Diagnosis of covid-19

Introduction

1. Problem domain

In the following we shall discuss about the admission of covid-19 patient based on the results of laboratory tests commonly collected among confirmed covid-19 cases during a visit to emergency room.

1. Relevance to current context

As we all know that the current situation due to covid-19 is worse ever seen in this modern era. Discovery of vaccination for this is the need of the hour. Although several countries are spending millions in finding the vaccine there hasn’t been any sign of impressive results. So it is very important in managing this epidemic so that its transmission is stopped.

1. Motivation

There is a general trend in the number of covid-19 cases observed and it is an exponential increase in the number of covid-19 cases. So it is very important to manage the bed in hospital for covid patients as all the patients is to be kept in isolation and with specific care. So classification of the intake covid patients would have a better impact in hospital management.

1. Problem statement

Predict confirmed COVID-19 cases among suspected cases.  
Based on the results of laboratory tests commonly collected for a suspected COVID-19 case during a visit to the emergency room, would it be possible to predict the test result for SARS-Cov-2 (positive/negative)?

Predict admission to general ward, semi-intensive unit or intensive care unit among confirmed COVID-19 cases.  
Based on the results of laboratory tests commonly collected among confirmed COVID-19 cases during a visit to the emergency room, would it be possible to predict which patients will need to be admitted to a general ward, semi-intensive unit or intensive care unit?

1. Objective

Objective is to classify the admission of the patient to general ward, semi-intensive unit and intensive care unit

Literature review

* 1. A report on various previous endeavours to address the problem as a whole(chronological with citattions)
  2. A report on various previous endeavours to address the objectives(chronological with citattions)
  3. Summary of their results, advantages and drawbacks, gaps unaddressed.

Materials and method

**Data set description and its analysis**

**Background**

The World Health Organization (WHO) characterized the COVID-19, caused by the SARS-CoV-2, as a pandemic on March 11

Until March 27, the state of São Paulo had recorded 1,223 confirmed cases of COVID-19, with 68 related deaths, while the county of São Paulo, with a population of approximately 12 million people and where Hospital Israelita Albert Einstein is located, had 477 confirmed cases and 30 associated death, as of March 23.

The data is taken from kaggle released by Israelita Albert Einstein Hospital,  São Paulo, brazil.

**Dataset**

dataset contains anonymized data from patients seen at the Hospital Israelita Albert Einstein, at São Paulo, Brazil, and who had samples collected to perform the SARS-CoV-2 RT-PCR and additional laboratory tests during a visit to the hospital.

No of features =111

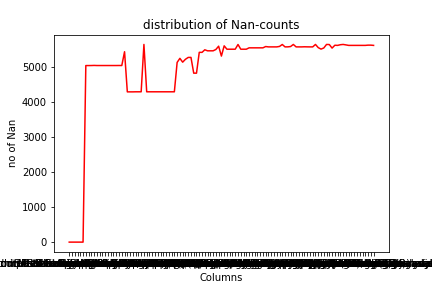
No of samples=5644

Percentage of NaN values = (551682 / 626484 ) \* 100 = 88.06%

**Methodology**

By considering number of NaN values of each  columns a distribution was plotted.

Fig – 1 shows the distribution of feature vs number of NaN



By analysing Fig –  1 we can say there only lesser columns which has  true values (ie. Not a missing value ).

So by considering threshold=1000 (ie minimum number of true values with respect to each column) we have the  maximum missing value persentage of 98% with 23 columns.

These 23 columns contain 16 columns with zero NaN values .

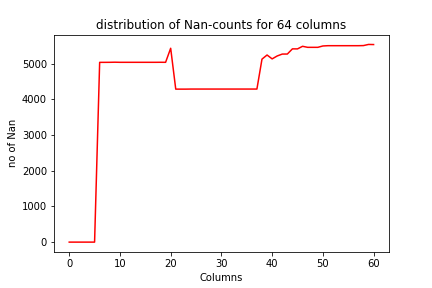


Fig – 2 shows the NaN value distribution of those 64 columns.

Coming to the binary features, these features are the several tests conducted on patients with two possible test outcome that are ‘detected’ and ‘not\_detected’.

These tests are :  ['RespiratorySyncytialVirus', 'InfluenzaA', 'InfluenzaB', 'Parainfluenza 1', 'CoronavirusNL63', 'Rhinovirus/Enterovirus', 'Coronavirus HKU1', 'Parainfluenza3', 'Chlamydophilapneumoniae', 'Adenovirus', 'Parainfluenza4', 'Coronavirus229E', 'CoronavirusOC43', 'Inf A H1N1 2009', 'Bordetella pertussis', 'Metapneumovirus', 'Parainfluenza 2']

  By assuming for all the missing values in the above mentioned columns are the particular tests not conducted on that patient . So we consider missing values for these binary features for value 2.

Then applying label encoding  we take:

‘not\_detected’  - 1

‘detected’         - 0

Next comes the Blood test results with over 14 columns having numeric values for all columns

i.e : ['Haemoglobin', 'Platelets', 'Mean platelet volume ', 'Red blood Cells',

'Lymphocytes', 'Mean corpuscular haemoglobin concentration (MCHC)',

'Leukocytes', 'Basophils', 'Mean corpuscular haemoglobin (MCH)',

'Eosinophils', 'Mean corpuscular volume (MCV)', 'Monocytes',

'Red blood cell distribution width (RDW)', 'Serum Glucose']

Multi-iterative Imputer was applied for each of these numeric columns , multi-iterative imputer is  a strategy for imputing missing values by modelling each feature with missing values as a function of other features in a round-robin fashion.

Task 1

Data set was split into X and  Y :

X = All above mentions columns were taken for value of X

Y= [‘SARS-Cov-2 exam result’]

Y has either positive or negative data, where out of 2101 rows 182 rows are positive and rest are negative

Random forest is a type of supervised machine learning algorithm based on [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning). Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The [random forest](https://en.wikipedia.org/wiki/Random_forest) algorithm combines multiple algorithm of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

Random forest classifier was applied and obtained an accuracy of 91%.

Task 2

Data set was split into X and  Y :

X = All above mentions columns were taken for value of X

Y= ['Patient admitted to regular ward (1=yes, 0=no)',

'Patient admitted to semi-intensive unit (1=yes, 0=no)',

'Patient admitted to intensive care unit (1=yes, 0=no)']

Random forest is a type of supervised machine learning algorithm based on [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning). Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The [random forest](https://en.wikipedia.org/wiki/Random_forest) algorithm combines multiple algorithm of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

Random forest classifier was applied and obtained an accuracy of 92.87%.