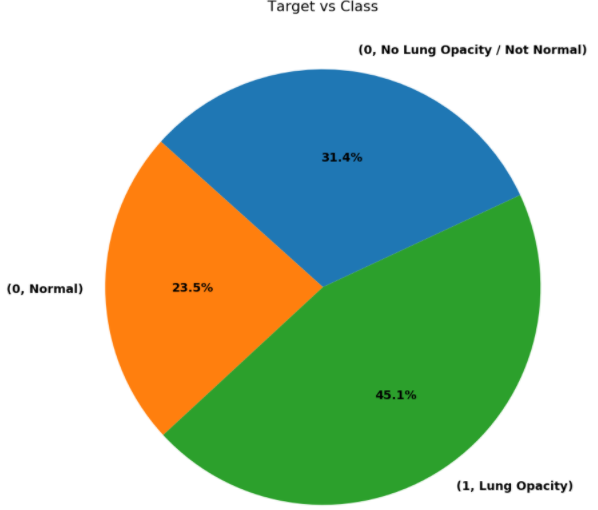
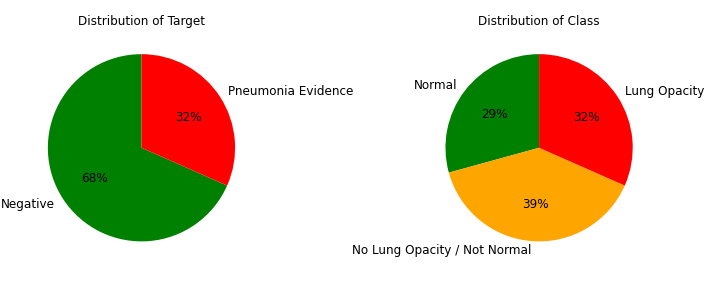


## **2.2 Data Observation:**

### **Unique Check**

There are 3 unique classes i.e. ***No Lung Opacity/Not Normal, Normal and Lung Opacity*** and 2 Target class ***1 and 0***. Target ***value 1*** indicates the presence of Pneumonia and Target ***value 0*** absence of Pneumonia.

39 % of Train Category is under No Lung Opacity/Not Normal which signifies that the lung in not normal but also there is no presence of Pneumonia. 29% of data is Normal Lung and 32% where Lung Opacity is found and has the presence of Pneumonia.



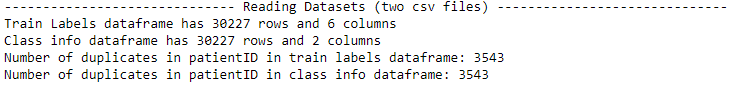
### **Null Check**

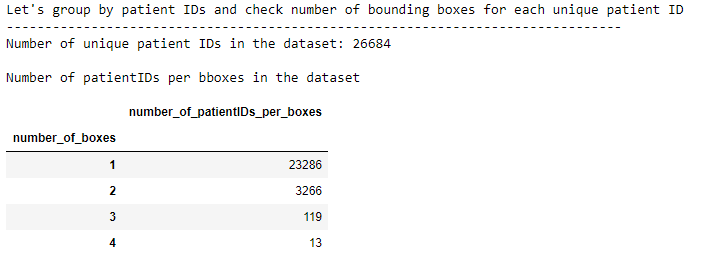
There are no boundary boxes provided for the patients who are under class Normal and No Lung Opacity/Not Normal. These records will have columns as Null in the dataset.



### **Duplicate Check**

There are duplicate records in both CSV files for some patients. All the duplicates are due to multiple masses found in the X-Ray for the patients. There are multiple boundary boxes provided for some patients. Each patients are associated to only one class. About 23,286 patient IDs (~87% of them) provided have 1 bounding boxes while 13 patients have 4 bounding boxes.



