**ANALYSIS OF COVID-19 BASED UPON SYMPTOMS**

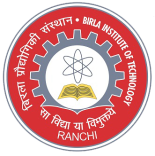
***A software design specification submitted in partial fulfillment of the requirements***

***for the award of the degree of***

**MASTER OF COMPUTER APPLICATIONS**

BY

**Meghaj Kumar Mallick (MCA/25017/18)**



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**BIRLA INSTITUTE OF TECHNOLOGY, MESRA**

**JAIPUR CAMPUS, JAIPUR**

**SP-2021**

**INDEX PAGE**

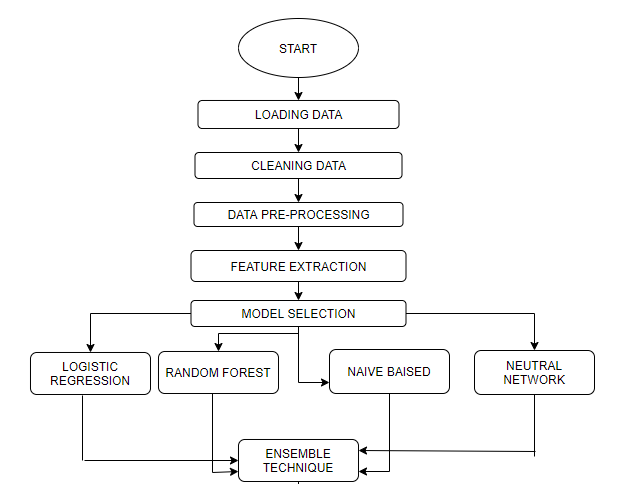
|  |  |  |
| --- | --- | --- |
| **SR NO.** | **TOPIC** | **PAGE NO.** |
| 1 | THE DATASET | 3 |
| 2 | SOFTWARE DESIGN | 3 |
| 3 | LOADING & CLEANING THE DATA |  |
| 4 | DATA PRE-PROCESSING |  |
| 5 | FEATURE EXTRACTION & MODEL SELECTION |  |
| 6 (A) | LOGISTIC REGRESSION ALGORITHM |  |
| 6 (B) | RANDOM FOREST ALGORITHM |  |
| 6 (C) | NAÏVE BAISED ALGORITHM |  |
| 6 (D) | NEURAL METWORK |  |
| 7 | ENSEMBLE TECHNIQUE |  |
| 8 | PREDICTION |  |

1. **DATASET**

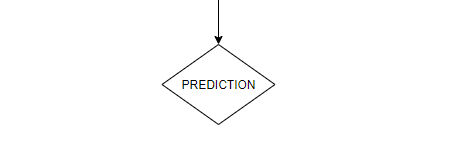
* Data collection was an essential and protracted process. Regardless the field of research, accuracy of the data collection is essential to maintain cohesion.
* These data will help to identify whether any person is having COVID-19 or not based upon some pre-defined standard symptoms by World Health Organization (WHO) and Ministry of Health & Family Welfare, India.
* The data-set is a combined multi-dimensional data. It contains fields with textual data and some with precise values. The data set for COVID-19 symptoms have been downloaded from world health organization.
* The attributes that were considered in the data-set for the machine learning model are presented in Table 1.

|  |  |  |
| --- | --- | --- |
| **SR. No.** | **Feature Name** | **Feature Description** |
| 1 | Person Gender | The gender of a person |
| 2 | Person Age | Classification of age according to WHO age group standard |
| 3 | Country Visited | List of country a person has visited. |
| 4 | Symptoms | According to WHO there are tiredness, difficulty in breathing, dry cough, sore throat, pain, nasal congestion, runny nose, diarrhoea etc. |
| 5 | Severity | Levels of severity are mild, moderate & severe. |
| 6 | Contact | Has the person came in contacted with COVID-19 patient |

1. **SOFTWARE DESIGN**



NAÏVE BAYES



1. **LOADING & CLEANING DATA**

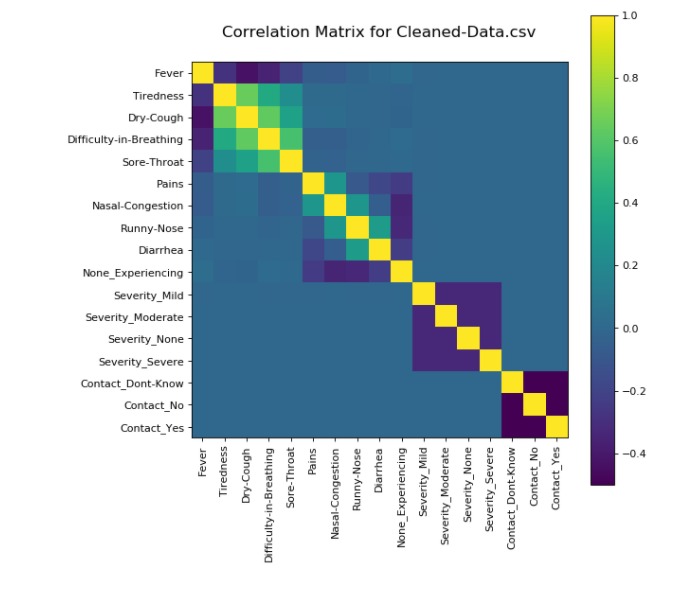
The main reason behind data processing is that **data almost never comes in a form that is ready for us** and our personal experience, a large amount of time spent on a data science project is on manipulating data

.

