**COVID -19 PREDICTION AND SYMPTOMS RELAVANCE**

**KIRUPA KRISHAN .G,**

**Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India.| School of Electronics and Communication.**

**SURVEY AND ANALYSIS**

**Abstract**

The main focus of this Research is to see how one can easily predict if he/she has been infected by COVID-19. Another aspect is deciding which COVID-19 symptom is more likely to show positive result of virus contamination. The virus has been declared as a pandemic and has affected more than 66,729,375 people across 220 countries and has also cost the lives of 1,535,982 people as of the time this paper is being written. Research still predicts that another second wave is to hit soon. So with my analysis we are able to predict if a person is infected by the covid-19 virus and to stipulate the relevance of a symptom with virus infection. The required objective is obtained using complex machine learning algorithms that are able to predict, up to an extent the probability that the person has covid-19 and also if the related covid-19 symptoms provided are relevant to the condition or not. The algorithm used is a LOGISTIC REGRESSION algorithm that is used as a classification tool to separate the data into binary results, which in our case is if the person has covid-19 or not (YES OR NO). The dataset used to train this machine learning algorithm is obtained from online resources\*(link given below) and a public survey. The machine learning model as of now has been able to predict the probability of virus contamination by 66.89% accuracy, further the model is able to relate if a given symptom is valid or not. With this we are able to conclude that the model is working fine and only fine tuning of the model is required in order to improve and enhance the accuracy and probability of exact results. The result could also be improved using ANN( Artificial Neural Networks) but as it is computational expensive and due to lack of resources the expected performance cannot be met. Further development will be made keeping current computational resources.

**Introduction**

COVID-19 is the disease caused by a new strain of coronavirus called SARS-CoV-2.  WHO first learned of this new virus on 31 December 2019, following a report of a cluster of cases of ‘viral pneumonia’ in Wuhan, People’s Republic of China. The virus from there on has been declared as a pandemic with over 66,729,375 reported cases across 220 countries and has also cost the lives of 1,535,982 humans. The goal has been to see that the time required to predict infection is reduced significantly and that the cost involved is also reduced. The model first uses a LogisticRegression machine learning model that is predominately used for classification type of problems (YES/NO). The dataset contains approx. 300,000 data from people around the world with features containing [Fever, Tiredness, Difficulty in breathing, Dry-Cough, Sore-Throat, Nasal-Congestion, Runny-Nose, Diarrhea, Age, Gender, Severity, Contact] all of these data were binary in nature and were encoded as YES-1 and NO-0.

**Discussion**

First the model was tested with the given data without any changes to the dataset and thus produced a very low probability of 35%-40%. With further changes in dataset such as reducing the data within certain limits and standardizing the data for better result, the probability range were significantly increased to 55%. The dataset from now was clean and needed no pre-processing hence the attention was set to changing the parameters of the LogisticRegression model produced the final probability to 66.9%, no further changes were able to increase the accuracy of the model. Thus to improve upon this **ANN(artificial neural network)** was used, most commonly ANN models perform better than the ML models, thus the hyper-parameters set for the ANN were set with default values. The ANN model finds the best result by reducing the loss of the given data, this is done using the **Gradient-Descent Method** with the **ADAM optimizer** and **Relu activation function for the hidden layers** and **Sigmoid activation for the output layer**. The neural network was set initially with **1 input layer , 4 hidden layers** (with 100 parameters) and **1 output layer** (with single parameter as it is binary in nature) and was set to run around **100 epochs**. The model was then fit and compiled, which upon completion gave the same 67.3% result. The problem was then clear within seconds, as the Gradient-Descent Method tries to reduce loss by finding the global minimum of the function. The major problem with this type of Model is that it is very easy to interpret the local minimum as the global minimum and hence the model fixes to this point. There is no easy way to solve this problem than to repeat the data collection in a different way, as the data collected was all categorical, thus result is clear that the model predicted all wrong values and no considerable predictions were made. One of the methods to solve the dataset situation will be to take nominal data for the given features instead of categorical as they tend to have very less local minimums.

**Conclusion**

Figures and Tables Figures and Tables can be included in the running script (if necessary). Figures should be in JPEG/TIF format with high resolution. All figures and tables should be inserted with a caption.

**Bilal H. Hungund K.**

MTech. Data Science Student at NMIMS University

Mumbai, Maharashtra, India

Page Limit: Maximum 3 pages