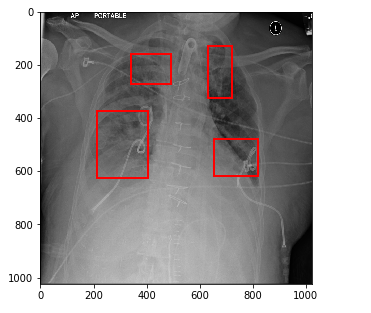
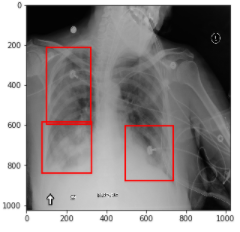


### **X-Ray with bounding boxes**

Different types of Image Data Provided are below:

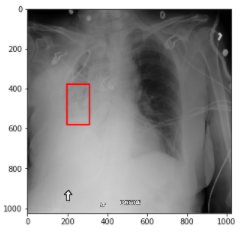
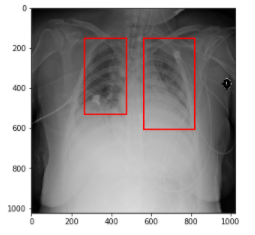
**Lung Opacity with 4 Bounding boxes**

**Lung Opacity with 3 Bounding boxes**

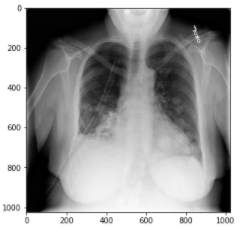
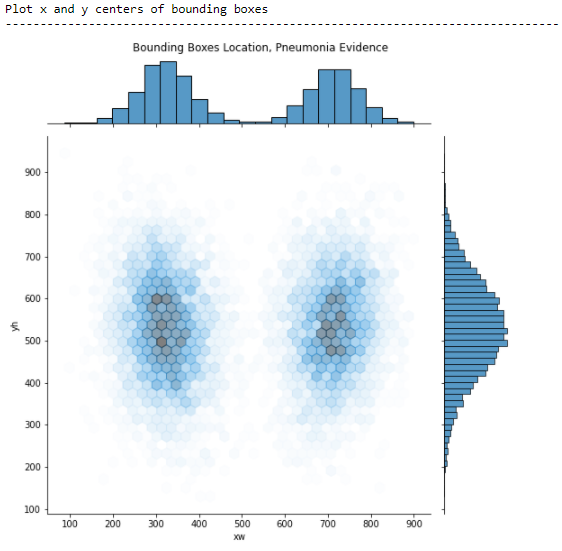
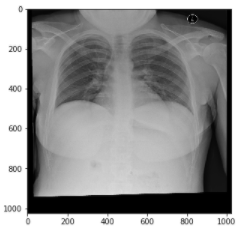
**Lung Opacity with 1 Bounding box**

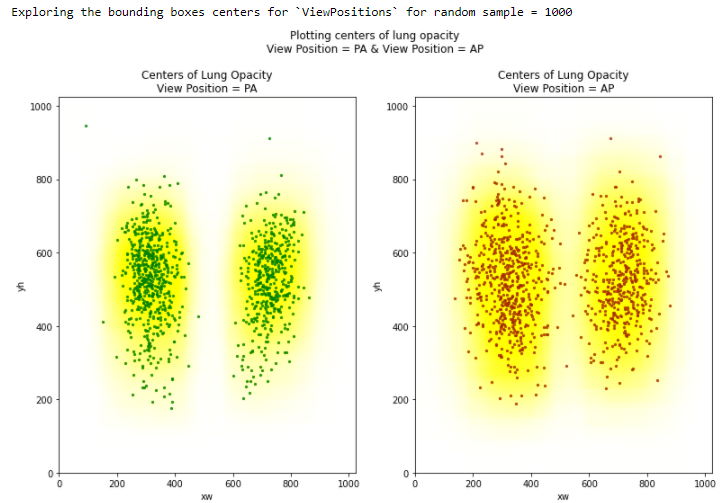
**Lung Opacity with 2 Bounding boxes**



**No Lung Opacity / Not Normal**

**Normal**

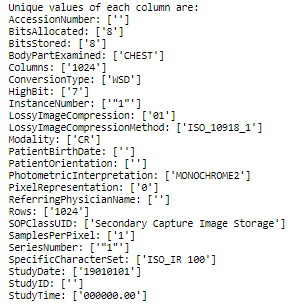


### **Metadata of Image**

There are few attributes in the Metadata which is same for all the records, these attributes will be not useful for the model hence can be dropped.