* The dataset is being

.

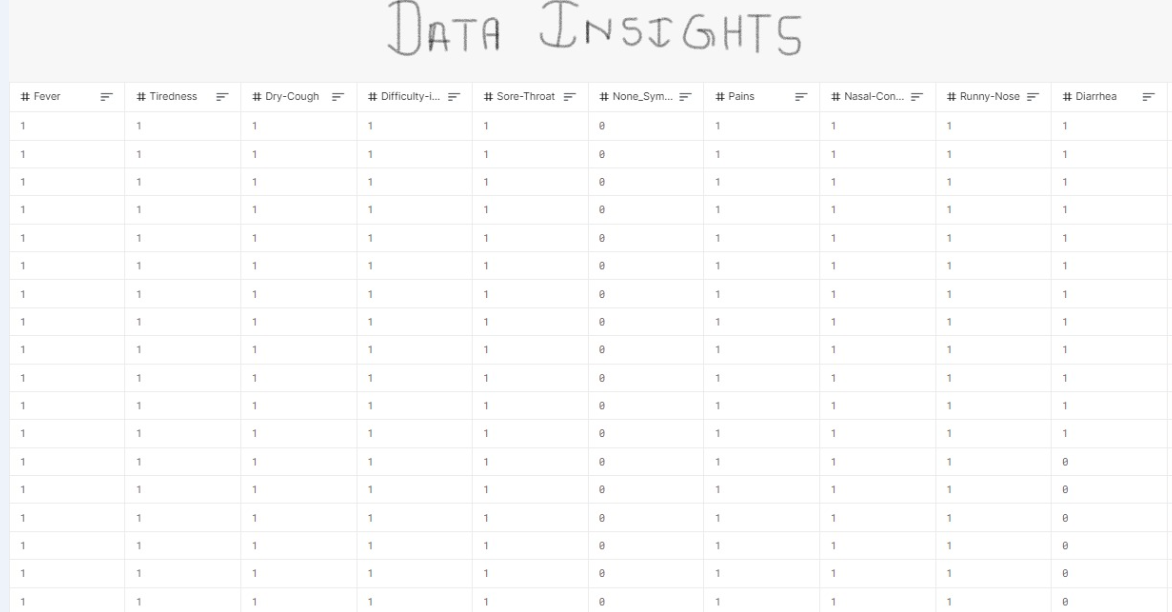
* While loading our dataset we have raw, separate values (.csv files). Where we have 3,16,000 records on patients varying from different regions of the world.
* We have approximately twenty-five plus features which are acting as symptoms & we have to classify the data into four major categories based on our feature.
* We have binary data for each and every dependent & independent variables



1. **DATA PRE-PROCESSING**

Data pre-processing occurs to be a curial step while implementing in every data science project. In our case due to binary data sight in every feature, we don’t require much pre-processing of our data.

* We have first tried to use feature engineering to handle our categorical data input. There are some null values after that we have gone for a basic dimensionally deduction by admitting the non reliable feature of the data.
* At last we have also gone through capturing feature i.e. very important for some analysis about the most relevant feature acting in our results.



1. **MODEL SELECTION**

Due to the crucial medical condition we cannot rely completely on a single model. Hence we have to go for the ensemble technique which can help in adapting the result from different models and finally improve our result.

We have to use several machine learning algorithms such as logistic regression, random forest, naïve bayes & artificial neural network to create a model. These model will help to analyse & predict the final result using the bagging approach of the ensemble technique.

1. **(A) LOGISTIC REGRESSION ALGORITHM**

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc



**6(B) RANDOM FOREST ALGORITHM**

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest.

Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.



Implementation in Python

First, start with importing necessary Python packages −

import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

**6 (C) NAÏVE BAYES ALGORITHMS**

Naïve Bayes algorithms is a classification technique based on applying Bayes’ theorem with a strong assumption that all the predictors are independent to each other. In simple words, the assumption is that the presence of a feature in a class is independent to the presence of any other feature in the same class.

In Bayesian classification, the main interest is to find the posterior probabilities i.e. the probability of a label given some observed features,

𝑃(𝐿 | 𝑓𝑒𝑎𝑡𝑢𝑟𝑒𝑠). With the help of Bayes theorem, we can express this in quantitative form as follows −

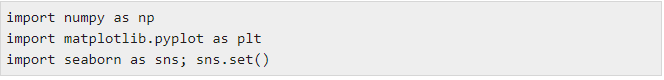
P(L|features)=P(L)P(features|L)P(features)P(L|features)=P(L)P(features|L)𝑃(𝑓𝑒𝑎𝑡𝑢𝑟𝑒𝑠)

Here, 𝑃(𝐿 | 𝑓𝑒𝑎𝑡𝑢𝑟𝑒𝑠) is the posterior probability of class.