Below is the summary of the analysis from Metadata:

1. Accession Number is a RIS generated number that identifies the order for the Study. For our data the value is unknown hence it is NULL.
2. Bits Allocated is the number of bits allocated for each pixel sample. Each sample is having 8 bits allocated
3. Bits Stored is the number of bits stored for each pixel sample. Each sample is having 8 bits allocated.
4. Body Part Examined is the text description of the part of the body examined. For the train data the value is Chest.
5. Columns and Rows indicate that the image is of size is 1024 x 1024.
6. Conversion Type describes the kind of image conversion. For our sample it is WSD(Work Station)
7. High Bit is the most significant bit for each pixel sample. Each data is having highest Bit as 7.
8. Instance Number is the number that identifies the image. The value is 1 for all the data.
9. Lossy Image Compression specifies whether an Image has undergone lossy compression. The value is 01 indicating that the image has been subjected to lossy compression.
10. Lossy Image Compression Method is the label for the lossy compression method(s) that have been applied to this image. Value is ISO\_10918\_1 for all the images.
11. Modality is the type of equipment that originally acquired the data used to create the images in this Series. Value is CR (Computed Radiography)
12. Patient Birth Date, Patient Orientation & Referring Physician Name these hold the Patient Birthdate, Patient Orientation is the direction of the rows and columns of the image, required if image does not require Image Orientation and Referring Physician Name refers to the Physician who referred. For the train data these values are NULL.
13. Photometric Interpretation specifies the intended interpretation of the pixel data. For the train data the value is MONOCHROME2 which means that pixel data represent a single monochrome image plane. The minimum sample value is intended to be displayed as black after any VOI gray scale transformations have been performed.
14. Pixel Representation is the data representation of the pixel samples all the records have this value as 0.
15. SOP Class UID uniquely identifies the SOP Class. This has the value Secondary Capture Image Storage for the train data.
16. Samples per Pixel is the number of samples (planes) in this image. The value is 1 for all the records.
17. Series Number is the number that identifies the Series. The value is 1 for all the records.
18. Specific Character Set Character Set that expands or replaces the Basic Graphic Set. Required if an expanded or replacement character set is used. The value is ISO\_IR 100 for all the records
19. Study Date, Study ID and Study Time these indicate when the study is done. The values for these are 19010101, NULL and 000000.00 respectively.



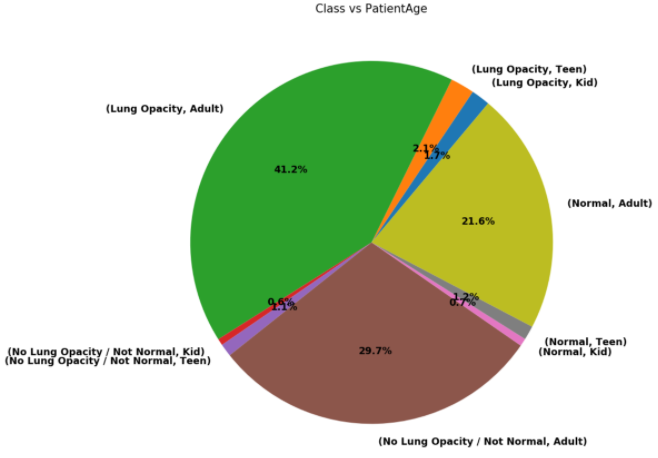
The other information of the DICOM file are Age, Sex, Series Description and View position.

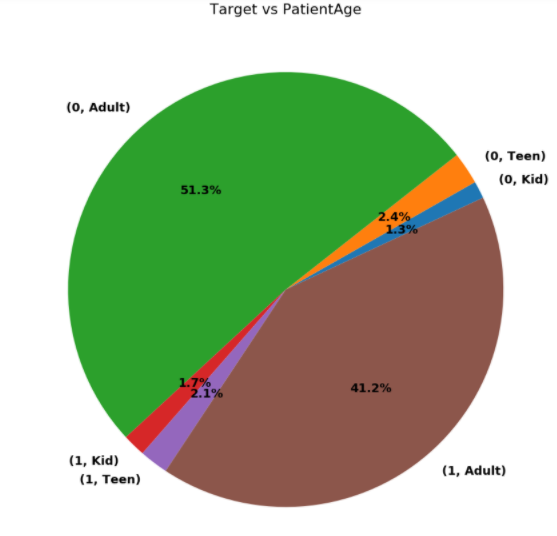
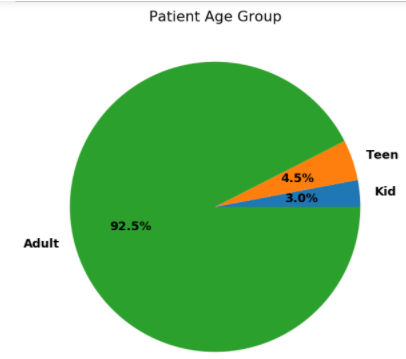
Analysis of these with the Target and Class variable is as below.

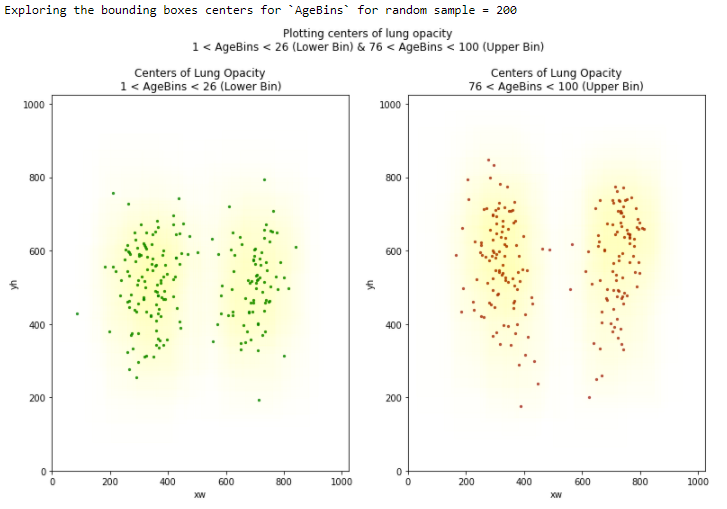
**Patient Age Category**: Patient Age is categorized into three category Kid (Age 0-13), Teen (Age 13 - 20) and Adult (Age 20 - 110).

The Train data has 92.5% Adult Patients out of which 51.3% has the Target class as 0 i.e. 21.6% are of class Normal and 29.7% are No Lung Opacity/ Not Normal. 41.2% Adult Patients are belonging to the Target class 1 i.e. Lung Opacity Class.

The Train data has 4.5% Teen Patients out of which 2.4% has the Target class as 0 i.e. 1.2% are of class Normal and 1.1% are No Lung Opacity/ Not Normal. 2.1% Teen Patients are belonging to the Target class 1 i.e. Lung Opacity Class.

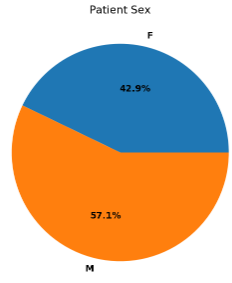
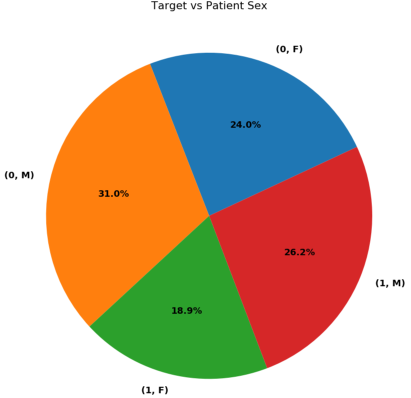
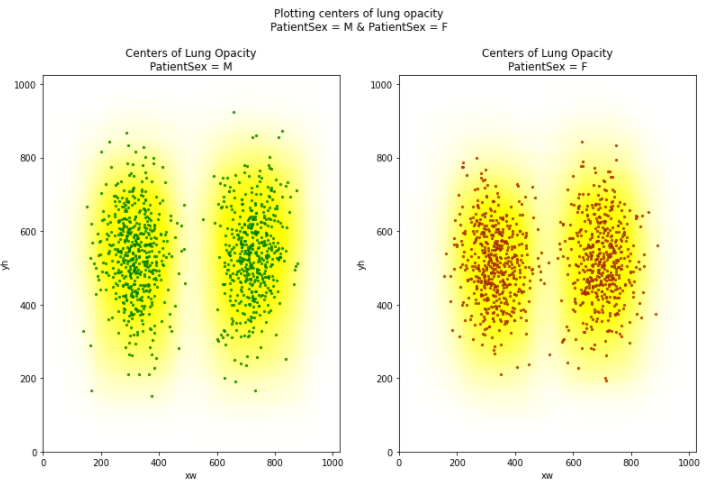
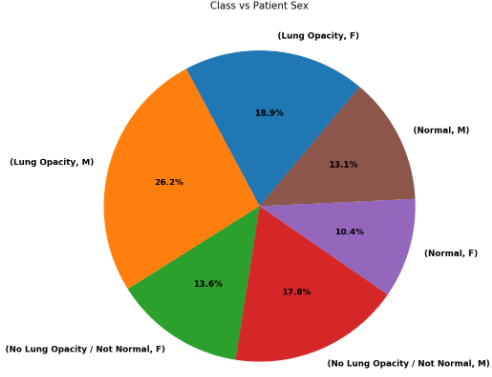
The Train data has 3.0% Kid Patients out of which 1.3% has the Target class as 0 i.e. 0.7% are of class Normal and 0.6% are No Lung Opacity/ Not Normal. 1.7% kid Patients are belonging to the Target class 1 i.e. Lung Opacity Class. 