𝑃(𝐿) is the prior probability of class.

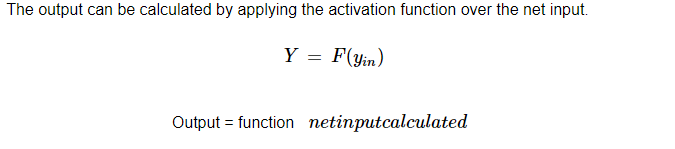
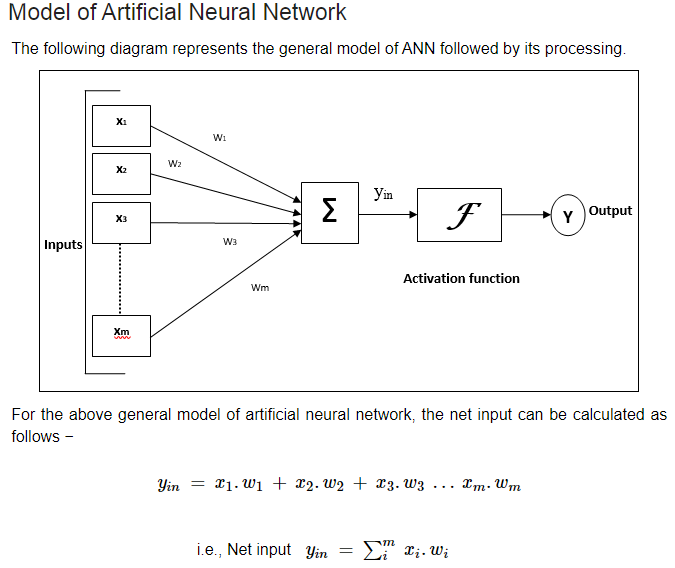
𝑃(𝑓𝑒𝑎𝑡𝑢𝑟𝑒𝑠 | 𝐿) is the likelihood which is the probability of predictor given class.

𝑃(𝑓𝑒𝑎𝑡𝑢𝑟𝑒𝑠) is the prior probability of predictor.



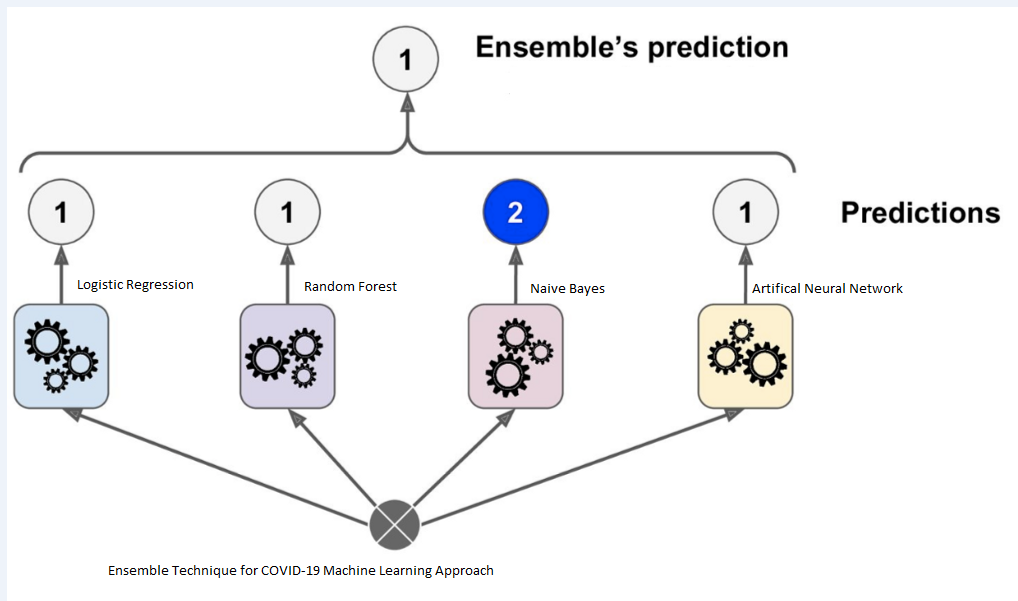
**6 (D) ARTIFICAIL NEURAL NETWORK ALGORITHM**

Artificial neural networks, usually simply called neural networks, are computing systems vaguely inspired by the biological neural networks that constitute animal brains. An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain.



1. **ENSEMBLE TECHNIQUE**

Ensemble technique is a machine learning technique that combines several base models in order to produce one optimal predictive model. In statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone



**Figure 1 shows the ensemble technique**

In the above figure ensemble technique is used to predict the result because multiple models are being used to get the result.

1. **PREDICTION**

Finally after evaluating the final result came out by the bagging approach is very accurate we will get our predictable result .Our data is in the binary format so the values will be stored in the array.