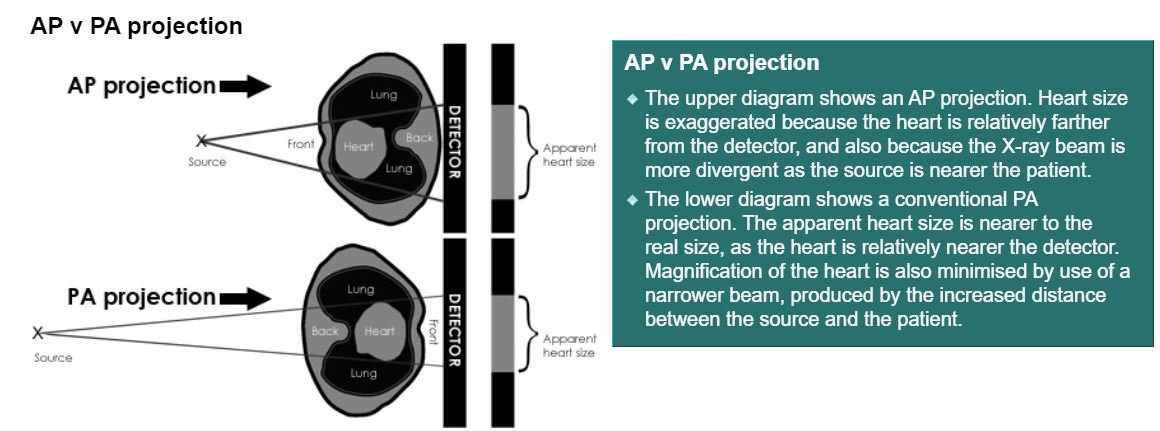
**Patient Sex**: 42.9% of Patient are Female among which 24.0% are of Target class 0 .i.e. 10.4% are of class Normal and 13.6 are of class No Lung Opacity/ Not Normal. 18.9% are of Target class 1 i.e. of class Lung Opacity.

57.1% of Patient are Male among which 31.0% are of Target class 0 .i.e. 13.1% are of class Normal and 17.8 are of class No Lung Opacity/ Not Normal. 26.2% are of Target class 1 i.e. of class Lung Opacity.

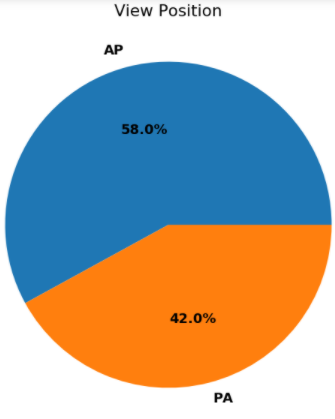
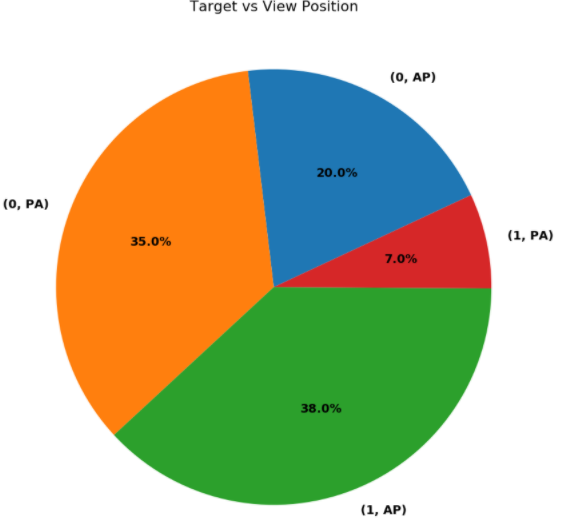
**SeriesDescription / ViewPosition :** Radiographic view of the image relative to the imaging subject's orientation. Understanding different View Positions As seen below, two View Positions that are in the training dataset are AP (Anterior/Posterior) and PA (Posterior/Anterior). These type of X-rays are mostly used to obtain the front-view. Apart from front-view, a lateral image is usually taken to complement the front-view.

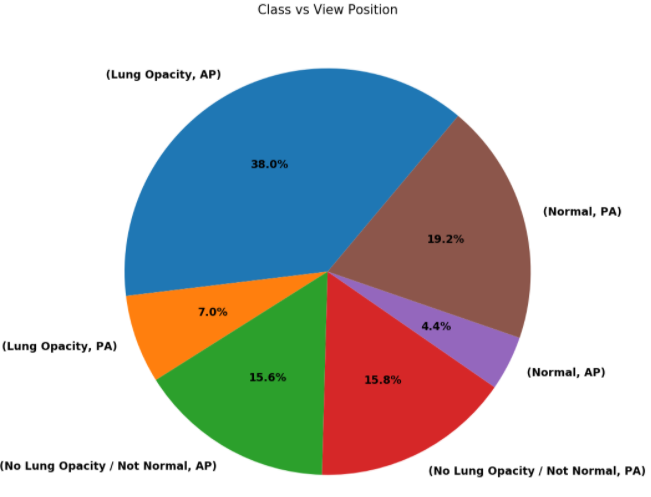
**Posterior/Anterior (PA):** In PA, X-Ray beam hits the posterior (back) part of the chest before the anterior (front) part. While obtaining the image patient is asked to stand with their chest against the film. **Anterior/Posterior (AP):** Attimes it's not possible for radiographers to acquire a PA chest X-ray. This is usually because the patient is too unwell to stand. AP projection images are of lower quality than PA images. Heart size is exaggerated (cardiothoracic ratio approximately 50%)

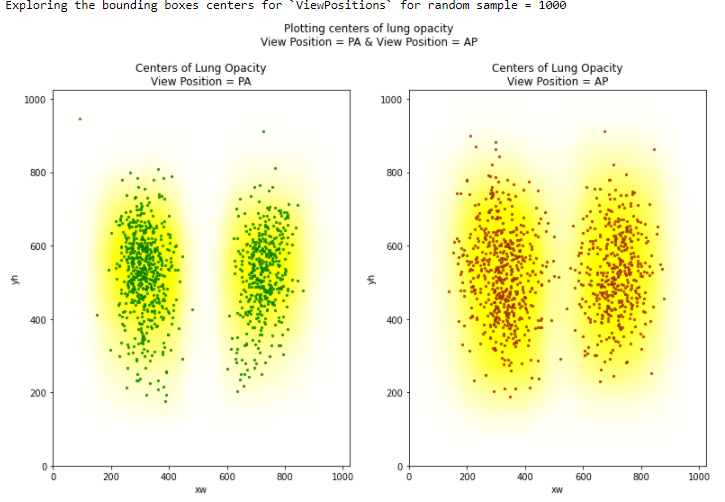


58.0% of Patient have View Position as AP among which 20.0% are of Target class 0 .i.e. 4.4% are of class Normal and 15.6 are of class No Lung Opacity/ Not Normal. 38.0% are of Target class 1 i.e. of class Lung Opacity.

42.0% of Patient have View Position as PA among which 35.0% are of Target class 0 .i.e. 19.2% are of class Normal and 15.8 are of class No Lung Opacity/ Not Normal. 7.0% are of Target class 1 i.e. of class Lung Opacity.

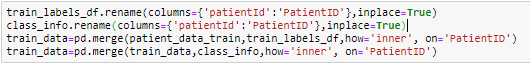




# **Data Preparation:**

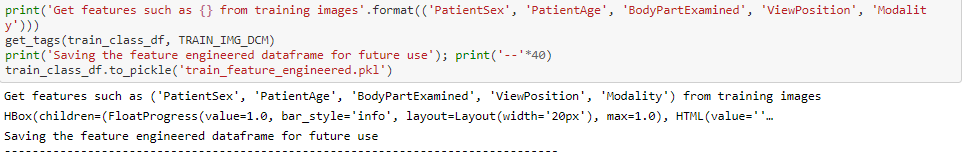
## **Merging the Data:**

The two CSV file as well as Metadata of all the train image must be read and stored into a data frame. These data frames is merged using Inner join on Patient Id. Below code Snippet is used to merge the data.



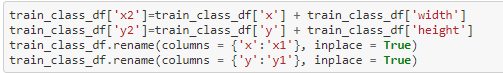
## **Creating the Pickle file.**

The Metadata of the Image has many attributes which will not contribute to the model hence only required column.

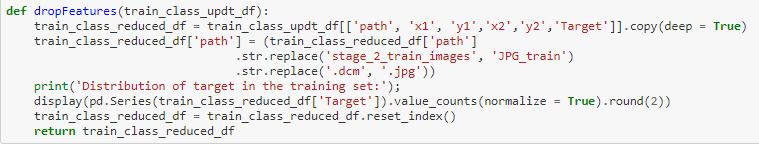


## **Deriving Attributes:**

The Bounding boxes for the Train data is specified using the x, y coordinates along with Height and Width information. The other coordinates can be derived using the below code snippet.

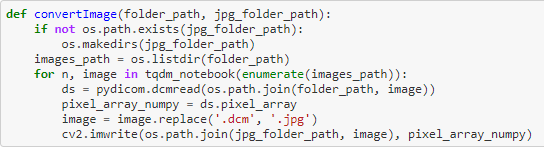


Also the Height and Width attribute can be dropped using the below code snippet.



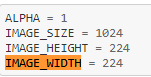
## **Image file Manipulation:**

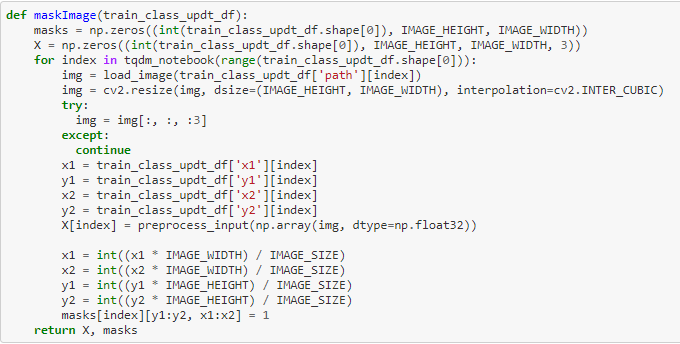
The DCM file will be converted to the JPG for faster processing. This is achieved using the below code snippet.



## **Image Masking:**

The size of the train images is 1024 × 1024 this needs to be modified to feed to the data to any standard model. Hence below code snippet is used to mask the train and test images.





## **Data Split for the Model:**

The train data has 68% of Target class 0 and 32% of Target class 1, there is huge imbalance in the data hence the data needs to be balanced for training the model.

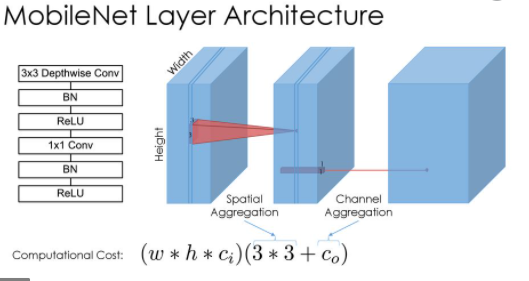
# **Algorithms and Techniques:**

## **4.1 Model 1: MobileNet**

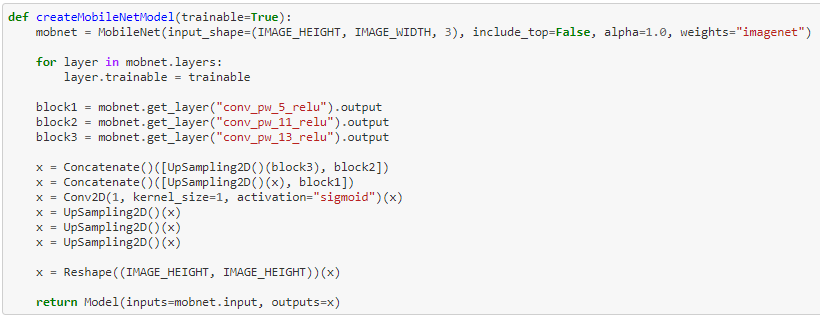
MobileNet is an efficient and portable CNN architecture that is used in real world applications. MobileNet primarily use depth wise separable convolutions in place of the standard convolutions used in earlier architectures to build lighter models.

Why MobileNet?

Computer vision networks have the responsibility to make a deeper network achieve higher accuracy. In short, the deeper the model, the harder it is to optimize. For compact applications, it becomes inconvenient to maintain the number of operations as the system has limited computation and power. MobileNet was introduced to mitigate these problems.



### **4.1.1 Summary of the Model:**



### **4.1.2 Model Performance:**

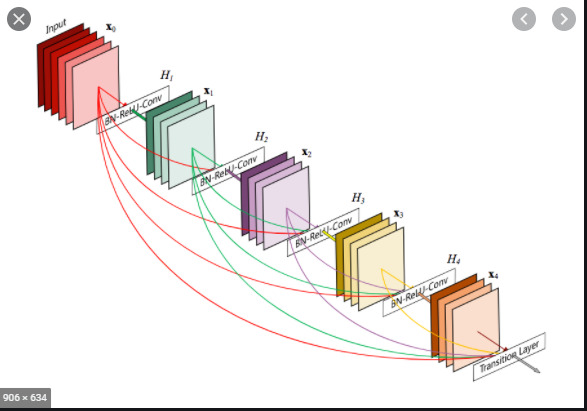
The model is giving the accuracy of 12% with 500 data samples.

## **Model 2: DenseNet**

DenseNet is one of the new discoveries in neural networks for visual object recognition. DenseNet is quite similar to ResNet with some fundamental differences. ResNet uses an additive method (+) that merges the previous layer (identity) with the future layer, whereas DenseNet concatenates (.) the output of the previous layer with the future layer.

Why Do We DenseNet?

DenseNet was developed specifically to improve the declined accuracy caused by the vanishing gradient in high-level neural networks. In simpler terms, due to the longer path between the input layer and the output layer, the information vanishes before reaching its destination.



### **Summary of the Model:**

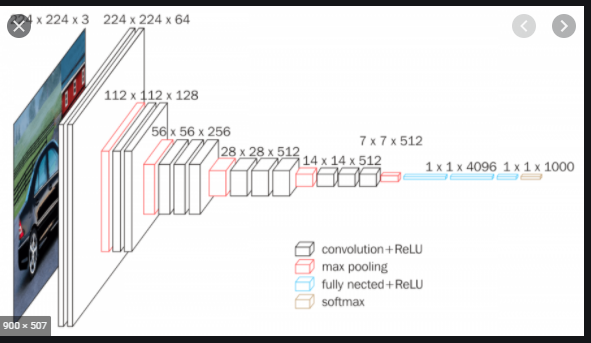


### **4.2.2 Model Performance:**

The model is giving the accuracy of 48% with 500 data samples.

## **Model 2: VGG 16**

VGG is an acronym for the Visual Geometric Group from Oxford University and VGG-16 is a network with 16 layers proposed by the Visual Geometric Group. These 16 layers contain the trainable parameters and there are other layers also like the Max pool layer but those do not contain any trainable parameters.



Why VGG16?

This model uses smaller filters and also can be used to build deeper networks.

### **Summary of the Model:**



### **Model Performance:**

The model is giving the accuracy of 46% with 500 data samples.